Student Name: Saeedreza Zouashkiani

Student ID: 400206262

Deep Learning Homework 1

1)

1-a)

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

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

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

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

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 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:

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

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

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

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

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

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

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.

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

Now inverting part ii, results in:

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

By comparing the first block matrix we get

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

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

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

5-iv)

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

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

6)

6-i)

Using question 5

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

6-ii)

Using Courant-Fischer's theorem

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

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

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

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

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 $V\subseteq W$

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

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

8-ii) MAP

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

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

Taking the log

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

Taking derivative with respect to μ and setting it to zero

10)

10-i)

$$\min_{\boldsymbol{x}}\|\boldsymbol{Ax}-\boldsymbol{b}\|=\min_{\boldsymbol{x}}(\boldsymbol{Ax}-\boldsymbol{b})^{\boldsymbol{T}}(\boldsymbol{Ax}-\boldsymbol{b})=\min_{\boldsymbol{x}}\left(\boldsymbol{x^TA^TAx}-\boldsymbol{x^TA^Tb}-\boldsymbol{b^TAx}-\boldsymbol{b^Tb}\right)$$

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

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

Therefore

$$\boldsymbol{x}=\left(\boldsymbol{A^TA}\right)^{-1}\boldsymbol{A^Tb}=\boldsymbol{A}^{\dagger}\boldsymbol{b}$$

10-ii)

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

$$\boldsymbol{x}=\boldsymbol{x}-\nu(\boldsymbol{A^TAx}-\boldsymbol{A^Tb})\rightarrow\nu(\boldsymbol{A^TAx}-\boldsymbol{A^Tb})=\boldsymbol{0}\rightarrow\boldsymbol{x}=\left(\boldsymbol{A^TA}\right)^{-1}\boldsymbol{A^Tb}$$

10-iii)

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

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

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

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

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

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

$(\boldsymbol{I}-\boldsymbol{\nu\Lambda})$ 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 ($k \to \infty$) we shall have

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

And to satisfy convergence of all modes we shall have

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