Student Name: Saeedreza Zouashkiani

Student ID: 400206262

Deep Learning Homework 1

1)

1-a)

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

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

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

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

1-b)

1-b-i)

$$P(X \ge a) \le \frac{E(X)}{a}$$

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

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

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

Therefore

$$P(X \ge a) \le \frac{E(X)}{a} \blacksquare$$

1-b-ii)

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

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 π.

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

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

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)

$$\frac{\partial \boldsymbol{a^T x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x^T a}}{\partial \boldsymbol{x}} = \frac{\partial(a_1x_1 + \cdots + a_nx_n)}{\partial \boldsymbol{x}} = [a_1 \dots a_n] = \boldsymbol{a^T}$$

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

2-ii)

$$\frac{\partial \boldsymbol{A A^{-1}}}{\partial \beta} = 0 = \frac{\partial \boldsymbol{A}}{\partial \beta}\boldsymbol{A^{-1}} + \boldsymbol{A}\frac{\partial \boldsymbol{A^{-1}}}{\partial \beta}$$

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

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

2-iii)

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

$$adj(\boldsymbol{A}) = \boldsymbol{C}^T$$

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

$$(\boldsymbol{A^{-1}})_{ij}^T = \frac{1}{\det(\boldsymbol{A})}C_{ij}$$

By cofactor expansion of **A**

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

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

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

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

3)

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

3-i) $tr(\boldsymbol{A}) = \lambda_1 + \cdots + \lambda_n$

The characteristic polynomial of A is defined as

$$p_{\boldsymbol{A}}(t) = \det(t\boldsymbol{I} - \boldsymbol{A}) = (-1)^n (t^n - (tr(\boldsymbol{A}) t^{n-1} + \cdots + (-1)^n \det(\boldsymbol{A}))$$

Also, the characteristic polynomial can be factorized as

$$p_{\boldsymbol{A}}(t) = (-1)^n (t - \lambda_1) \ldots (t - \lambda_n)$$

So, by comparing terms we get

$$tr(\boldsymbol{A}) = \lambda_1 + \cdots + \lambda_n$$

3-ii) $\det(\boldsymbol{A}) = \lambda_1 \ldots \lambda_n$

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

4)

4-i)

If A has full column rank, the inverse of $\boldsymbol{A^T A}$ exists

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

$$\boldsymbol{A} = \boldsymbol{A A^\dagger A}$$

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

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

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

4-ii)

If **A** has full row rank, the inverse of $\boldsymbol{AA^T}$ exists

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

Or

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

5)

5-i)

We start by eliminating the block matrix under A

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

Then by eliminating the element above D

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

Therefore, by combining the two

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

Thus

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

Which is the LDU decomposition of M.

Therefore

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

5-ii)

Like the last part, using block LDU decomposition when D is invertible, we get

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

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

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.

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

$$= \begin{bmatrix} \boldsymbol{I_n} & \boldsymbol{-A^{-1}B} \\ \boldsymbol{0} & \boldsymbol{I_k} \end{bmatrix} \begin{bmatrix} \boldsymbol{A^{-1}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\left(D - CA^{-1}B\right)^{-1}} \end{bmatrix} \begin{bmatrix} \boldsymbol{I_n} & \boldsymbol{0} \\ \boldsymbol{-CA^{-1}} & \boldsymbol{I_k} \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{A^{-1} + A^{-1}B\left(D - CA^{-1}B\right)^{-1}CA^{-1}} & \boldsymbol{-A^{-1}B\left(D - CA^{-1}B\right)^{-1}} \\ \boldsymbol{-\left(D - CA^{-1}B\right)^{-1}CA^{-1}} & \boldsymbol{\left(D - CA^{-1}B\right)^{-1}} \end{bmatrix}$$

Now inverting part ii, results in:

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

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

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

By comparing the first block matrix we get

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

If we substitute $\boldsymbol{D}$ with $\boldsymbol{-D}$

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

5-iv)

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

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

6)

6-i)

Using question 5

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

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

6-ii)

Using Courant-Fischer's theorem

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

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

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 $\max_{\substack{\boldsymbol{x}\in V_A, \\ \|\boldsymbol{x}\|=1}} < \boldsymbol{A}\boldsymbol{x}, \boldsymbol{x} >$, $\max_{\substack{\boldsymbol{x}\in V_B, \\ \|\boldsymbol{x}\|=1}} < \boldsymbol{B}\boldsymbol{x}, \boldsymbol{x} >$ and let $W = V_A \cap V_B$ be their intersection. W has dimension of at least i.

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

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

6-iii)

Let $\boldsymbol{M} = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{0} & 0 \end{pmatrix}$ and $\boldsymbol{N} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{y} \\ \boldsymbol{y}^* & a \end{pmatrix}$. Another inequality from Weyl states that:

$$\lambda_i(\boldsymbol{M}+\boldsymbol{N}) = \lambda_i(\boldsymbol{A}) \geq \lambda_j(\boldsymbol{N}) + \lambda_{i-j+1}(\boldsymbol{M}) \quad (1)$$

Using part i:

$$p_{\boldsymbol{M}}(t) = t p_{\boldsymbol{B}}(t) \quad (2)$$
$$p_{\boldsymbol{N}}(t) = (t-a)\det(t\boldsymbol{I}) - \boldsymbol{y}^* \det(t\boldsymbol{I})(t\boldsymbol{I})^{-1}\boldsymbol{y} = t^{n-1}(t(t-a) - \boldsymbol{y}^*\boldsymbol{y}) \quad (3)$$

Therefore $\lambda_{i+1}(\boldsymbol{M}) = \lambda_i(\boldsymbol{B})$ and $\lambda_1(\boldsymbol{M}) = 0$. Also $\lambda_i(\boldsymbol{N}) = 0, \forall i \neq n+1, i \neq 1$ and it has a negative and positive eigenvalue which correspond to $\lambda_1(\boldsymbol{N})$ and $\lambda_{n+1}(\boldsymbol{N})$ respectively.

Using Weyl inequalities and (2), (3):

$$\lambda_i(\boldsymbol{A}) \leq \lambda_{i+j}(\boldsymbol{M}) + \lambda_{n+1-j}(\boldsymbol{N})$$
$$\lambda_i(\boldsymbol{A}) \leq \lambda_{i+1}(\boldsymbol{M}) + \lambda_{n+1-1}(\boldsymbol{N}) = \lambda_{i+1}(\boldsymbol{M}) = \lambda_i(\boldsymbol{B}) \quad (4)$$

And:

$$\lambda_i(\boldsymbol{A}) \geq \lambda_1(\boldsymbol{N}) + \lambda_i(\boldsymbol{M}) = \lambda_i(\boldsymbol{M}) = \lambda_{i-1}(\boldsymbol{B}) \quad (5)$$

Using (4), (5):

$$\lambda_1(\boldsymbol{A}) \leq \lambda_1(\boldsymbol{B}) \leq \lambda_2(\boldsymbol{A}) \leq \lambda_2(\boldsymbol{B}) \leq \cdots \leq \lambda_n(\boldsymbol{A}) \leq \lambda_n(\boldsymbol{B}) \leq \lambda_{n+1}(\boldsymbol{A})$$

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

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

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

Taking derivative of log likelihood function with respect to μ 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 μ

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

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 μ and setting it to zero

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

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

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

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

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

We will at first prove part ii

ii)

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

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

$$\boldsymbol{x_a}=\boldsymbol{S_a}\boldsymbol{x}$$

Therefore, by applying linear transformation

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

Therefore $\boldsymbol{x_a}$ is a normal distribution with

$$E(\boldsymbol{x_a})=\boldsymbol{\mu_a}$$

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

i)

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

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

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

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

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

According to question 5

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

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

Plugging the inverse into (1):

$$p(\boldsymbol{x_a}|\boldsymbol{x_b}) = \frac{1}{\sqrt{(2\pi)^{n-n_b}}}\sqrt{\frac{|\boldsymbol{\Sigma_{bb}}|}{|\boldsymbol{\Sigma}|}}\exp\Big(-\frac{1}{2}((\boldsymbol{x_a}-\boldsymbol{\mu_a})^{\boldsymbol{T}}\big(\boldsymbol{\Sigma_{aa}}-\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}\boldsymbol{\Sigma_{ba}}\big)^{-1}(\boldsymbol{x_a}-\boldsymbol{\mu_a})$$
$$-2(\boldsymbol{x_a}-\boldsymbol{\mu_a})^{\boldsymbol{T}}\big(\boldsymbol{\Sigma_{aa}}-\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}\boldsymbol{\Sigma_{ba}}\big)^{-1}\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}(\boldsymbol{x_b}-\boldsymbol{\mu_b})+(\boldsymbol{x_b}-\boldsymbol{\mu_b})^{\boldsymbol{T}}(\boldsymbol{\Sigma_{bb}^{-1}}$$
$$+\boldsymbol{\Sigma_{bb}^{-1}}\boldsymbol{\Sigma_{ba}}\big(\boldsymbol{\Sigma_{aa}}-\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}\boldsymbol{\Sigma_{ba}}\big)^{-1}\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}})(\boldsymbol{x_b}-\boldsymbol{\mu_b}))+\frac{1}{2}(\boldsymbol{x_b}-\boldsymbol{\mu_b})^{\boldsymbol{T}}\boldsymbol{\Sigma_{\mathbf{bb}}^{-1}}(\boldsymbol{x_b}-\boldsymbol{\mu_b})\Big)$$
$$=\frac{1}{\sqrt{(2\pi)^{n-n_b}}}\sqrt{\frac{|\boldsymbol{\Sigma_{bb}}|}{|\boldsymbol{\Sigma}|}}\exp\Big(-\frac{1}{2}((\boldsymbol{x_a}-\boldsymbol{\mu_a})^{\boldsymbol{T}}\big(\boldsymbol{\Sigma_{aa}}-\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}\boldsymbol{\Sigma_{ba}}\big)^{-1}(\boldsymbol{x_a}-\boldsymbol{\mu_a})$$
$$-2(\boldsymbol{x_a}-\boldsymbol{\mu_a})^{\boldsymbol{T}}\big(\boldsymbol{\Sigma_{aa}}-\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}\boldsymbol{\Sigma_{ba}}\big)^{-1}\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}(\boldsymbol{x_b}-\boldsymbol{\mu_b})$$
$$+(\boldsymbol{x_b}-\boldsymbol{\mu_b})^{\boldsymbol{T}}\boldsymbol{\Sigma_{bb}^{-1}}\boldsymbol{\Sigma_{ba}}\big(\boldsymbol{\Sigma_{aa}}-\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}\boldsymbol{\Sigma_{ba}}\big)^{-1}\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}(\boldsymbol{x_b}-\boldsymbol{\mu_b}))\Big)$$
$$=\frac{1}{\sqrt{(2\pi)^{n-n_b}}}\sqrt{\frac{|\boldsymbol{\Sigma_{bb}}|}{|\boldsymbol{\Sigma}|}}\exp\Big(-\frac{1}{2}\big[(\boldsymbol{x_a}-\boldsymbol{\mu_a})-\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}(\boldsymbol{x_b}-\boldsymbol{\mu_b})\big]^{\boldsymbol{T}}\big(\boldsymbol{\Sigma_{aa}}$$
$$-\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}\boldsymbol{\Sigma_{ba}}\big)^{-1}\big[(\boldsymbol{x_a}-\boldsymbol{\mu_a})-\boldsymbol{\Sigma_{ab}}\boldsymbol{\Sigma_{bb}^{-1}}(\boldsymbol{x_b}-\boldsymbol{\mu_b})\big]\Big)$$

Also using the determinant derived in question 5

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

Thus

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

Which is the probability distribution of multivariate normal distribution

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

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

10)

10-i)

$$\min_{\boldsymbol{x}}\|\boldsymbol{Ax}-\boldsymbol{b}\| = \min_{\boldsymbol{x}}(\boldsymbol{Ax}-\boldsymbol{b})^{\boldsymbol{T}}(\boldsymbol{Ax}-\boldsymbol{b}) = \min_{\boldsymbol{x}}\big(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{Ax}-\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{b}-\boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{Ax}-\boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{b}\big)$$

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

$$\frac{\partial \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{Ax}-\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{b}-\boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{Ax}}{\partial \boldsymbol{x}}|_{\boldsymbol{x}=\boldsymbol{x}^*} = 0 = 2\boldsymbol{x}^{*\boldsymbol{T}}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A}-2\boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{A}$$

Therefore

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

To prove that $\boldsymbol{x}^*$ has the smallest norm 2 among all solutions, suppose that there exists $\boldsymbol{x}$ such that $\boldsymbol{Ax} = \boldsymbol{b}$. Therefore $\boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}^*) = \boldsymbol{0}$. This leads to the conclusion that $(\boldsymbol{x} - \boldsymbol{x}^*) \in \boldsymbol{K}(\boldsymbol{A})$, but the least square solution $\boldsymbol{x}^* \in \boldsymbol{K}(\boldsymbol{A})^{\perp}$. Therefore:

$$\|\boldsymbol{x}\|_2^2 = \|\boldsymbol{x} - \boldsymbol{x}^* + \boldsymbol{x}^*\|_2^2 = \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 + \|\boldsymbol{x}^*\|_2^2 \geq \|\boldsymbol{x}^*\|_2^2$$

10-ii)

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

$$\boldsymbol{x} = \boldsymbol{x} - \nu(\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^T\boldsymbol{b}) \rightarrow \nu(\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}^T\boldsymbol{b}) = \boldsymbol{0} \rightarrow \boldsymbol{x} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{b}$$

10-iii)

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

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

Define $\boldsymbol{y_t} = \boldsymbol{x_t} - \boldsymbol{x}^*$

$$\boldsymbol{y_{t+1}} = \boldsymbol{y_t} - \nu\left(\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{y_t} = \left(\boldsymbol{I} - \nu\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{y_t}$$

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

$$\boldsymbol{Q}^T\boldsymbol{y_{t+1}} = \boldsymbol{z_{t+1}} = \boldsymbol{Q}^T\left(\boldsymbol{I} - \nu\boldsymbol{A}^T\boldsymbol{A}\right)\boldsymbol{y_t} = \boldsymbol{Q}^T\left(\boldsymbol{Q}\boldsymbol{Q}^T - \nu\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T\right)\boldsymbol{y_t} = \boldsymbol{Q}^T\boldsymbol{Q}(\boldsymbol{I} - \nu\boldsymbol{\Lambda})\boldsymbol{Q}^T\boldsymbol{y_t} = (\boldsymbol{I} - \nu\boldsymbol{\Lambda})\boldsymbol{z_t}$$

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

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

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

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

And to satisfy convergence of all modes we shall have

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