Yosef Haim Abraham
206208258

Introduction to Machine Learning (67577)

Exercise 1
Estimation Theory & Mathematical Background

Second Semester, 2023

Contents

1 Submission Instructions **2**

2 Theoretical Part **2**

2.1 Mathematical Background 2
2.1.1 Linear Algebra 2
2.1.2 Multivariate Calculus 2
2.1.3 convexity 3
2.2 Estimation Theory 3

3 Practical Part **3**

3.1 Univariate Gaussian Estimation 3
3.2 Multivariate Gaussian Estimation 4

1 Submission Instructions

Please make sure to follow the general submission instructions available on the course website. In addition, for the following assignment, submit a single `ex1_ID.tar` file containing:

- An `Answers.pdf` file with the answers for all theoretical and practical questions (include plotted graphs *in* the PDF file).
- The following python files (without any directories): `gaussian_estimators.py`, `fit_gaussian_estimators.py`

The `ex1_ID.tar` file must be submitted in the designated Moodle activity prior to the date specified *in the activity*.

- Late submissions will not be accepted and result in a zero mark.
- Plots included as separate files will be considered as not provided.

2 Theoretical Part

2.1 Mathematical Background

2.1.1 Linear Algebra

Based on Recitation 1

1. Prove that orthogonal matrices are isometric transformations. That is, let $T : V \mapsto W$ be some linear transformation and A the corresponding matrix. Show that if A is an orthogonal matrix then $\forall x \in V \ ||Ax|| = ||x||$.
2. Calculate the SVD of the following matrix A. That is, find the matrices $U, \Sigma, V^\top$ where U, V are orthogonal matrices and Σ diagonal.

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix}$$

Recall, that to find the SVD of A we can calculate $A^\top A$ to deduce V,Σ and then calculate $AA^\top$ to deduce U. Equivalently, once we deduced V, Σ we can fine U using the equality $AV = U\Sigma$.

3. Show that the outer product of two vectors $\mathbf{v} \in \mathbb{R}^n, \mathbf{u} \in \mathbb{R}^m$, which is denoted by $\mathbf{v} \otimes \mathbf{u}$ or $\mathbf{v} \cdot \mathbf{u}^\top$ is a matrix $A \in \mathbb{R}^{n \times m}$ with $rank(A) = 1$. That is, show that all rows (or columns) in A are linearly dependent.
4. Show that for any otrthonormal basis $(\mathbf{u}_1, ..., \mathbf{u}_n)$ and any aribtrary vector $\mathbf{x} \in \mathbb{R}_n$ such that $\mathbf{x} = \sum_{i=1}^{n} a_i \cdot \mathbf{u}_i$, it holds that $a_i = \langle \mathbf{x}, \mathbf{u}_i \rangle$ for any $i \in [1, n]$. That is, show that the i'th coefficient of representing $\mathbf{x}$ in the basis $(\mathbf{u}_1, ..., \mathbf{u}_n)$, is the inner product between $\mathbf{x}$ and $\mathbf{u}_i$.

2.1.2 Multivariate Calculus

Based on Recitation 2

5. Let $x \in \mathbb{R}^n$ be a fixed vector and $U \in \mathbb{R}^{n \times n}$ a fixed orthogonal matrix. Calculate the Jacobian of the function $f : \mathbb{R}^n \to \mathbb{R}^n$:

$$f(\sigma) = U \cdot \text{diag}(\sigma) U^\top x$$

Where $\text{diag}(\sigma)$ is an $n \times n$ matrix where

$$\text{diag}(\sigma)_{ij} = \begin{cases} \sigma_i & i = j \\ 0 & i \neq j \end{cases}$$

6. Use the chain rule to calculate the gradient of $h(\sigma) = \frac{1}{2}||f(\sigma) - y||^2$
7. Calculate the Jacobian of the softmax function $S : \mathbb{R}^d \to [0,1]^k$

$$S(\mathbf{x})_j = \frac{e^{x_j}}{\sum_{l=1}^{k} e^{x_l}}$$

8. Let $f : \mathbb{R}^d \to \mathbb{R}$ be defined as $f(x,y) = x^3 - 5xy - y^5$. Calculate the Hessian of f.

2.1.3 convexity

Based on Recitation 2

9. Prove that the intersection $C := \bigcap_{i \in I} C_i$ for $\{C_i : i \in I\}$ a collection of convex sets is convex.
10. Prove that the vector sum $C_1 + C_2 := \{c_1 + c_2 : c_1 \in C_1, c_2 \in C_2\}$ of two convex sets is convex.
11. Prove that the set $\lambda C := \{\lambda c : c \in C\}$ is convex, for any convex set C, and every scalar λ.

2.2 Estimation Theory

Based on Lecture 1

12. Let $x_1, x_2, \ldots \overset{iid}{\sim} \mathcal{P}$ be a sample of infinity size drawn from some probability distribution function $\mathcal{P}$ with finite expectation and variance. Show that the sample mean estimator $\hat{\mu}_n = \frac{1}{n}\sum x_i$ calculated over the first n samples is a consistent estimator. Hint: for any given fixed value of $n \in \mathbb{N}$ bound from above the probability of deviating more than ε.
13. Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$ be m observations sampled i.i.d from a multivariate Gaussian with expectation of $\mu \in \mathbb{R}^d$ and a covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. Provide an expression for the log-likelihood function of $\mathcal{N}(\mu, \Sigma)$. Develop the expression as much as you can. Hint: follow the approach used to derive the likelihood function for the univariate case.

3 Practical Part

Before starting the practical part please make sure to have cloned/downloaded the IML.HUJI GitHub repository and set up a working virtual environment. Write the necessary code in the files specified in the questions.

3.1 Univariate Gaussian Estimation

Based on lecture 1

Implement the `UnivariateGaussian` class in the `learners.gaussian_estimators.py` file. Follow details specified in class and function documentation.

1. Using `numpy.random.normal` draw 1000 samples $x_1, \ldots, x_{1000} \overset{iid}{\sim} \mathcal{N}(10, 1)$ and fit a univariate Gaussian. Print the estimated expectation and variance. Output format should be `(expectation, variance)`.
2. Over previously drawn samples, fit a series of models of increasing samples size: 10, 20,...,100, 110,...1000. Plot the absolute distance between the estimated- and true value of the expectation, as a function of the sample size. Provide meaningful axis names and title.
3. Compute the PDF of the previously drawn samples using the model fitted in question 1. Plot the empirical PDF function under the fitted model. That is, create a scatter plot with the ordered sample values along the x-axis and their PDFs (using the `UnivariateGaussian.pdf` function) along the y-axis. Provide meaningful axis names and title. What are you expecting to see in the plot?

3.2 Multivariate Gaussian Estimation

Based on Lecture 1

Implement the `Multivariate` class in the `learners.gaussian_estimators.py` file. Follow details specified in class and function documentation.

NOTICE: When implementing the `log_likelihood` function you are required to use the expression developed in the q13 above. That is, the expression for $\ell(\mu, \Sigma | \mathbf{x}_1, \ldots, \mathbf{x}_m)$.

4. Using `numpy.random.multivariate_normal` draw 1000 samples $\mathbf{x}_1, \ldots, \mathbf{x}_{1000} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$

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

   Fit a multivariate Gaussian and print the estimated expectation and covariance matrix. Print each in a separate line.

5. Using the samples drawn in the question above calculate the log-likelihood for models with expectation $\mu = [f_1, 0, f_3, 0]^\top$ and the true covariance matrix defined above, where f_1, f_3 get values returned from `np.linspace(-10, 10, 200)`. Plot a heatmap of $f1$ values as rows, f_3 values as columns and the color being the calculated log likelihood. Provide meaningful axis names and title. What are you able to learn from the plot?

6. Of all values tested in question 5, which model (pair of values for feature 1 and 3) achieved the maximum log-likelihood value? Round to 3 decimal places

1. Prove that orthogonal matrices are isometric transformations. That is, let $T : V \mapsto W$ be some linear transformation and A the corresponding matrix. Show that if A is an orthogonal matrix then $\forall x \in V \ ||Ax|| = ||x||$.

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

$A^TA = Id \Leftarrow$ A אורתוג'

$$= x^T x = \langle x, x\rangle = \|x\|^2$$

כנדרש. ✓

2. Calculate the SVD of the following matrix A. That is, find the matrices $U, \Sigma, V^\top$ where U, V are orthogonal matrices and Σ diagonal.

$$A = \overbrace{\begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix}}^{r} \Big\} m$$

Recall, that to find the SVD of A we can calculate $A^\top A$ to deduce V, Σ and then calculate $AA^\top$ to deduce U. Equivalently, once we deduced V, Σ we can fine U using the equality $AV = U\Sigma$.

נחשב:

$$A^TA = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix}$$

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

ל- A^TA ול- AA^T אותם ערכים עצמיים לכן ניקח את אלה של AA^T:

$\lambda_1 = 6, \lambda_2 = 2$ כאשר הווקטורים העצמיים המתאימים הם $u_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, u_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$.

אז $U = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

נסתכל על $A^TA-2I=0$ ונדרג את A^TA-2I:

$$\begin{bmatrix}0&0&2\\0&0&-2\\2&-2&2\end{bmatrix}\to\begin{bmatrix}1&-1&1\\0&0&1\\0&0&0\end{bmatrix}\to\begin{bmatrix}1&-1&0\\0&0&1\\0&0&0\end{bmatrix}$$

$$\Leftarrow \quad \begin{aligned}x-y&=0\\z&=0\end{aligned}$$

$$\Leftarrow \quad \ker(A^TA-2I)=\{(t,t,0)\mid t\in\mathbb{R}\}$$

$$=\operatorname{Span}\{(1,1,0)\}$$

נסתכל על $A^TA-6I=0$ ונדרג את המטריצה:

$$\begin{bmatrix}-4&0&2\\0&-4&-2\\2&-2&-2\end{bmatrix}\to\begin{bmatrix}1&0&-\frac{1}{2}\\0&1&\frac{1}{2}\\1&-1&-1\end{bmatrix}\to\begin{bmatrix}1&0&-\frac{1}{2}\\0&1&\frac{1}{2}\\0&-1&-\frac{1}{2}\end{bmatrix}$$

$$\to\begin{bmatrix}1&0&-\frac{1}{2}\\0&1&\frac{1}{2}\\0&0&0\end{bmatrix}$$

$$\Leftarrow \quad y+\frac{1}{2}z=0\ ,\ x-\frac{z}{2}=0$$

$$\Leftarrow \quad \ker(A^TA-6I)=\left\{\left(\frac{t}{2},-\frac{t}{2},t\right)\ t\in\mathbb{R}\right\}$$

$$= \mathrm{Span}\left(\{(1,-1,2)\}\right)$$

לכן $\lambda = 0$ עבור A^TA נקבל:

$$\begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix} \to \begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 0 & -2 & 2 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\Rightarrow \mathrm{Ker}(A^TA) = \{(-t,t,t) \mid t \in \mathbb{R}\} = \mathrm{Sp}\{(-1,1,1)\}$$

אז נוכל

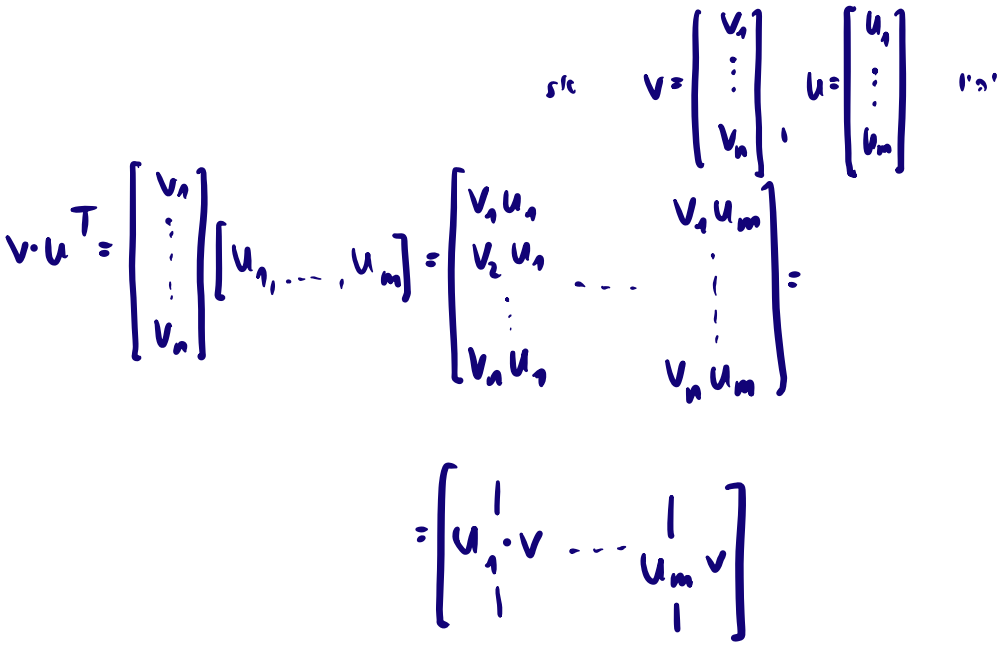
$$V = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{3}} \\ \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} \end{bmatrix}$$

ולכן

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & \sqrt{2} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{3}} \\ \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} \end{bmatrix}$$

זהו כיוון ה SVD של A.

3. Show that the outer product of two vectors $\mathbf{v} \in \mathbb{R}^n, \mathbf{u} \in \mathbb{R}^m$, which is denoted by $\mathbf{v} \otimes \mathbf{u}$ or $\mathbf{v} \cdot \mathbf{u}^\top$ is a matrix $A \in \mathbb{R}^{n \times m}$ with $rank(A) = 1$. That is, show that all rows (or columns) in A are linearly dependent.



4. Show that for any otrthonormal basis $(\mathbf{u}_1, ..., \mathbf{u}_n)$ and any aribtrary vector $\mathbf{x} \in \mathbb{R}_n$ such that $\mathbf{x} = \sum_{i=1}^{n} a_i \cdot \mathbf{u}_i$, it holds that $a_i = \langle \mathbf{x}, \mathbf{u}_i \rangle$ for any $i \in [1, n]$. That is, show that the i'th coefficient of representing $\mathbf{x}$ in the basis $(\mathbf{u}_1, ..., \mathbf{u}_n)$, is the inner product between $\mathbf{x}$ and $\mathbf{u}_i$.

$$x = \sum_{i=1}^{n} \langle x, u_i \rangle u_i$$

$$\sum_{i=1}^{n} \langle x, u_i \rangle u_i = \sum_{i=1}^{n} \left\langle \sum_{j=1}^{n} a_j u_j, u_i \right\rangle u_i =$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} a_j \langle u_j, u_i \rangle u_i =$$

$$= \sum_{i=1}^{n} a_i \|u_i\|^2 \cdot u_i = \sum_{i=1}^{n} a_i \cdot u_i = x$$

$\langle u_i, u_j \rangle = 0$ if $i \neq j$

$\|u_i\|^2 = 1$

2.2 Estimation Theory

Based on Lecture 1

12. Let $x_1, x_2, \ldots \overset{iid}{\sim} \mathcal{P}$ be a sample of infinity size drawn from some probability distribution function $\mathcal{P}$ with finite expectation and variance. Show that the sample mean estimator $\hat{\mu}_n = \frac{1}{n}\sum x_i$ calculated over the first n samples is a consistent estimator. Hint: for any given fixed value of $n \in \mathbb{N}$ bound from above the probability of deviating more than ε.

$$\mathbb{E}\hat{N}_n = \frac{1}{n}\sum \mathbb{E}X_i = N$$

$$P\left(|\hat{N}_n - N| \geq \varepsilon\right) \leq \frac{Var(\hat{N}_n)}{\varepsilon^2} =$$

$$= \frac{Var\left(\frac{1}{n}\sum X_i\right)}{\varepsilon^2} = \frac{1}{\varepsilon^2}\cdot\frac{1}{n^2}\cdot\sum_{i=1}^{n} Var(X_i) =$$

$$= \frac{1}{\varepsilon^2}\cdot\frac{1}{n^2}\cdot n \cdot Var(X_n) = \frac{Var(X_n)}{\varepsilon^2 \cdot n} \xrightarrow{n\to\infty} 0$$

13. Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$ be m observations sampled i.i.d from a multivariate Gaussian with expectation of $\mu \in \mathbb{R}^d$ and a covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. Provide an expression for the log-likelihood function of $\mathcal{N}(\mu, \Sigma)$. Develop the expression as much as you can. Hint: follow the approach used to derive the likelihood function for the univariate case.

$$\theta = (\hat{\mu}, \hat{\Sigma}) \in \mathbb{R}^d \times \mathbb{R}^{d \times d}_{\succeq 0}$$

$$\mathcal{L}(\theta | x_1, \ldots, x_m) = f_\theta(x_1, \ldots, x_m) =$$

$$\overset{x_1 \ldots x_m \text{ בלתי}}{\downarrow} = \prod_{i=1}^{m} f_\theta(x_i) = \prod_{i=1}^{m} \frac{1}{\sqrt{(2\pi)^d \cdot |\hat{\Sigma}|}} \cdot \exp\left(-\frac{1}{2}(x_i - \hat{\mu})^T \cdot \hat{\Sigma}^{-1} \cdot (x_i - \hat{\mu})\right) =$$

$$= \frac{1}{((2\pi)^d \cdot |\hat{\Sigma}|)^{\frac{m}{2}}} \cdot \exp\left\{\sum_{i=1}^{m} -\frac{1}{2}(x_i - \hat{\mu})^T \cdot \hat{\Sigma}^{-1} \cdot (x_i - \hat{\mu})\right\}$$

נלקח לוג משני הצדדים ונקבל:

$$\mathcal{L}^{\log}(\theta | x_1, \ldots, x_m) = \sum_{i=1}^{m} -\frac{1}{2}(x_i - \hat{\mu})^T \cdot \hat{\Sigma}^{-1} \cdot (x_i - \hat{\mu}) - \frac{m}{2} \cdot \log\left((2\pi)^d \cdot |\hat{\Sigma}|\right)$$

$\blacksquare$

5. Let $x \in \mathbb{R}^n$ be a fixed vector and $U \in \mathbb{R}^{n \times n}$ a fixed orthogonal matrix. Calculate the Jacobian of the function $f : \mathbb{R}^n \to \mathbb{R}^n$:

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

Where $\operatorname{diag}(\sigma)$ is an $n \times n$ matrix where

$$\operatorname{diag}(\sigma)_{ij} = \begin{cases} \sigma_i & i = j \\ 0 & i \neq j \end{cases}$$

$$\text{כלומר} \quad U = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{bmatrix}$$

$$f(\sigma) = U \cdot \operatorname{diag}(\sigma) \cdot U^T \cdot x = \begin{bmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \cdot \begin{bmatrix} — & u_1^T & — \\ & \vdots & \\ — & u_n^T & — \end{bmatrix} \cdot x =$$

$$= \begin{bmatrix} | & & | \\ \sigma_1 u_1 & \cdots & \sigma_n u_n \\ | & & | \end{bmatrix} \cdot \begin{bmatrix} \langle u_1, x \rangle \\ \vdots \\ \langle u_n, x \rangle \end{bmatrix} = \langle u_1, x \rangle \cdot \sigma_1 u_1 + \ldots + \langle u_n, x \rangle \cdot \sigma_n u_n$$

$$f_j(\sigma) = \sigma_1 \cdot u_1^j \cdot \langle u_1, x \rangle + \ldots + \sigma_n \cdot u_n^j \cdot \langle u_n, x \rangle \quad \Leftarrow$$

$$\frac{\partial f_j}{\partial x_i}(\sigma) = u_i^j \cdot \langle u_i, x \rangle \quad \Leftarrow$$

$$Df|_\sigma = \begin{bmatrix} u_1^1 \cdot \langle u_1, x \rangle & \cdots & u_n^1 \cdot \langle u_n, x \rangle \\ \vdots & \ddots & \vdots \\ u_1^n \langle u_1, x \rangle & \cdots & u_n^n \cdot \langle u_n, x \rangle \end{bmatrix} = \quad \text{וזה}$$

$$= \begin{bmatrix} | & & | \\ \langle u_1, x \rangle \cdot u_1 & \cdots & \langle u_n, x \rangle \cdot u_n \\ | & & | \end{bmatrix} =$$

$$= \begin{bmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{bmatrix} \cdot \begin{bmatrix} \langle u_1, x \rangle & & 0 \\ & \ddots & \\ 0 & & \langle u_n, x \rangle \end{bmatrix} =$$

$$= U \cdot \operatorname{diag}\left(\begin{bmatrix} \langle u_1, x \rangle \\ \vdots \\ \langle u_n, x \rangle \end{bmatrix} \right) = U \cdot \operatorname{diag}([x]_{B_U})$$

6. Use the chain rule to calculate the gradient of $h(\sigma) = \frac{1}{2}||f(\sigma) - y||^2$

$g_2(\sigma) = f(\sigma) - y$, $g_1(\sigma) = \frac{1}{2}\|\sigma\|^2$

$g_2 : \mathbb{R}^n \to \mathbb{R}^n$, $g_1 : \mathbb{R}^n \to \mathbb{R}$

$h(\sigma) = (g_1 \circ g_2)(\sigma)$

$$Dh_{|\sigma} = Dg_{1|g_2(\sigma)} \cdot Dg_{2|\sigma}$$

$$Dg_{1|\sigma} = 2 \cdot \frac{1}{2} \cdot \sigma^T = \sigma^T$$

$$Dg_{2|\sigma} = U \cdot \operatorname{diag}\left(\begin{bmatrix} \langle u_1, x \rangle \\ \vdots \\ \langle u_n, x \rangle \end{bmatrix}\right)$$

$$Dh_{|\sigma} = (f(\sigma) - y)^T \cdot U \cdot \operatorname{diag}\left(\begin{bmatrix} \langle u_1, x \rangle \\ \vdots \\ \langle u_n, x \rangle \end{bmatrix}\right)$$

$$\nabla h(\sigma) = \operatorname{diag}\left(\begin{bmatrix} \langle u_1, x \rangle \\ \vdots \\ \langle u_n, x \rangle \end{bmatrix}\right) \cdot U^T \cdot (f(\sigma) - y) \quad <=$$

7. Calculate the Jacobian of the softmax function $S : \mathbb{R}^d \to [0,1]^k$

$$S(\mathbf{x})_j = \frac{e^{x_j}}{\sum_{l=1}^{k} e^{x_l}}$$

נגדיר $f_i : \mathbb{R}^d \to \mathbb{R}$ ע"י $f_i(x) = e^{x_i}$, ונסמן $h(x) = \sum_{l=1}^{d} f_l(x)$

אם $i=j$ אז הראינו (בתרגול) ש:

$$\frac{\partial S_j}{\partial x_i}(x) = S_j(x)\left(1 - S_i(x)\right)$$

אם $i \neq j$ אז

$$\frac{\partial S_j}{\partial x_i}(x) = \frac{\partial}{\partial x_i}\frac{f_j(x)}{h(x)} = \frac{\overset{0 \ (i \neq j)}{\cancel{\frac{\partial f_j(x)}{\partial x_i}}} \cdot h(x) - \frac{\partial h(x)}{\partial x_i} \cdot f_j(x)}{h^2(x)}$$

$$= \frac{-f_i(x) \cdot f_j(x)}{h^2(x)} = -\frac{f_i(x)}{h(x)} \cdot \frac{f_j(x)}{h(x)} = -S_i(x) \cdot S_j(x)$$

לכן סה"כ

$$J_x(S) = \begin{bmatrix} S_1(x)(1 - S_1(x)) & \cdots & -S_d(x) \cdot S_1(x) \\ -S_1(x) \cdot S_2(x) & \ddots & \vdots \\ \vdots & & -S_d(x) \cdot S_{d-1}(x) \\ -S_1(x) \cdot S_d(x) & \cdots & S_d(x)(1 - S_d(x)) \end{bmatrix}$$

8. Let $f:\mathbb{R}^d\to\mathbb{R}$ be defined as $f(x,y)=x^3-5xy-y^5$. Calculate the Hessian of f.

$$J_{(x,y)}(f)=\begin{bmatrix}3x^2-5y & -5x-5y^4\end{bmatrix}$$ נגזרת ראשונה:

$$H_{(x,y)}(f)=\begin{bmatrix}6x & -5\\ -5 & -20y^3\end{bmatrix}$$ נגזרת שניה:

■

2.1.3 convexity

Based on Recitation 2

9. Prove that the intersection $C:=\bigcap_{i\in I}C_i$ for $\{C_i : i\in I\}$ a collection of convex sets is convex.
10. Prove that the vector sum $C_1+C_2:=\{c_1+c_2 : c_1\in C_1, c_2\in C_2\}$ of two convex sets is convex.
11. Prove that the set $\lambda C:=\{\lambda c : c\in C\}$ is convex, for any convex set C, and every scalar λ.

9. יהיו $x,y\in C$ ו- $t\in[0,1]$ אזי לכל $i\in I$

$x,y\in C_i$ קמור ולכן $tx+(1-t)y\in C_i$ וזה נכון לכל

$i\in I$ לכן $tx+(1-t)y\in C$ ■

10. יהיו $x,y\in C$ אז $x=x_1+x_2$ ו- $y=y_1+y_2$

ו- $x_1,y_1\in C_1$ ו- $x_2,y_2\in C_2$.

ויהי $t\in[0,1]$

$$= t x_1 + t x_2 + (1-t) y_1 + (1-t) y_2 = (t x + (1-t) y$$

$$= \underset{\text{(קמור)}}{C_1} + \underset{\text{(קמור)}}{C_2}$$

$\Leftarrow$ כיוון $\blacksquare$ $t x + (1-t) y \in C$

הוכחה: יהי $\lambda \in \mathbb{R}$ כך ש $\lambda x = \lambda x'$! $\lambda y = \lambda y'$

כלומר $x, y \in C$ כך ש $x', y' \in C$.

עבור $t \in [0,1]$ כך ש

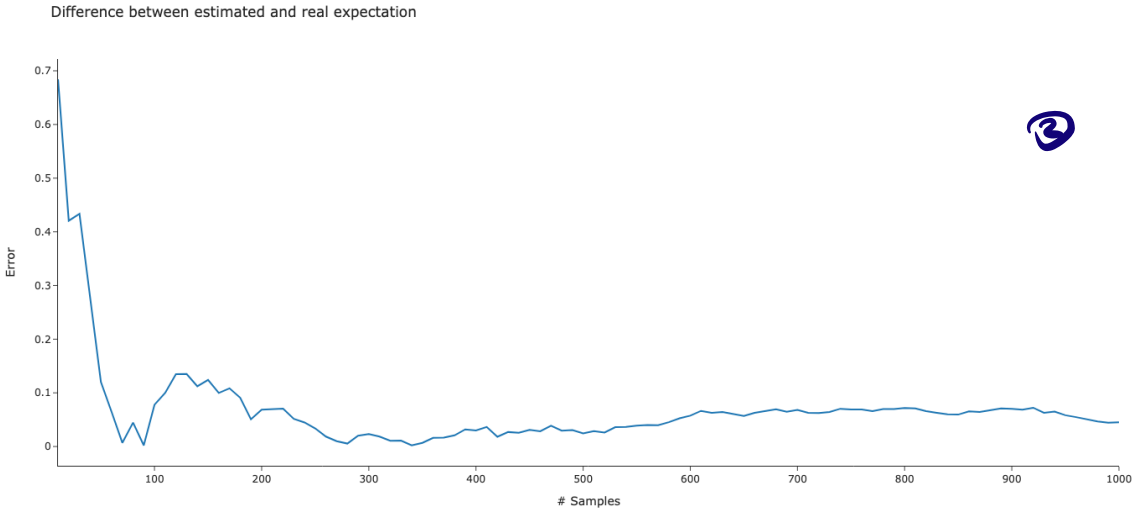
$$= \lambda (t x' + (1-t) y') = \underset{\text{(קמור)}}{\lambda C}$$

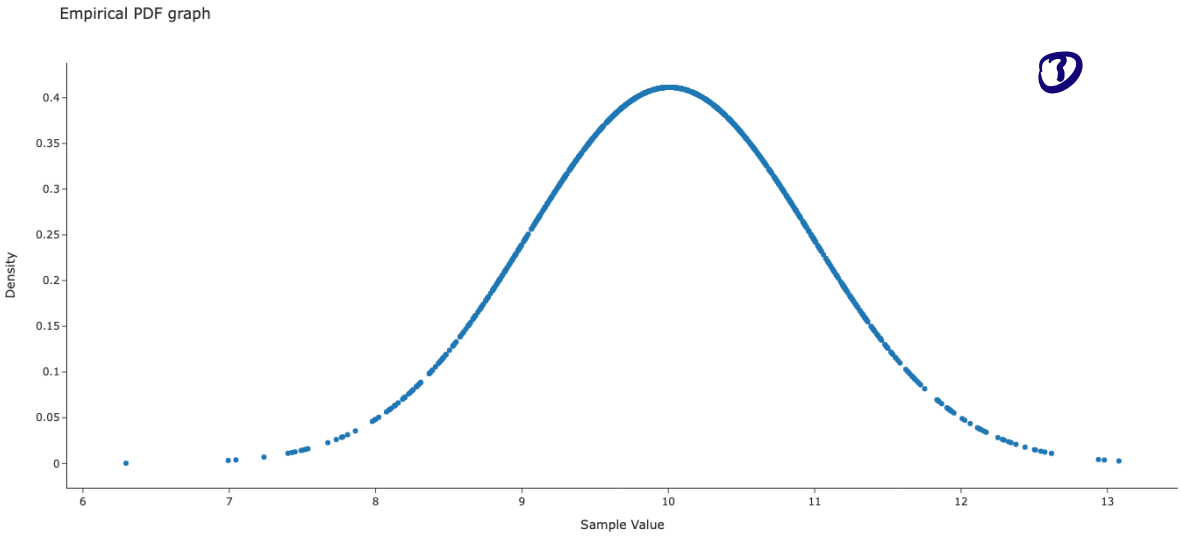
$$= \lambda (t x + (1-t) y) + \lambda (t-1) x' + \lambda (1-t) y' = \lambda (t x + (1-t) y)$$

$\Rightarrow$ $\lambda x + (1-t) y \in \lambda C$

כלומר $\blacksquare$

1. Using `numpy.random.normal` draw 1000 samples $x_1,\ldots,x_{1000} \overset{iid}{\sim} \mathcal{N}(10,1)$ and fit a univariate Gaussian. Print the estimated expectation and variance. Output format should be `(expectation, variance)`.
2. Over previously drawn samples, fit a series of models of increasing samples size: 10, 20,...,100, 110,...1000. Plot the absolute distance between the estimated- and true value of the expectation, as a function of the sample size. Provide meaningful axis names and title.
3. Compute the PDF of the previously drawn samples using the model fitted in question 1. Plot the empirical PDF function under the fitted model. That is, create a scatter plot with the ordered sample values along the x-axis and their PDFs (using the `UnivariateGaussian.pdf` function) along the y-axis. Provide meaningful axis names and title. What are you expecting to see in the plot?





כמו שציפינו, יש יותר נקודות פיזור הנמצא על הישר
הממוצע ופחות נקודות ככל שמתרחקים.

4. Using `numpy.random.multivariate_normal` draw 1000 samples $\mathbf{x}_1, \ldots, \mathbf{x}_{1000} \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$

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

   Fit a multivariate Gaussian and print the estimated expectation and covariance matrix. Print each in a separate line.
5. Using the samples drawn in the question above calculate the log-likelihood for models with expectation $\mu = [f_1, 0, f_3, 0]^\top$ and the true covariance matrix defined above, where f_1, f_3 get values returned from `np.linspace(-10, 10, 200)`. Plot a heatmap of $f1$ values as rows, f_3 values as columns and the color being the calculated log likelihood. Provide meaningful axis names and title. What are you able to learn from the plot?
6. Of all values tested in question 5, which model (pair of values for feature 1 and 3) achieved the maximum log-likelihood value? Round to 3 decimal places