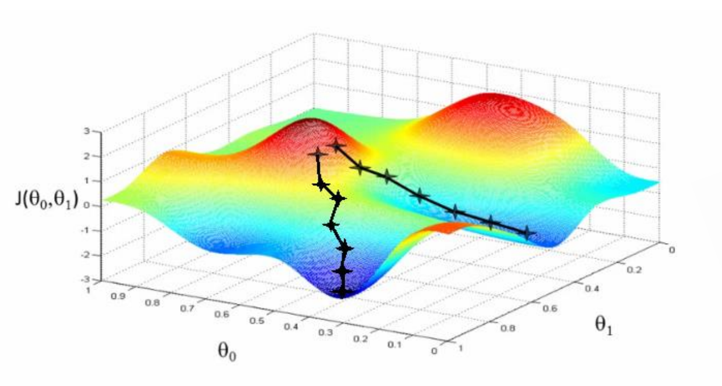
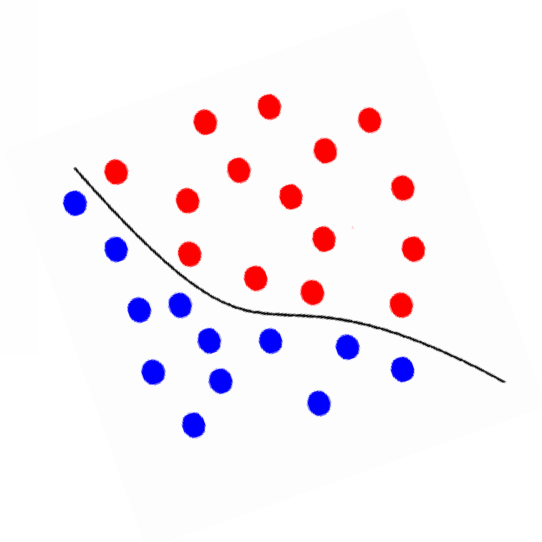
לא לשכוח להפעיל הקלטה!

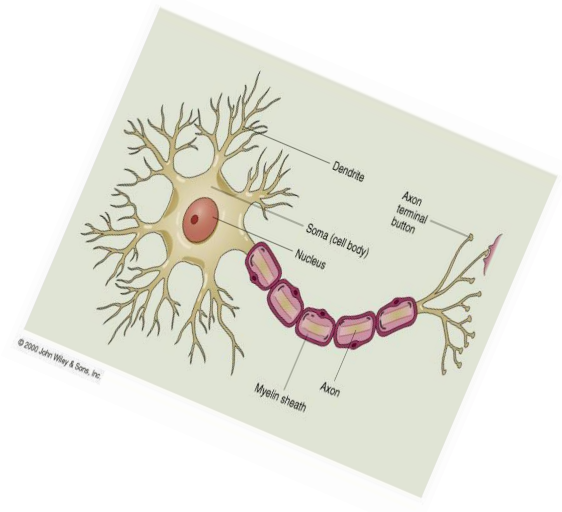
מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

מכון דוידסון לחינוך מדעי
DAVIDSON INSTITUTE OF SCIENCE EDUCATION

Gradient descent







שיעור סינכרוני

מה אנחנו צריכים כבר לדעת?

- ✓ Python – הגדרת משתנים, מטריצות, וקטורים, פונקציות מובנות, כתיבת פונקציות, גרפים
- ✓ מושגי יסוד בלמידת מכונה
- ✓ למידה מונחית Supervised learning
- ✓ למידה לא מונחית Unsupervised learning

Gradient descent

Gradient descent היא שיטה הנעזרת בגרדיאנט (מידע מקומי) על מנת למצוא את המינימום של פונקציה.

קישור

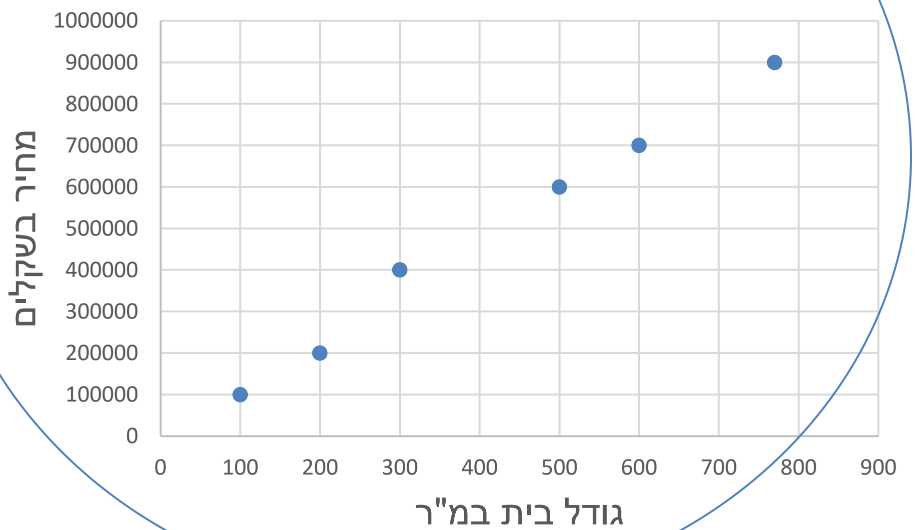
https://towardsdatascience.com/a-visual-explanation-of-gradient-descent-methods-momentum-adagrad-rmsprop-adam-f898b102325c

למידה מונחית - מאפיינים

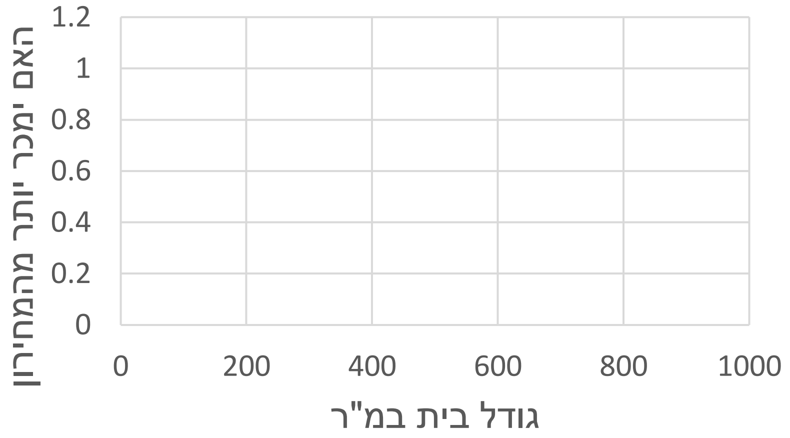
התשובה הנכונה ניתנת – Data Set

יש קשר בין משתני ה INPUT למשתנה ה OUTPUT

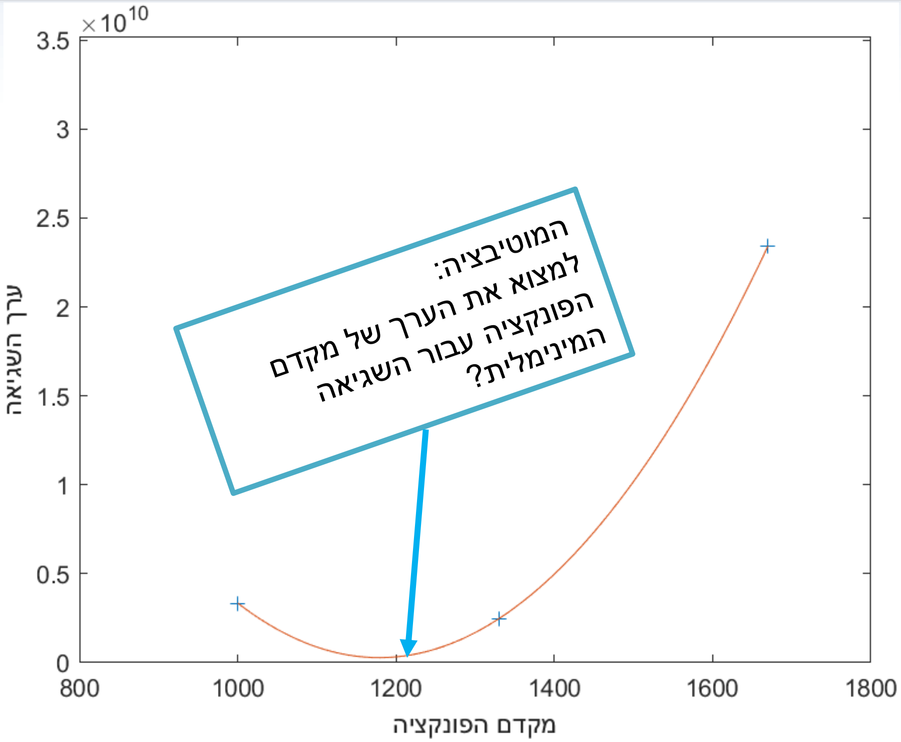
רגרסיה
Regression
פלט רציף

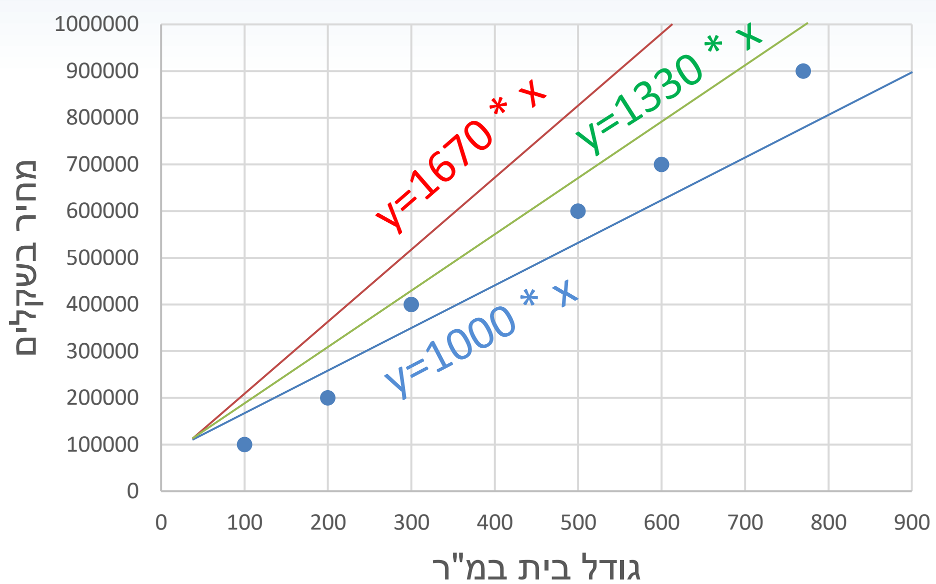


סיווג
classification
פלט - קטגוריות



תזכורת – רגרסיה לינארית

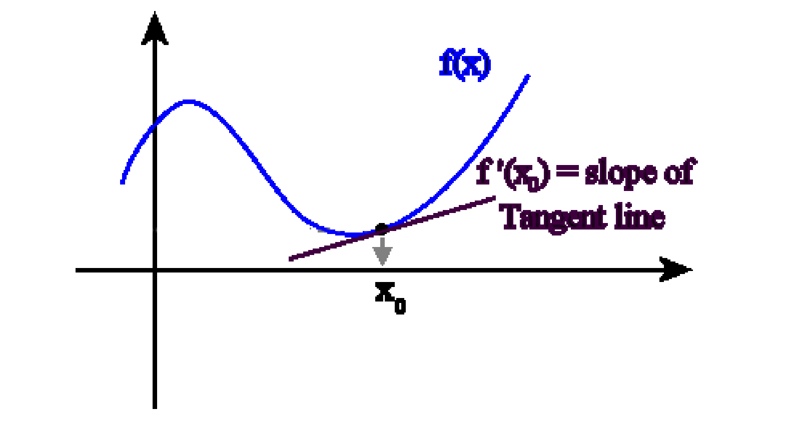


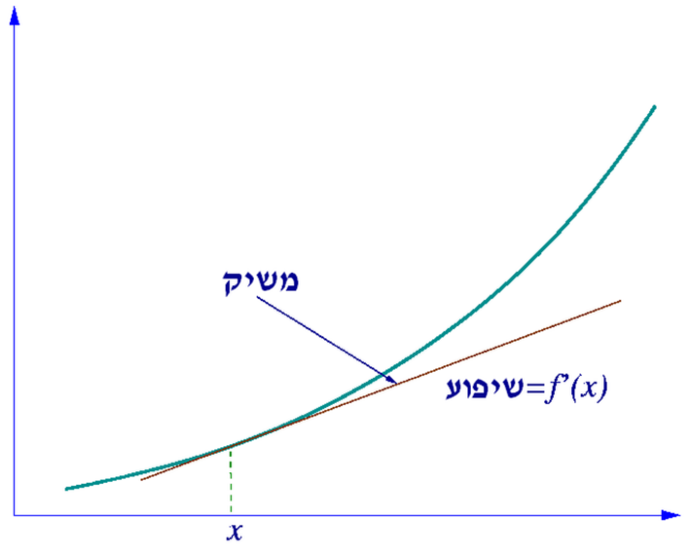


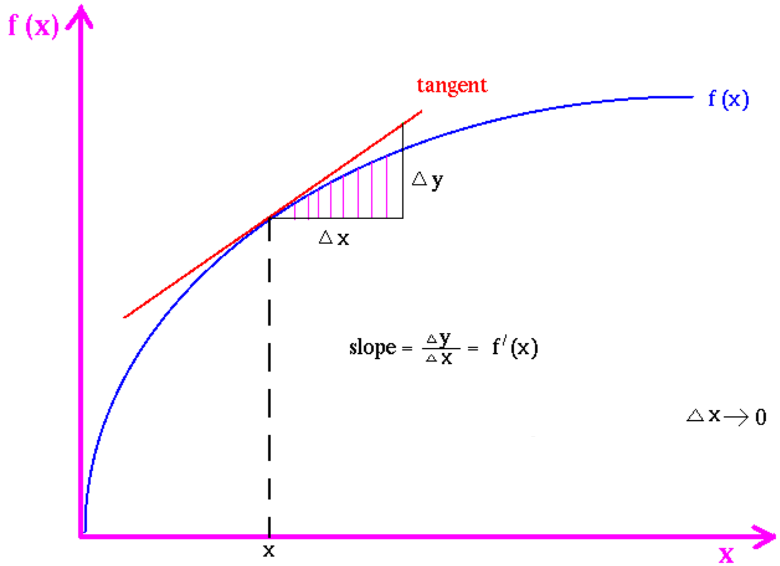
- ❑ מחשבים עבור כל קו את ערך הטעות
- ❑ ערך **הטעות** – סכימה של כל ריבועי ההפרשים בין הערך הרצוי למצוי לחלק ל 2 לחלק למספר הדגימות
- ❑ משרטטים גרף של הטעות כפונקציה של המקדם: (1000,1330,1670)

רקע מתמטי – Gradient descent

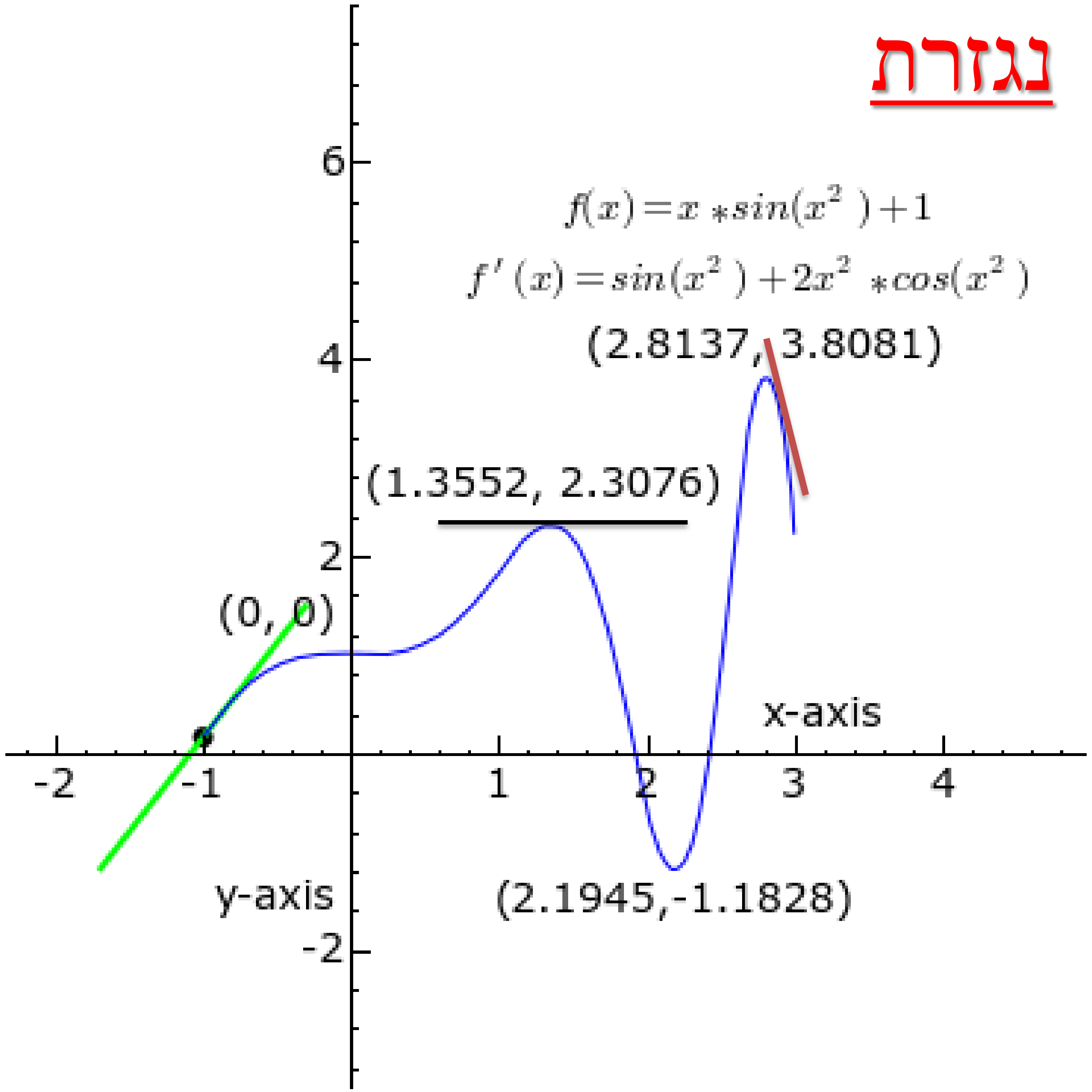
- נגזרת







נגזרת

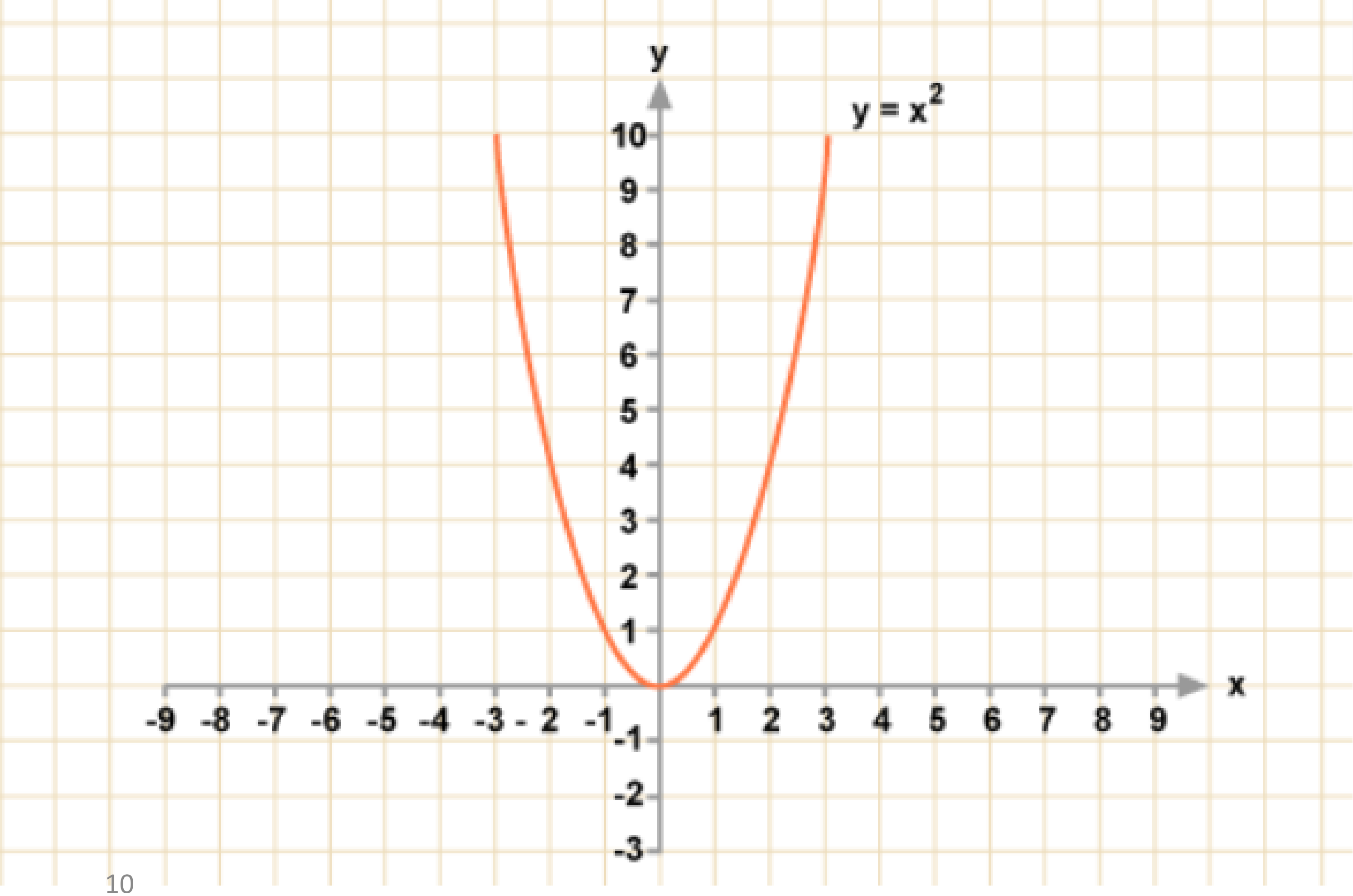


$f(x) = x * sin(x^2) + 1$
$f'(x) = sin(x^2) + 2x^2 * cos(x^2)$
(2.8137, 3.8081)
(1.3552, 2.3076)
(0, 0)
x-axis
y-axis
(2.1945,-1.1828)
6
4
2
-2
-2
-1
1
2
3
4

'נגזרת descent'

- על מנת להגיע למינימום של הפונקציה f(x):
  - נתחיל מערך x רנדומלי.
  - עבור מספר צעדים מסוים:
    - נחשב את הנגזרת של הפונקציה בנקודה זו.
    - אם הנגזרת חיובית – נקטין את x (נוסיף ל-x את מינוס הנגזרת).
    - אם הנגזרת שלילית – נגדיל את x (נוסיף ל-x את מינוס הנגזרת).
- כלומר – נלך בכיוון ההפוך לנגזרת.
- המטרה – למצוא את המשתנה x בנקודת המינימום של הפונקציה.

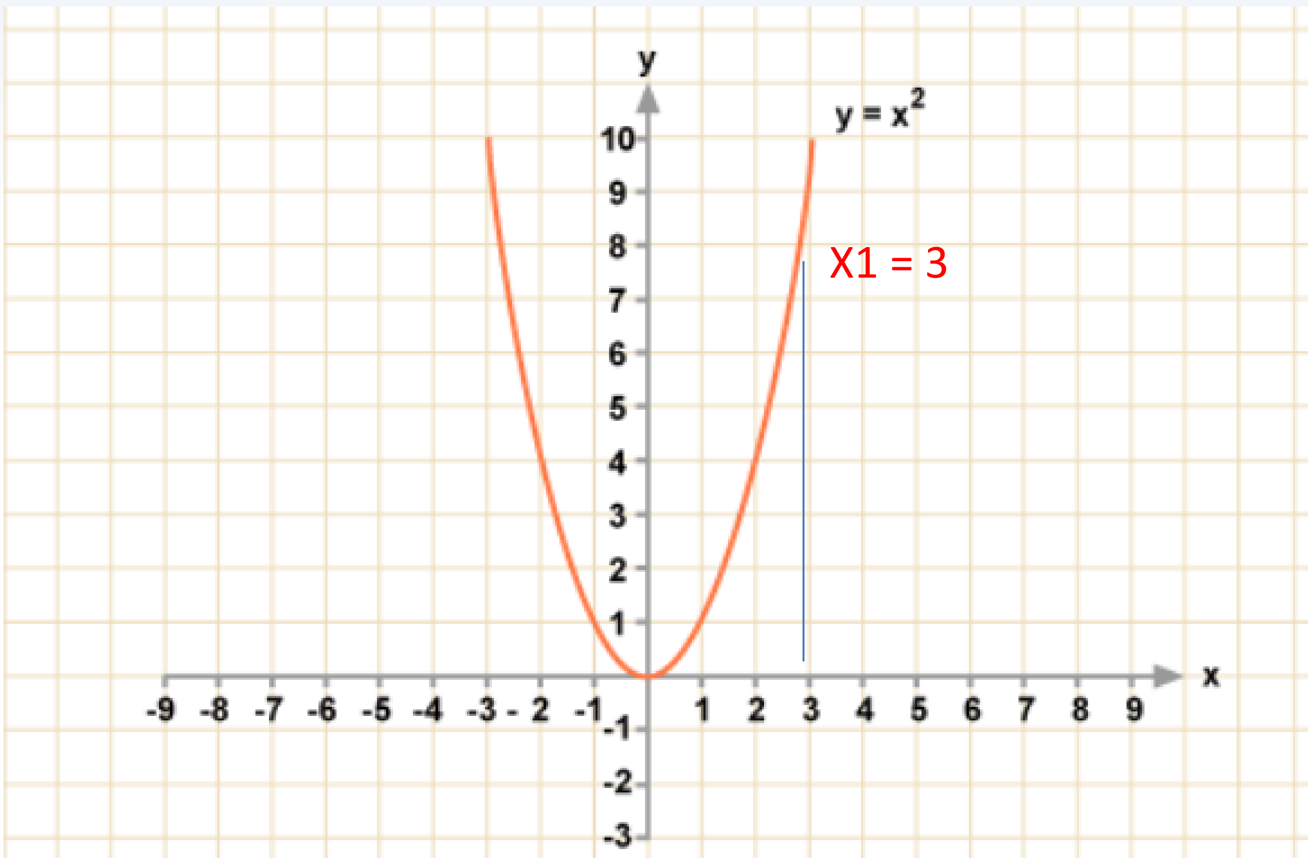
נתונה הפונקציה   y = x^2
איך נמצא את המינימום בשיטת Gradient descent



- שלב 1 – איתחול משתנים:

נבחר נקודה רנדומלית, למשל X1 = 3

נבחר את גודל הצעד (Learning Rate) lr=0.01



- שלב 2 - שינוי בערכי X עד שמגיעים לערך X בנקודת המינימום:

לפי החישוב: X(i+1) = X(i) – lr * dy/dx(i)

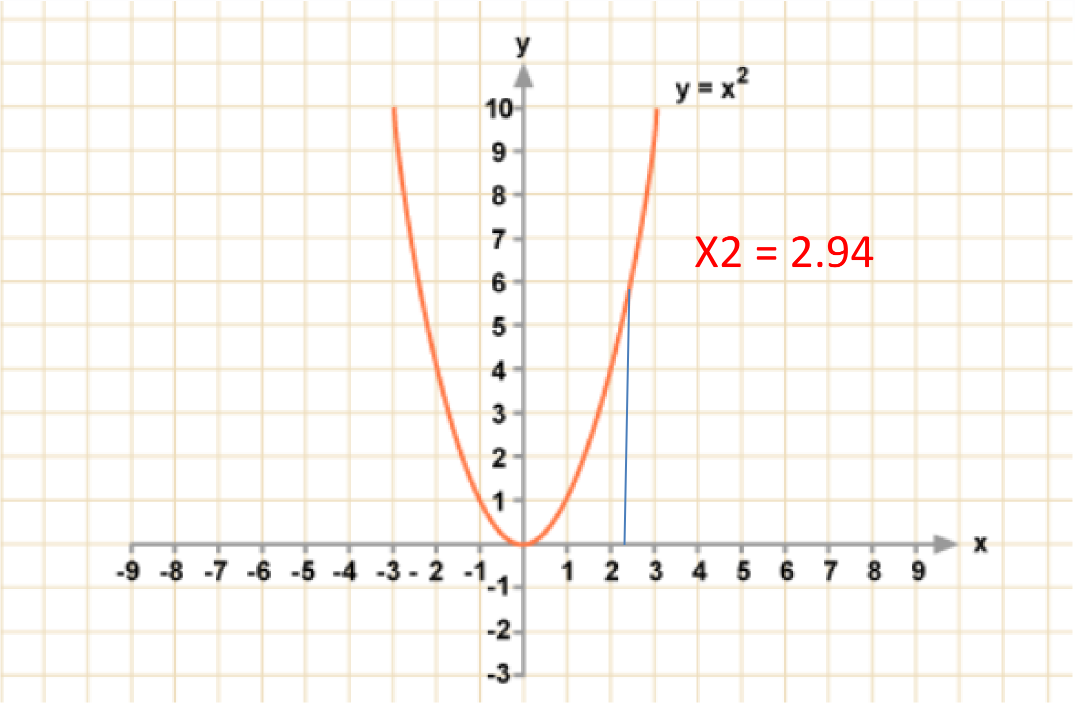
איטרציה #1

– שלב 2.1 - חישוב הנגזרת עבור X1

dy/dx=2*X1=2*3=6

– שלב 2.1 – חישוב הערך של X2

X2=X1-0.01*6=3-0.01*6=2.94



- שלב 2 - שינוי בערכי X עד שמגיעים לערך X בנקודת המינימום:

לפי החישוב: X(i+1) = X(i) – lr * dy/dx(i)

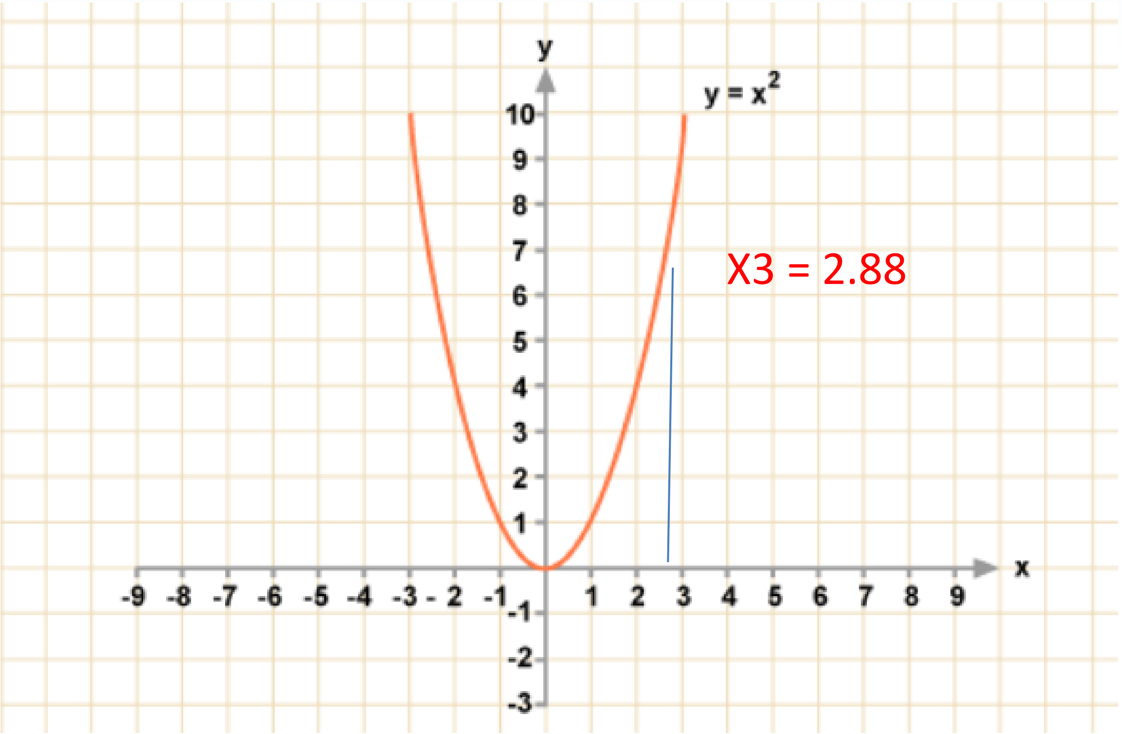
איטרציה #2

– שלב 2.1 - חישוב הנגזרת עבור X2

dy/dx=2*X2=2*2.94=5.88

– שלב 2.1 – חישוב הערך של X2

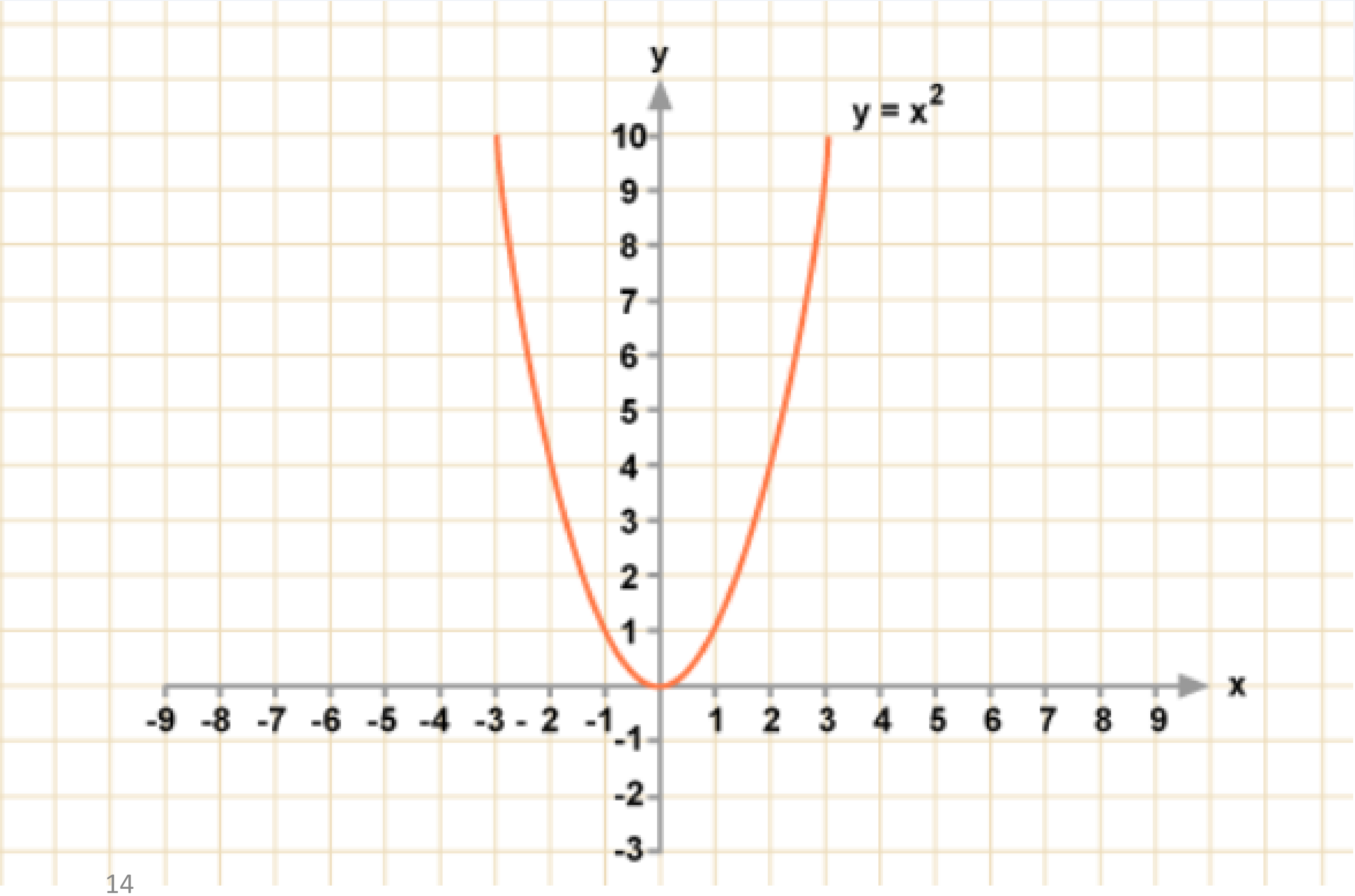
X3=X2-0.01*5.88=2.94-0.01*5.88=2.88



נתונה הפונקציה $y = x^2$

עד מתי נבצע את האיטרציות?

על מה משפיע משתנה ה Learning Rate?



תרגול

א. מצא את נקודת המינימום עבור

הפונקציה: $f(x) = x^2 - 10x + 5$

הנגזרת - 2x-10

ב. שרטטו את הפונקציה

ג. על אותה מערכת צירים שרטטו את

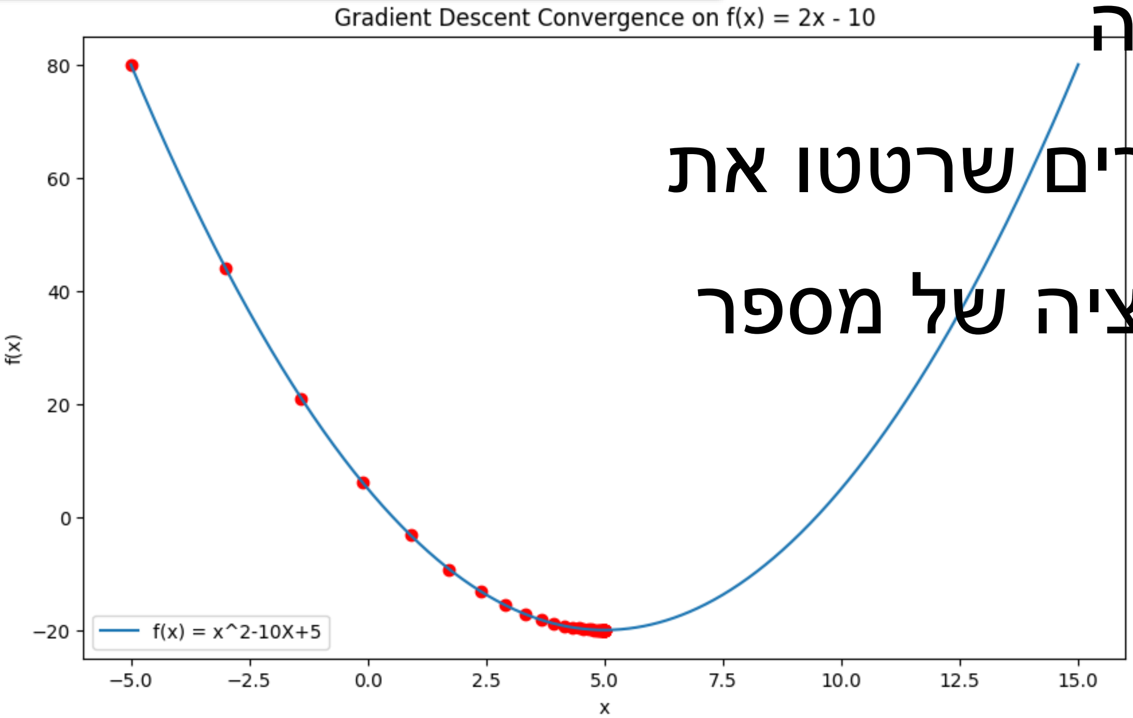
ההשתנות של x כפונקציה של מספר

האיטרציות

מקסימום האיטרציות 40

lr=0.1

נקודה התחלתית X1 = -5



תרגול

שאלות:

- ❑ מהי נקודת המינימום שהתקבלה?_________________
- ❑ האם היא נקודת המינימום הצפויה?________________
- ❑ מה קורה לערכי המשתנה x מהרגע שהגעתם/ן לנקודת המינימום? (קטנים/גדלים/לא משתנים)_____________ מדוע?_________________________________
- ❑ מה קרה כאשר קצב הלמידה היה קטן? האם הגעתם/ן לנקודת המינימום של הפונקציה? מדוע? האם ניתן להגיע עם קצב למידה זה למינימום של הפונקציה?_____________________________
- ❑ מה קרה כאשר קצב הלמידה היה גדול? האם הגעתם/ן לנקודת המינימום של הפונקציה? מדוע? האם ניתן להגיע עם קצב למידה זה למינימום של הפונקציה? _____________

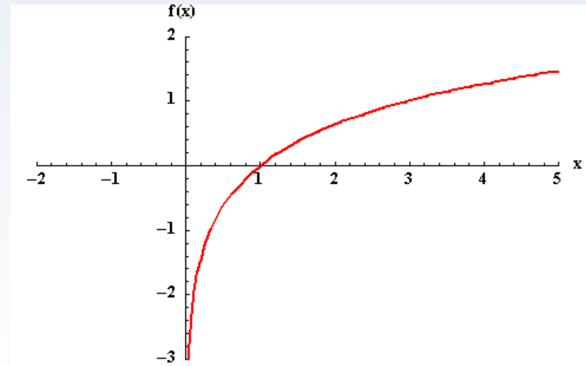
____

שאלות:

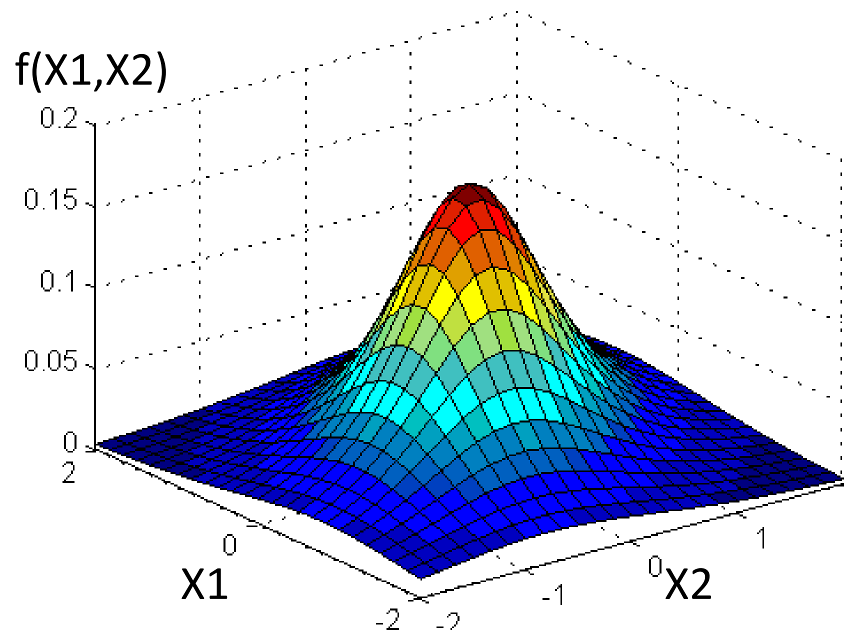
- ❑ מהי נקודת המינימום שהתקבלה? **5**
- ❑ האם היא נקודת המינימום הצפויה? **5**
- ❑ מה קורה לערכי המשתנה x מהרגע שהגעתם/ן לנקודת המינימום? (קטנים/גדלים/**לא משתנים**)
מדוע? **הנגזרת בנקודה הזו שואפת לאפס**
- ❑ שנו את קצב הלמידה: learning rate=0.001. האם הגעתם/ן לנקודת המינימום של הפונקציה? מדוע? האם ניתן להגיע עם קצב למידה זה למינימום של הפונקציה? **לא מגיעים**
למינימום קצב הלמידה קטן מדי למספר האיטרציות
- ❑ שנו את קצב הלמידה: learning rate=1. האם הגעתם/ן לנקודת המינימום של הפונקציה? מדוע? האם ניתן להגיע עם קצב למידה זה למינימום של הפונקציה? **לא מגיעים**
למינימום קצב הלמידה גדול, כל איטרציה מדלגת על מיקום המינימום

פונקציה מרובת-משתנים

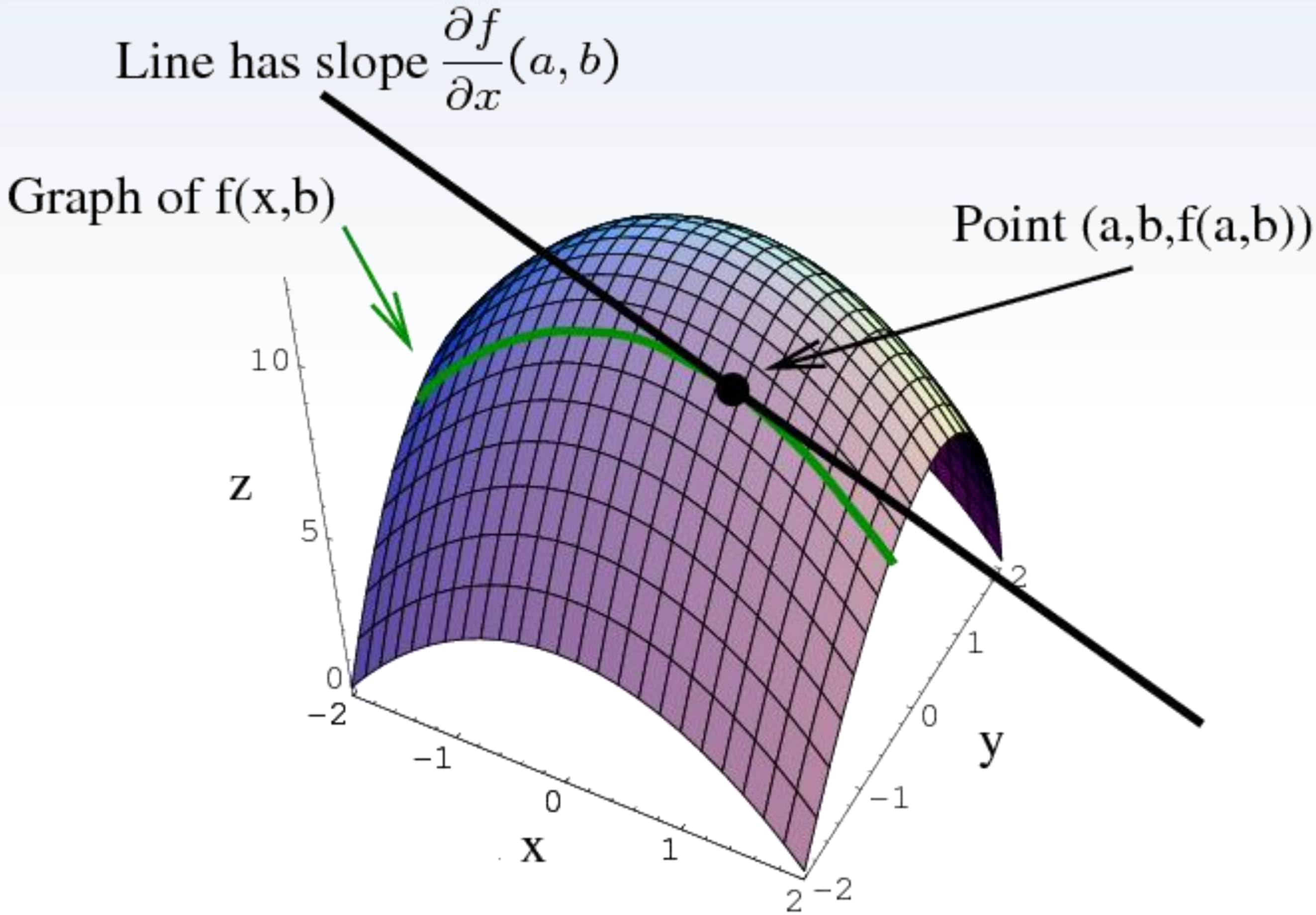
$$f(x)$$



$$f(x_1, ..., x_n)$$



נגזרת חלקית


Line has slope $\frac{\partial f}{\partial x}(a,b)$
Graph of f(x,b)
Point (a,b,f(a,b))
10
5
0
z
−2
−1
0
1
2
x
−2
−1
0
1
2
y

נגזרת חלקית

- נגזרת חלקית - גזירת פונקציה מרובת משתנים לפי אחד המשתנים.
- המשתנים האחרים נחשבים קבועים עבור הגזירה.

$$z = f(x, y) = x^2 + xy + y^2$$

$$\frac{\partial z}{\partial x} = 2x + y$$

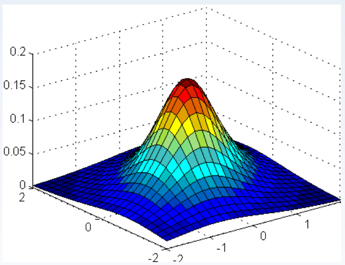
גרדיאנט

- אם f היא פונקציה גזירה של n משתנים, הגרדיאנט הוא **וקטור** באורך n של הנגזרות החלקיות של f.

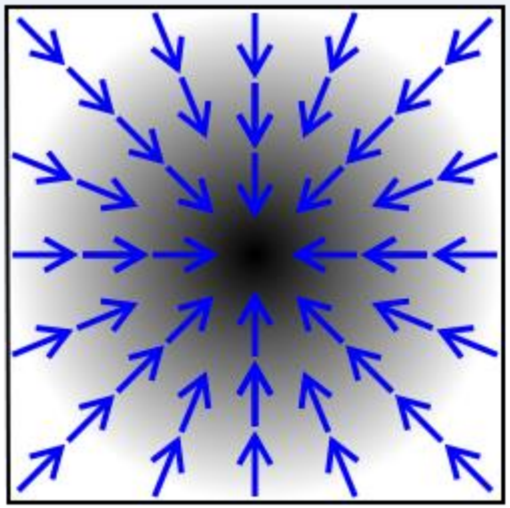
$$f(x_1, \dots, x_n)$$

$$\left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right)$$

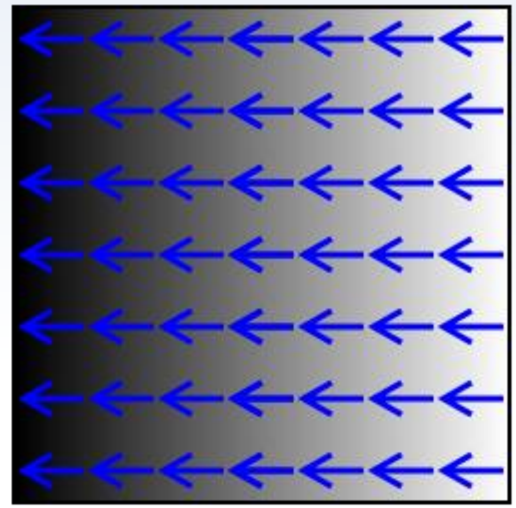
גרדיאנט - משמעות גיאומטרית



X2

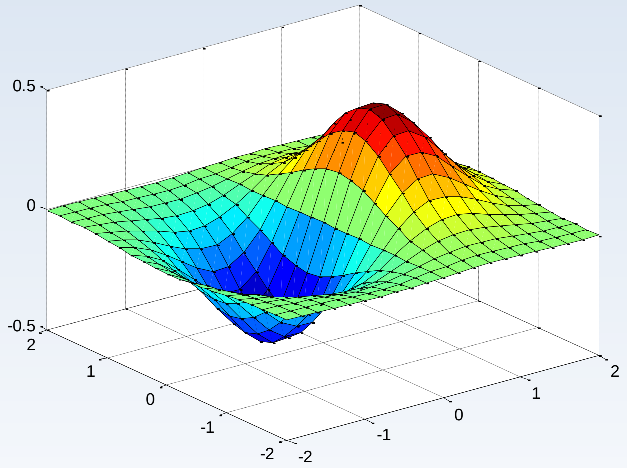


X1

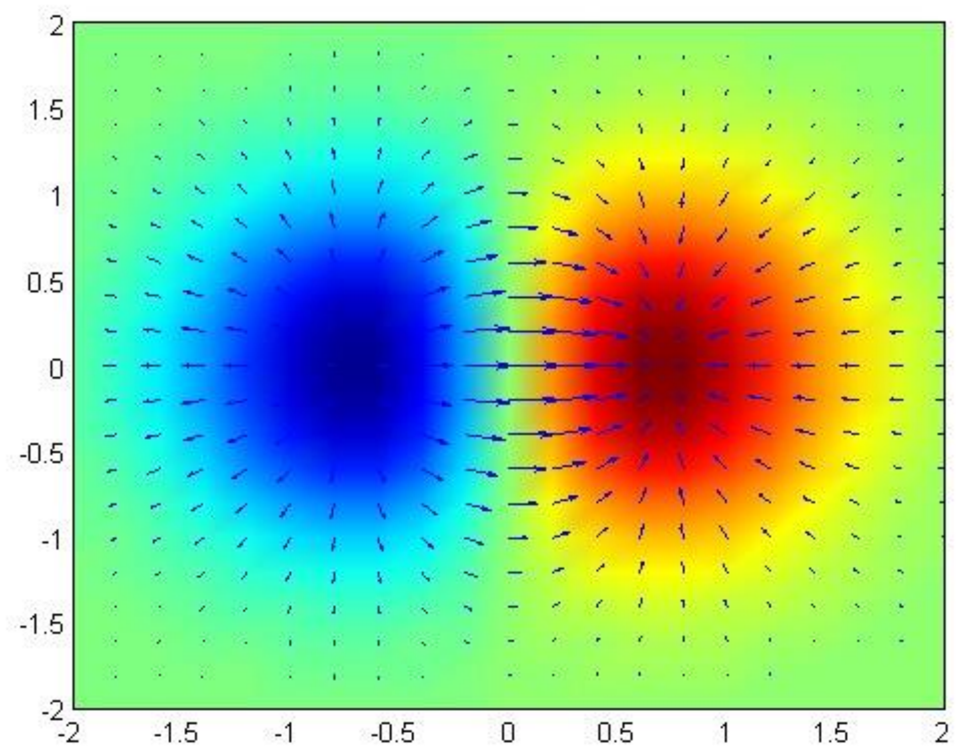


ניתן לחשב את ערך הגרדיאנט של הפונקציה בכל נקודה ונקודה (עבור כל ערך של משתנים).

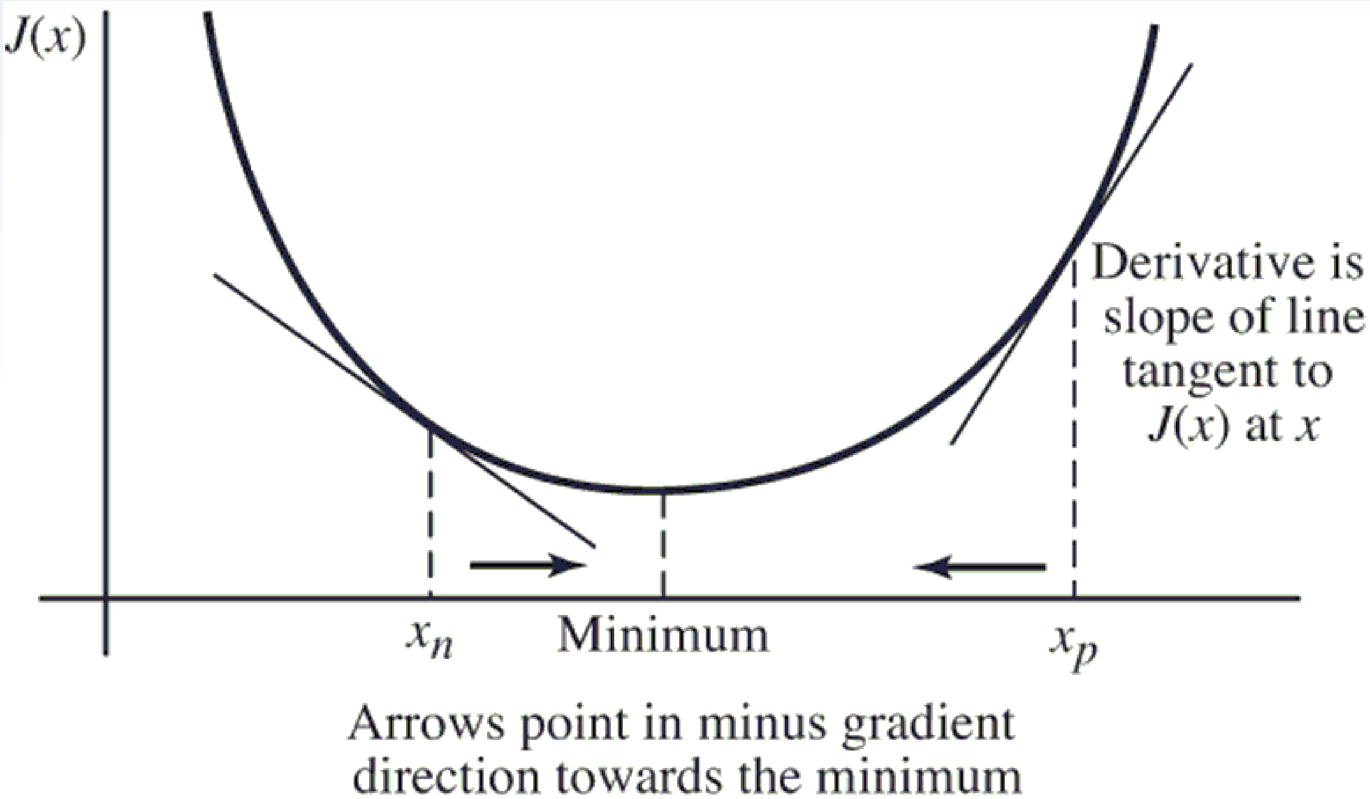
גרדיאנט



- הגרדיאנט מצביע לכיוון הגידול המקסימלי של הפונקציה.
- הליכה בכיוון הגרדיאנט תוביל להגדלת ערכי הפונקציה. הליכה נגד כיוון הגרדיאנט תוביל להקטנת ערכי הפונקציה.
- מעשית – עבור הפונקציה מרובת המשתנים, נשנה כל משתנה בהתאם לנגזרת החלקית שלו.



Gradient descent



כלל העדכון של המשקלות לפי gradient descent

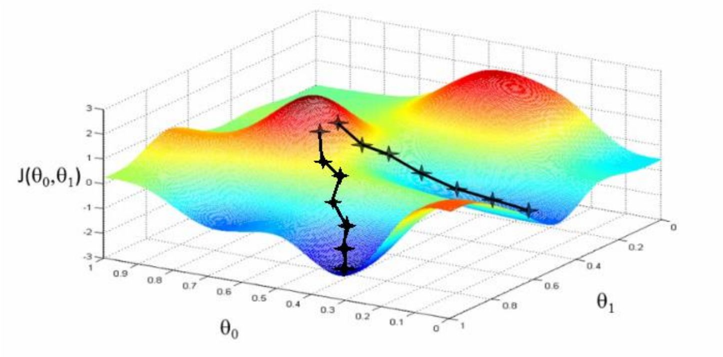
- על מנת לרדת בכיוון ההפוך לגרדיאנט, נעדכן <u>**כל משתנה**</u> בגודל השווה למינוס הנגזרת החלקית של פונקצית השגיאה ביחס למשקולת זו:

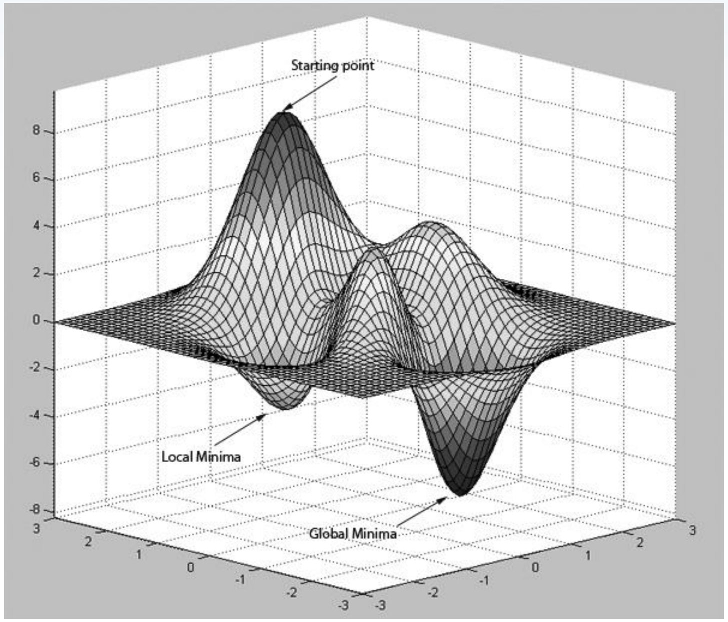
$$\Delta x_i = \varepsilon * -\frac{\partial f}{\partial x_i} \qquad \Delta \vec{x} = \varepsilon * - gradient$$

- גודל הצעד בכיוון ההפוך לכיוון הגרדיאנט נקבע ע"י קצב הלמידה (**learning rate - ε**).
- העדכון יתבצע בכל המשתנים **בבת אחת** (לא נעדכן משתנה אחד ואז נחשב שוב את הגרדיאנט ונעדכן את השני וכו', אלא באותה נקודה נחשב את הגרדיאנט עבור כל המשתנים ונעדכן את כולם).
- **נגדיר מספר מקסימלי של צעדים**.
- נסו לכתוב למה שווה וקטור המשתנים החדש בכתיבה וקטורית.

Gradient descent

- קיימת סכנה להתכנסות למינימום לוקאלי.

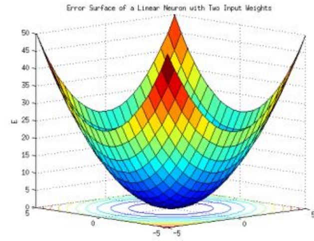




- מהו הגרדיאנט בנקודת מינימום (לוקאלי או גלובאלי?).
כיצד יתקדם האלגוריתם?

חשיבות קצב הלמידה

- אפילו אם קיים רק מינימום יחיד במשטח השגיאה, אנחנו עלולים לפספס אותו אם נקבע קצב למידה גבוה מדי.
- לעומת זאת, אם נקבע קצב למידה נמוך מדי, מספר האיטרציות (הצעדים) הדרוש בשביל להגיע למינימום עלול להיות גבוה מאד (למשל, גבוה יותר ממספר האיטרציות המקסימלי שהגדרנו).



סיכום מעשי – Gradient descent

- על מנת להגיע למינימום של הפונקציה f(x(1),...,x(n)):
  - נתחיל מנקודה מסוימת, שהיא וקטור המשתנים, המכיל את הערכים הרנדומלים של x(1),...x(n).
  - עבור מספר צעדים מסוים שהגדרנו:
    - נחשב את וקטור הגרדיאנט של הפונקציה בנקודה זו.
    - נוסיף לוקטור המשתנים את מינוס וקטור הגרדיאנט (נעדכן את כל המשתנים **<u>בבת אחת</u>**).
    - נפסיק כאשר אחד מהתנאים הבאים התממש:
      - הגענו למספר הצעדים המקסימלי.
      - הגענו לערך של הפונקציה f המקובל עלינו.

  - מהי תשובת האלגוריתם?