שיר רשקוביץ

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 $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ 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 $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$, 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

2.1 Mathematical Background

2.1.1 Linear Algebra

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

(1

$$||Ax||^2 = (Ax)^t Ax \underset{\text{transpose חוקי}}{=} x^t A^t Ax \underset{\text{A אורתוגונלית ולכן } A^tA=I_n}{=} x^t x = ||x||^2$$

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 $AV = U\Sigma$.

(2 בכדי למצוא את ה- SVD של A, נמצא את ה- EVD של A^tA:

$$A^tA = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 2 \end{pmatrix}\begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{pmatrix}$$

נמצא את הע"ע של A^TA כלומר את השורשים של הפולינום האופייני של A^tA:

$$\det(\lambda I_n - A^tA) = \det\begin{pmatrix} \lambda-2 & 0 & -2 \\ 0 & \lambda-2 & 2 \\ -2 & 2 & \lambda-4 \end{pmatrix} = (\lambda-2)\begin{pmatrix} \lambda-2 & 2 \\ 2 & \lambda-4 \end{pmatrix} - 2\begin{pmatrix} 0 & -2 \\ \lambda-2 & 2 \end{pmatrix} = (\lambda-2)\left((\lambda-2)(\lambda-4)-4\right)$$

$$\underbrace{-2(0-(-2)(\lambda-2))}_{\underbrace{-2\cdot 2(\lambda-2)}_{-4(\lambda-2)}} \underset{\text{הוצאת גורם משותף}}{=} (\lambda-2)(\lambda^2-6\lambda) = \underbrace{(\lambda-2)\lambda(\lambda-6)}_{\text{הפולינום האופייני של } A^tA} \Rightarrow$$

לכן הע"ע הם: 2,0,6.

נמצא את הפתרונות למע' המשוואות ההומוגנית של $\lambda I_n - A^tA$ לכל λ בכדי למצוא ו"ע ששייכים לע"ע λ:

עבור λ=2:

$$\begin{pmatrix} 0 & 0 & -2 \\ 0 & 0 & 2 \\ -2 & 2 & -2 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & -1 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & -2 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & -1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & -1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

$x_2=t$, $x_3=0$, $x_1=t$ אז $V_2 = \left\{\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} t \,\middle|\, t\in\mathbb{R}\right\}$, ולכן $\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ ו"ע של הע"ע 2.

עבור $\lambda=0$:

$$\begin{pmatrix}-2&0&-2\\0&-2&2\\-2&2&-4\end{pmatrix}\longrightarrow\begin{pmatrix}1&0&1\\0&1&-1\\1&-1&2\end{pmatrix}\longrightarrow\begin{pmatrix}1&0&1\\0&1&-1\\0&-1&1\end{pmatrix}\longrightarrow\begin{pmatrix}1&0&1\\0&1&-1\\0&0&0\end{pmatrix}$$

אז $x_3=t$, $x_1=-t$, $x_2=t$ אז נבחר $\begin{pmatrix}-1\\1\\1\end{pmatrix}$ ו״ע של 0.

עבור $\lambda=6$:

$$\begin{pmatrix}4&0&-2\\0&4&2\\-2&2&2\end{pmatrix}\longrightarrow\begin{pmatrix}2&0&-1\\0&2&1\\-1&1&1\end{pmatrix}\longrightarrow\begin{pmatrix}0&2&1\\0&0&0\\-1&1&1\end{pmatrix}\longrightarrow\begin{pmatrix}1&-1&-1\\0&1&\frac12\\0&0&0\end{pmatrix}$$

$x_1=\frac12 t$, $x_2=-\frac12 t$, $x_3=t$ אז ניקח ו״ע $\begin{pmatrix}1\\-1\\2\end{pmatrix}$ של 6.

סה״כ מצאנו $\begin{pmatrix}1\\-1\\2\end{pmatrix},\begin{pmatrix}1\\1\\0\end{pmatrix},\begin{pmatrix}-1\\1\\1\end{pmatrix}$ ו״ע של הע״ע $6,2,0$ בהתאמה. מכיוון ש-A^tA מטריצה סימטרית, אז ו״ע של ע״ע שונים ניצבים ולכן הוקטורים הללו מנורמלים מהווים בסיס אורתונורמלי ל-$\mathbb{R}^3$.

$$\underbrace{(v_1,v_2,v_3)=\left(\begin{pmatrix}1\\-1\\2\end{pmatrix}\cdot\frac{1}{\sqrt6},\begin{pmatrix}1\\1\\0\end{pmatrix}\cdot\frac{1}{\sqrt2},\begin{pmatrix}-1\\1\\1\end{pmatrix}\cdot\frac{1}{\sqrt3}\right)}_{\text{בסיס אורתונורמלי ל-}\mathbb{R}^3}\Leftarrow\left\|\begin{pmatrix}1\\-1\\2\end{pmatrix}\right\|=\sqrt{1+1+4}=\sqrt6,\quad\left\|\begin{pmatrix}1\\1\\0\end{pmatrix}\right\|=\sqrt2,\quad\left\|\begin{pmatrix}-1\\1\\1\end{pmatrix}\right\|=\sqrt3$$

אז מתקיים $A^tA=\begin{pmatrix}|&|&|\\v_1&v_2&v_3\\|&|&|\end{pmatrix}\begin{pmatrix}6&0&0\\0&2&0\\0&0&0\end{pmatrix}\underbrace{\begin{pmatrix}|&|&|\\v_1&v_2&v_3\\|&|&|\end{pmatrix}^t}_{=V^T}$.

לכן נסיק ש- $V^T=\begin{pmatrix}1/\sqrt6&-1/\sqrt6&2/\sqrt6\\1/\sqrt2&1/\sqrt2&0\\-1/\sqrt3&1/\sqrt3&1/\sqrt3\end{pmatrix}$, והערכים הסינגולריים הם השורשים של הע״ע של

A^tA שאינם 0, כלומר $\sqrt2,\sqrt6$. ולכן $\Sigma=\begin{pmatrix}\sqrt6&0&0\\0&\sqrt2&0\end{pmatrix}$ (Σ באותם מימדים כמו A).

נמצא את AA^t כדי למצוא את הוקטורים הסינגולריים השמאליים-

$$AA^t=\begin{pmatrix}1&1&0\\1&-1&2\end{pmatrix}\begin{pmatrix}1&1\\1&-1\\0&2\end{pmatrix}=\begin{pmatrix}2&0\\0&6\end{pmatrix}$$

נבחין ש- (e_2,e_1) בסיס אורתונורמלי של ו״ע (בסדר יורד לפי הע״ע) מכיוון ש- $AA^te_2=6e_2$, $AA^te_1=2e_1$, $e_1\perp e_2$, $\|e_1\|=\|e_2\|=1$. לכן $U=(e_2\ e_1)=\begin{pmatrix}0&1\\1&0\end{pmatrix}$

אז: $A=\underbrace{\begin{pmatrix}0&1\\1&0\end{pmatrix}}_{U}\underbrace{\begin{pmatrix}\sqrt6&0&0\\0&\sqrt2&0\end{pmatrix}}_{\Sigma}\underbrace{\begin{pmatrix}1/\sqrt6&-1/\sqrt6&2/\sqrt6\\1/\sqrt2&1/\sqrt2&0\\-1/\sqrt3&1/\sqrt3&1/\sqrt3\end{pmatrix}}_{V^T}$.

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.

(3

Definition 1.4 For two vectors $\mathbf{v} \in \mathbb{R}^n, \mathbf{u} \in \mathbb{R}^m$, the *outer product* of $\mathbf{v}$ and $\mathbf{u}$, which is denoted by $\mathbf{v} \otimes \mathbf{u}$ or $\mathbf{v}\mathbf{u}^\top$ is an $n \times m$ matrix with entries:

מכפלה חיצונית $\quad n\times 1 \quad 1\times m \Rightarrow n\times m$

$$[\mathbf{v} \otimes \mathbf{u}]_{ij} = v_i \cdot u_j, \quad \mathbf{v} \otimes \mathbf{u} = \begin{bmatrix} v_1u_1 & v_1u_2 & \cdots & v_1u_m \\ \vdots & \vdots & \ddots & \vdots \\ v_nu_1 & v_nu_2 & \cdots & v_nu_m \end{bmatrix} = A$$

נשים לב שמהגדרה, $A = \begin{pmatrix} v_1 \cdot u^t \\ v_2 \cdot u^t \\ \vdots \\ v_n \cdot u^t \end{pmatrix}$ ולכן כל השורות הן כפל בסקלר אחת של השנייה, ולכן תלויות ליניארית.

למשל: $v_2 u^t = v_1 u^t \cdot \frac{v_2}{v_1}$. בנוסף, אם $u,v \neq 0$ אז קיימת קורדינטה $v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ v_i $(1\le i\le n)$ ששונה מ-0, וכן $u\neq 0$, ולכן

$v_i u^t \neq 0$ ולכן $rank(A)=1$.

4. Show that for any otrthonormal basis $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ 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 $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$, is the inner product between $\mathbf{x}$ and $\mathbf{u}_i$.

(4 יהי $1 \le i \le n$. אזי -

$$\langle x, u_i \rangle = \langle \sum_{j=1}^n a_j u_j, u_i \rangle \underset{\text{לינאריות}}{=} \sum_{j=1}^n a_j \langle u_j, u_i \rangle = a_i$$

הבסיס אורתונורמלי ולכן עבור $1 \le i \neq j \le n$ המכפלה הפנימית שווה 0 כי הם ניצבים. ועבור $j=i$ המכפלה הפנימית שווה 1 כי v_i מנורמל

תזכורת:

Definition 2.4 Let $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ where $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))^\top$. The *Jacobian* of f is the $m \times d$ matrix of all partial derivatives:

המטריצה מכלילה את מושג הנגזרת הק"מ עבור פונקציות סקלריות במשתנה יחיד.

$$J_\mathbf{x}(\mathbf{f}) := \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_d} \\ \vdots & & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_d} \end{bmatrix}$$

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

$$f(\sigma) = \underbrace{\underset{n\times n}{U} \cdot \underbrace{\underset{n\times n}{\operatorname{diag}(\sigma)} \underset{n\times n}{U^\top}}_{n\times n} \underset{n\times 1}{x}}_{n\times 1}$$

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

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

5) נסיון

$$U = \begin{pmatrix} | & & | \\ u_1 & \dots & u_n \\ | & & | \end{pmatrix}$$

$$f(\sigma) = U\operatorname{diag}(\sigma)U^T x = U\begin{pmatrix} \sigma_1 & & 0 \\ & \sigma_2 & \\ 0 & & \ddots \ \sigma_n \end{pmatrix}\begin{pmatrix} u_1^t x \\ \vdots \\ u_n^t x \end{pmatrix} = U\begin{pmatrix} \sigma_1 u_1^t x \\ \vdots \\ \sigma_n u_n^t x \end{pmatrix} = \sum_{i=1}^{n} \overbrace{(\sigma_i \underset{1\times n}{u_i^t} \underset{n\times 1}{x})}^{\text{סקלר}} u_i =$$

$$= \sum_{i=1}^{n} u_i (u_i^t \sigma_i x) = \sum_{i=1}^{n} \sigma_i u_i u_i^t x$$

$$[J_\sigma(f)]_{i,j} = \frac{\partial f_j(\sigma)}{\partial \sigma_i} = [u_i u_i^t x]_j \qquad \Longrightarrow \qquad J_\sigma(f) = \begin{pmatrix} | & & | \\ u_1 & \dots & u_n \\ | & & | \end{pmatrix}\begin{pmatrix} u_1^T x & & 0 \\ & \ddots & \\ 0 & & u_n^t x \end{pmatrix} = U\operatorname{diag}(U^T x)$$

הקורדינטה ה-j של $u_i u_i^t x$

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

Definition 2.3 Let $f : \mathbb{R}^d \to \mathbb{R}$. The *gradient* of f at $\mathbf{x}$ is the vector of partial derivatives:

וקטור הנגזרות החלקיות

$$\nabla f(\mathbf{x}) := \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_d}\right)^{T}$$

6

$$h(\sigma) = \frac{1}{2}\|f(\sigma)-y\|^2 = \frac{1}{2}\langle f(\sigma)-y, f(\sigma)-y\rangle = \frac{1}{2}\left(\|f(\sigma)\|^2+\|y\|^2-2\langle f(\sigma),y\rangle\right) = \frac{1}{2}\|f(\sigma)\|^2+\frac{1}{2}\|y\|^2-\langle f(\sigma),y\rangle =$$

$$= \frac{1}{2}\|f(\sigma)\|^2 - f(\sigma)^t y + \frac{1}{2}\|y\|^2$$

כלל השרשרת

$$\nabla h(\sigma) = \frac{\partial h(\sigma)}{\partial \sigma} = \frac{\partial h(\sigma)}{\partial f(\sigma)}\frac{\partial f(\sigma)}{\partial \sigma} = \frac{\partial\left(\frac{1}{2}\|f(\sigma)\|^2 - f(\sigma)^t y + \frac{1}{2}\|y\|^2\right)}{\partial f(\sigma)} \cdot \frac{\partial f(\sigma)}{\partial \sigma} = (f(\sigma)-y)^t \cdot \frac{\partial f(\sigma)}{\partial \sigma} =$$

$$= \underbrace{(f(\sigma)-y)^t}_{n\times n} \underbrace{J_\sigma(f)}_{n\times h} \underset{\text{מתרגיל 6}}{=} (f(\sigma)-y)^t U \mathrm{diag}(U^T x)$$

עבור x,y וקטורים ב- R^n עבור $n \in N$.

$$\frac{\partial x^t y}{\partial x_i} = \frac{\sum_i x_i y_i}{\partial x_i} = y_i \Rightarrow \frac{\partial x^t y}{\partial x} = y, \quad \frac{\partial x^t x}{\partial x_i} = \frac{\sum_i x_i^2}{\partial x_i} = 2x_i \Rightarrow \frac{\partial \|x\|^2}{\partial x} = 2x$$

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

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

7) נשתמש לפי כלל המנה: $\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$

$$[J_x(s)]_{i,j} = \frac{\partial s(x)_i}{\partial x_j} = \frac{\partial}{\partial x_j}\frac{e^{x_i}}{\sum_{\ell=1}^{k} e^{x_\ell}} = \begin{cases} \frac{e^{x_i}\sum_{\ell=1}^{k} e^{x_\ell} - e^{x_i}\cdot e^{x_j}}{\left(\sum_{\ell=1}^{k} e^{x_\ell}\right)^2} \overset{\text{הוצאת גורם משותף}}{=} \frac{e^{x_i}}{\sum_{\ell=1}^{k} e^{x_\ell}} \cdot \frac{\left(\sum_{\ell=1}^{k} e^{x_\ell} - e^{x_j}\right)}{\sum_{\ell=1}^{k} e^{x_\ell}} \overset{i=j}{=} \\ s(x)_i \cdot \left(\frac{\sum_{\ell=1}^{k} e^{x_\ell}}{\sum_{\ell=1}^{k} e^{x_\ell}} - \frac{e^{x_i}}{\sum_{\ell=1}^{k} e^{x_\ell}}\right) = s(x)_i \cdot (1 - s(x)_i) & i=j \\ \frac{0\cdot\sum_{\ell=1}^{k} e^{x_\ell} - e^{x_i}\cdot e^{x_j}}{\left(\sum_{\ell=1}^{k} e^{x_\ell}\right)^2} = -\frac{e^{x_i}}{\sum_{\ell=1}^{k} e^{x_\ell}} \cdot \frac{e^{x_j}}{\sum_{\ell=1}^{k} e^{x_\ell}} = -s(x)_i \cdot s_j(x) = s(x)_i (-s(x)_j) & i\neq j \end{cases}$$

כלומר $[J_x(f)]_{i,j} = s(x)_i \cdot (\delta_{ij} - s(x)_j)$ עם $\delta_{ij} = \begin{cases} 1 & i=j \\ 0 & i\neq j \end{cases}$.

ובכתיב מטריצוני: $J_x(f) = diag(s) - ss^T$ מכיוון ש- $[ss^T]_{i,j} = s(x)_i \cdot s(x)_j$ וכן $[diag(s)]_{i,j} = \begin{cases} s(x)_i & i=j \\ 0 & i\neq j \end{cases}$

וכן $s(x)_i - s(x)_i s(x)_j = s(x)_i (1 - s(x)_j)$, $0 - s(x)_i s(x)_j = -s(x)_i s(x)_j$

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.

8)

$\frac{\partial f}{\partial x} = 3x^2 - 5y$ $\qquad \frac{\partial f}{\partial y} = -5x - 5y^4$ $\qquad \frac{\partial f}{\partial x \partial y} = -5$

$\frac{\partial f}{\partial^2 x} = 6x$ $\qquad \frac{\partial f}{\partial^2 y} = -20y^3$ $\qquad \frac{\partial f}{\partial y \partial x} = -5$

אז מתקיים:

$$H_{(x,y)}(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial^2 x} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial^2 y} \end{pmatrix} = \begin{pmatrix} 6x & -5 \\ -5 & -20y^3 \end{pmatrix}$$

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 $u,v\in\bigcap_{i\in I} C_i$. נראה שמתקיים $\alpha u+(1-\alpha)v\in C$ עבור $\alpha\in[0,1]$.

אז מכוון שלכל $i\in I$ מתקיים $\bigcap_{i\in I} C_i\subseteq C_i$ מהגדרת החיתוך. אז $u,v\in C_i$ ולכן $\alpha u+(1-\alpha)v\in C_i$ מכך ש-C_i קבוצה קמורה. לכן מהגדרת החיתוך $\alpha u+(1-\alpha)v\in C$.

10) יהיו $u,v\in C_1+C_2$ אז קיימים $c_1,c_3\in C_1$, $c_2,c_4\in C_2$ כך ש- $u=c_1+c_2$, $v=c_3+c_4$. אז מכך ש- C_1,C_2 קבוצות קמורות מתקיים לכל $v\in C$

$$\alpha u+(1-\alpha)v=\alpha(c_1+c_2)+(1-\alpha)(c_3+c_4)=(\alpha c_1+(1-\alpha)c_3)+(\alpha c_2+(1-\alpha)c_4)$$

נבחין ש- $\alpha c_1+(1-\alpha)c_3\in C_1$, $\alpha c_2+(1-\alpha)c_4\in C_2$ כי הקבוצות קמורות. לכן מהגדרה זוהי קבוצה קמורה.

11) יהיו $u,v\in\lambda C$. אז קיימים $c_1,c_2\in C$ כך ש- $u=\lambda c_1$, $v=\lambda c_2$. אז-

$$\alpha u+(1-\alpha)v=\lambda\alpha c_1+(1-\alpha)\lambda c_2=\lambda(\underbrace{\alpha c_1+(1-\alpha)c_2}_{\in C})\Rightarrow \alpha u+(1-\alpha)v\in\lambda C$$

כי C קבוצה קמורה   מהגדרה

2.2 Estimation Theory

Based on Lecture 1 בוחן

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.

(12)

נסמן μ כתוחלת של $x\sim P$. לכל $\varepsilon>0$:

$$P(|\hat{\mu}_n - \mu| > \varepsilon) \le P(|\mu_n - E[\mu_n]| \ge \varepsilon) \le \frac{var[\mu_n]}{\varepsilon^2} = \frac{var[x]}{n\varepsilon^2} \xrightarrow[n\to\infty]{} 0 \Rightarrow$$

מהגדרה הראנו μ_n אומד עקבי.

$$E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\cdot nE(x_i) = E(x) \quad (x\sim P)$$

א"ש צ'בישב - ניתן להפעיל כי קיימת תוחלת ושונות סופית. ו- $\varepsilon>0$

$$var\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} var[x_i] = \frac{var[x]}{n}$$

ב"ת $x_1,\ldots,x_n$; $x\sim P$

(13) נסמן

$$X = \begin{pmatrix} | & & | \\ x_1 & \cdots & x_m \\ | & & | \end{pmatrix} \in M_{d\times m}$$

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

$$L(\mu,\Sigma|x_1,\ldots,x_m) = f_\mu(x_1,\ldots,x_m) = \prod_{i=1}^{m} f_\mu(x_i) = \prod_{i=1}^{m} \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}\underset{1\times d}{(x_i-\mu)^t}\underset{d\times d}{\Sigma^{-1}}\underset{d\times 1}{(x_i-\mu)}\right)$$

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

$e^{x+y} = e^x \cdot e^y$

כעת נמצא את $\log L(\mu,\Sigma|x_1,\ldots,x_m)$:

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

$\log x\cdot y = \log x + \log y$
$\log a^r = r\log a$
$\log(\exp(x)) = x$

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

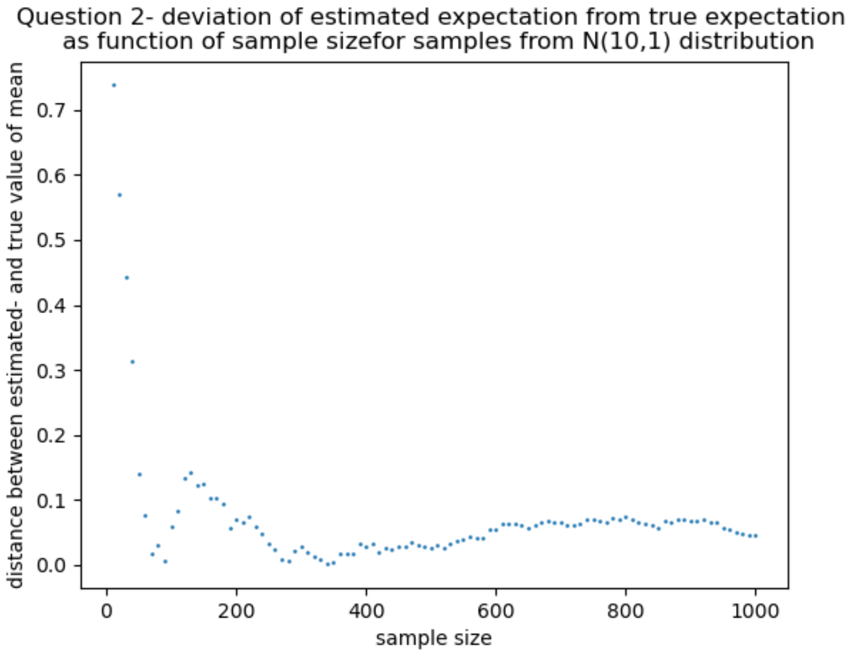
$\log x\cdot y = \log x + \log y$
$\log a^r = r\log a$

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

$r\log a = \log a^r$

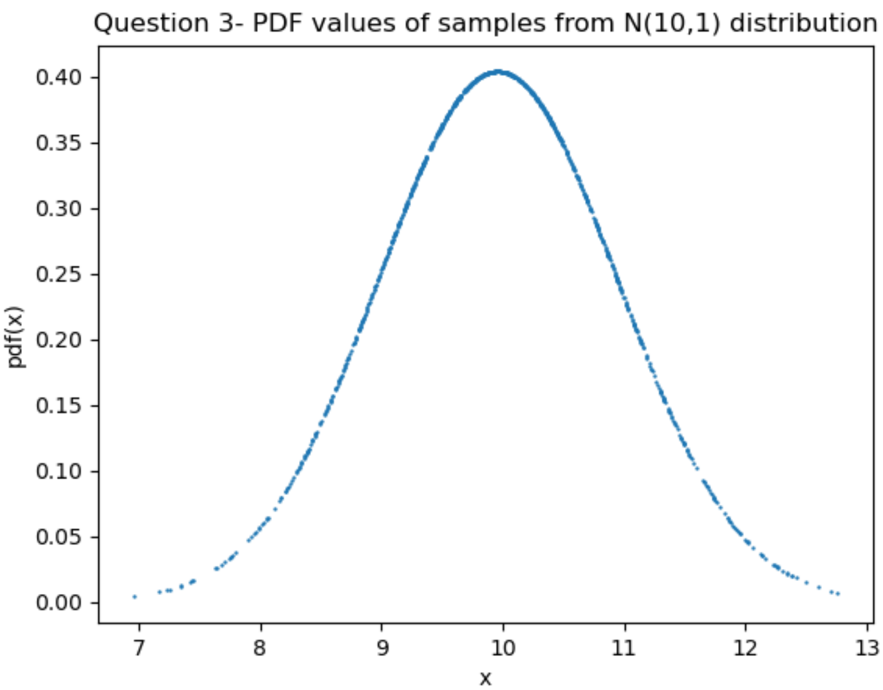
(9.955, 0.975) (1

(2



what are you expecting to see in the plot?

(3 הגרלנו את וקטור ה-X מתוך התפלגות N(10,1) ולכן נצפה לראות צורת פעמון באופן דומה לפונקציות צפיפות נורמליות, עם מרכוז סביב 10 (התוחלת) וצפיפות של נקודות בעיקר סביב התוחלת. זאת מכיוון שכזכור מקורס הסתברות שלקחנו בקירוב 68% מהדגימות בהתפלגות נורמלית נמצאות במרחק סטיית תקן אחת לכל היותר מהתוחלת, ובערך 95% במרחק שתי סטיות תקן מהתוחלת לכל היותר. מכיוון שהשונות קטנה (1) נצפה לראות נקודות בעיקר בתחום [9,11].



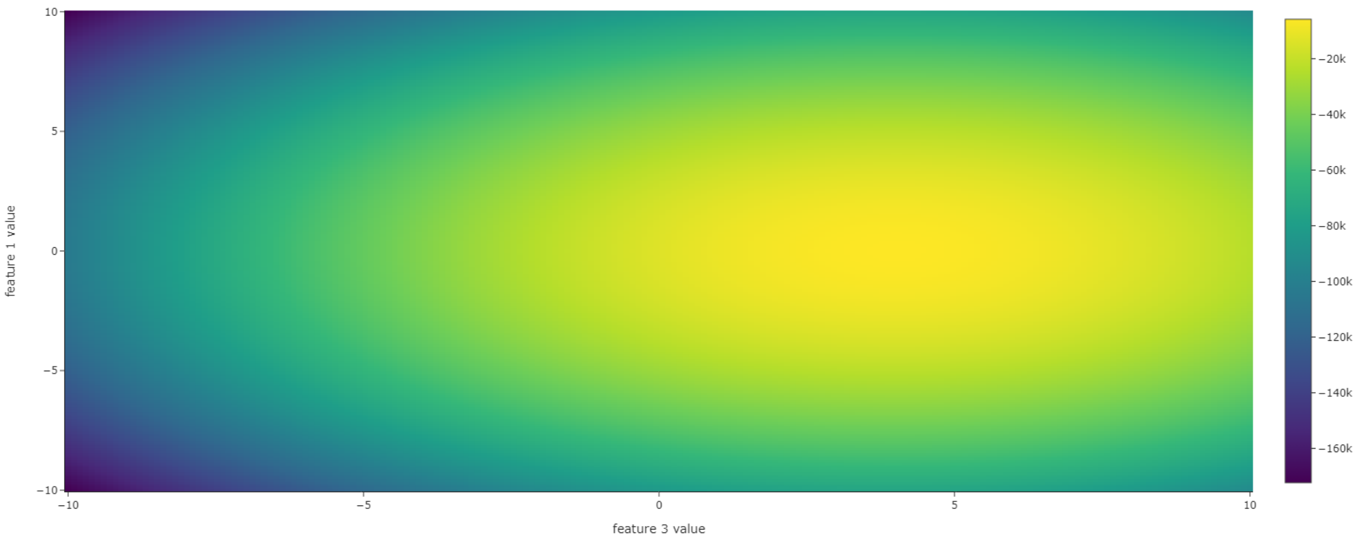
(2

mean =
cov =

```
[-0.023 -0.043  3.993 -0.02 ]
[[ 0.917  0.166 -0.03   0.463]
 [ 0.166  1.974 -0.006  0.046]
 [-0.03  -0.006  0.98  -0.02 ]
 [ 0.463  0.046 -0.02   0.973]]
```

(5

Question 5- Log likelihood of Multivariate Gaussian as function of mean's features 1,3



What are you able to learn from the plot?

נבחין שהאזור הצהוב, בו הlog- likelihood גבוה יותר, הוא האזור שבו הfeatures קרובים יותר לערך האמת שלהם בתוחלת (f1= 0, f3= 4).
כלומרניתן ללמוד מכך ששערוך לפי עיקרון הלמידה של מקסום פונקציית הlog-likelihood היא בחירה טובה (נזכור שהגענו לחפש מקסימום לפונקציית ה- log-likelihood מתוך מטרה לחפש את ה- MLE).

(6

```
features 1,3 that maximize the log-likelihood are:
f1= -0.05 ,f3= 3.97
```