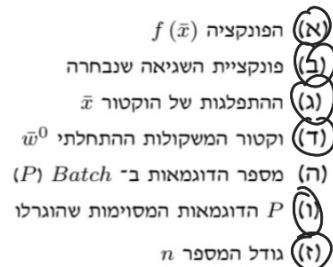


# תרגיל 3 - יותם גרדוש

שאלות הקדמה

1. נתונה פונקציה  $(\bar{x})_f$ , אשר מקבלת כקלט מושtnה מקרי רב-ימדי  $\bar{x}$ . נני פרספטרון לינארי שמנסה לשערך את הפונקציה  $(\bar{x})_f$  בעזרת אלגוריתם למידות *Batch*, עבורו פוקנציית השיאנה כלשטי ( $\bar{a}$  בהכרח ריבועית). לאחר  $a$  צעדי עדכון, הפרפסptrון למד את וקטור המשקלות  $\bar{w}$ . מה מhabאים שפע על ערכיה של שגיאת ההכללה  $(\bar{w})_g$ ?



לעומת נסלה נטולן:

לעתים קוראים בפראט שפה כפולה, כלומר שפה בה מתקיימת מילה אחת המבוצעת בשתי צורות שונות. במקרה של שפה כפולה, מילון אחד יספיק למסור כל המידע הדרוש על כל מילה.

Yesh  $\int (1/f) \times \text{voltage} \propto \text{current} \propto \text{time}$  so for Eq vcc  $\int (1/f) \times \text{voltage} \propto \text{current} \propto \text{time}$  (2). Eq 1) is for

וְאֵלֶיךָ יִתְהַלֵּךְ וְאֵלֶיךָ יִתְהַלֵּךְ כִּי תְּבִיא אֶת־  
מִזְרָחֶךָ וְאֶת־מִזְרָחֶךָ תְּבִיא אֶל־מִזְרָחֶךָ וְאֶל־מִזְרָחֶךָ תְּבִיא אֶת־מִזְרָחֶךָ

נ<sup>א</sup>ק פִּידֵּה מִלְּוַיָּה וּפִּתְּמַנְּה כְּשֶׁגַּדְתָּךְ קְרַבְתָּךְ לְעַמְּךָ וְלְמִזְרָחָךְ (ס)

γΓΩΝ ήσαν οι πρώτες πόλεις της Ελλάδας.

2. נתונות שתי פונקציות מה  $\mathbb{R}^3$  ל-  $\mathbb{R}$  שמודדרות כך:  $.g(\bar{x}) = 3x_1^2 + 2x_2 + 1$ ,  $f(\bar{x}) = x_1^2 + 2x_2 - x_3$ . מה ניתן לומר על וקטורי הגרדיאנט של שתי הפונקציות הללו בנקודה מסוימת  $\bar{x}_0$ ?  
הינו ש-  $\bar{x}_0 \neq \bar{0}$ .

- (א) זהים בכל הכניסות שלהם
  - (ב) זהים בשתי כניסה ו疏散ways בכניסה אחת
  - (ג) זהים בכניסה אחת ו疏散ways בשתי כניסה
  - (ד) שונים בכל הכניסות שלהם

నా g, f సాధించి  $\mathbb{R}^n \rightarrow \mathbb{R}$  లల్లి f సాధించి  $\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$  అనుమతి ఇప్పి

**3.** נתונה רשות נוירוגנים ארכיטקטורתית לא ידועה שמנסה ללמידה פונקצייה לא ליניארית בעזרת מידת online. מה נכון לנוכח תיכון יותר מושבנה אחת נוכנה

- (ג) ככל שהרשות רואה יותר דוגמאות כך שניית הכלכלה שלה בהכרח תקתו (שנigkeit הכלכלה היא פונקציה מונוטונית יורדת של מספר הדוגמאות).

(ב) עבור קצב למידה גדול יותר, לרשות ייקח הרבה זמן להתקרוב לוקטור משקלות אופטימי (כזה שביאם למינימום המקומי של שניית הכלכלה), וגם הפלקטואציות סביבו יהיו גדולות יותר.

(ג) נניח שאנו מאמנים את הרשות 100 פעמים שונות (בכל פעם על אלף דוגמאות) ואנו בוחנים את שניית הכלכלה. אם חשוב לנו שלא תהיה שונות גדולה בין הביצועים (שנigkeit הכלכלה) של הרשות ביראיליציות השונות, כדאי לנו לבחור קצב למידה קטן.

• **Ways to increase the efficiency of the system** - (1) Reducing friction - (2) Reducing air resistance - (3) Reducing weight - (4) Reducing heat loss

לפניהם נתקל בהנתקה (הנתקה מהנתקה)

$$\bar{a} = \begin{bmatrix} 2 \\ 4 \\ 8 \\ 16 \end{bmatrix} \text{ נתונה הפונקציה } \mathbb{R} \rightarrow \mathbb{R} \text{ אשר מוגדרת כך: } f(\bar{x}) = \|\bar{x}\|^2 + \bar{a}^T \bar{x} \text{ כאשר}$$

חלק א'

$$f(\bar{x}_0) \text{ חשבו את } \bar{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \text{ עבור 1.}$$

$$\|\bar{x}_0\| = 4 \quad | \text{ס} \quad \|x\|^2 = \sum_{i=1}^n x_i^2 \Leftarrow \|x\| = \sqrt{\sum_{i=1}^n x_i^2} \quad \text{גזרה גור}$$

$$\Rightarrow f(\bar{x}) = \|x\|^2 + \bar{a}^T x = 4 + \sum_{i=1}^4 a_i x_i = 4 + 2 + 4 + 8 + 16 = 34$$

2. חשבו את הגרדיינט של הפונקציה,  $(\bar{x}) f(\bar{x})$ , ואת ערכו של הגרדיינט בנקודה  $\bar{x}_0$ , כלומר את  $\nabla f(\bar{x}_0)$ .

$$f(\bar{x}) = x_1^2 + x_2^2 + x_3^2 + x_4^2 + 2x_1 + 4x_2 + 8x_3 + 16x_4$$

$$\frac{\partial f}{\partial x_1} = 2x_1 + 2, \quad \frac{\partial f}{\partial x_2} = 2x_2 + 4, \quad \frac{\partial f}{\partial x_3} = 2x_3 + 8, \quad \frac{\partial f}{\partial x_4} = 2x_4 + 16$$

$$\Rightarrow \nabla f(\bar{x}) = \begin{bmatrix} \partial x_1 + 2 \\ \partial x_2 + 4 \\ \partial x_3 + 8 \\ \partial x_4 + 16 \end{bmatrix} = 2x + a \Rightarrow \nabla f(\bar{x}) \Big|_{\bar{x}} = \begin{bmatrix} 2 \cdot 1 + 2 \\ 2 \cdot 1 + 4 \\ 2 \cdot 1 + 8 \\ 2 \cdot 1 + 16 \end{bmatrix} = \begin{bmatrix} 4 \\ 6 \\ 10 \\ 18 \end{bmatrix}$$

3. כעת נרצה לבדוק את השינוי בפונקציה  $f$  בעקבות תזוזה מי- $\bar{x}_0$  ב- $\bar{\varepsilon}$  (כאשר  $\bar{\varepsilon}$  וקטור של ערכים קטנים), עבור וקטורי  $\bar{\varepsilon}$  שונים שווים באותו גודל אך מצביעים לכיוונים שונים. ככלומר:

$$\Delta f = f(\bar{x}_0 + \bar{\varepsilon}) - f(\bar{x}_0)$$

הוקטורים  $\bar{\varepsilon}$  שאומרים לנו מהו השינוי בפונקציה  $f$  בעקבות תזוזה מי- $\bar{x}_0$ ?

$$\bar{\varepsilon}^1 = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.1 \\ 0.2 \end{bmatrix}, \bar{\varepsilon}^2 = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.2 \\ 0.1 \end{bmatrix}, \bar{\varepsilon}^3 = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.1 \\ 0.1 \end{bmatrix}$$

עבור כל אחד משלושת הוקטורים, חשבו את השינוי בפונקציה:  $\Delta f$ .  
רמז: ישנו איברים מסוימים ל- $\Delta f$  עבור הוקטורים השונים, אותם תוכל לחשב רק פעם אחת, ולאחר מכן מכן לחשב רק את האיברים השונים בין המקרים.

$$A = \begin{pmatrix} 2 \\ 4 \\ 8 \\ 16 \end{pmatrix}, f(\bar{x}) = \|\bar{x}\|^2 + \bar{x}^T \bar{x}$$

הנואגה  $f$

$$\begin{pmatrix} 0.1 \\ 0.2 \\ 0.1 \\ 0.2 \end{pmatrix} = \bar{\varepsilon}^1$$

$$f(\bar{x}_0 + \bar{\varepsilon}^1) - f(\bar{x}_0) = f\left(\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 0.1 \\ 0.2 \\ 0.1 \\ 0.2 \end{pmatrix}\right) - 34 = f\left(\begin{pmatrix} 1.1 \\ 1.2 \\ 1.1 \\ 1.2 \end{pmatrix}\right) - 34 = 2 \cdot 1.1^2 + 2 \cdot 1.2^2 + 1.1 \cdot 2 + 1.2 \cdot 4 + 1.1 \cdot 8 + 1.2 \cdot 16 - 34$$

$$= 2 \cdot 4.2 + 2 \cdot 8.8 + 1.1(2+8) + 1.2(4+16) - 34 = 6.3$$

$$\begin{pmatrix} 0.1 \\ 0.2 \\ 0.2 \\ 0.1 \end{pmatrix} = \bar{\varepsilon}^2$$

$$f(\bar{x}_0 + \bar{\varepsilon}^2) - f(\bar{x}_0) = f\left(\begin{pmatrix} 1.1 \\ 1.2 \\ 1.2 \\ 1.1 \end{pmatrix}\right) - 34 = 2 \cdot 1.1^2 + 2 \cdot 1.2^2 + 1.1 \cdot (2+16) + 1.2 \cdot (4+8) - 34 = 5.5$$

$$\begin{pmatrix} 0.2 \\ 0.2 \\ 0.1 \\ 0.1 \end{pmatrix} = \bar{\varepsilon}^3$$

$$f(\bar{x}_0 + \bar{\varepsilon}^3) - f(\bar{x}_0) = f\left(\begin{pmatrix} 1.2 \\ 1.2 \\ 1.1 \\ 1.1 \end{pmatrix}\right) - 34 = 2 \cdot 1.1^2 + 2 \cdot 1.2^2 + 1.1 \cdot (8+16) + 1.2 \cdot (2+4) - 34 = 4.9$$

4. עבור כל אחד מהוקטורים בסעיף הקודם, חשבו את הזווית בין הגרדיאנט בנקודה שלנו:  $\bar{\nabla} f(\bar{x}_0)$ , אותו חישבתם בסעיף ב'. נזכיר כי הזווית בין שני וקטורים  $\bar{v}, \bar{u}$ , מוגדרת כך:

$$\theta = \cos^{-1} \left( \frac{\bar{v} \cdot \bar{u}}{\|\bar{v}\| \|\bar{u}\|} \right)$$

$$\bar{v} = \nabla f(\bar{x}_0) = \begin{bmatrix} 4 \\ 6 \\ 10 \\ 18 \end{bmatrix}, \text{ אקראי } \rightarrow \text{ וקטור } \bar{v} \text{ הוא } \bar{v} = \sqrt{4^2 + 6^2 + 10^2 + 18^2} = \sqrt{476}$$

$$\Rightarrow \|\bar{v}\| = \sqrt{0.1^2 + 0.2^2 + 0.1^2 + 0.3^2} = \sqrt{0.1}$$

$$\begin{aligned} \bar{v} &= \begin{pmatrix} 0.1 \\ 0.2 \\ 0.1 \\ 0.3 \end{pmatrix} \Rightarrow \theta = \cos^{-1} \left( \frac{\begin{pmatrix} 4 \\ 6 \\ 10 \\ 18 \end{pmatrix} * \begin{pmatrix} 0.1 \\ 0.2 \\ 0.1 \\ 0.3 \end{pmatrix}}{\sqrt{476} \cdot \sqrt{0.1}} \right) = \cos^{-1} \left( \frac{0.1 \cdot 4 + 0.2 \cdot 6 + 0.1 \cdot 10 + 0.3 \cdot 18}{\sqrt{476} \cdot \sqrt{0.1}} \right) = \cos^{-1} \left( \frac{6.2}{\sqrt{476} \cdot \sqrt{0.1}} \right) \\ &= \cos^{-1}(0.898) = 26.019^\circ \end{aligned}$$

$$\begin{aligned} \bar{v} &= \begin{pmatrix} 0.1 \\ 0.2 \\ 0.2 \\ 0.1 \end{pmatrix} \Rightarrow \theta = \cos^{-1} \left( \frac{\begin{pmatrix} 4 \\ 6 \\ 10 \\ 18 \end{pmatrix} * \begin{pmatrix} 0.1 \\ 0.2 \\ 0.2 \\ 0.1 \end{pmatrix}}{\sqrt{476} \cdot \sqrt{0.1}} \right) = \cos^{-1} \left( \frac{5.4}{\sqrt{476} \cdot \sqrt{0.1}} \right) = \cos^{-1}(0.783) = 38.492^\circ \end{aligned}$$

$$\begin{aligned} \bar{v} &= \begin{pmatrix} 0.1 \\ 0.2 \\ 0.1 \\ 0.1 \end{pmatrix} \Rightarrow \theta = \cos^{-1} \left( \frac{\begin{pmatrix} 4 \\ 6 \\ 10 \\ 18 \end{pmatrix} * \begin{pmatrix} 0.2 \\ 0.1 \\ 0.1 \\ 0.1 \end{pmatrix}}{\sqrt{476} \cdot \sqrt{0.1}} \right) = \cos^{-1} \left( \frac{4.8}{\sqrt{476} \cdot \sqrt{0.1}} \right) = \cos^{-1}(0.696) = 45.915^\circ \end{aligned}$$

5. הסבירו את המגמה שהתקבלה בין הഫישים  $f$  ו- $\Delta$  שחושו בסעיף ג' לבין האזויות שחושו בסעיף ד'.

הנשאלה מ' הטענה ש $\lim_{x \rightarrow x_0} f(x) = L$  אם ורק אם  $\forall \epsilon > 0 \exists \delta > 0$  כך ש-  
 $|f(x) - L| < \epsilon$  עבור כל  $x$  אשר  $|x - x_0| < \delta$ .

6. איזו בחירה של  $\bar{c}$  (לא מותך שלושת הוקטורים הנתונים, אלא באופן כללי) תיתן את השינוי  $f$  הגדול ביותר האפשרי? (mobli לשנות את גודלו של  $\bar{c}$ ).

הנתקה מ- $\bar{x}$  מוגדרת כ-

\delta = \min\_{x \in S} d(x, \bar{x})

$\alpha$  និង  $\lambda$  ជាលំដាប់និងនៅក្នុង  $\alpha \cdot \lambda^{\max} = \nabla f(\bar{x})$  ដូច  $\alpha \in \mathbb{R}$  និង  $\lambda \in \mathbb{R}$

$$\Rightarrow \|\alpha \mathcal{E}^{\max}\| = \|\nabla f(x^0)\| \Leftrightarrow |\alpha| \|\mathcal{E}^{\max}\| = \|\nabla f(x^0)\| = |\alpha| \sqrt{0.1} = \sqrt{476} \Rightarrow \alpha = \sqrt{476}$$

אנו מודים לך על הביקור בדף זה.

$$\Rightarrow \left\{ \begin{matrix} m_0 x \\ = \end{matrix} \right. \frac{1}{\sqrt{4760}} \nabla f(\bar{x}^0) = \frac{1}{\sqrt{4760}} \begin{bmatrix} 4 \\ 6 \\ 10 \\ 18 \end{bmatrix}$$

7. רשמו בקצחה את הקשר בין שני השיעיפים הקודמים לבין האלגוריתם של למידת גרדיאנט.

1. עברו  $\bar{x}, 0, \bar{a}$  כלשהם, רשמו קירוב ליניארי (טור טילור מסדר ראשון) לערך של  $(\bar{x} + \bar{a})^f$ , כאשר  $\bar{x}$  וקטור של ערכים קבועים.

$$f(\bar{x}_0 + \bar{\varepsilon}) \approx f(x) + f'(x) \cdot \bar{\varepsilon} \quad \text{in } \mathbb{R}^n$$

$$Df(x) = Ax + b \quad \text{on } \mathbb{R}^n \quad \text{and} \quad f'(x) = Df(x) \quad \text{for} \quad f(x) = \|x\|^2 + bx$$

$$\Rightarrow f(\bar{x}_0 + \bar{\zeta}) \approx \| \bar{x}_0 \|^2 + a^T \bar{x}_0 + (2\bar{x}_0 + a) \cdot \bar{\zeta}$$

השו אותו לררך המדוקים  $f(\bar{x}_0 + \bar{\varepsilon}^1)$  על ידי הוכח בפוקציה המקורית. ורשמו את הפרש הערכבים.

(1)  $f_1(x) = \frac{1}{x}$  הינה פונקציה מוגדרת ב- $\mathbb{R} \setminus \{0\}$ .

$$\Rightarrow \left( \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}^+ \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \left( 2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix} \right) \cdot \begin{pmatrix} 0.1 \\ 0.2 \\ 0.1 \\ 0.5 \end{pmatrix} \right) = 34 + 14 \cdot 0.1 + 24 \cdot 0.2 = 40.2$$

$$\text{f}(x_0 + \bar{\epsilon}) - f(x_0) = 6.3 \rightarrow \text{ijk} \quad (3) \quad \text{f}'(0N) : f \rightarrow x \rightarrow$$

$$f(\bar{x}_1 + \bar{\epsilon}) = 6.3 + f(x_0) = 6.3 + 34 = 40.3$$

3. חזו על הסעיף הקודם עבור  $\bar{\varepsilon}^1 = 2\bar{\varepsilon}^1$ , שמקיים:  $\bar{\varepsilon}^{1 \text{ new}} = 2\bar{\varepsilon}^1$ . האם כוונת הערכה הליניארית לפונקציה טובה יותר או טובה פחותה? הסבירו.

$$\begin{aligned} \mathbf{x}^{\text{new}} &= \mathbf{x}^1 + \begin{pmatrix} 0.1 \\ 0.4 \\ 0.2 \\ 0.4 \end{pmatrix} \Rightarrow f(x_0) + \nabla f(x_0) \cdot \begin{pmatrix} 0.1 \\ 0.4 \\ 0.2 \\ 0.4 \end{pmatrix} = 34 + 14 \cdot 0.2 + 24 \cdot 0.4 = 46.4 \end{aligned}$$

$$f(\mathbf{x}_0 + \mathbf{z}^{\text{new}}) = \left\| \begin{pmatrix} 1.1 \\ 1.4 \\ 1.2 \\ 1.4 \end{pmatrix} \right\|^2 + \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{16} \\ \frac{1}{4} \end{pmatrix}^T \begin{pmatrix} 1.1 \\ 1.4 \\ 1.2 \\ 1.4 \end{pmatrix} = 2 \cdot 1.1^2 + 2 \cdot 1.4^2 + 10 \cdot 1.2^2 + 20 \cdot 1.4^2 = 46.8$$

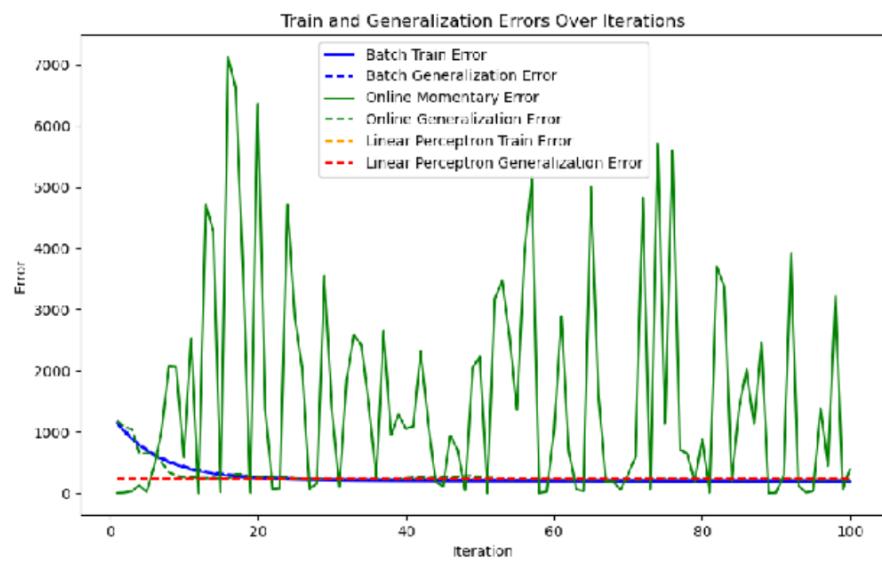
לעתה נזכיר את הערך הנוכחי  $\mathbf{z}^{\text{new}}$  בז'רנו. מילוי הערך  $\mathbf{z}^{\text{new}}$  מוביל להגדרה נסובסית של מטרית  $A$ . מילוי הערך  $\mathbf{z}^{\text{new}}$  מוביל לrang  $A$  ש�ה. מילוי הערך  $\mathbf{z}^{\text{new}}$  מוביל לrang  $A$  ש�ה.

4. רשמו בקצרה את הקשר בין שני הסעיפים הקודמים לבין האלגוריתם של למידת גראדיאנט.

הקשר בין הסעיפים הוא בכך שאלגוריתם גראדיאנט מנסה למצוא את הערך המינימלי של פונקציית האנרגיה, בעוד שאלגוריתם גראדיאנט מנסה למצוא את הערך המינימלי של פונקציית האנרגיה. אלגוריתם גראדיאנט מנסה למצוא את הערך המינימלי של פונקציית האנרגיה, בעוד שאלגוריתם גראדיאנט מנסה למצוא את הערך המינימלי של פונקציית האנרגיה. אלגוריתם גראדיאנט מנסה למצוא את הערך המינימלי של פונקציית האנרגיה, בעוד שאלגוריתם גראדיאנט מנסה למצוא את הערך המינימלי של פונקציית האנרגיה.

## 2.1

### השוואה בין האלגוריתמים השונים



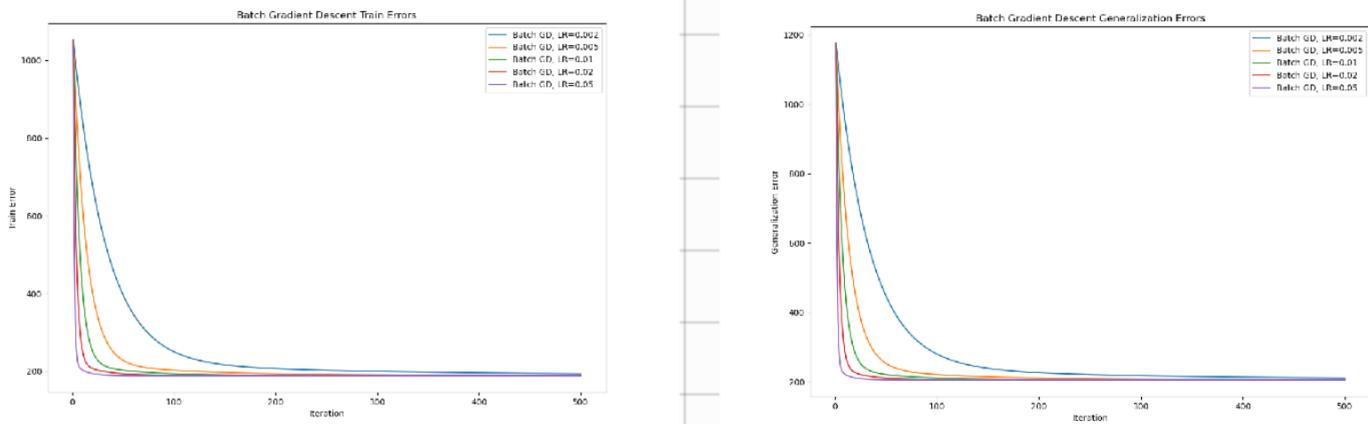
ב-**Batch** אלגוריתם ה-**Generalization Error** מוגברת ביחס ל-**Train Error**, כלומר ה-**Generalization Error** לא מוגברת אוניברסלית. סביר לנו כי ה-**Generalization Error** מוגברת רק על ידי אוסף נתונים אחד בלבד.

ב-**Online** אלגוריתם ה-**Generalization Error** מוגברת ביחס ל-**Train Error**, כלומר ה-**Generalization Error** מוגברת אוניברסלית. סביר לנו כי ה-**Generalization Error** מוגברת על ידי אוסף נתונים אחד בלבד.

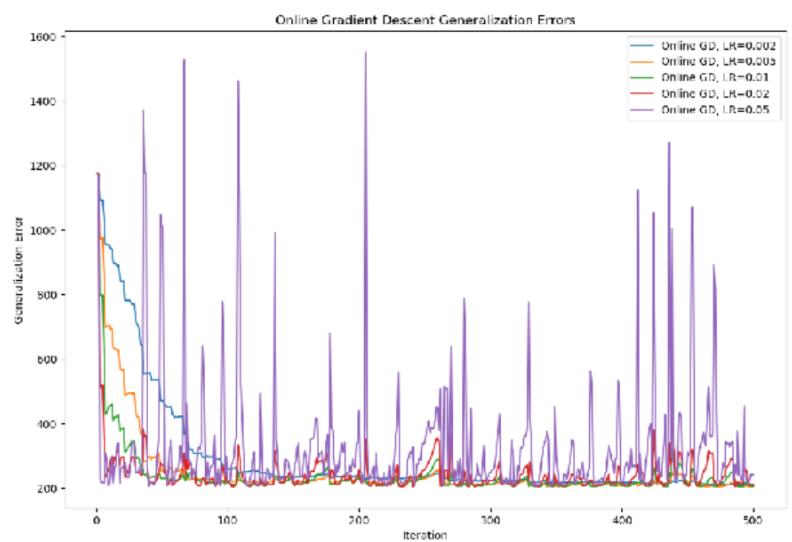
ב-**inverse cor matrix** אלגוריתם ה-**Generalization Error** מוגברת ביחס ל-**Train Error**, כלומר ה-**Generalization Error** מוגברת אוניברסלית. סביר לנו כי ה-**Generalization Error** מוגברת על ידי אוסף נתונים אחד בלבד.

ב-**Linear Perceptron** אלגוריתם ה-**Generalization Error** מוגברת ביחס ל-**Train Error**, כלומר ה-**Generalization Error** מוגברת אוניברסלית. סביר לנו כי ה-**Generalization Error** מוגברת על ידי אוסף נתונים אחד בלבד.

השפעת קצב הלמידה 2.2



הנתקה מארון הרים ורבים נתקיימו מפגשים אונליין. מפגשים אלו הובילו לפתיחת מילויים אונליין, ורבים מהתושבים נתקיימו מפגשים אונליין.



לעומת ה-**Batch** פוליפיקט, ה-**online** פוליפיקט משלב פוליפיקט ב-**הוילג'ר** ו-**הפלט** ב-

ה'גרן ז'רגן פ'ינט נ'מ'ר ו'נ'ה ו'ל'ע'ל ג'פ'ס'ה א'כ'ל'ו' ו'נ'ז'ה ו'נ'ה ג'א'ל'ג' ו'ג'ק'ן (ה'ג'ק'ן)

## чисוביות וקומבינטורית - תרגיל 3

להגשה עד: 08/02/2024

שימו לב: בתחילת התרגיל מופיעה כמה שאלות הקדמה אמריקאיות. יש להגיש את התשובות אליהן, עם משפט נימוק קצר לכל שאלה. לאחר מכן, שאלה 1 היא שאלה אנליטית ו שאלה 2 היא שאלת תכונות.

### שאלות הקדמה

1. נתונה פונקציה  $f(\bar{x})$ , אשר מקבלת כקלט משתנה מקרי רב-ממדי  $\bar{x}$ . נניח פרטוטו לינארי שמנסה לשערך את הפונקציה  $f(\bar{x})$  בעזרת אלגוריתם למידה, *Batch*, ובעור פונקציית שנייה  $\bar{x}$  (לא בהכרח ריבועית). לאחר  $n$  עדכונים, הפרטוטו למד את וקטורי המשקלות  $\bar{w}$ . מה מהבאים יופיע על ערכיה של שגיאת הכלכלה  $(\bar{w})_g$  סמנו את כל התשובות הכנות

- (א) הפונקציה  $f(\bar{x})$
- (ב) פונקציית השגיאה שנבחרה
- (ג) ההתפלגות של הוקטור  $\bar{x}$
- (ד) וקטורי המשקלות ההתחלתי  $\bar{w}^0$
- (ה) מספר הדוגמאות ב- *Batch*  $P$
- (ו) הדוגמאות המסוימות שהוגרלו
- (ז) גודל המספר  $n$

2. נתונות שתי פונקציות מרץ  $\mathbb{R}^3 \rightarrow \mathbb{R}$  שמודדות כך:  $g(\bar{x}) = 3x_1^2 + 2x_2 + 1$ ,  $f(\bar{x}) = x_1^2 + 2x_2 - x_3$ . מה ניתן לומר על וקטורי הגרדיאנט של שתי הפונקציות האלה בנקודה מסוימת  $\bar{x}_0$ ? מה הניחו ש-  $\bar{x}_0 \neq 0$ .

- (א) זרים בכל הכניסות שליהם
- (ב) זרים בשתי כניסה ושונים בכניסה אחת
- (ג) זרים בכניסה אחת ושונים בשתי כניסות
- (ד) שונים בכל הכניסות שלהם

3. נתונה רשת נוירונים ארכיטקטורה לא ידועה שמנסה ללמידה פונקציה לא ליניארית בעזרת למידה online. מה נכון לומר? תיתכן יותר מTESLA אחת נכונה

- (א) ככל שהרשת רואה יותר דוגמאות כך שגיאת הכלכלה היא פונקציה מונוטונית יורדת של מספר הדוגמאות.
- (ב) עבור קצב למידה גדול יותר, לרשות יי'יך הרבה זמן להתרחב לוקטור משקלות אופטימי (זה שambil למיינום לוקלי של שגיאת הכלכלה), וגם הפלקטואציות סיבובי היו גדלות יותר.
- (ג) נניח שאנו מאמנים את הרשת 100 פעמים שונות (בכל פעם על אלף דוגמאות) ואז בוחנים את שגיאת הכלכלה. אם חשוב לנו שלא תהיה שונות גדולה בין הביצועים (שגיאות הכלכלה) של הרשת בראיליזציה השונה, כדאי לנו לבחור קצב למידה קבוע.

$$\bar{a} = \begin{bmatrix} 2 \\ 4 \\ 8 \\ 16 \end{bmatrix} \quad \text{notונה הפונקציה } f(\bar{x}) = \|\bar{x}\|^2 + \bar{a}^T \bar{x}, \text{ אשר מוגדרת כך: } f(\bar{x}) : \mathbb{R}^4 \rightarrow \mathbb{R}$$

חלק א'

$$.f(\bar{x}_0) \text{ חשבו את } \bar{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \text{ עבר}$$

. $\bar{\nabla} f(\bar{x}_0)$ , ואת ערכו של הגרדיינט בנקודה  $\bar{x}$ , כולל את (2).

3. כעת נרצה לבדוק את השינוי בפונקציה  $f$  כתוצאה מזיהום מ- $\bar{x}$ .  
 שהם באוטו גודל אך מצביעים לכיוונים שונים. ככלומר:

$$\Delta f = f(\bar{x}_0 + \bar{\varepsilon}) - f(\bar{x}_0)$$

הוקטורים  $\vec{e}$  שאוטם נבחן הם (שים לב שגם מבנים מדויקים הם באותו הגודל אך בכיוונים שונים):

$$\bar{\varepsilon}^1 = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.1 \\ 0.2 \end{bmatrix}, \bar{\varepsilon}^2 = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.2 \\ 0.1 \end{bmatrix}, \bar{\varepsilon}^3 = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.1 \\ 0.1 \end{bmatrix}$$

עבורו כל אחד משלוחות הוקטורים, חשבו את השינוי בפונקציה:  $\Delta f$ .  
 רמז: ישים איברים מסווגפים ל- $\Delta f$  עבור הוקטוריים השונים, אוטם תוכלו לחשב רק פעם אחת, ולאחר מכן מכן לחשב רק את האיברים השונים בין המקרים.

4. עבור כל אחד מהוקטורים בסעיף הקודם, חשבו את האזוטה ביןו לבין הגרדיינט בנקודה שלנו:  $(\bar{x}_0 f)'$ , אותו חישבתם בסעיף ב'. נזכיר כי האזוטה בין שני וקטורים  $\bar{u}$ ,  $\bar{v}$ , מוגדרת כך:

$$\theta = \cos^{-1} \left( \frac{\bar{v} \cdot \bar{u}}{||\bar{v}|| ||\bar{u}||} \right)$$

5. הסבירו את המגמה שהתקבלה בין הפרשנים  $f$  ו- $\Delta$  שחושו בסעיף ג' לבין הזרויות שהושבו בסעיף ד'.

6. איזו בירה ש-ל-ן (לא מותך לשולש הוקטורים הנתונים, אלא באופן כללי) תיתן את השינוי  $\Delta f$  הגדול ביותר האפשר? (ambil לשנות את גודלו של  $\Delta$ ).

7. רשמו בקצרה את הקשר בין שני הטעיפים הקודמים לבין האלגוריתם של למידת גרדיאנט.

חלק ב'

1. עברור  $\bar{x}, 0, \bar{a}$  כלשהם, רשמו קירוב ליניארי (טור טילול מסדר ראשון) לערך של  $(\bar{e} + 0(\bar{x}), f)$ , כאשר  $\bar{e}$  וקטור של ערכאים קבועים.

$$\bar{\varepsilon}^1 = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.1 \\ 0.2 \end{bmatrix} \quad \text{היעזר בקירוב הליניארי שכתבתם בסעיף הקודם וחשבו את הקירוב לערך של } f(\bar{x}_0 + \bar{\varepsilon}^1), \text{ עבור } \bar{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \text{ ו.}$$

השו אותו לערך המדוק  $f(\bar{x}_0 + \bar{\varepsilon}^1)$  על ידי הצבה בפונקציה המקורית, ורשמו את הפרש הערכיהם.

3. חזרו על הסעיף הקודם עבור  $\bar{\varepsilon}^{new}$ , שמיים:  $\bar{\varepsilon}^1 = 2\bar{\varepsilon}^{new}$ . האם CUT הינה נכונה לפונקציה טוביה יותר או טוביה פחות? הסבירו.

4. רשמו בקצרה את הקשר בין שני הסעיפים הקודמים לבין האלגוריתם של למידת גרדיאנט.

בשאלה זו תשוו בין מספר אלגוריתמי למידה שהוצעו בשיעור לפתרון של בעיית למידה מפוקחת. הפונקציה שננסה למדוד היא:

$$y = 1 + x + x^2 + x^3$$

כאשר  $(5, 5) \sim x$ , כלומר  $x$  מתפלג בהתפלגות איחידה רציפה בין  $5 - \bar{x}$ . בכל הסעיפים הבאים, עליים ללמידה פרטפרון לינארי עם סוף - קלומר וקטור המשקלות  $\bar{w}$  יהיה דומימדי, וקטור האינפוט  $\bar{x}$  יהיה דו-ימדי מהצורה  $\bar{x} = \begin{bmatrix} x \\ 1 \end{bmatrix}$ . שימוש לבן העזרות שבסוף השאלה.

## 2.1 השוואת בין האלגוריתמים השונים

ראשית, צרו 100 דוגמאות מההתפלגות הנתונה, בהן השתמשו בכל אחד מהסעיפים הבאים (**אותן דוגמאות** לכל האלגוריתמים). בסעיף זה השתמשו בקצב לימוד  $\eta = 0.01$ .

- **למידת גרדיאנט Batch:** כתבו פונקציה שimplements את אלגוריתם למידת Batch. הריצו את האלגוריתם 100 פעמיים עדכון, וחשבו את שגיאת האימון ואת שגיאת הכלכלה לאחר כל עדכון (ראו הערבה בסוף על חישוב שגיאת הכלכלה).
- **למידת גרדיאנט Online:** כתבו פונקציה שimplements את אלגוריתם למידת Online. הריצו את האלגוריתם 100 פעמיים עדכון (כך של דוגמא מוצגת פעמיות אחת בדיקון), וחשבו את שגיאת הרגעים ואת שגיאת הכלכלה לאחר כל עדכון.
- **היפוך מטריצת הקורלציה:** מצאו את הפתרון שמתקיים  $U' \bar{w} = C\bar{w}$ , וחשבו את שגיאת האימון ושגיאת ה经济学家.

2. הציגו את התוצאות מההעapr הקודם על גרף אחד, כולם החיגו את שגיאת האימון (או השגיאת הרגעים) ואת שגיאת ה经济学家 של מספר עדכון לא שכוח להסיף מקרה ברור (גרף). דנו בהבדלים בין הביצועים והמאפיינים של האלגוריתמים השונים, והסבירו את התוצאות. (עבור אלגוריתם למידה שלא מבצע 'עדכון', כולם שיטת היפוך מטריצת הקורלציה, הציגו קווים קבועים עבור שגיאת האימון וה经济学家 המקבילים לציר הזמן של עדכון).

## 2.2 השפעת קצב הלמידה

1. בסעיף זה נרצה לראות את ההשפעה של קצב הלמידה על אלגוריתמי למידת גרדיאנט. צרו 500 דוגמאות מההתפלגות הנתונה. עבור כל אחד מקבבי הלמידה הבאים:  $\eta = 0.002, 0.005, 0.01, 0.02, 0.05$ , השתמשו בדוגמאות שיצרתם:

- **למידת גרדיאנט Batch:** עבור כל קצב לימוד, הריצו את האלגוריתם 500 פעמיים עדכון, וחשבו את שגיאת האימון ואת שגיאת ה经济学家 לאחר כל עדכון.
- **למידת גרדיאנט Online:** עבור כל קצב לימוד, הריצו את האלגוריתם 500 פעמיים עדכון (כך של דוגמא מוצגת פעמיות אחת בדיקון), וחשבו את שגיאת ה经济学家 לאחר כל עדכון.

2. צרו 3 גרפים:

- גרף שמציג את שגיאת האימון של למידת גרדיאנט batch כפונקציה של מספר עדכון, בחלוקת לקצביה הלמידה השונים.
- גרף שמציג את שגיאת经济学家 של למידת גרדיאנט batch כפונקציה של מספר עדכון, בחלוקת לקצביה הלמידה השונים.
- גרף שמציג את שגיאת经济学家 של למידת גרדיאנט online כפונקציה של מספר עדכון, בחלוקת לקצביה הלמידה השונים.

3. דנו בהשפעה של קצב הלמידה על השגיאות בשני האלגוריתמים. האם קיבלתם מה שציפיתם? מהו החסרון ומהו הייתرون של קצב לימוד?

### הערות:

- על מנת לחשב את שגיאת经济学家 בדיקון, חשבו את ממוצע השגיאה על פני כל התחים והנתון,  $5 - \bar{x}$  עד  $5$ , בקפיצות של 0.01. אמנם זהה לא שגיאת经济学家 בדיקון, אבל זהו קירוב טוב.

- עבור אלגוריתמי גרדיאנט, אתחלו את  $\bar{w}$  באופן שרירותי (למשל  $\begin{bmatrix} 1 \\ 1 \end{bmatrix} = w$ , או הגרילו ערכים באקראוי). השתמשו באותו סט של נתונים שני אלגוריתמי הלמידה, וגם עבור ההשוואה בין קצבים הלמידה השונים. בambilים אחרים, לאורך כל התרגילים התכנוניים השתמשו באותו סט ערכים התחלתיים של  $\bar{w}$ .