קורס: מבוא למערכות מידע 67577

תרגיל: תרגיל 1

מגישה: שלי רס

ת"ז: 211975453

Introduction to Machine Learning (67577)

Exercise 1
Estimation Theory & Mathematical Background

Second Semester, 2023

Contents

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 $f : \mathbb{R}^n \to \mathbb{R}^n$:

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

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

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

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 $C \coloneqq \bigcap_{i \in I} C_i$ for $\{C_i : i \in I\}$ a collection of convex sets is convex.
10. Prove that the vector sum $C_1 + C_2 \coloneqq \{c_1 + c_2 : c_1 \in C_1, c_2 \in C_2\}$ of two convex sets is convex.
11. Prove that the set $\lambda C \coloneqq \{\lambda c : c \in C\}$ 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* (find the definition in the course book, page 14, Definition 1.1.10 under "Consistency"). 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

חלק תאורטי

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 $\forall x \in V \ \ ||Ax|| = ||x||$.

צ"ל: מטריצה אורתוגונלית היא איזומטרית, כלומר אם $T:V \to W$ הנ"ל ו-A
המטריצה המתאימה לה, אזי אם A אורתוגונלית מתקיים: $||Ax|| = ||x|| \ \ \forall x \in V$.

הוכחה:

$$||Ax||^2 \underset{\text{הגדרת נורמה}}{=} \langle Ax | Ax \rangle = (Ax)^T \cdot (Ax) \underset{T \text{ תכונות}}{=} x^T A^T \cdot (Ax) \underset{\text{כי } A \text{ אורתוגונלית}}{=} x^T \cdot I \cdot x = x^T \cdot x = \langle x | x \rangle = ||x||^2$$

קיבלנו כי $||Ax||^2 = ||x||^2$, מכאן $||Ax|| = ||x|| \ \ \forall x \in V$ כנדרש.

∎

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.

2.

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

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 $AV = U\Sigma$.

נמצא את ה- SVD של A:

נתחיל מכך שנמצא את ה- EVD של A^TA:

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

נמצא את הפ"א של A^TA:

$$\det(A^TA - \lambda I) = \det\left(\begin{bmatrix} 2-\lambda & 0 & 2 \\ 0 & 2-\lambda & -2 \\ 2 & -2 & 4-\lambda \end{bmatrix}\right) = -\lambda^3 + 8\lambda^2 - 12\lambda = 0 \iff \lambda_1 = 0,\ \lambda_2 = 2,\ \lambda_3 = 6$$

כעת נמצא את הו"ע:

נמצא צ"ל וקטורים $V = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$ שמ"מ: $(A^TA - \lambda I)\cdot V = 0$

- עבור $\lambda = 0$:

$$(A^TA - \lambda I) = (A^TA - 0 \cdot I) = A^TA = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 2v_1 + 2v_3 \\ 2v_2 - 2v_3 \\ 2v_1 - 2v_2 + 4v_3 \end{pmatrix} = 0 \iff \begin{bmatrix} v_1 = -v_3 \\ v_2 = v_3 \end{bmatrix} = v_1 = -v_2$$

נקח הוקטור $\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$ הינו וע שמ"ל ל- $\lambda = 0$.

- עבור $\lambda = 2$:

$$(A^TA - \lambda I) = (A^TA - 2 \cdot I) = \begin{bmatrix} 2-2 & 0 & 2 \\ 0 & 2-2 & -2 \\ 2 & -2 & 4-2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & -2 \\ 2 & -2 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & -2 \\ 2 & -2 & 2 \end{bmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} 2v_3 \\ -2v_3 \\ 2v_1 - 2v_2 + 2v_3 \end{pmatrix} = 0 \iff \begin{bmatrix} v_3 = 0 \\ 2v_1 - 2v_2 + 2v_3 = 0 \end{bmatrix} \iff \begin{bmatrix} v_3 = 0 \\ v_1 = v_2 \end{bmatrix}$$

המשך שאלה 2

ולכן הוקטור $\begin{pmatrix}1\\1\\0\end{pmatrix}$ הינו ו"ע ששייך ל- $\lambda=2$.

- עבור $\lambda=6$:

$$(A^TA-\lambda I)=(A^TA-6\cdot I)=\begin{bmatrix}2-6 & 0 & 2\\ 0 & 2-6 & -2\\ 2 & -2 & 4-6\end{bmatrix}=\begin{bmatrix}-4 & 0 & 2\\ 0 & -4 & -2\\ 2 & -2 & -2\end{bmatrix}$$

$$\begin{bmatrix}-4 & 0 & 2\\ 0 & -4 & -2\\ 2 & -2 & -2\end{bmatrix}\begin{pmatrix}v_1\\v_2\\v_3\end{pmatrix}=\begin{pmatrix}-4v_1+2v_3\\-4v_2-2v_3\\2v_1-2v_2-2v_3\end{pmatrix}=0 \iff \begin{cases}v_1=-v_2\\2v_1-2v_2-2v_3=0\end{cases}$$

ולכן הוקטור $\begin{pmatrix}1\\-1\\2\end{pmatrix}$ הינו ו"ע ששייך ל- $\lambda=6$.

ננרמל את הוקטורים ונקבל: $v_1=\begin{pmatrix}1/\sqrt6\\-1/\sqrt6\\2/\sqrt6\end{pmatrix}$, $v_2=\begin{pmatrix}1/\sqrt2\\1/\sqrt2\\0\end{pmatrix}$, $v_3=\begin{pmatrix}1/\sqrt3\\-1/\sqrt3\\1/\sqrt3\end{pmatrix}$.

מכאן: $V=\begin{bmatrix}1/\sqrt6 & 1/\sqrt2 & 1/\sqrt3\\ -1/\sqrt6 & 1/\sqrt2 & -1/\sqrt3\\ 2/\sqrt6 & 0 & 1/\sqrt3\end{bmatrix}$, $D=\begin{bmatrix}6 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 0\end{bmatrix}$ ונקבל $A^TA=VDV^T$.

כעת ראינו בתרגול שניתן לחשב את המטריצה U של ה- SVD ע"י הנוסחה: $u_i=\frac{Av_i}{\sqrt{\lambda_i}}$

נחשב:

$$u_1=\frac{Av_1}{\sqrt{\lambda_1}}=\frac{1}{\sqrt6}\cdot\begin{pmatrix}0\\6/\sqrt6\end{pmatrix}=\begin{pmatrix}0\\1\end{pmatrix}$$

$$u_2=\frac{Av_2}{\sqrt{\lambda_2}}=\frac{1}{\sqrt2}\cdot\begin{pmatrix}2/\sqrt2\\0\end{pmatrix}=\begin{pmatrix}1\\0\end{pmatrix}$$

(התמונה ממימד 2 ולכן אין u_3)

מכאן, $U=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}$ ו- $\Sigma=\begin{bmatrix}\sqrt6 & 0 & 0\\ 0 & \sqrt2 & 0\end{bmatrix}$ ולבסוף נקבל $A=U\Sigma V^T$

כלומר, ה- SVD של A כנדרש.

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.

3.

3. כי המכפלה החיצונית של הוקטורים $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{u} \in \mathbb{R}^m$ היא מטריצה $A \in \mathbb{R}^{n \times m}$ כך ש-

$rank(A) = 1$.

נוכיח כי כל העמודות של A הינן תלויות לינ׳.

$$A = \mathbf{v} \otimes \mathbf{u} = \begin{bmatrix} v_1 u_1 & v_1 u_2 & \cdots & v_1 u_m \\ \vdots & \vdots & \ddots & \vdots \\ v_n u_1 & v_n u_2 & \cdots & v_n u_m \end{bmatrix}, \quad [\mathbf{v} \otimes \mathbf{u}]_{ij} = v_i u_j$$

יהי $v_i u_j$ עמודת של A שהיא כך ש- $j \neq i$, נראה כי היא תלויה לינ׳:

לפי ההגדרה של מכפלה חיצונית אזי"ס $\exists k$: $u_i \cdot v_k = x(v_i)$, $u_k v_k = x(v_j)$.

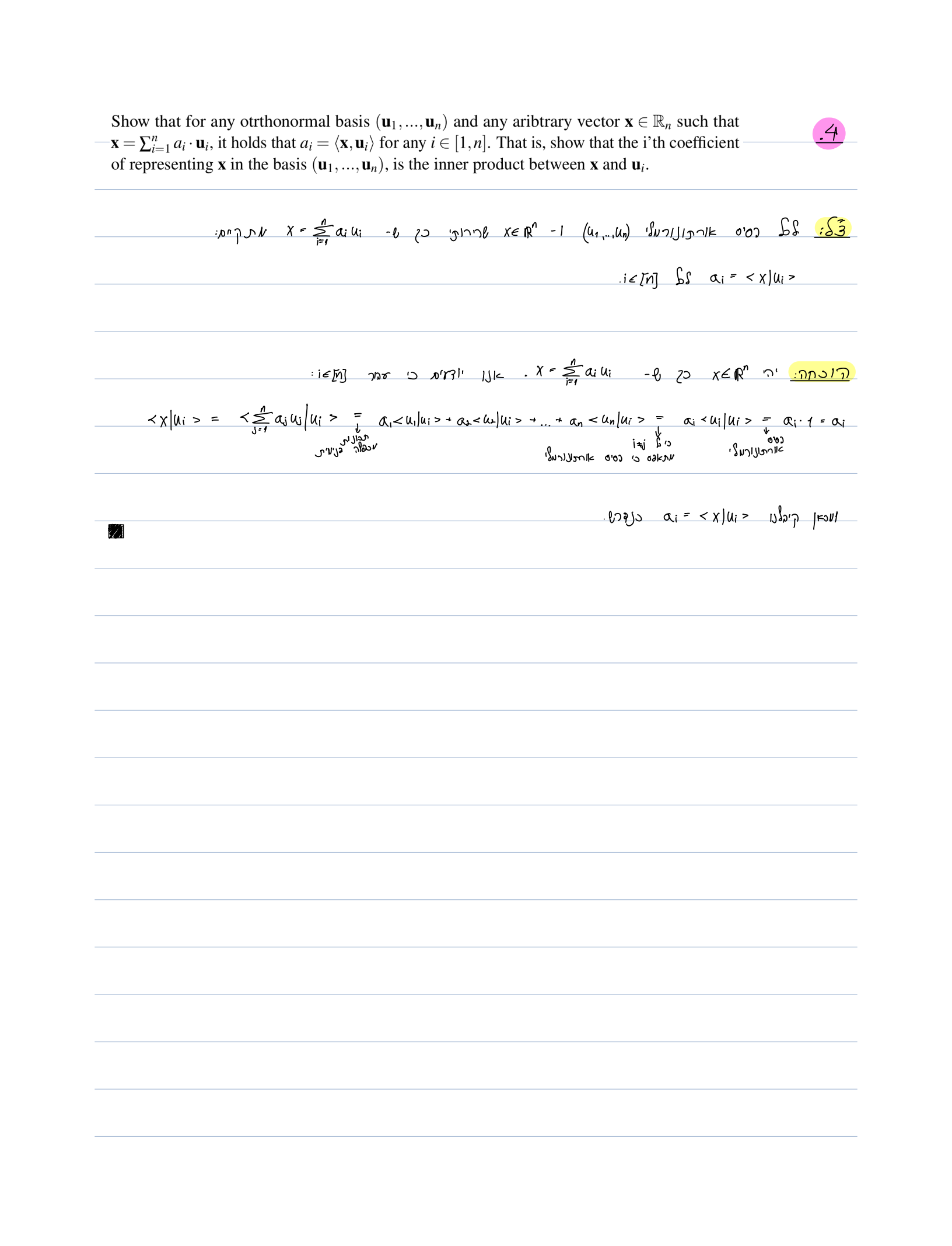
מכאן קיבלנו שאזי"ס $u_i \cdot v = v_i$, $u_k \cdot v = v_k$ לפי קיבלנו כי v_i ו- v_k תלויים לינ׳,

לפיכך $rank(A) = 1$ כנדרש.

■

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 $\mathbf{u}_i$.

7.



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 $f : \mathbb{R}^n \to \mathbb{R}^n$:

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

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

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

לפי הנוסחה מתרגול 2: $(J_x(f))_{i,j} = \frac{\partial f_i(x)}{\partial x_j}$

הראינו בתרגול 1 את השוויון $(UDU^T)x = \sum_{k=1}^{n} \lambda_k \langle x|u_k\rangle u_k$

כש $(u_1,\dots,u_n)$ הינו בסיס אורתונורמלי למטריצה U השייכות לע״ע $\lambda_1,\dots,\lambda_n$ בהתאמה ו- $D=\mathrm{diag}(\sigma)$

כלומר $f(\sigma) = (U \cdot \mathrm{diag}(\sigma) U^T)x = \sum_{k=1}^{n} \lambda_k \langle x|u_k\rangle u_k$. בנוסף משוואה זו מתקיימת כשהערכים

על האלכסון של $D=\mathrm{diag}(\sigma)$ הינם הע״ע $\lambda_1,\dots,\lambda_n$. לכן לפי ההגדרה הנתונה של $\mathrm{diag}(\sigma)$ מתקיים:

$$f(\sigma) = (U \cdot \mathrm{diag}(\sigma) U^T)x = \sum_{k=1}^{n} \sigma_k \langle x|u_k\rangle u_k.$$

מכאן מתקיים:

$$(J_\sigma(f))_{i,j} = \frac{\partial f_i(\sigma)}{\partial \sigma_j} = \frac{\partial \left(\sum_{k=1}^{n} \sigma_k \langle x|u_k\rangle u_k\right)_i}{\partial \sigma_j} = [\langle x|u_j\rangle u_j]_i = [(x^T \cdot u_j)\cdot u_j]_i$$

כש $k \neq j$ אזי הנגזרת היא 0 בגלל שמתייחסים כקבועים, אחרת ניקח את המקדם של σ_j

ומכאן נובע כי $J_\sigma(f) = U \cdot \mathrm{diag}(U^T x)$.

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

אנו יודעים כי $f(\sigma)=(U\cdot diag(\sigma)\,V^t)\,x$ כש- $f:\mathbb{R}^n\to\mathbb{R}^n$

נגדיר $T(\sigma)=f(\sigma)-y$, $S(x)=\frac{1}{2}\|x\|^2$.

מכאן: $h(\sigma)=(S\circ T)(\sigma)=S(f(\sigma)-y)=\frac{1}{2}\|f(\sigma)-y\|^2$

לפי כלל השרשרת עם יעקוביאן מתקיים: $J_x(S\circ T)_{i,j}=\frac{\partial}{\partial g_k(x)}S_i(T(x))\cdot\frac{\partial T_k(x)}{\partial x_j}$ כלומר:

$J_x(S\circ T)=J_{T(x)}(S)\cdot J_x(T)$.

נחשב את הנגזרות:

ראינו בתרגול כי הנגזרת של $\|x\|^2$ הינה $2x^t$, ולכן $S'(x)=\left(\frac{1}{2}\|x\|^2\right)'=\frac{1}{2}\cdot 2x^t=x^t$

כלומר $J_{T(x)}(S)=(T(x))^t$.

בשאלה 5 הראנו כי $J_\sigma(f)=U\cdot diag(V^tx)$, ולכן:

$J_x(T)=T'(\sigma)=(f(\sigma)-y)'=f'(\sigma)=U\cdot diag(V^tx)$

(y קבוע)

מכאן: $J_\sigma(h)=J_\sigma(S\circ T)=J_{T(\sigma)}(S)\cdot J_\sigma(T)=(T(x))^t\cdot U\cdot diag(V^tx)$

h הינה נורמה ולכן מחזירה ערך ב- $\mathbb{R}$, מכאן מתקיים $\nabla h(\sigma)=(J_\sigma(h))^t$.

$\nabla h(\sigma)=(J_\sigma(h))^t=\left((T(x))^t\cdot U\cdot diag(V^tx)\right)^t=(U\cdot diag(V^tx))^t\cdot T(x)=(J_\sigma(f))^t\cdot f(\sigma)-y$

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

.7

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

← אמור להיות $S:\mathbb{R}^d \to [0,1]^d$

$$S(x)_j = \frac{e^{x_j}}{\sum_{k=1}^{d} e^{x_k}}$$

נגדיר: $g_i(x) = e^{x_i}$, $h(x) = \sum_{k=1}^{d} g_k(x)$ ונקבל כי $S(x)_j = \frac{g_j(x)}{h(x)}$.

הראינו בתרגול כי כש $i=j$ מתקיים: $\frac{\partial}{\partial x_j} \frac{e^{x_i}}{\sum_{k=1}^{d} e^{x_k}} = S_i(1-S_j)$

נחשב את המקרה ש- $i \neq j$:

$$(J_x(S))_{i,j} = \frac{\partial}{\partial x_j} S_i = \frac{\partial}{\partial x_j} \frac{e^{x_i}}{\sum_{k=1}^{d} e^{x_k}} \underset{\text{לפי איך שהגדרנו}}{=} \frac{\partial}{\partial x_j} \frac{g_i}{h} \underset{\text{נגזרת של מנה}}{=} \frac{g' \cdot h - h' \cdot g}{h^2}$$

נחשב את הנגזרות של g ו- h כש $i \neq j$:

$$\frac{\partial}{\partial x_j} g_i = \frac{\partial}{\partial x_j} e^{x_i} \underset{\text{כי } i \neq j}{=} 0$$

$$\frac{\partial}{\partial x_j} h = \frac{\partial}{\partial x_j} \sum_{k=1}^{d} g_k(X) = e^{x_j} \longrightarrow \text{לכל } i$$

מכאן:

$$(J_x(S))_{i,j} = \frac{g_i' \cdot h - h' \cdot g_i}{h^2} = \frac{0 \cdot h - e^{x_j} \cdot g_i}{h^2} = \frac{-e^{x_j} \cdot e^{x_i}}{\left(\sum_{k=1}^{d} g_k(X)\right)^2} = \frac{-e^{x_j}}{\sum_{k=1}^{d} g_k(X)} \cdot \frac{e^{x_i}}{\sum_{k=1}^{d} g_k(X)} = -S_j \cdot S_i$$

ומשם נקבל:

$$(J_x(S))_{i,j} = \begin{cases} S_i(1-S_j) & i = j \\ -S_j \cdot S_i & i \neq j \end{cases}$$

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

Prove that the intersection $C := \bigcap_{i \in I} C_i$ for $\{C_i : i \in I\}$ a collection of convex sets is convex.

6.

הוכחה: יהי $x, y \in C = \bigcap_{i \in I} C_i$ ויהי $\lambda \in [0,1]$ מתקיים $\lambda x + (1-\lambda) y \in C$.

יהי $i \in I$ כלשהו, אז מהגדרת החיתוך, $x, y \in C_i$ מתקיים $\lambda x + (1-\lambda) y \in C_i$.

כלומר C קמורה כנדרש.

■

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. 10.

צ"ל: יהיו $v, u \in C_1 + C_2$ ויהי $\lambda \in [0,1]$ מתקיים $\lambda u + (1-\lambda) v \in C_1 + C_2$.

נזכור כי C_1 ו- C_2 קמורות.

יהיו $u, v \in C_1 + C_2$, אזי קי"א $u_1, u_2 \in C_1$ ו- $v_1, v_2 \in C_2$ כך ש- $u = u_1 + u_2$, $v = v_1 + v_2$.

יהי $\lambda \in [0,1]$ מתקיים:

$$\lambda u + (1-\lambda) v = \lambda(u_1 + u_2) + (1-\lambda)(v_1 + v_2) = \lambda v_2 + \lambda v_1 + (1-\lambda) u_1 + (1-\lambda) u_2 = \lambda v_2 + (1-\lambda) u_1 + \lambda v_1 + (1-\lambda) u_2$$

אך יודע כי $\lambda v_2 + (1-\lambda) u_1 \in C_2$, $\lambda v_1 + (1-\lambda) u_2 \in C_1$ כי C_2 ו- C_1 קמורות.

מכאן, $\forall \lambda \in [0,1]$: $\lambda u + (1-\lambda) v = \lambda v_1 + (1-\lambda) u_1 + \lambda v_2 + (1-\lambda) u_2 \in C_1 + C_2$

מ.ש.ל, $C_1 + C_2$ קמורה בהכרח.

■

Prove that the set $\lambda C := \{\lambda c : c \in C\}$ is convex, for any convex set C, and every scalar λ. 11.

דוגמה: יהי $\lambda \in \mathbb{R}$ ויהי $\alpha \in [0,1]$ ויהיו $\gamma, \eta \in \mathbb{R}$ כך ש- $\gamma \cdot \lambda + (1-\alpha)\cdot \eta$ מתקיים $\gamma \cdot \lambda + (1-\alpha)\cdot \eta$.

הוכחה: יהי $\alpha \in [0,1]$ מתקיים:

יהיו $\lambda \in \mathbb{R}$, $\bar{\lambda}, \bar{\eta} \in C$ כך ש- $\lambda = \bar{\lambda}\cdot\lambda$, $\eta = \bar{\eta}\lambda$. מכאן שהקבוצה C קמורה.

$$\gamma\lambda \in (\bar{\eta}\cdot(\alpha-1) + \bar{\lambda}\alpha)\lambda = \bar{\eta}\lambda\cdot(\alpha-1) + \bar{\lambda}\lambda\cdot\alpha = \eta\cdot(\alpha-1) + \lambda\cdot\alpha$$

$$\downarrow$$

$$\gamma \ni \bar{\eta}\cdot(\alpha-1) + \bar{\lambda}\cdot\alpha$$

כי C קמורה.

לפיכך $\gamma\lambda$ הוא קמור כלשהו.

הוכחה: יהי $\alpha \in [0,1]$ מתקיים $\gamma\lambda \ni \eta\cdot(\alpha-1) + \lambda\cdot\alpha$.

■

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* (find the definition in the course book, page 14, Definition 1.1.10 under "Consistency"). Hint: for any given fixed value of $n \in \mathbb{N}$ bound from above the probability of deviating more than ε. **.12**

צ"ל: לכל $0<\varepsilon$ $\lim_{n\to\infty} P(|\hat{\mu}_n - \mu| > \varepsilon) = 0$.

נסמן $var(x_i) = \sigma^2$. לפי ההגדרה $\hat{\mu}_n$ הינו אומדן בלתי מוטה (unbiased estimator)

ולכן $E(\hat{\mu}_n) = \mu$. מכאן $|\hat{\mu}_n - \mu| = |\hat{\mu}_n - E(\hat{\mu}_n)|$ ולכן נשתמש באי שוויון צ'בישב

לחסום מלמעלה:
$$P(|\hat{\mu}_n - \mu| > \varepsilon) = P(|\hat{\mu}_n - E(\hat{\mu}_n)| > \varepsilon) \leq \frac{Var(\hat{\mu}_n)}{\varepsilon^2}$$
(אי"ש צ'בישב)

הראנו בספר כי מתקיים $Var(\hat{\mu}_n) = \frac{\sigma^2}{n}$ ולכן $P(|\hat{\mu}_n - \mu| > \varepsilon) \leq \frac{Var(\hat{\mu}_n)}{\varepsilon^2} = \frac{\frac{\sigma^2}{n}}{\varepsilon^2}$

כלומר: $P(|\hat{\mu}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2 \cdot n}$.

מתקיים $\frac{\sigma^2}{\varepsilon^2 \cdot n} \xrightarrow[n\to\infty]{} 0$,
$$0 \leftarrow 0 \leq P(|\hat{\mu}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2 \cdot n} \xrightarrow[n\to\infty]{} 0$$
(תכונות של הסתברות)

ומכאן מסנדוויץ' $\lim_{n\to\infty} P(|\hat{\mu}_n - \mu| > \varepsilon) = 0$ כנדרש.

∎

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.

13.

נתון: $x_1, \ldots, x_m \overset{i.i.d}{\sim} N(\mu, \Sigma)$ (בגלל זהות ההתפלגות).

נמצא את ה- log-likelihood של $N(\mu, \Sigma)$. נסמן $\theta = (\mu, \Sigma)$.

נזכר בפונקציית הצפיפות של $N(\mu, \Sigma)$. אז:

$$\mathcal{L}(\theta \mid x_1, \ldots, x_m) = \underbrace{f_\theta(x_1, \ldots, x_m)}_{\text{הצפיפות המשותפת}} = \underbrace{\prod_{i=1}^m f_\theta(x_i)}_{\text{כי בלתי תלויים}} = \underbrace{\prod_{i=1}^m \frac{1}{\sqrt{(2\pi)^d \cdot |\Sigma|}} \exp\left(-\frac{1}{2}(x_i - \mu)^t \cdot \Sigma^{-1}(x_i - \mu)\right)}_{\text{נוסחה מעמוד 23}} =$$

$$= \underbrace{\left(\frac{1}{\sqrt{(2\pi)^d \cdot |\Sigma|}}\right)^m \cdot \exp\left(\sum_{i=1}^m \left(-\frac{1}{2}(x_i - \mu)^t \cdot \Sigma^{-1}(x_i - \mu)\right)\right)}_{\exp(x) \cdot \exp(y) = \exp(x+y)} =$$

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

כדי למצוא את ה- log-likelihood function של $N(\mu, \Sigma)$ (likelihood - function), נסיק log על פונקציה זו:

$$\log\left(\left((2\pi)^d \cdot |\Sigma|\right)^{-\frac{m}{2}} \cdot \exp\left(\sum_{i=1}^m \left(-\frac{1}{2}(x_i - \mu)^t \cdot \Sigma^{-1}(x_i - \mu)\right)\right)\right) \overset{\text{תכונות log}}{=} \log\left((2\pi)^{-\frac{dm}{2}} \cdot |\Sigma|^{-\frac{m}{2}}\right) + \log\left(\exp\left(\sum_{i=1}^m \left(-\frac{1}{2}(x_i - \mu)^t \cdot \Sigma^{-1}(x_i - \mu)\right)\right)\right) =$$

$$= \log\left((2\pi)^{-\frac{dm}{2}}\right) + \log\left(|\Sigma|^{-\frac{m}{2}}\right) + \sum_{i=1}^m \left(-\frac{1}{2}(x_i - \mu)^t \cdot \Sigma^{-1}(x_i - \mu)\right) =$$

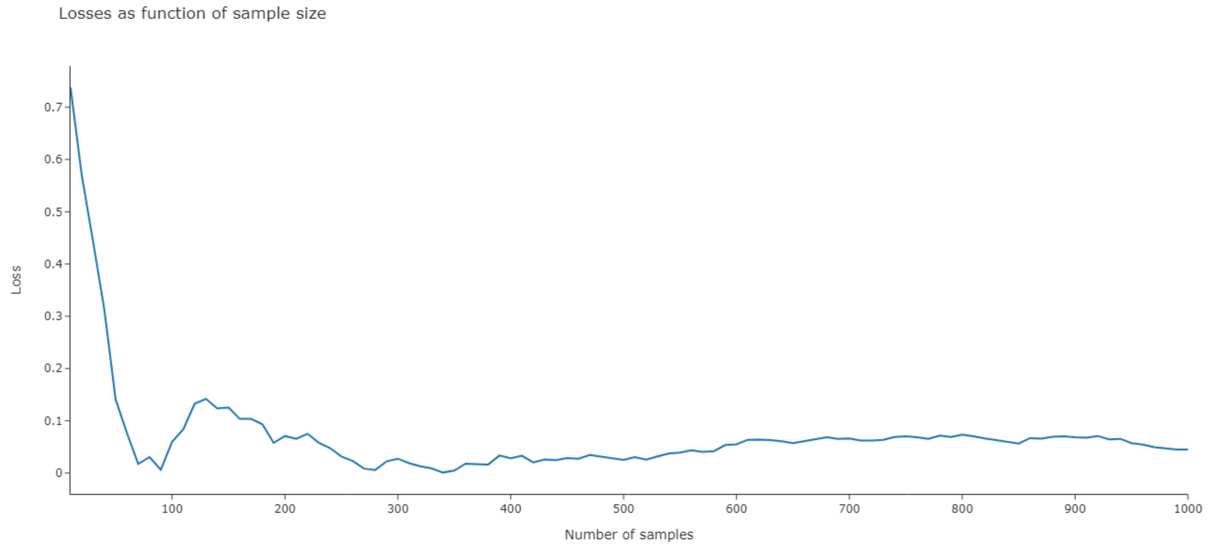
$$= -\frac{m}{2}\left(d \log(2\pi) + \log(|\Sigma|)\right) + \sum_{i=1}^m \left(-\frac{1}{2}(x_i - \mu)^t \cdot \Sigma^{-1}(x_i - \mu)\right)$$

$\Downarrow$

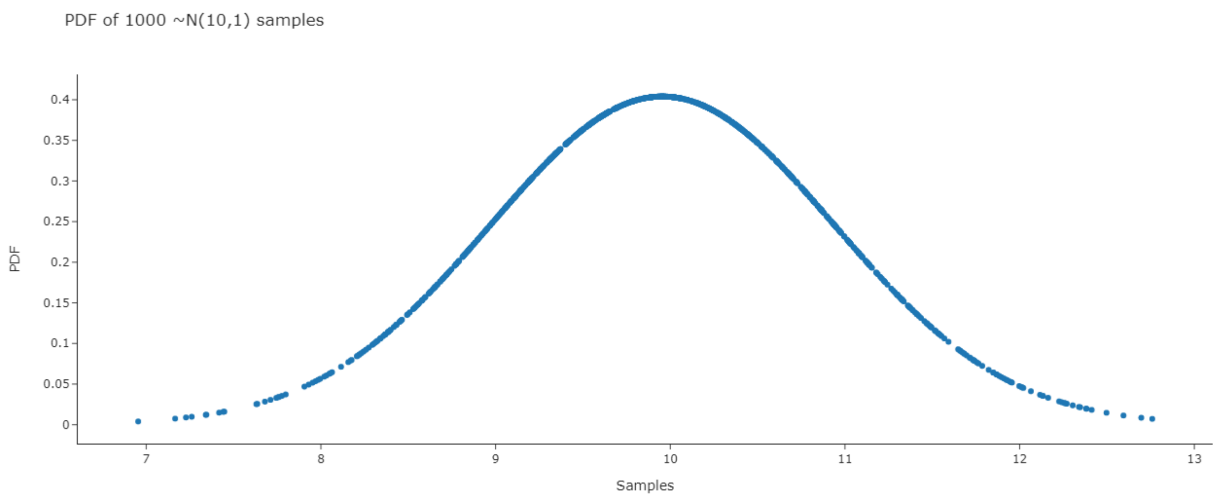
ה- log-likelihood function הנדרש.

חלק מעשי

גרף line משאלה 2:



גרף Scatter משאלה 3:



5 הנקודות heatmap של

Likelihoods as function of f3 and f1