IML Exercise 1 Answers

2. Theoretical part

2.1. Mathematical Background

2.1.1. Linear Algebra

1. For any orthogonal matrix $A \in M_{n\times n}$ the linear transformation by A is isometric. i.e.:

$$\forall x \in V_n : \|Ax\|_2 = \|x\|_2$$

Proof: Orthogonal Matrix is made of columns that are orthonormal vectors, i.e. they are $\{v_1, v_2, \dots, v_n\} : \forall i \in [n] \quad \|v_i\| = 1$
and $\forall i \in [n] : \langle v_i, v_i \rangle = 1, \ \ \forall i, j \in [n]\ i \neq j : \langle v_i, v_j \rangle = 0$ vectors are mutually perpendicular.
and the $span(\{v_1, v_2, \dots, v_n\}) = \mathbb{R}^n$
And for the reciprocal matrix $A^\top A = AA^\top = I = A^{-1}A = AA^{-1}$

~~So as a result, $\forall x \in \mathbb{R}^n$ can be represented by this base: $x = \sum a_i v_i$~~

For simplicity we'll prove the square of each norm $\|Ax\|_2^2 = \|x\|_2^2$ Since the norms are in $\mathbb{R}^+$ even after taking the root we know it is not negative.

$\|x\|^2 = \langle x, x \rangle = x^\top x \qquad \cancel{= (\sum a_i v_i)^\top (\sum a_j v_j) = \sum_i \sum_j a_i a_j v_i^\top v_j =}$

$\|Ax\|^2 = \langle Ax, Ax \rangle = (Ax)^\top (Ax) = x^\top A^\top A x = x^\top x$

So $\|Ax\|^2 = \|x\|^2$ and hence: $\|Ax\| = \|x\|$

2. We'll calculate SVD for $A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix} = U\Sigma V^\top$

We can start the decomposition by wither calculating $A^\top A$ or $AA^\top$

Smaller and easier: $AA^\top = U\Sigma V^\top V \Sigma^\top U^\top \ = \ U\Sigma\Sigma^\top U^\top$

$$AA^\top = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix} \times \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 6 \end{bmatrix}$$

$$= I_{2\times 2} \times \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & \sqrt{6} & 0 \end{bmatrix} \times \begin{bmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{6} \\ 0 & 0 \end{bmatrix} \times I_{2\times 2} = \ U\Sigma\Sigma^\top U^\top$$

So: $U = I_{2\times 2} = U^\top \quad \Sigma = \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & \sqrt{6} & 0 \end{bmatrix}$

Now for computing V we will make EVD that we already know it's e.vals and we need only to compute the e.vecs:

$$A^\top A = V\Sigma^\top U^\top U \Sigma V^\top \ = \ V\Sigma^\top \Sigma V^\top = V \times \begin{bmatrix} 2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 0 \end{bmatrix} \times V^\top$$

$$A^{\top}A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 2 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix} = V \times \begin{bmatrix} 2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 0 \end{bmatrix} \times V^{\top} = A'$$

For $\lambda_1 = 2 : \text{null}(A' - \lambda I)$ is by $\begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix} - \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & -2 \\ 2 & -2 & 2 \end{bmatrix}$

$$\begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & -2 \\ 2 & -2 & 2 \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{matrix} z = 0 \\ z = 0 \\ x + \cancel{z} = y \end{matrix} \text{ So the e.vec: } \begin{bmatrix} \alpha \\ \alpha \\ 0 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \\ 0 \end{bmatrix}$$

Normalized by: $1 = \|e.vec\|_2 = \sqrt{\alpha^2 + \alpha^2 + 0} = \sqrt{2} \cdot \alpha \;\Rightarrow\; \alpha = 1/\sqrt{2}$

For $\lambda_2 = 6 : \text{null}(A' - \lambda I)$ is by $\begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix} - \begin{bmatrix} 6 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 6 \end{bmatrix} = \begin{bmatrix} -4 & 0 & 2 \\ 0 & -4 & -2 \\ 2 & -2 & -2 \end{bmatrix}$

$$\text{null}\begin{bmatrix} -4 & 0 & 2 \\ 0 & -4 & -2 \\ 2 & -2 & -2 \end{bmatrix} = \begin{matrix} r_1 \\ r_2 \\ r_3 + 0.5r_1 \end{matrix} \;\text{null}\begin{bmatrix} -4 & 0 & 2 \\ 0 & -4 & -2 \\ 0 & -2 & -1 \end{bmatrix} = \begin{matrix} r_1 \\ r_2 \\ r_3 - 0.5r_2 \end{matrix} \;\text{null}\begin{bmatrix} -4 & 0 & 2 \\ 0 & -4 & -2 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} -4 & 0 & 2 \\ 0 & -4 & -2 \\ 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{matrix} 2x = z \\ -2y = z \\ 0 = 0 \end{matrix} \text{ So the e.vec: } \begin{bmatrix} \alpha \\ -\alpha \\ 2\alpha \end{bmatrix} = \begin{bmatrix} 1/\sqrt{6} \\ -1/\sqrt{6} \\ \sqrt{4/6} \end{bmatrix}$$

Normalized by: $1 = \|e.vec\|_2 = \sqrt{\alpha^2 + \alpha^2 + 4\alpha^2} = \sqrt{6} \cdot \alpha \;\Rightarrow\; \alpha = \frac{1}{\sqrt{6}}$

For $\lambda_3 = 0 : \text{null}(A' - \cancel{\lambda I})$ is by $\begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix}$

$$\text{null}\begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix} = \begin{matrix} r_1 \\ r_2 \\ r_3 - r_1 \end{matrix} \;\text{null}\begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 0 & -2 & 2 \end{bmatrix} = \begin{matrix} r_1 \\ r_2 \\ r_3 + r_2 \end{matrix} \;\text{null}\begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{matrix} -x = z \\ y = z \\ 0 = 0 \end{matrix} \text{ So the e.vec: } \begin{bmatrix} -\alpha \\ \alpha \\ \alpha \end{bmatrix} = \begin{bmatrix} -1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix}$$

Normalized by: $1 = \|e.vec\|_2 = \sqrt{\alpha^2 + \alpha^2 + \alpha^2} = \sqrt{3} \cdot \alpha \;\Rightarrow\; \alpha = 1/\sqrt{3}$

So collecting overall e.vecs: $V = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{6} & -1/\sqrt{3} \\ 1/\sqrt{2} & -1/\sqrt{6} & 1/\sqrt{3} \\ 0 & \sqrt{2/3} & 1/\sqrt{3} \end{bmatrix}$

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix} = U\Sigma V^{\top} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & \sqrt{6} & 0 \end{bmatrix} \times \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{6} & -1/\sqrt{6} & \sqrt{2/3} \\ -1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{bmatrix}$$

If we want e.vals sorted in the Σ matrix it will look like

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix} = U\Sigma V^\top = \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}}_{row\ swap} \times \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & \sqrt{2} & 0 \end{bmatrix} \times \begin{bmatrix} 1/\sqrt{6} & -1/\sqrt{6} & \sqrt{2/3} \\ 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ -1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{bmatrix}$$

3. We'll prove the power-iteration algorithm convergence to $\pm v_1$ (e.vec of λ_1) when $\forall i \in \{2..n\}:\ \lambda_1 > \lambda_i$ and initial selected $b_0 = \sum_{i=1}^{n} a_i v_i$ has $a_1 \neq 0$

First we'll prove that $b_{k+1} = \frac{C_0^{k+1} b_0}{\|C_0^{k+1} b_0\|}$ $\forall k \in \mathbb{R}^+$ by recursion

For $b_1 = \frac{C_0^1 b_0}{\|C_0^1 b_0\|}$ by definition, Assuming $b_k = \frac{C_0^k b_0}{\|C_0^k b_0\|}$:

$$b_{k+1} = \frac{C_0^1 b_k}{\|C_0^1 b_k\|} = \frac{C_0^1 \frac{C_0^k b_0}{\|C_0^k b_0\|}}{\left\| C_0^1 \frac{C_0^k b_0}{\|C_0^k b_0\|} \right\|} = \frac{C_0^1 C_0^k b_0}{\|C_0^k b_0\|} \cdot \frac{1}{\left(\frac{1}{\|C_0^k b_0\|}\right) \|C_0^1 C_0^k b_0\|} = \frac{C_0^{k+1} b_0}{\|C_0^{k+1} b_0\|}$$

Now $C_0^m = (A^\top A)^m = (V\Sigma^\top \Sigma V^\top)^m = \underbrace{(V\Sigma^\top \Sigma \cancel{V^\top})(\cancel{V}\Sigma^\top \Sigma \cancel{V^\top}) \ldots (\cancel{V}\Sigma^\top \Sigma V^\top)}_{m\ times} = VD^m V^\top$

when $D \stackrel{\text{def}}{=} \Sigma^\top \Sigma$ and is a diagonal matrix with $\lambda_1 \ldots \lambda_n$ on its diagonal.

Now let's examine: $\boldsymbol{C_0^{k+1} b_0} = VD^{k+1}V^\top \sum a_i v_i = VD^{k+1} \sum\sum a_i v_j^\top v_i \hat{e}_j =$

$$VD^{k+1} \sum\sum a_i \delta_{ij} \hat{e}_j = VD^{k+1} \sum a_i \hat{e}_i = V \sum \lambda_i^{k+1} a_i \hat{e}_i = V \sum \lambda_i^{k+1} a_i \hat{e}_i = \sum \lambda_i^{k+1} a_i V \hat{e}_i =$$

$$\sum \lambda_i^{k+1} a_i v_i$$

$$\left\|C_0^{k+1} b_0\right\|^2 = \left\langle \sum \lambda_i^{k+1} a_i v_i , \sum \lambda_j^{k+1} a_j v_j \right\rangle = \sum\sum a_i a_j \lambda_i^{k+1} \lambda_j^{k+1} \langle v_i, v_j \rangle =$$

$$= \sum\sum a_i a_j \lambda_i^{k+1} \lambda_j^{k+1} \delta_{ij} = \sum a_i^2 \lambda_i^{2k+2}$$

So:

$$b_{k+1} = \frac{C_0^{k+1} b_0}{\|C_0^{k+1} b_0\|} = \frac{\sum \lambda_i^{k+1} a_i v_i}{\left(\sum a_i^2 \lambda_i^{2k+2}\right)^{1/2}} \underbrace{\xrightarrow{k \to \infty}}_{*} \frac{\lambda_1^{k+1} a_1}{|\lambda_1^{k+1} a_1|} v_1 = \begin{Bmatrix} +v_1 & if\ a_1 > 0 \\ -v_1 & if\ a_1 < 0 \end{Bmatrix}$$

* - Since $\lambda_1 >$ other λs it dominants when k goes to infinity.

2.1.2 Multivariate calculus

4. For $x \in \mathbb{R}^n$ fixed, $U \in \mathbb{R}^{n\times n}$ fixed orthogonal matrix, $f: \mathbb{R}^n \to \mathbb{R}^n$

$$f(\sigma) = U \operatorname{diag}(\sigma)\, U^\top x$$

Since U is orthogonal matrix it's columns are vectors that spans R^n so we can express x:

$x = \sum a_i u_i$ or $U \times \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} = x$ and to get this $\bar{a}$ use $\bar{a} = U^\top x$

$$f(\sigma) = U \operatorname{diag}(\sigma)\, U^\top \sum a_i u_i = U \operatorname{diag}(\sigma) \sum\sum a_i u_j^\top u_i \widehat{e_j} =$$

$$= U \operatorname{diag}(\sigma) \sum a_i \widehat{e_i} = U \sum \sigma_i a_i \widehat{e_i} = \sum \sigma_i a_i u_i$$

So the Jacobian:

$$J_\sigma(f(\sigma)) = \begin{bmatrix} f_1(\sigma) \\ \vdots \\ f_n(\sigma) \end{bmatrix} \times \begin{bmatrix} \frac{\partial}{\partial \sigma_1} & \dots & \frac{\partial}{\partial \sigma_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(\sigma)}{\partial \sigma_1} & \dots & \frac{\partial f_1(\sigma)}{\partial \sigma_n} \\ \vdots & & \vdots \\ \frac{\partial f_n(\sigma)}{\partial \sigma_1} & \dots & \frac{\partial f_n(\sigma)}{\partial \sigma_n} \end{bmatrix} =$$

We'll examine on specific index k, l:

$$J_\sigma(f(\sigma))_{kl} = \left(J_\sigma\left(\sum \sigma_i a_i u_i\right)\right)_{kl} = \frac{\partial\left(\sum \sigma_i a_i u_i\right)_k}{\partial \sigma_l} = (a_l \overline{u_l})_k = a_l U_{kl}$$

$$J_\sigma(f(\sigma)) = \begin{bmatrix} a_1 U_{11} & \dots & a_n U_{1n} \\ \vdots & & \vdots \\ a_1 U_{n1} & \dots & a_n U_{nn} \end{bmatrix} = U \times \begin{bmatrix} a_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & a_n \end{bmatrix} = U \operatorname{diag}(\bar{a})$$

$$J_\sigma(f(\sigma)) = U \operatorname{diag}(U^\top x)$$

5. For $h: \mathbb{R}^n \to \mathbb{R}$ $h(\sigma) = \frac{1}{2}\|f(\sigma) - y\|^2$ we want to find $\nabla_x(h(x))$

We got: $f: \mathbb{R}^n \to \mathbb{R}^n$ $f(\sigma) = U \operatorname{diag}(\sigma)\, U^\top x$

Define: $g: \mathbb{R}^n \to \mathbb{R}$ $g(x) \stackrel{\text{def}}{=} \frac{1}{2}\|x\|^2$ $\nabla_x(g): \mathbb{R} \to \mathbb{R}^n$

We'll leaned in the recitation that for $g(x) \stackrel{\text{def}}{=} \frac{1}{2}\|x\|^2$: $\nabla_x(g(x)) = \nabla_x\left(\frac{1}{2}\|x\|^2\right) = x$

Short proof: $\nabla_x\left(\frac{1}{2}\|x\|^2\right)_i = \frac{\partial \frac{1}{2}\|x\|^2}{\partial x_i} = \frac{1}{2}\frac{\partial x^\top x}{\partial x_i} = \frac{1}{2}\frac{\partial \sum x_k^2}{\partial x_i} = \frac{1}{2}(2x_i + 0) = x_i$

$$\nabla_\sigma(h(\sigma)) = \nabla_\sigma(g(f(\sigma) - y)) = \nabla_{f(\sigma)}(g(f(\sigma) - y)) \times \mathrm{J}_\sigma(f(\sigma)) =$$

$$= \underbrace{f(\sigma)}_{\nabla_{f(\sigma)}\left(\frac{1}{2}\|f(\sigma)\|^2\right) = f(\sigma)} \times \underbrace{U \operatorname{diag}(U^\top x)}_{\text{const in } w} = \underbrace{(U \operatorname{diag}(\sigma)\, U^\top x)^\top}_{f(\sigma) \in \mathbb{R}^{1\times n}} \times \underbrace{U \operatorname{diag}(U^\top x)}_{\in \mathbb{R}^{n\times n} \text{ const in } w}$$

$$= x^\top U \operatorname{diag}(\sigma)\, \cancel{U^\top U} \operatorname{diag}(U^\top x)$$

$$= x^\top U \operatorname{diag}(\sigma) \operatorname{diag}(U^\top x) \quad \text{of } \mathbb{R}^{1\times n}$$

6. The soft-max function is $S:\mathbb{R}^d \to [0,1]^k \quad \forall j \in [d]: \ S(x)_j = \frac{e^{x_j}}{\sum_{l=1}^{k} e^{x_l}}$

   We'll assume $k = d$ as was answered in the exercise forum.

   The Q's is $J_x\big(S(x)\big) = ?$

   By $\left(\frac{g(x)}{f(x)}\right)' = \frac{g'(x)f(x)-f'(x)g(x)}{f^2(x)}$

   We have $\frac{\partial S(x)_j}{\partial x_i} = \frac{1}{\left(\sum_{l=1}^{k} e^{x_l}\right)^2} \cdot \left(\frac{\partial e^{x_j}}{\partial x_i}\sum_{l=1}^{k} e^{x_l} - \frac{\partial \sum_{l=1}^{k} e^{x_l}}{\partial x_i} e^{x_j}\right) =$

$$= \frac{1}{\left(\sum_{l=1}^{k} e^{x_l}\right)^2} \cdot \left(\delta_{ij} e^{x_j} \sum_{l=1}^{k} e^{x_l} - e^{x_i} e^{x_j}\right) =$$

$$= \frac{\delta_{ij} e^{x_j}}{\sum_{l=1}^{k} e^{x_l}} - \frac{e^{x_i} e^{x_j}}{\left(\sum_{l=1}^{k} e^{x_l}\right)^2} =$$

$$= \frac{e^{x_j}}{\sum_{l=1}^{k} e^{x_l}}\left(\delta_{ij} - \frac{e^{x_i}}{\sum_{l=1}^{k} e^{x_l}}\right) =$$

$$S(x)_j\big(\delta_{ij} - S(x)_i\big)$$

   So the Jacobian is $J_x\big(S(x)\big) = \begin{bmatrix} S(x)_1 - S(x)_1^2 & -S(x)_1 S(x)_2 & \cdots & -S(x)_1 S(x)_d \\ -S(x)_1 S(x)_2 & S(x)_2 - S(x)_2^2 & \cdots & -S(x)_2 S(x)_d \\ \vdots & \vdots & \ddots & \vdots \\ -S(x)_1 S(x)_d & -S(x)_2 S(x)_d & \cdots & S(x)_d - S(x)_d^2 \end{bmatrix}$

   Note it is a symmetric matrix.

   It is positive on the diagonal and negative on the rest of the entries.

7. The function $f: \mathbb{R}^2 \to \mathbb{R} \quad f(x,y) = x^3 - 5xy - y^5$, we'll find the Hessian of f

$$\nabla f(x,y) = \left[\frac{\partial f(x,y)}{\partial x} \ , \frac{\partial f(x,y)}{\partial y}\right] = [3x^2 - 5y \ , -5y^4 - 5x]$$

$$\mathrm{H}[f(x,y)] = \begin{bmatrix} \frac{\partial^2 f(x,y)}{\partial x^2} & \frac{\partial^2 f(x,y)}{\partial x \partial y} \\ \frac{\partial^2 f(x,y)}{\partial y \partial x} & \frac{\partial^2 f(x,y)}{\partial y} \end{bmatrix} = \begin{bmatrix} 6x & -5 \\ -5 & -20y^3 \end{bmatrix}$$

2.2. Estimation Theory

8. $x_1, x_2, \ldots \sim^{\text{i.i.d}} \mathcal{P}$ with $\mathbb{E}(\mathcal{P}) = \mu\,, Var(\mathcal{P}) = \sigma^2$ finite.

   We look on first $n \in \mathbb{N}$ and the mean estimator $\hat{\mu}_n = \frac{1}{n}\sum x_i$

   This estimator is unbiased as we saw in the lecture and $\mathbb{E}(\hat{\mu}_n) = \mu$

   We'll show it is also consistent meaning the probability is concentrated around the expected value with $\mathbb{P}(|\mu - \hat{\mu}_n| > \varepsilon) \underbrace{\to}_{n\to\infty} 0$

   By Chebyshev:

$$\mathbb{P}(|\mu - \hat{\mu}_n| > \varepsilon) < \frac{Var(\hat{\mu}_n)}{\varepsilon^2}$$

   As we saw in the lecture

$$Var(\hat{\mu}_n) = Var\left(\frac{1}{n}\sum x_i\right) = \frac{1}{n^2} Var\left(\sum x_i\right) = \frac{1}{n^2}\sum_{i=j}^{|\cdot|=n} Var(x_i) + \underbrace{\sum_{i\neq j}^{|\cdot|=n^2-n} Cov(x_i, x_j)}_{\text{0 bacuse it's i.i.d}}$$

$$Var(\hat{\mu}_n) = \frac{n}{n^2}\sigma^2$$

   So for every finite given $\mu\,, \sigma\,, \varepsilon$ the bounding is: $\mathbb{P}(|\mu - \hat{\mu}_n| > \varepsilon) < \frac{\sigma^2}{n\cdot\varepsilon^2} \underbrace{\to}_{n\to\infty} 0$

9. The sample set: $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_m \sim^{\text{i.i.d}} \mathcal{N}(\mu, \Sigma)$ when $\mu \in \mathbb{R}^d\,, \Sigma \in \mathbb{R}^{\mathrm{d}\times\mathrm{d}}$ are finite.

   We saw in the lecture the PDF of a sample is:

$$f(X) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{1}{2}(X-\mu)^\top \Sigma^{-1}(X-\mu)\right)$$

   The likelihood of the m i.i.d. samples set will be:

$$\mathcal{L}(\mu, \Sigma | \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m) = \prod_{i=1}^{m} \mathcal{L}(\mu, \Sigma | \boldsymbol{x}_i)$$

$$\mathcal{L}(\mu, \Sigma | \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m) = \prod_{i=1}^{m}\left((2\pi)^d|\Sigma|\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}_i-\mu)^\top \Sigma^{-1}(\boldsymbol{x}_i-\mu)\right)$$

   So the log-likelihood will be:

$$\ell(\mu, \Sigma | \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m) = \sum_{i=1}^{m} -\frac{\mathrm{d}}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(\boldsymbol{x}_i-\mu)^\top \Sigma^{-1}(\boldsymbol{x}_i-\mu)$$

   While for a single sample:

$$\ell(\mu, \Sigma | \boldsymbol{x}_i) = -\frac{\mathrm{d}}{2}\log(2\pi) - \frac{1}{2}\log(|\Sigma|) - \frac{1}{2}(\boldsymbol{x}_i-\mu)^\top \Sigma^{-1}(\boldsymbol{x}_i-\mu)$$

3. Practical part

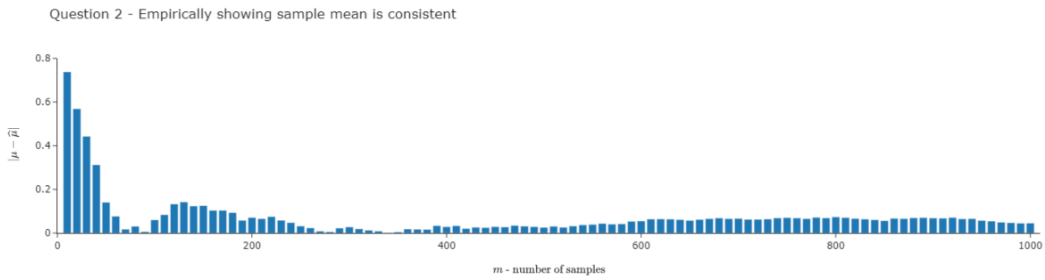
3.1. Univariate Gaussian Estimation

1. From a 1000 samples of normal distribution of $\mathcal{N}(10,1)$ with `np.random.seed(0)`, we got estimated mean , variance (unbiased estimator) of:

   `(9.954743292509804, 0.9752096659781323)`

   Calculated by: $\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \hat{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \hat{\mu}_m)^2$

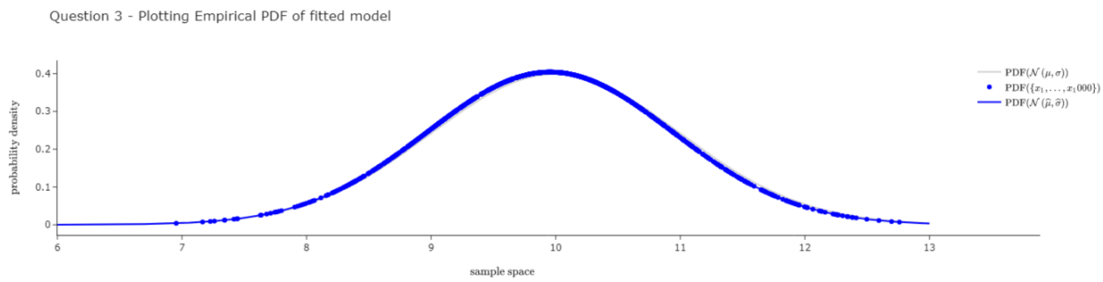
2. When sample set size is increasing from 10 to 1000 only on the samples set we already took on Q1, the consistency is demonstrated.



3. The Probability-Density-Function of the values in the data set of 1000 samples of $\mathcal{N}(10,1)$ with `np.random.seed(0)`, is compared here versus the ideal PDF.

   We can see we under-estimated the variance (0.975) and the ideal was slightly higher (1.0)

   The PDF of the sample points are on the estimated normal distribution model.



3.2. Multivariate Gaussian Estimation

4. From a 1000 samples of normal distribution of $\mathcal{N}(\mu, \Sigma)$ when

$$\mu = [0,0,4,0] \quad \Sigma = \begin{bmatrix} 1 & 0.2 & 0 & 0.5 \\ 0.2 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 1 \end{bmatrix}$$

with `np.random.seed(0)`,

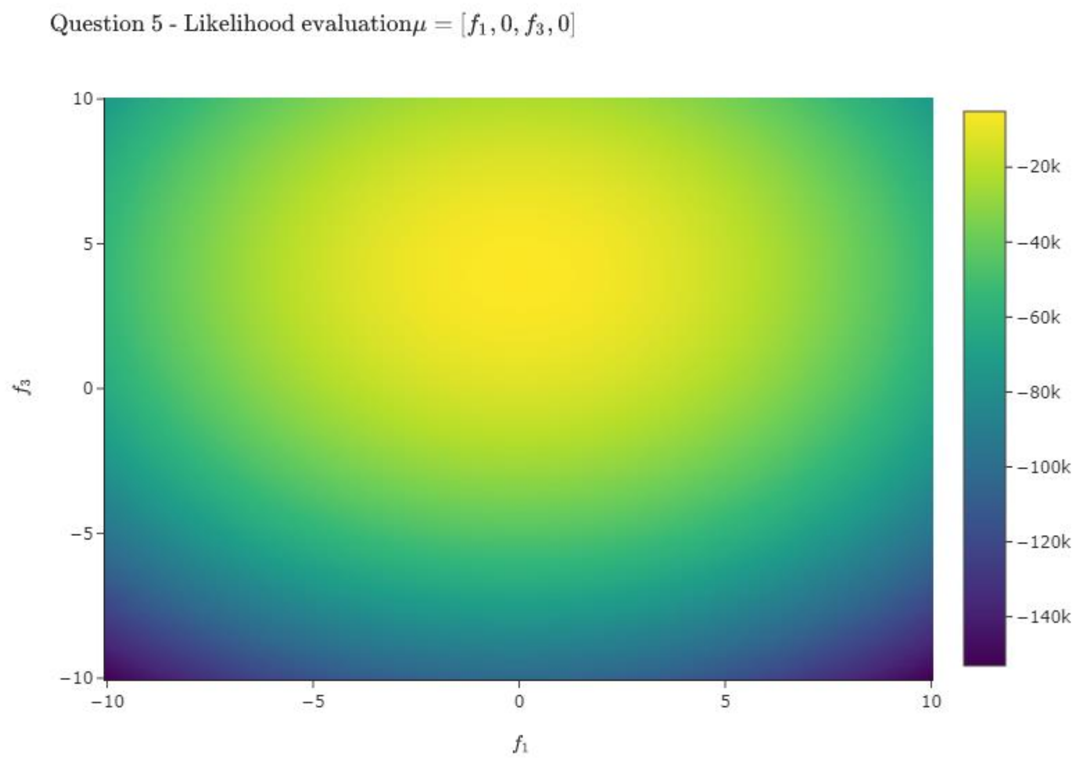
We estimate expected value by: $\hat{\mu}_{m_i} = \frac{1}{m}\sum_{k=1}^{m} x_{i_k}$

We estimate variance value by: $\hat{\sigma}^2_{m_{ij}} = \frac{1}{m-1}\sum_{k=1}^{m}\left(x_{j_k} - \hat{\mu}_{m_j}\right)\left(x_{i_k} - \hat{\mu}_{m_i}\right)$

we got estimated mean , variance (unbiased estimator) of:

```
Estimated mu vector is
[-0.02282878 -0.04313959  3.9932571  -0.02038981]
Estimated Covariance (Sigma) matrix is
[[ 0.91667608  0.16634444 -0.03027563  0.46288271]
 [ 0.16634444  1.9741828  -0.00587789  0.04557631]
 [-0.03027563 -0.00587789  0.97960271 -0.02036686]
 [ 0.46288271  0.04557631 -0.02036686  0.9725373 ]]
```

5. With the same covariance Matrix and same samples as in Q4, we scan the most probable $\mu = [f_1, 0, f_3, 0]$ vector.
We expect to get the result of $\widehat{f_1} \cong 0$, $\widehat{f_3} \cong 4$ and this is indeed the point with the highest log-likelihood.



6. The highest probable f_1, f_3 are

`(f1,f3) = (-0.05, 3.97)`

(Note it comes from an estimation with resolution of $\frac{10-(-10)}{200-1} \cong \frac{1}{10}$ and at the center of the range we have ~ … -0.15, -0.05, 0.05, 0.15 …)