מבוא למערכות לומדות (67577) | תרגיל 1 חלק תאורטי

שם: נמרוד בר גיורא | ת"ז: 207090622

1 חלק תאורטי

אלגברה ליניארית

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

הוכחה: תהי A מטריצה אורתוגונלית המייצגת העתקה $T : V \to W$.
נסמן $\dim(V) = n$ ו־ $\dim(W) = m$, כלומר $A \in M_{m\times n}(\mathbb{R})$. מתקיים לכל $x \in V$ ש־

$$\|Ax\| = \sqrt{\langle Ax \mid Ax\rangle} \overset{\text{הגדרת הצמוד}}{=} \sqrt{\langle x \mid A^* Ax\rangle} \overset{A\in M_{m\times n}(\mathbb{R})}{=} \sqrt{\langle x \mid A^\top Ax\rangle} \overset{\text{אורתוגונלית } A}{=} \sqrt{\langle x \mid I_n x\rangle} = \sqrt{\langle x \mid x\rangle} = \|x\|$$

כנדרש. ■

שאלה 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.

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

נתונה המטריצה $A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix}$. נמצא מטריצות U, Σ, V כך ש־ U, V אורתוגונליות מסדר 3×3 ו־ 2×2 בהתאמה ו־ Σ אלכסונית מסדר 2×3 כך ש־ $A = U\Sigma V^\top$.
נחשב את $A^\top A$ ונלכסן אותה אורתוגונלית:

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

נחשב את הפולינום האופייני:

$$\chi_{A^\top A}(\lambda) = \det \begin{bmatrix} \lambda - 2 & 0 & -2 \\ 0 & \lambda - 2 & 2 \\ -2 & 2 & \lambda - 4 \end{bmatrix} = (\lambda - 2)\det \begin{bmatrix} \lambda - 2 & 2 \\ 2 & \lambda - 4 \end{bmatrix} - 2\det \begin{bmatrix} 0 & \lambda - 2 \\ -2 & 2 \end{bmatrix} =$$
$$= (\lambda - 2)\left((\lambda - 2)(\lambda - 4) - 4\right) - 4(\lambda - 2) = (\lambda - 2)\left((\lambda - 2)(\lambda - 4) - 8\right) =$$
$$= (\lambda - 2)\left(\lambda^2 - 6\lambda + 8 - 8\right) = (\lambda - 2)(\lambda - 6)\lambda$$

לכן הערכים העצמיים הם $\{0,2,6\}$. נמצא להם וקטורים עצמיים. עבור הע"ע 0:

$$A^\top A-0I=\begin{bmatrix}2&0&2\\0&2&-2\\2&-2&4\end{bmatrix}\overset{\text{דירוג}}{\Longrightarrow}\begin{bmatrix}1&0&1\\0&1&-1\\0&0&0\end{bmatrix}$$

לכן הוקטור $\mathbb{R}^3 \ni \begin{bmatrix}-1\\1\\1\end{bmatrix}$ הוא ו"ע של $A^\top A$ עם ע"ע 0. ננרמל אותו ונקבל את הוקטור

$$\frac{1}{\sqrt{3}}\begin{bmatrix}-1\\1\\1\end{bmatrix}$$

עבור הע"ע 2:

$$A^\top A-2I=\begin{bmatrix}0&0&2\\0&0&-2\\2&-2&2\end{bmatrix}\overset{\text{דירוג}}{\Longrightarrow}\begin{bmatrix}1&-1&0\\0&0&1\\0&0&0\end{bmatrix}$$

לכן הוקטור $\begin{bmatrix}1\\1\\0\end{bmatrix}$ הוא ו"ע של $A^\top A$ עם ע"ע 2. ננרמל אותו ונקבל את הוקטור:

$$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\\0\end{bmatrix}$$

עבור הע"ע 6:

$$A^\top A-6I=\begin{bmatrix}-4&0&2\\0&-4&-2\\2&-2&-2\end{bmatrix}\overset{\text{דירוג}}{\Longrightarrow}\begin{bmatrix}1&0&-1/2\\0&1&1/2\\0&0&0\end{bmatrix}$$

לכן הוקטור $\begin{bmatrix}1\\-1\\2\end{bmatrix}$ הוא ו"ע של $A^\top A$ עם ע"ע 6. ננרמל ונקבל את הוקטור:

$$\frac{1}{\sqrt{6}}\begin{bmatrix}1\\-1\\2\end{bmatrix}$$

ולכן בסיס אורתונורמלי של $\mathbb{R}^3$ שמלכסן את $A^\top A$ הוא:

$$\left(\frac{1}{\sqrt{6}}\begin{bmatrix}1\\-1\\2\end{bmatrix},\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\\0\end{bmatrix},\frac{1}{\sqrt{3}}\begin{bmatrix}-1\\1\\1\end{bmatrix}\right)$$

נסמן

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

אז V מטריצה אורתוגונלית כי עמודותיה הן בסיס א"נ של $\mathbb{R}^3$.
את המטריצה $\Sigma \in M_{2\times 3}(\mathbb{R})$ נגדיר להיות:

$$\Sigma = \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & \sqrt{2} & 0 \end{bmatrix}$$

כי היא מקיימת ש־ $\Sigma^\top \Sigma = \begin{bmatrix} 6 & & \\ & 2 & \\ & & 0 \end{bmatrix}$. עכשיו כפי שראינו בתרגול מתקיים ש־

$$U\Sigma = AV = \begin{bmatrix} 0 & \sqrt{2} & 0 \\ \sqrt{6} & 0 & 0 \end{bmatrix}$$

ולכן

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

היא אורתוגונלית למשל כי עמודותיה הן הבסיס הסטנדרטי של $\mathbb{R}^2$, והוא בסיס א"נ.
עם המטריצות שמצאנו מתקיים ש־ $A = U\Sigma V^\top$.

שאלה 3

3. In this question we prove the Power-Iteration algorithm for finding the SVD of a matrix. Let $A \in \mathbb{R}^{m\times n}$ and define $C_0 = A^\top A$. Denote $\lambda_1 \geq \ldots \geq \lambda_n$ the eigenvalues of C_0, with the corresponding normalized eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

Let us assume the $\lambda_1 > \lambda_2$. Define $b_k \in \mathbb{R}$ as follows:

$$b_0 = \sum_{i=1}^{n} a_i v_i, \quad b_{k+1} = \frac{C_0 b_k}{\|C_0 b_k\|}$$

where $a_1 \neq 0$. Show that: $\lim_{k\to\infty} b_k = \pm v_1$.

הוכחה: נתונה מטריצה $A \in M_{m\times n}(\mathbb{R})$. נסמן $C_0 = A^\top A$, $\lambda_1 \geqslant \ldots \geqslant \lambda_n$ הע"ע של C_0, $v_1, \ldots, v_n \in \mathbb{R}^n$ הו"ע המתאימים מנורמלים. נניח ש־ $\lambda_1 > \lambda_2$ ונגדיר:

$$b_0 = \sum_{i=1}^{n} a_i v_i \quad , \quad b_{k+1} = \frac{C_0 b_k}{\|C_0 b_k\|}$$

עם $a_1 \neq 0$. נשים לב שלפי הגדרת b_0 ומכך ש־ $v_1, \ldots, v_n$ ו"ע של C_0 נובע ש־

$$C_0 b_0 = C_0 \sum_{i=1}^{n} a_i v_i = \sum_{i=1}^{n} a_i C_0 v_i = \sum_{i=1}^{n} a_i \lambda_i v_i$$

נוכיח באינדוקציה שלכל $k \in \mathbb{N}$ מתקיים ש־

$$b_k = \frac{\sum_{i=1}^{n} a_i \lambda_i^k v_i}{\left\| \sum_{i=1}^{n} a_i \lambda_i^k v_i \right\|}$$

בסיס: לפי ההגדרה ולפי החישוב לעיל:

$$b_1 = \frac{C_0 b_0}{\|C_0 b_0\|} = \frac{\sum_{i=1}^{n} a_i \lambda_i v_i}{\left\| \sum_{i=1}^{n} a_i \lambda_i v_i \right\|}$$

כנדרש.

צעד: נניח נכונות עבור $k-1$ ונוכיח עבור k:

$$b_k \overset{\text{לפי ההגדרה}}{=} \frac{C_0 b_{k-1}}{\|C_0 b_{k-1}\|} \overset{\text{מהנחת האינדוקציה}}{=} \frac{C_0 \frac{\sum_{i=1}^n a_i\lambda_i^{k-1}v_i}{\left\|\sum_{i=1}^n a_i\lambda_i^{k-1}v_i\right\|}}{\left\|C_0 \frac{\sum_{i=1}^n a_i\lambda_i^{k-1}v_i}{\left\|\sum_{i=1}^n a_i\lambda_i^{k-1}v_i\right\|}\right\|} = \frac{\frac{\sum_{i=1}^n a_i\lambda_i^{k-1}C_0v_i}{\left\|\sum_{i=1}^n a_i\lambda_i^{k-1}v_i\right\|}}{\left\|\frac{\sum_{i=1}^n a_i\lambda_i^{k-1}C_0v_i}{\left\|\sum_{i=1}^n a_i\lambda_i^{k-1}v_i\right\|}\right\|} = \frac{\frac{\sum_{i=1}^n a_i\lambda_i^{k-1}\lambda_iv_i}{\left\|\sum_{i=1}^n a_i\lambda_i^{k-1}v_i\right\|}}{\left\|\frac{\sum_{i=1}^n a_i\lambda_i^{k-1}\lambda_iv_i}{\left\|\sum_{i=1}^n a_i\lambda_i^{k-1}v_i\right\|}\right\|} =$$

$$= \frac{\frac{\sum_{i=1}^n a_i\lambda_i^{k}v_i}{\left\|\sum_{i=1}^n a_i\lambda_i^{k-1}v_i\right\|}}{\frac{1}{\left\|\sum_{i=1}^n a_i\lambda_i^{k-1}v_i\right\|}\cdot\left\|\sum_{i=1}^n a_i\lambda_i^{k}v_i\right\|} = \frac{\sum_{i=1}^n a_i\lambda_i^{k}v_i}{\left\|\sum_{i=1}^n a_i\lambda_i^{k}v_i\right\|}$$

וזה מוכיח את צעד האינדוקציה.

נחשב את הנורמה $\left\|\sum_{i=1}^n a_i\lambda_i^k v_i\right\|$ לכל $k\in\mathbb{N}$:

$$\left\|\sum_{i=1}^n a_i\lambda_i^k v_i\right\| = \sqrt{\left\langle \sum_{i=1}^n a_i\lambda_i^k v_i \mid \sum_{j=1}^n a_j\lambda_j^k v_j\right\rangle} = \sqrt{\sum_{i=1}^n\sum_{j=1}^n a_ia_j\lambda_i^k\lambda_j^k\langle v_i \mid v_j\rangle} \overset{\text{כי זו סדרת וקטורים א"נ}}{=} \sqrt{\sum_{i=1}^n \left(a_i\lambda_i^k\right)^2}$$

ועכשיו בעזרת הזהות שהוכחנו באינדוקציה וע"י חילוק המונה והמכנה ב־λ_1^k נקבל ש־:

$$b_k = \frac{\sum_{i=1}^n a_i\lambda_i^k v_i}{\left\|\sum_{i=1}^n a_i\lambda_i^k v_i\right\|} = \frac{\sum_{i=1}^n a_i\lambda_i^k v_i}{\sqrt{\sum_{i=1}^n \left(a_i\lambda_i^k\right)^2}} \overset{\cdot\frac{\lambda_1^k}{\lambda_1^k}}{=} \frac{\sum_{i=1}^n a_i\left(\frac{\lambda_i}{\lambda_1}\right)^k v_i}{\sqrt{\sum_{i=1}^n \left(a_i\left(\frac{\lambda_i}{\lambda_1}\right)^k\right)^2}}$$

מכיוון ש־$\lambda_1 > \lambda_2 \geqslant \lambda_3 \geqslant \ldots \geqslant \lambda_n$ אז לכל $2 \leqslant i \leqslant n$ מתקיים ש־$\frac{\lambda_i}{\lambda_1} < 1$ ולכן:

$$\left(\frac{\lambda_i}{\lambda_1}\right)^k \xrightarrow[k\to\infty]{} 0$$

ולכן:

$$b_k = \frac{\sum_{i=1}^n a_i\left(\frac{\lambda_i}{\lambda_1}\right)^k v_i}{\sqrt{\sum_{i=1}^n \left(a_i\left(\frac{\lambda_i}{\lambda_1}\right)^k\right)^2}} \xrightarrow[k\to\infty]{} \frac{a_1v_1}{\sqrt{a_1^2}} = \frac{a_1}{|a_1|}v_1 = \pm v_1$$

כנדרש. ■

אינפי רב־מימדי

שאלה 4

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

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

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

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

נסמן לכל $1 \leqslant i, j \leqslant n$ את הכניסה ה־i,j של U ב־ $u_j^i \in \mathbb{R}$ (כאשר i מסמן את אינדקס השורה ו־j את אינדקס העמודה), ולכל $1 \leqslant i \leqslant n$ נסמן את העמודה ה־i של U ב־ $u_i \in \mathbb{R}^n$. נשים לב שלכל $\sigma \in \mathbb{R}^n$ מתקיים:

$$f(\sigma) = U \operatorname{diag}(\sigma) U^\top x = \begin{bmatrix} | & & | \\ u_1 & \dots & u_n \\ | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} \begin{bmatrix} - & u_1^\top & - \\ & \vdots & \\ - & u_1^\top & - \end{bmatrix} x = \begin{bmatrix} | & & | \\ \sigma_1 u_1 & \dots & \sigma_n u_n \\ | & & | \end{bmatrix} \begin{bmatrix} u_1^\top x \\ \vdots \\ u_n^\top x \end{bmatrix} =$$

$$= u_1^\top x \sigma_1 u_1 + \ldots + u_n^\top x \sigma_n u_n = \sigma_1 u_1^\top x u_1 + \ldots + \sigma_n u_n^\top x u_n = \sum_{\ell=1}^{n} \sigma_\ell u_\ell^\top x u_\ell \in \mathbb{R}^n$$

נשים לב שלכל $1 \leqslant i \leqslant n$, הכניסה ה־$i$ של הוקטור שקיבלנו היא סכום הכניסות ה־i של הוקטורים שהוא הסכום שלהם, כלומר:

$$\left[\sum_{\ell=1}^{n} \sigma_\ell u_\ell^\top x u_\ell\right]_i = \sum_{\ell=1}^{n} \left[\sigma_\ell u_\ell^\top x u_\ell\right]_i = \sum_{\ell=1}^{n} \sigma_\ell u_\ell^\top x u_\ell^i$$

ונשים לב שבנוסחא הזו $u_\ell^\top x u_\ell^i \in \mathbb{R}$ (הוא סקלר). לכן נוכל להגדיר לכל $1 \leqslant i \leqslant n$ $f_i : \mathbb{R}^n \to \mathbb{R}$ ע"י:

$$\forall \sigma \in \mathbb{R}^n : \quad f_i(\sigma) = \sum_{\ell=1}^{n} \sigma_\ell u_\ell^\top x u_\ell^i$$

ונקבל ש־

$$f(\sigma) = \begin{bmatrix} f_1(\sigma) \\ \vdots \\ f_n(\sigma) \end{bmatrix} = \begin{bmatrix} \sum_{\ell=1}^{n} \sigma_\ell u_\ell^\top x u_\ell^1 \\ \vdots \\ \sum_{\ell=1}^{n} \sigma_\ell u_\ell^\top x u_\ell^n \end{bmatrix}$$

עכשיו לכל $1 \leqslant i, j \leqslant n$ נקבל ש־

$$\frac{\partial}{\partial x_j} f_i(\sigma) = \frac{\partial}{\partial x_j} \sum_{\ell=1}^{n} \sigma_\ell u_\ell^\top x u_\ell^i = \sum_{\ell=1}^{n} \frac{\partial}{\partial x_j} \left(\sigma_\ell u_\ell^\top x u_\ell^i\right) = \frac{\partial}{\partial x_j} \left(\sigma_j u_j^\top x u_j^i\right) = u_j^\top x u_j^i = \langle u_j \mid x \rangle u_j^i$$

ולכן:

$$J_\sigma(f) = \begin{bmatrix} \langle u_1 \mid x \rangle u_1^1 & \dots & \langle u_n \mid x \rangle u_n^1 \\ \vdots & \ddots & \vdots \\ \langle u_1 \mid x \rangle u_1^n & \dots & \langle u_n \mid x \rangle u_n^n \end{bmatrix} = \begin{bmatrix} | & & | \\ \langle u_1 \mid x \rangle u_1 & \dots & \langle u_n \mid x \rangle u_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ u_1 & \dots & u_n \\ | & & | \end{bmatrix} \begin{bmatrix} \langle u_1 \mid x \rangle & & \\ & \ddots & \\ & & \langle u_n \mid x \rangle \end{bmatrix} =$$

$$= U \cdot \operatorname{diag}\left(\begin{bmatrix} \langle u_1 \mid x \rangle \\ \vdots \\ \langle u_1 \mid x \rangle \end{bmatrix} \right) = U \cdot \operatorname{diag}\left([x]_{\mathcal{B}_U}\right)$$

כאשר השיוויון האחרון נובע מכך ש־U אורתוגונלית, ולכן עמודותיה הן בסיס אורתונורמלי $\mathcal{B}_U = (u_1, \ldots, u_n)$ של $\mathbb{R}^n$, ולומדים בליניארית ש־$\begin{bmatrix} \langle u_1|x\rangle \\ \vdots \\ \langle u_1|x\rangle \end{bmatrix}$ הוא בדיוק עמודת הקוארדינטות של x ביחס לבסיס האורתונורמלי הזה.

שאלה 5

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

נגדיר $g : \mathbb{R}^n \to \mathbb{R}$ ע"י:

$$\forall \sigma \in \mathbb{R}^n : \quad g(\sigma) = \frac{1}{2}\|\sigma - y\|^2$$

ונקבל ש־ $h(\sigma) = \frac{1}{2}\|f(\sigma) - y\|^2 = g(f(\sigma)) = (g \circ f)(\sigma)$. לפי כלל השרשרת מתקיים ש־

$$J_\sigma(h) = J_{f(\sigma)}(g) \cdot J_\sigma(f)$$

בשאלה הקודמת ראינו ש־ $J_\sigma(f) = U \cdot \operatorname{diag}\left([x]_{\mathcal{B}_U}\right)$, כאשר $x \in \mathbb{R}^n$ ו־$U \in \mathbb{R}^{n \times n}$ אורתוגונלית, ו־$\mathcal{B}_U$ בסיס א"נ של $\mathbb{R}^n$ המורכב מעמודותיה של U.

נחשב את $J_\sigma(g)$:

לכל $1 \leqslant i \leqslant n$ הנגזרת של $g(\sigma)$ לפי σ_i היא:

$$\begin{aligned}\frac{\partial}{\partial \sigma_i} g(\sigma) &= \frac{\partial}{\partial \sigma_i}\left(\frac{1}{2}\|\sigma - y\|^2\right) = \frac{\partial}{\partial \sigma_i}\left(\frac{1}{2}\sum_{\ell=1}^{n}(\sigma_\ell - y_\ell)^2\right) = \frac{\partial}{\partial \sigma_i}\left(\frac{1}{2}\sum_{\ell=1}^{n}\left(\sigma_\ell^2 - 2\sigma_\ell y_\ell + y_\ell^2\right)\right) = \\ &= \frac{1}{2}\sum_{\ell=1}^{n}\frac{\partial}{\partial \sigma_i}\left(\sigma_\ell^2 - 2\sigma_\ell y_\ell + y_\ell^2\right) = \frac{1}{2}(2\sigma_i - 2y_i) = \sigma_i - y_i\end{aligned}$$

ולכן:

$$\nabla g(\sigma) = \begin{bmatrix} \sigma_1 - y_1 \\ \vdots \\ \sigma_n - y_n \end{bmatrix} = \sigma - y$$

ולכן כפי שראינו - $J_\sigma(g) = (\nabla g(\sigma))^\top = (\sigma - y)^\top$. נציב בכלל השרשרת ונקבל ש־

$$J_\sigma(h) = (f(\sigma) - y)^\top U \operatorname{diag}\left([x]_{\mathcal{B}_U}\right)$$

שאלה 6

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

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

כמו בתרגול, נגדיר לכל $i \in [k]$ את הפונקציה $g_i : \mathbb{R}^d \to \mathbb{R}$ ע"י:

$$\forall x \in \mathbb{R}^d : \quad g_i(x) = e^{x_i}$$

ואת הפונקציה $h : \mathbb{R}^d \to \mathbb{R}$ ע"י:

$$\forall x \in \mathbb{R}^d : \quad h(x) = \sum_{\ell=1}^{k} e^{x_\ell}$$

נקבל שלכל $i \in [k]$ מתקיים ש- $S_i(x) = \frac{g_i(x)}{h(x)}$.
לכן לכל $i \in [k]$ ולכל $j \in [n]$ מתקיים ש-

$$\frac{\partial}{\partial x_j} S_i(x) = \frac{\partial}{\partial x_j} \frac{e^{x_i}}{\sum\limits_{\ell=1}^{k} e^{x_\ell}} = \frac{\partial}{\partial x_j} \frac{g_i(x)}{h(x)} = \frac{\frac{\partial g_i(x)}{\partial x_j} h(x) - \frac{\partial h(x)}{\partial x_j} g_i(x)}{h^2(x)} \quad \bigstar$$

נניח תחילה ש-$k \leqslant d$.
נשים לב שלכל $j \in [d]$ הנגזרת החלקית של $h(x)$ ביחס ל-x_j היא:

$$\frac{\partial}{\partial x_j} h(x) = \begin{cases} e^{x_j} & j \leqslant k \\ 0 & j > k \end{cases}$$

בתרגול ראינו שעבור $i = j$ מתקיים ש-

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

במקרה שבו $j \neq i$, עם $j \in [d]$ ו- $i \in [k]$, נקבל שהנגזרת החלקית של $g_i(x)$ ביחס ל-x_j היא:

$$\frac{\partial}{\partial x_j} g_i(x) = \frac{\partial}{\partial x_j} e^{x_i} = 0$$

ולכן אם $k \leqslant j$ אז:

$$\frac{\partial}{\partial x_j} S_i(x) \overset{\bigstar}{=} \frac{\frac{\partial g_i(x)}{\partial x_j} h(x) - \frac{\partial h(x)}{\partial x_j} g_i(x)}{h^2(x)} = \frac{0 - e^{x_j} \cdot e^{x_i}}{\left(\sum\limits_{\ell=1}^{k} e^{x_\ell}\right)^2} = -\frac{e^{x_j}}{\sum\limits_{\ell=1}^{k} e^{x_\ell}} \cdot \frac{e^{x_i}}{\sum\limits_{\ell=1}^{k} e^{x_\ell}} = -S_j(x) S_i(x)$$

ואם $k > j$ אז:

$$\frac{\partial}{\partial x_j} S_i(x) = \frac{0 - 0}{\left(\sum\limits_{\ell=1}^{k} e^{x_\ell}\right)^2} = 0$$

ולכן מטריצת היעקוביאן של S היא:

$$\mathbb{R}^{k\times d} \ni J_x\left(S\right) = \begin{bmatrix} S_1\left(1-S_1\right) & -S_2S_1 & \dots & -S_kS_1 & 0 & \dots & 0 \\ -S_1S_2 & S_2\left(1-S_2\right) & \dots & -S_kS_2 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -S_1S_k & -S_2S_k & \dots & S_k\left(1-S_k\right) & 0 & \dots & 0 \end{bmatrix}$$

במקרה שבו $k \leqslant d$, לכל $j \in [d]$ נקבל ש־ $j \in [k]$ ולכן $\frac{\partial}{\partial x_j} h\left(x\right) = e^{x_j}$, ולכן לכל $i \in [k]$, $j \in [d]$ שונים:

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

ונקבל שמטריצת היעקוביאן היא:

$$J_x\left(S\right) = \begin{bmatrix} S_1\left(1-S_1\right) & -S_2S_1 & \dots & -S_dS_1 \\ -S_1S_2 & S_2\left(1-S_2\right) & \dots & -S_dS_2 \\ \vdots & \vdots & \ddots & \vdots \\ -S_1S_d & -S_2S_d & \dots & S_d\left(1-S_d\right) \\ -S_1S_{d+1} & -S_2S_{d+1} & \dots & -S_dS_{d+1} \\ \vdots & \vdots & \ddots & \vdots \\ -S_1S_k & -S_2S_k & \dots & -S_dS_k \end{bmatrix}$$

שאלה 7

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

נתחיל בחישוב הנגזרות החלקיות של f:

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

ונמשיך עם הנגזרות השניות החלקיות:

$$\frac{\partial^2}{\partial^2 x} f\left(x,y\right) = \frac{\partial}{\partial x}\left(3x^2 - 5y\right) = 6x$$
$$\frac{\partial^2}{\partial x \partial y} f\left(x,y\right) = \frac{\partial}{\partial y}\left(3x^2 - 5y\right) = -5$$
$$\frac{\partial^2}{\partial^2 y} f\left(x,y\right) = \frac{\partial}{\partial y}\left(-5x - 5y^4\right) = -20y^3$$
$$\frac{\partial^2}{\partial y \partial x} f\left(x,y\right) = \frac{\partial}{\partial x}\left(-5x - 5y^4\right) = -5$$

ולכן:

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

תאוריית אומדן

שאלה 8

> 8. 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.

הוכחה: נסמן ב־μ את התוחלת של $\mathcal{P}$ וב־γ את השונות שלה.

לפי ההגדרה, האומד $\hat{\mu}_n$ מנסה לאמוד את התוחלת של $\mathcal{P}$ ־ μ, ולכן הוא קונסיסטנטי אם ורק אם לכל $\varepsilon > 0$ מתקיים ש־

$$\lim_{n \to \infty} \mathbb{P}\left(|\mu_n - \mu| \geqslant \varepsilon\right) = 0$$

נחשב לכל $n \in \mathbb{N}$ את התוחלת של $\hat{\mu}_n$ (הוא הרי מ"מ):

$$\mathbb{E}\left(\hat{\mu}_n\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \overset{\text{ליניאריות התוחלת}}{\overset{\downarrow}{=}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(x_i\right) \overset{x_i \overset{\text{iid}}{\sim} \mathcal{P}}{\overset{\downarrow}{=}} \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} \cdot n \cdot \mu = \mu$$

מכיוון ש־ $x_i \overset{\text{iid}}{\sim} \mathcal{P}$ אז בפרט הם ב"ת ולכן נוכל לחשב גם את השונות של $\hat{\mu}_n$ לכל $n \in \mathbb{N}$ בקלות:

$$\operatorname{Var}\left(\hat{\mu}_n\right) = \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{Var}\left(x_i\right) = \frac{1}{n^2} \cdot n \cdot \gamma = \frac{\gamma}{n}$$

יהי $\varepsilon > 0$. לפי אי־שיוויון צ'בישב:

$$\mathbb{P}\left(|\hat{\mu}_n - \mu| \geqslant \varepsilon\right) \leqslant \frac{\operatorname{Var}\left(\hat{\mu}_n\right)}{\varepsilon^2} = \frac{\frac{\gamma}{n}}{\varepsilon^2} = \frac{\gamma}{n\varepsilon^2} \xrightarrow[n \to \infty]{} 0$$

ומכיוון שהסתברות היא אי־שלילית לפי הגדרתה אז נובע מסנדוויץ' ש־

$$\lim_{n \to \infty} \mathbb{P}\left(|\mu_n - \mu| \geqslant \varepsilon\right) = 0$$

וזה נכון לכל $\varepsilon > 0$. לכן $\hat{\mu}_n$ קונסיסטנטי כנדרש. ■

שאלה 9

> 9. 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 $\Sigma \in \mathbb{R}^{d \times d}$. Derive the log-likelihood function of $\mathcal{N}(\mu, \Sigma)$. Hint: follow the approach used to derive the likelihood function for the univariate case.

ראינו בכיתה שפונקציית הצפיפות של מ"מ X המתפלג $\mathcal{N}(\mu, \Sigma)$ עם $\mu \in \mathbb{R}^d$ ו־ $\Sigma \in M_{d \times d}(\mathbb{R})$ היא:

$$\forall x \in \mathbb{R}^d: \quad f_X(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

בנוסף ראינו שעבור מ"מ $X \sim \mathcal{P}(\theta)$ עם פונקציית צפיפות f, פונקציית ה־likelihood שלו היא:

$$\mathcal{L}(\theta \mid x) = f_\theta(x)$$

ופונקציית ה־log-likelihood היא $\log\left(f_\theta(x)\right)$.

ולכן עבור התצפיות $x_1,\ldots,x_m \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu,\Sigma\right)$ הנתונות מתקיים לכל $i\in[m]$ ש־

$$\mathcal{L}\left(\theta \mid x_i\right)=f_\theta\left(x_i\right)=\frac{1}{\sqrt{\left(2\pi\right)^d\left|\Sigma\right|}}\exp\left(-\frac{1}{2}\left(x_i-\mu\right)^\top\Sigma^{-1}\left(x_i-\mu\right)\right)$$

מכיוון שהם בלתי תלויים ובעלי התפלגות זהה:

$$\begin{aligned}\mathcal{L}\left(\theta \mid x_1,\ldots,x_m\right)=f_\theta\left(x_1,\ldots,x_m\right)=&\prod_{i=1}^m f_\theta\left(x_i\right)=\prod_{i=1}^m\left(\frac{1}{\sqrt{\left(2\pi\right)^d\left|\Sigma\right|}}\exp\left(-\frac{1}{2}\left(x_i-\mu\right)^\top\Sigma^{-1}\left(x_i-\mu\right)\right)\right)=\\ =&\frac{1}{\left(\left(2\pi\right)^d\left|\Sigma\right|\right)^{\frac{m}{2}}}\prod_{i=1}^m\exp\left(-\frac{1}{2}\left(x_i-\mu\right)^\top\Sigma^{-1}\left(x_i-\mu\right)\right)=\\ =&\frac{1}{\left(\left(2\pi\right)^d\left|\Sigma\right|\right)^{\frac{m}{2}}}\exp\left(-\frac{1}{2}\sum_{i=1}^m\left(x_i-\mu\right)^\top\Sigma^{-1}\left(x_i-\mu\right)\right)\end{aligned}$$

ולכן פונקציית ה־log-likelihood של $\mathcal{N}\left(\mu,\Sigma\right)$ היא:

$$\begin{aligned}&\log\left(\frac{1}{\left(\left(2\pi\right)^d\left|\Sigma\right|\right)^{\frac{m}{2}}}\exp\left(-\frac{1}{2}\sum_{i=1}^m\left(x_i-\mu\right)^\top\Sigma^{-1}\left(x_i-\mu\right)\right)\right)=\\ &=\log\left(\frac{1}{\left(\left(2\pi\right)^d\left|\Sigma\right|\right)^{\frac{m}{2}}}\right)+\log\left(\exp\left(-\frac{1}{2}\sum_{i=1}^m\left(x_i-\mu\right)^\top\Sigma^{-1}\left(x_i-\mu\right)\right)\right)=\\ &=-\frac{m}{2}\log\left(\left(2\pi\right)^d\left|\Sigma\right|\right)-\frac{1}{2}\sum_{i=1}^m\left(x_i-\mu\right)^\top\Sigma^{-1}\left(x_i-\mu\right)\end{aligned}$$

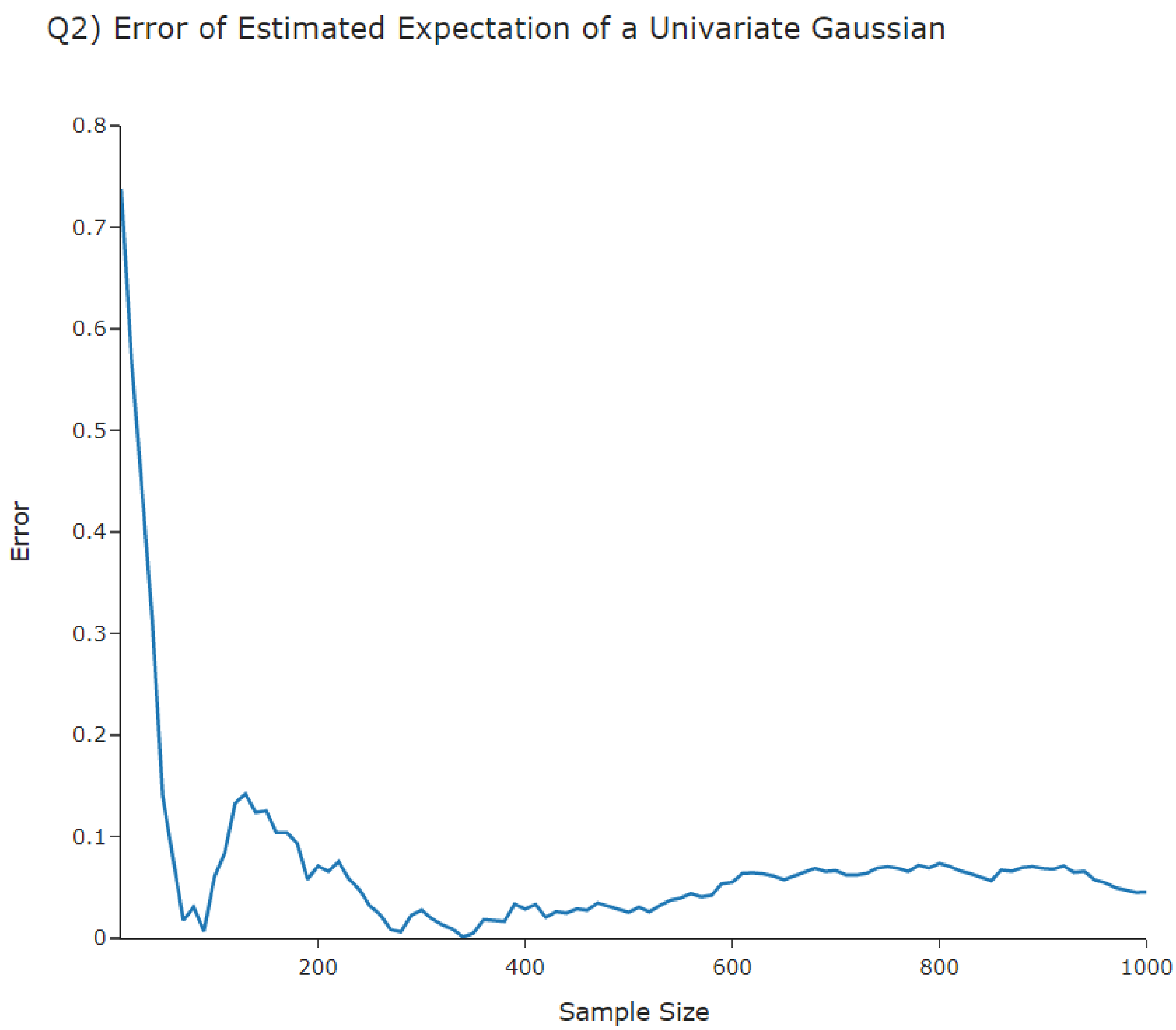
2 חלק מעשי

שאלה 1

הערכים של התוחלת ושל השונות שחושבו מודפסים בקוד כפי שכתוב בהוראות.

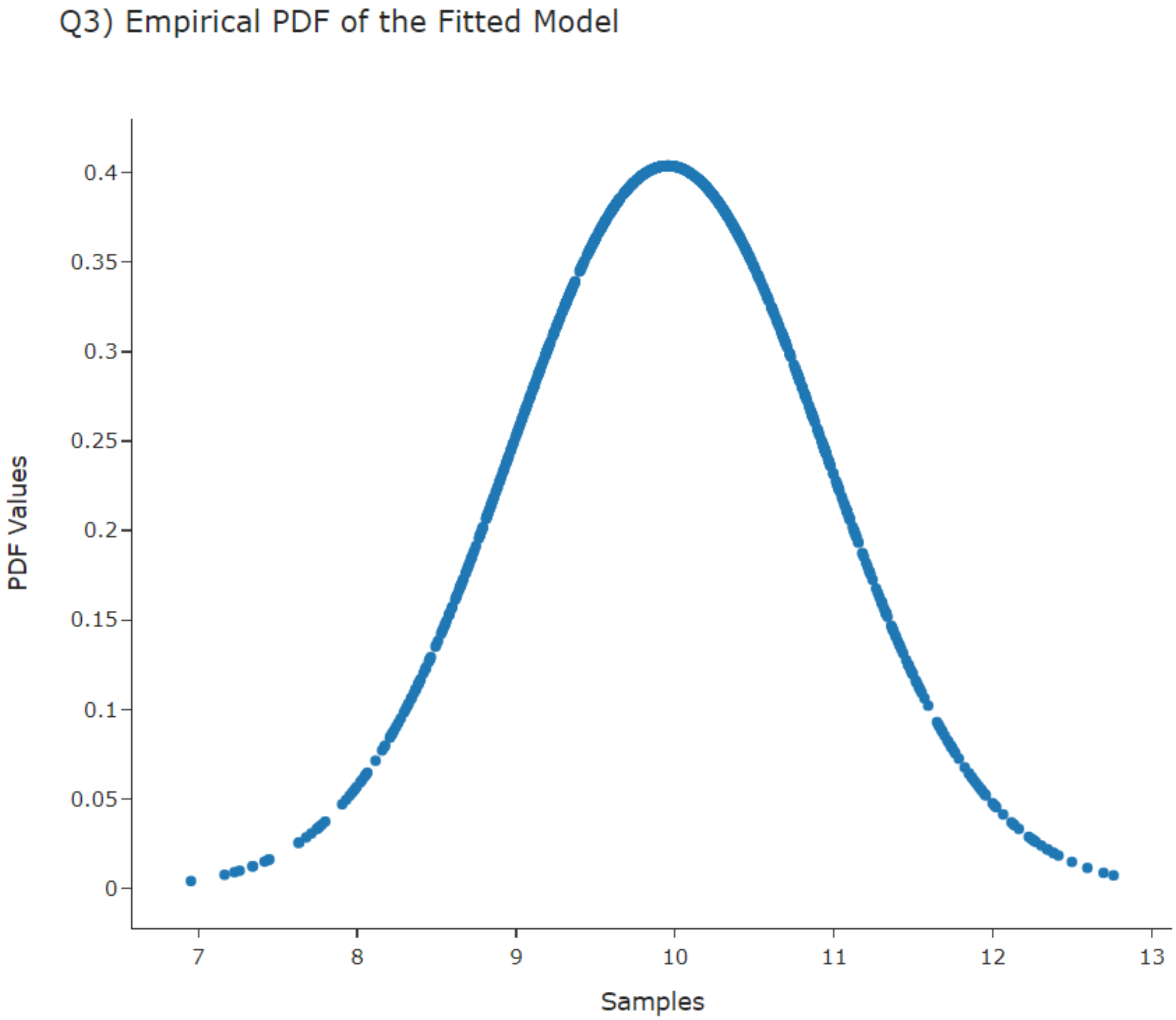
שאלה 2

גרף המציג את גודל השגיאה באומדן התוחלת ביחס למספר הדגימות:



שאלה 3

גרף המציג את פונקציית הצפיפות שחושבה ע"י המודל:

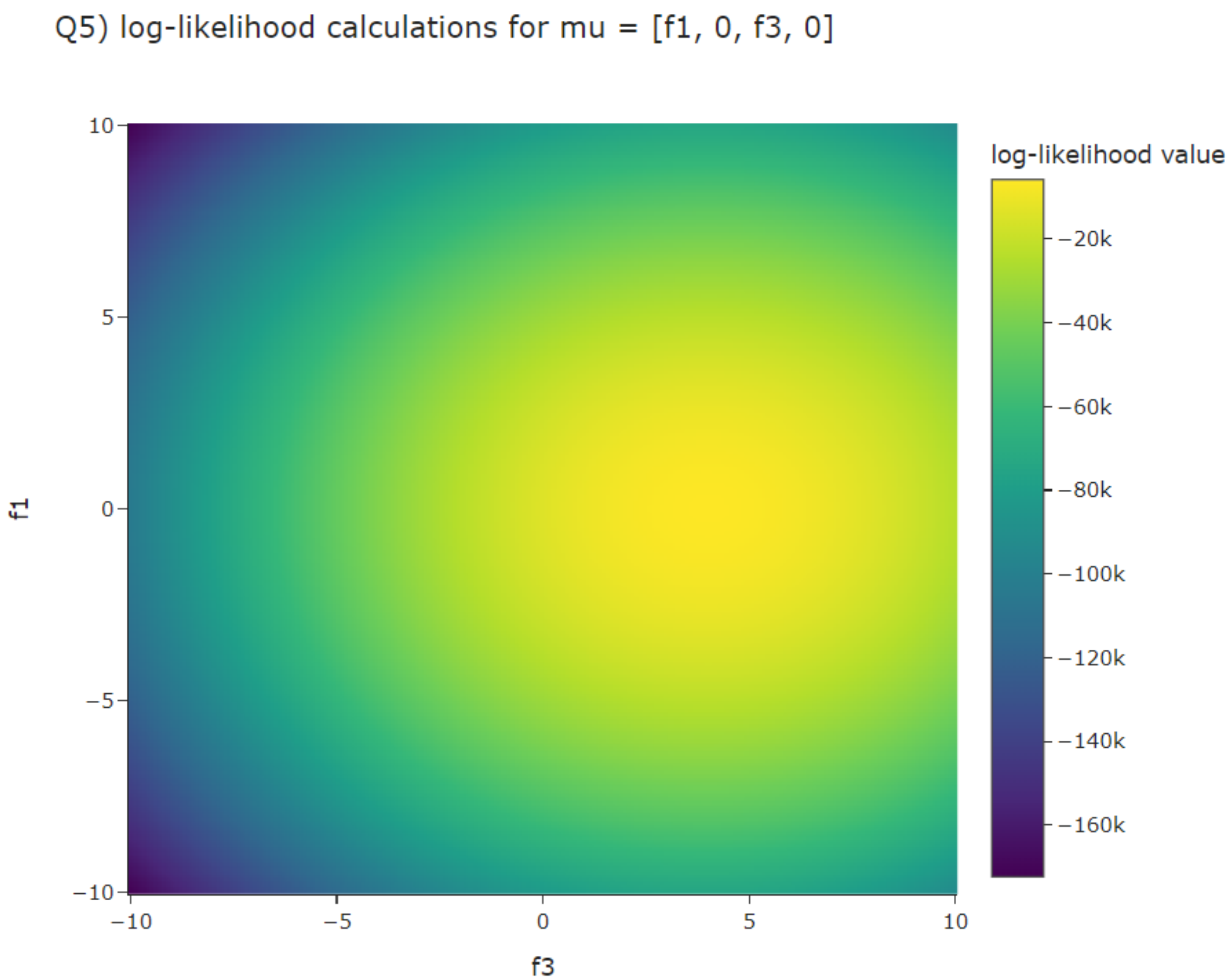


שאלה 4

הערכים של וקטור התוחלת ושל מטריצת השונות המשותפת שחושבו מודפסים בקוד כפי שכתוב בהוראות.

שאלה 5

גרף (HeatMap) המציג את ערך פונקציית ה־log-likelihood:



שאלה 6

מתוך כל הערכים שחושבו בשאלה הקודמת, המודל שהשיג ערך מקסימלי עבור פונקציית ה־log-likelihood הוא:

$$\mu = \begin{bmatrix} f_1 & 0 & f_3 & 0 \end{bmatrix} = \begin{bmatrix} -0.050 & 0 & 3.970 & 0 \end{bmatrix}$$

והערך המקסימלי שהתקבל הוא -5806.003.