*machine learning*

# Linear regression

Lecture VI

פיתוח:
ד"ר יהונתן שלר
משה פרידמן

תודות לד"ר יונתן רובין
שעזר בהכנת המצגת

# מוטיבציה -אנחנו רוצים קורת גג לזוג צעיר

האם האפשרות הזו באה בחשבון?

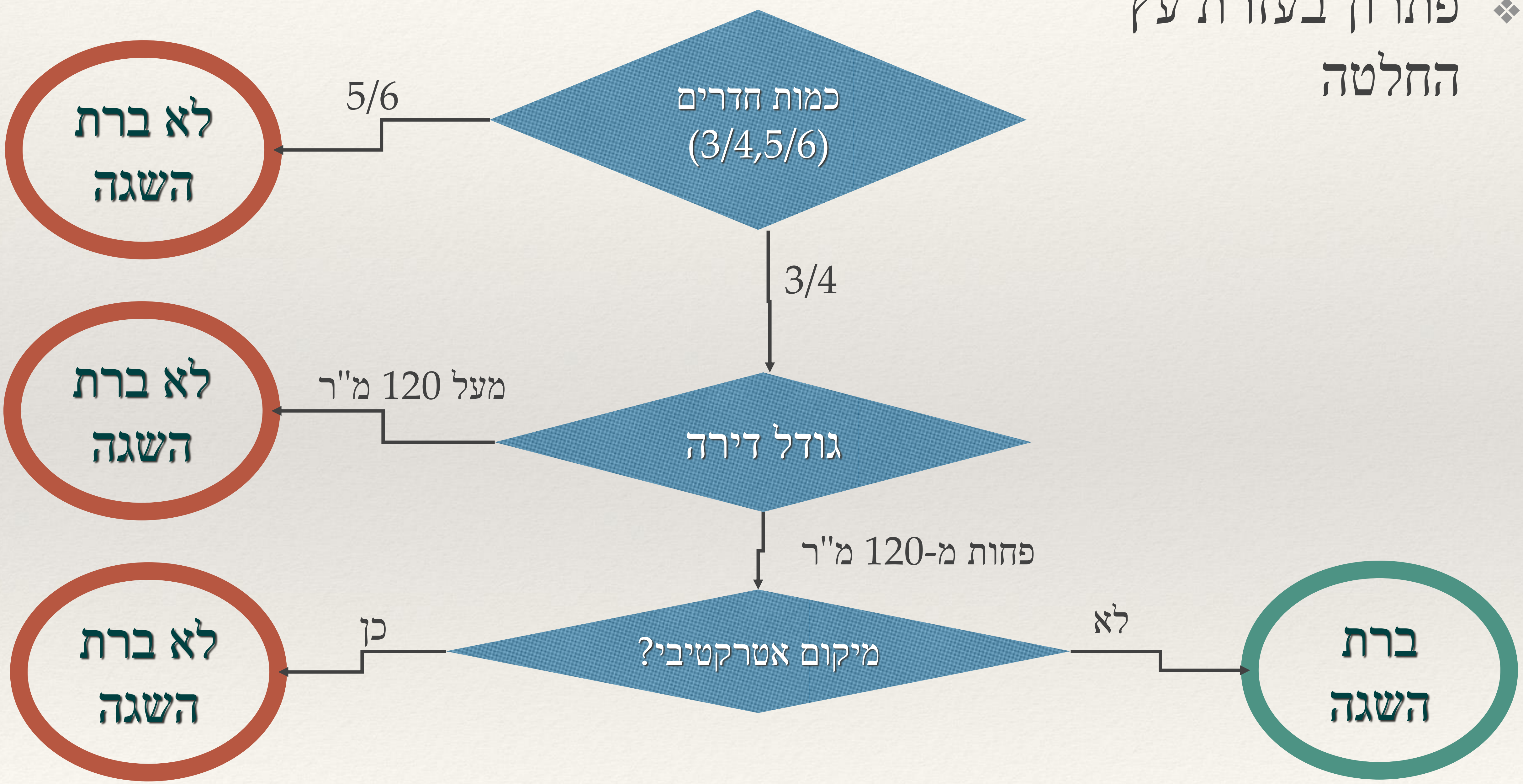
מדוע?

# מה ההגדרה של דירה יקרה?

- נניח שמדובר בדירה
- מה משפיעה על המחיר?
  - כמות חדרים
  - מיקום
  - גודל
  
  ...
- קונים דירה לזוג הצעיר
- שאלה חדשה – האם הדירה ברת השגה?

# מידול בעזרת סיווג: האם הדירה ברת השגה?

- פתרון בעזרת עץ החלטה



# בעיות במידול הסיווג לבעיה – מאפיינים

המאפיינים – קיבוץ המשתנים

- המיקום
  - מגוון גדול של אפשרויות

- כמות החדרים
  - הקיבוץ מאבד מידע
  - לא כל החדרים שווים

# בעיות במידול הסיווג לבעיה – בעיה בהגדרת המשימה

הגדרת הבעיה

- הגדרה סובייקטיבית בדירה ברת השגה

המשימה האמתית:

- מהו מחיר הדירה?

זו אינה בעיית סיווג

# מוטיבציה – מחיר הדירה

הבעיה אותה רוצים לפתור:

- חיזוי מחיר הדירה
- רגרסיה (regression) - בעיית למידת מכונה בה אנו רוצים לחזות מספר רציף (במקרה זה מחיר הדירה)

נסתכל תחילה על 2 משתנים:

- גודל הדירה, ומחיר הדירה.

| Size in meter$^2$ (x) | Price ($) in 1000's (y) |
|---|---|
| 2104 | 460 |
| 1416 | 232 |
| 1534 | 315 |
| 852 | 178 |
| … | … |

# סוג הבעיה

רגרסיה (regression) - בעיית למידת מכונה אנחנו רוצים לחזות מספר רציף (במקרה זה מחיר הדירה)

- שייכת ללמידה מונחת (supervised learning)
- בעיית סיווג, אשר גם שייכת לבעיות למידה מונחת (ושאותה למדנו בשיעורים הקודמים), חוזה קטגוריה ולא ערך

# פתרון בעיית רגרסיה

סוג הבעיה - רגרסיה (regression) - בעיית למידת מכונה אנחנו רוצים לחזות מספר רציף

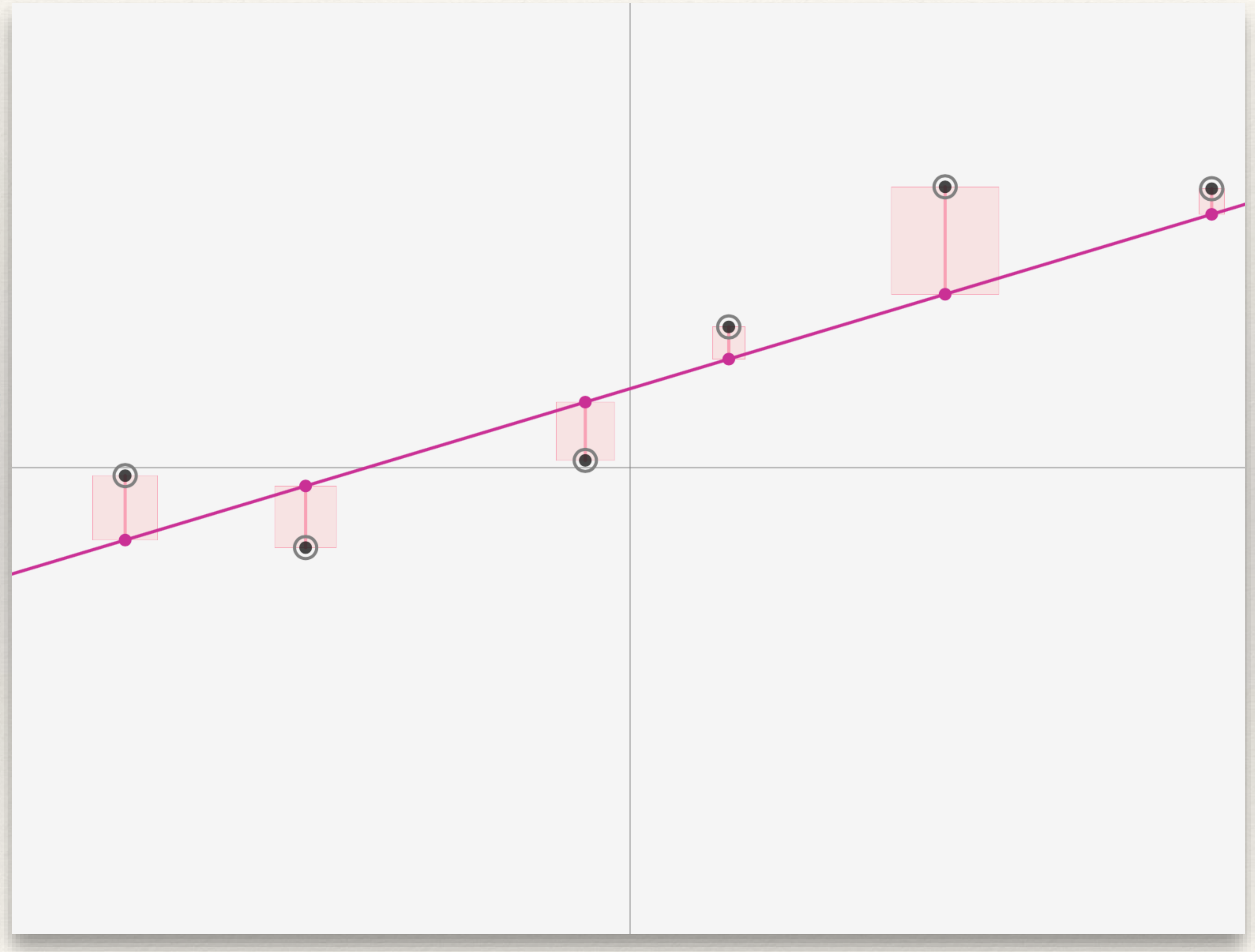
השיטה – מציאת פונקציה רציפה שעבור וקטור המאפיינים, תחזה את הערך

- בדוגמה שלנו –פונקציה שבהנתן המאפיינים תחזה את מחיר הדירה

האלגוריתם שנלמד: רגרסיה לינארית (linear regression)

# רגרסיה לינארית (linear regression)

# רגרסיה לינארית



<u>בעיית רגרסיה</u> - בעיית למידת מכונה אנחנו רוצים לחזות מספר רציף (במקרה זה מחיר הדירה)

<u>רגרסיה לינארית</u>
אלגוריתם בו הקשר בין וקטור המאפיינים, לערך אותו רוצים לחזות הוא פונקציה לינארית (בין המאפיינים לערך שנרצה לנבא)

# רגרסיה לינארית

ברגרסיה לינארית – הקשר בין וקטור המאפיינים, לערך אותו רוצים לחזות הוא פונקציה לינארית

מקרה פשוט (כמו בדוגמה) : יש רק מאפיין אחד בווקטור המאפיינים

| Size in meter$^2$ (x) | Price in K$ (y) |
|---|---|
| 2104 | 460 |
| 1416 | 232 |
| 1534 | 315 |
| 852 | 178 |
| … | … |

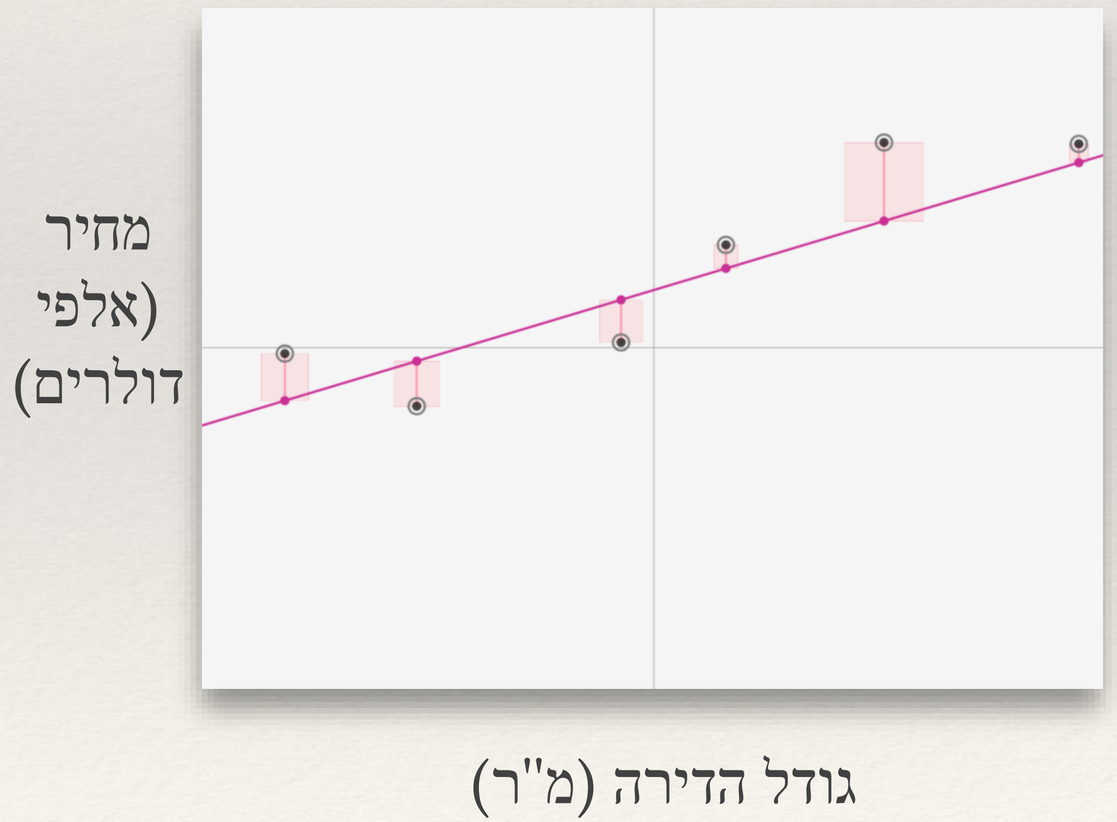
# מודלים לינארים - שאלת סקר
# רגרסיה לינארית – לעומת מפריד לינארי

איזה גרף שייך למודל הסיווג ואיזה שייך למודל רגרסיה:

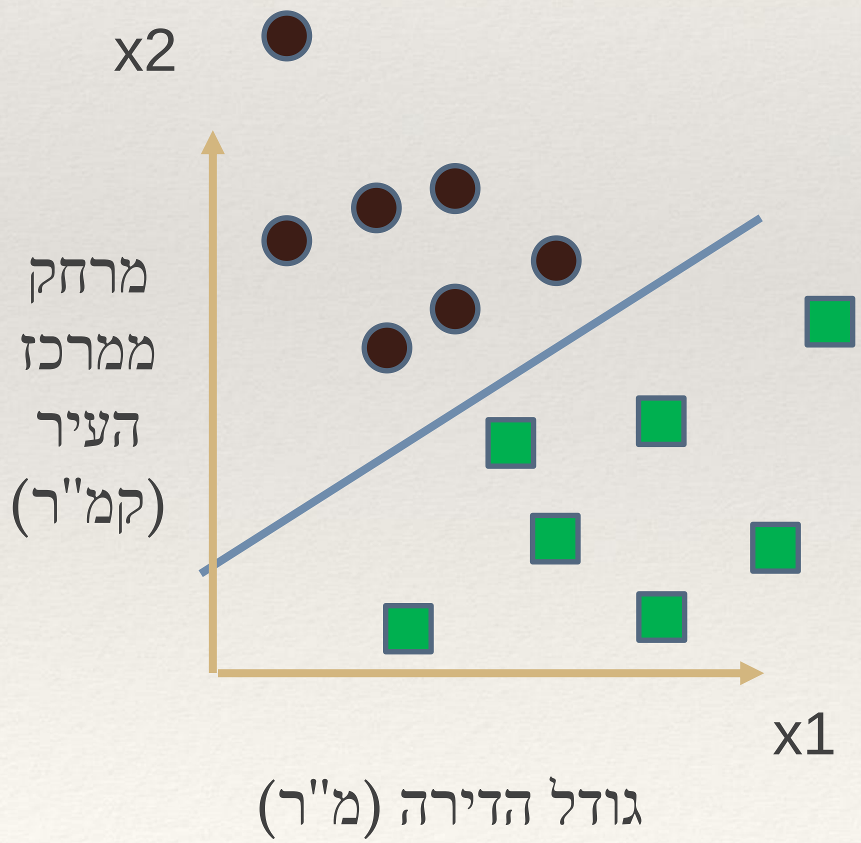
1. א – סיווג, ב- רגרסיה
2. א – רגרסיה ב- סיווג
3. א – סיווג ב- סיווג
4. א – רגרסיה ב- רגרסיה

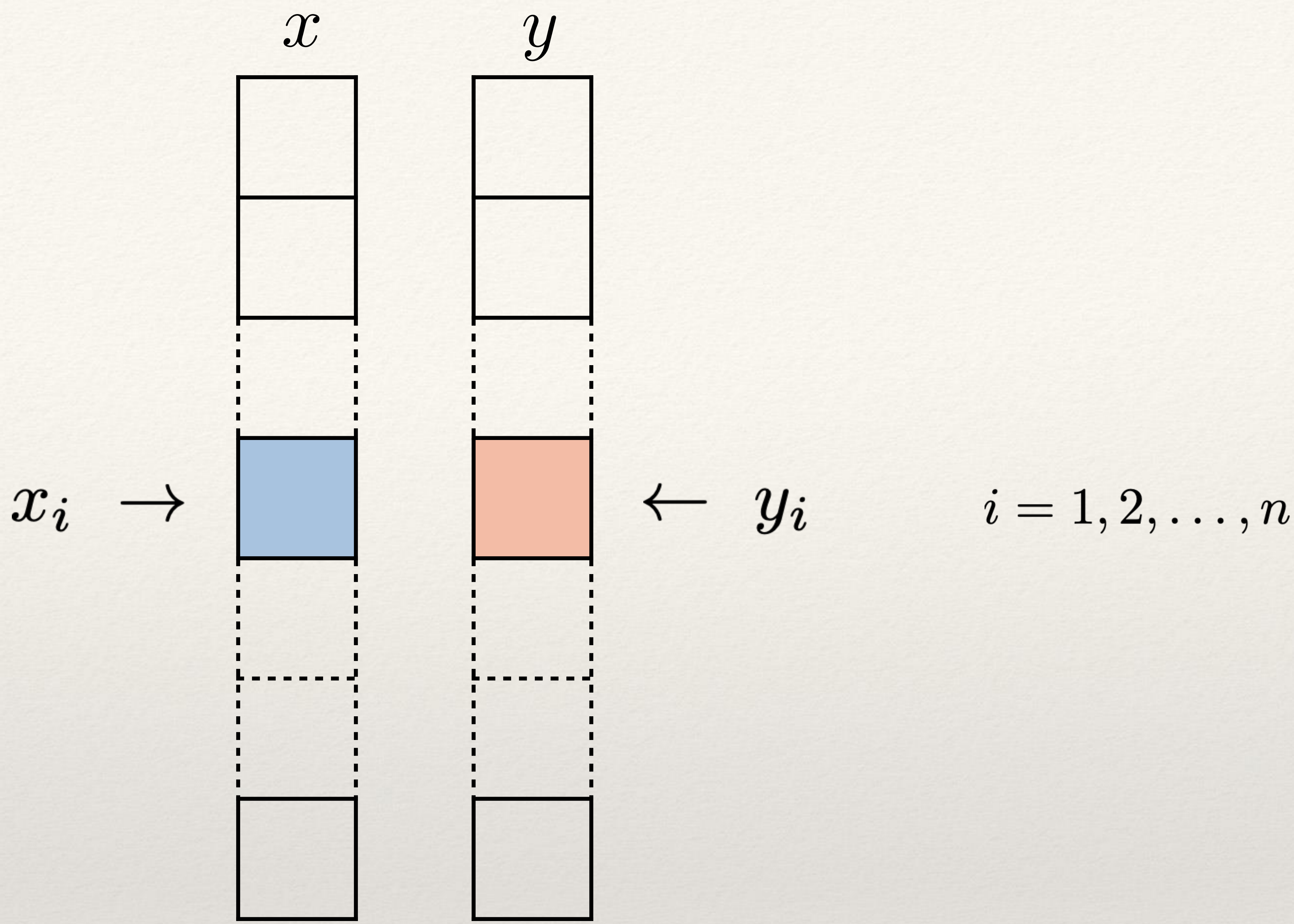
**תשובה: 2 - א – רגרסיה ב- סיווג**

<u>מודל א</u>




<u>מודל ב</u>






המודל הלינארי נראה כך:

$$\hat{y} = w_0 + w_1 x$$

# רגרסיה לינארית (1-D)

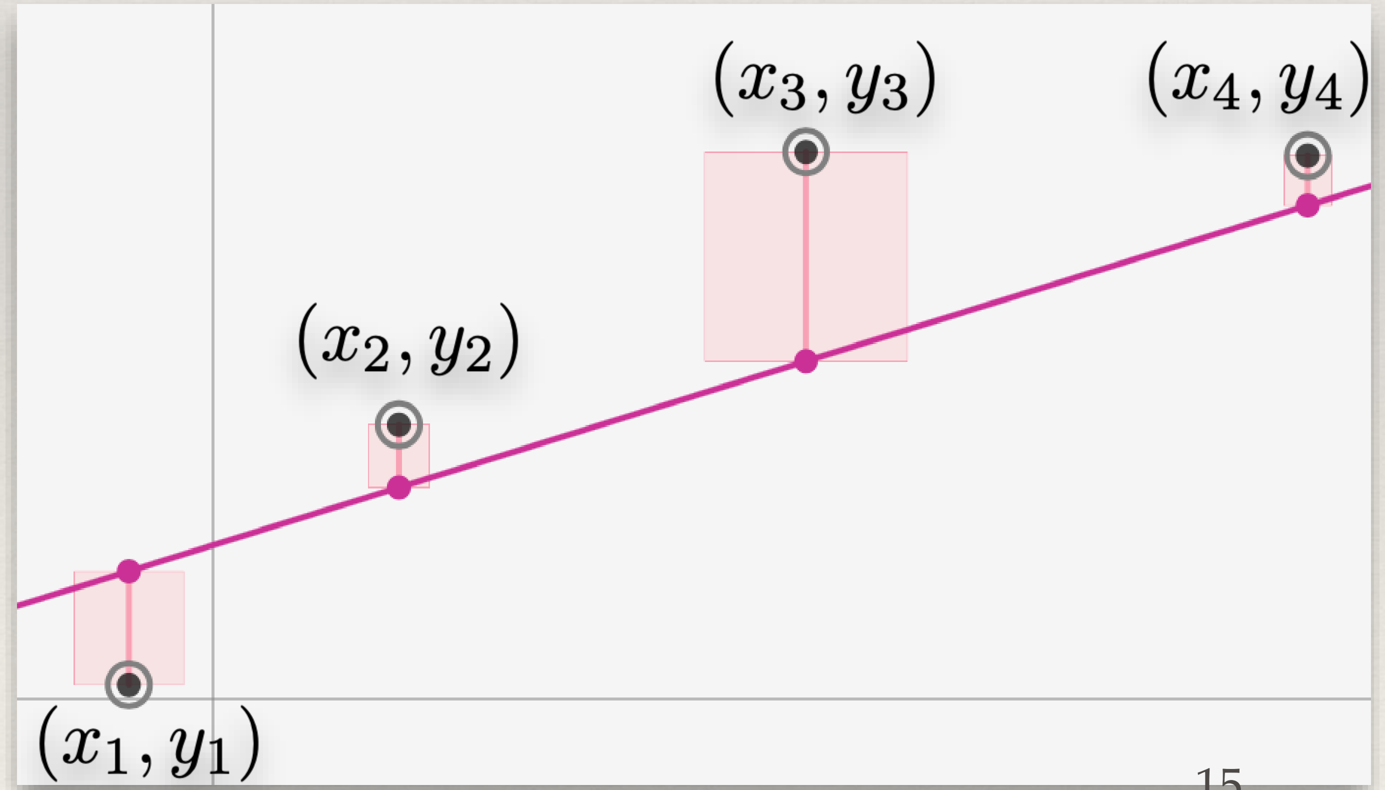
*data-set*

$$\left\{(x_i, y_i)\right\}_{i=1}^{n} = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

| input | model | prediction |
|---|---|---|
| $x$ $\rightarrow$ | $f(x; \vec{w})$ $\rightarrow$ | $\hat{y}$ |

**רגרסיה לינארית עם מאפיין יחיד -**
מבטאת את הקשר בין המאפיין (היחיד),
לערך אותו רוצים לחזות

לכן, ניתן לבטא ע"י קו ישר במישור




**המודל הלינארי נראה כך:**

$$\hat{y} = f(x; \vec{w}) = w_0 + w_1 x$$

**כיצד נעשה זאת?**
**נראה בהמשך ...**

# שיערוך מודל רגרסיה
# (regression model evaluation)

# שיערוך מודל סיווג - תזכורת

| | Predicted Yes | Predicted No |
|---|---|---|
| **Actual Yes** | True Positive (TP) | False Negative (FN) |
| **Actual No** | False Positive (FP) | True Negative (TN) |

$$\text{Precision} = \frac{\#TP}{\#TP+\#FP} \qquad \text{Recall} = \frac{\#TP}{\#TP+\#FN}$$

$$\text{Accuracy} = \frac{\#TP+\#TN}{TP+\#TN+\#FP+\#FN} \qquad \text{Error} = 1\text{- Accuracy}$$

$$f_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

בעיה – אף אחד משיטות השערוך לא מתאימה לרגרסיה מדוע?

# שיערוך מודל רגרסיה

פתרון שערוך מודל רגרסיה: מדידת הטעות ברמת הדוגמה הבודדת.

עבור כל instance בדיקה i, נמדוד את המרחק בין $y_i$ ל- $\widehat{y_i}$

ברמת ה-test set נשווה בין וקטור התוצאות המתויגיות $\vec{y}$ לבין וקטור תוצאות הסיווג של המודל - $\vec{\hat{y}}$, עבור כל n הדוגמאות ב-test set

**SAE (Sum of Absolute Error):**

בדומה לפונקציית מרחק מנהטן ב-KNN (פה נשווה בין וקטור התוצאות המתויגיות $\vec{y}$ ווקטור תוצאות הסיווג של המודל - $\vec{\hat{y}}$) = $SAE = \sum_{I=1}^{n} |y_i - \hat{y}_i|$

- שימו לב, שבצעם מדובר במרחק מנהטן בין וקטור הערכים הצפויים ווקטור הערכים שחזה המודל.

**MAE (Mean of Absolute Error):** מקובל למצע את מדד SAE, ע"י חלוקה בכמות הדוגמאות ב- test, נקבל $MAE = \frac{1}{n} \sum_{I=1}^{n} |y_i - \hat{y}_i|$

# שיערוך מודל רגרסיה – מדדים נוספים

**SSE (Sum of Squared Error ):**

סכום ההפרשים הריבועי, בין התוצאות האמתיות לתוצאות החיזוי = $SSE = \sum_{I=1}^{n}(y_i - \hat{y}_i)^2$

- שימו לב שבצם מדובר במרחק אוקלידי בריבוע, בין וקטור הערכים הצפויים ווקטור הערכים שחזה המודל..

**MSE (Mean of Squared Error ):** מקובל למצע את מדד MSE, ע"י חלוקה בכמות הדוגמאות ב- test, נקבל

$$MSE = \frac{1}{n}\sum_{I=1}^{n}(y_i - \hat{y}_i)^2$$

- מדד זה דומה למדד השונות (variance), אשר מודד את ממוצע המרחק מהממוצע

**RMSE (Root Mean of Squared Error):** אם נוציא שורש לערך ה-MSE, נקבל:

$$RMSE = \sqrt{\frac{1}{n}\sum_{I=1}^{n}(y_i - \hat{y}_i)^2}.$$

- מדד זה דומה למדד סטיית התקן
- כפי ששונות וסטיית תקן הם מדד לפיזור הערכים ב-מ"מ (במאפיין), מדדים כמו RMSE, מודדים, אם התוצאות הצפויות צמודות לישר החיזוי, או מפוזרות ורחוקות יותר.

# שיערוך מודל רגרסיה

פתרון: מדידת הטעות ברמת הדוגמה הבודדת.

עבור כל instance בדיקה i, נמדוד את המרחק בין $y_i$ ל- $\hat{y}_i$

ברמת ה-test set נשווה בין וקטור התוצאות המתויגיות $\vec{y}$ לבין וקטור תוצאות הסיווג של המודל - $\vec{\hat{y}}$, עבור כל n הדוגמאות ב-test set

**(Sum of Squared Error ) SSE** =: $SSE = \sum_{I=1}^{n}(y_i - \hat{y}_i)^2$

כעת, נסמן את ממוצע הערכים המתויג, כ- $\bar{y}$, כלומר $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

**(Sum of Squared Total) SST**: = $SST = \sum_{I=1}^{n}(y_i - \bar{y})^2$ (מדד השונות – variance, הוא ממוצע של SST)

**R-SQARE** - $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{MSE}{\sigma^2}$,

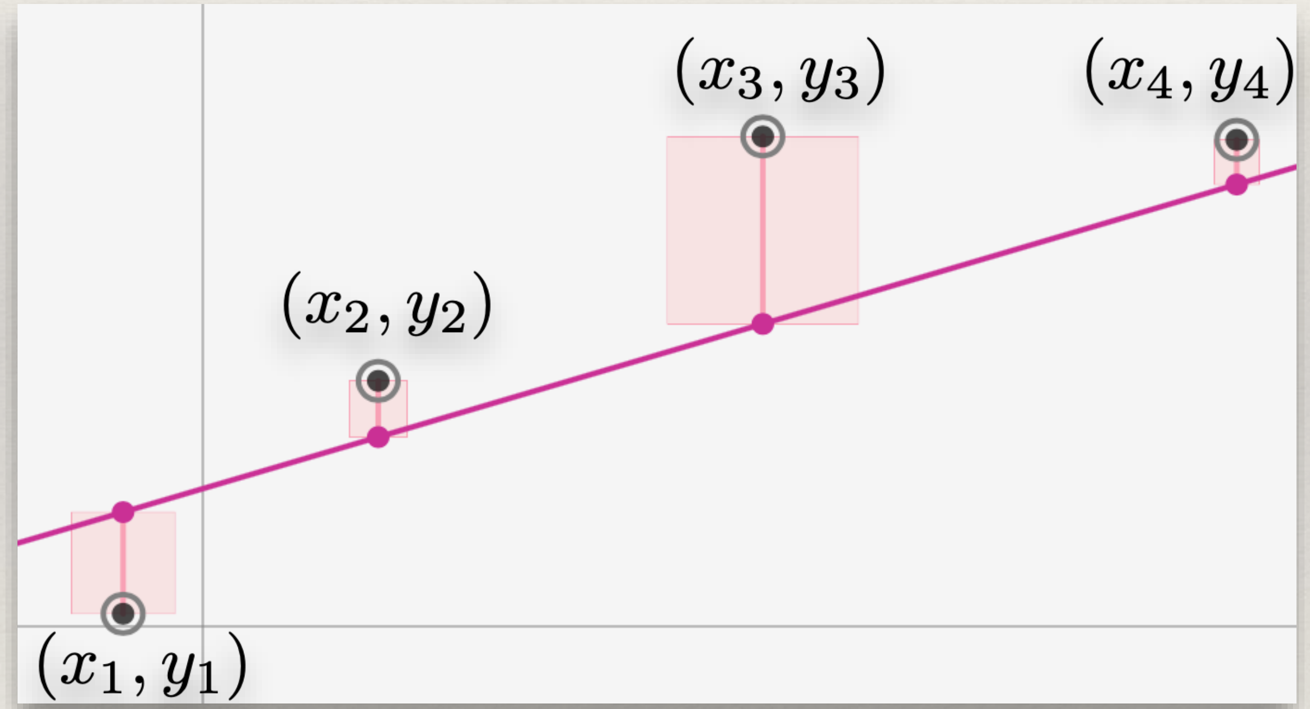
- ערך של $R^2 = 1$, מתאר ניבוי מושם של המודל;
- ערך של $R^2 = 0$, חוסר התאמה מלא

# חזרה לרגרסיה לינארית (1-D)

| prediction | model | input |
|---|---|---|

<u>המטרה:</u> $x \rightarrow f(x; \vec{w}) \rightarrow \hat{y}$

<u>המודל הלינארי:</u> $\hat{y} = f(x; \vec{w}) = w_0 + w_1 x$



# שאלת סקר

1. אילו שיטות שיערוך מיועדים לבעיית רגרסיה?

<u>תשובות אפשרויות:</u>

א. Accuracy

ב. Euclidean Distance

ג. Variance

ד.SAE

<u>תשובה</u> – ד.SAE (Sum Absolute Error).

הערות:

- מדד accuracy, בודק דיוק בחיזוי של מודל סיווג
- בעצם - Variance (שונות) באוכלוסיה, זהה למדד MSE, אולם כשנתייחס לשערוך רגרסיה, נשתמש ב- MSE
- Euclidean Distance דומה ל- SSE

# שאלת סקר

שאלה: מדוע לא משתמשים במדד MSE, להערכת ביצועיו של מודל רגרסיה?

<u>תשובות אפשריות</u>:
א. מכיוון שללא שימוש במקדם למידה מתאים, לא נוכל להכליל בעזרת מדד ה-MSE.
ב. מכיוון שמדד ה- MSE, הינו מדד סיפורי לאיכות, ואנחנו צריכים שיטה מתמטית מדוייקת.
ג. ההנחה אינה נכונה, מדד ה- MSE, הינו אחד המדדים השימושיים להערכת ביצועיו של מודל רגרסיה.
ד. מכיוון שהתוצאה ברמת הדוגמה יכולה להיות נכונה או לא נכונה, ולכן מדד זה אינו מתאים.

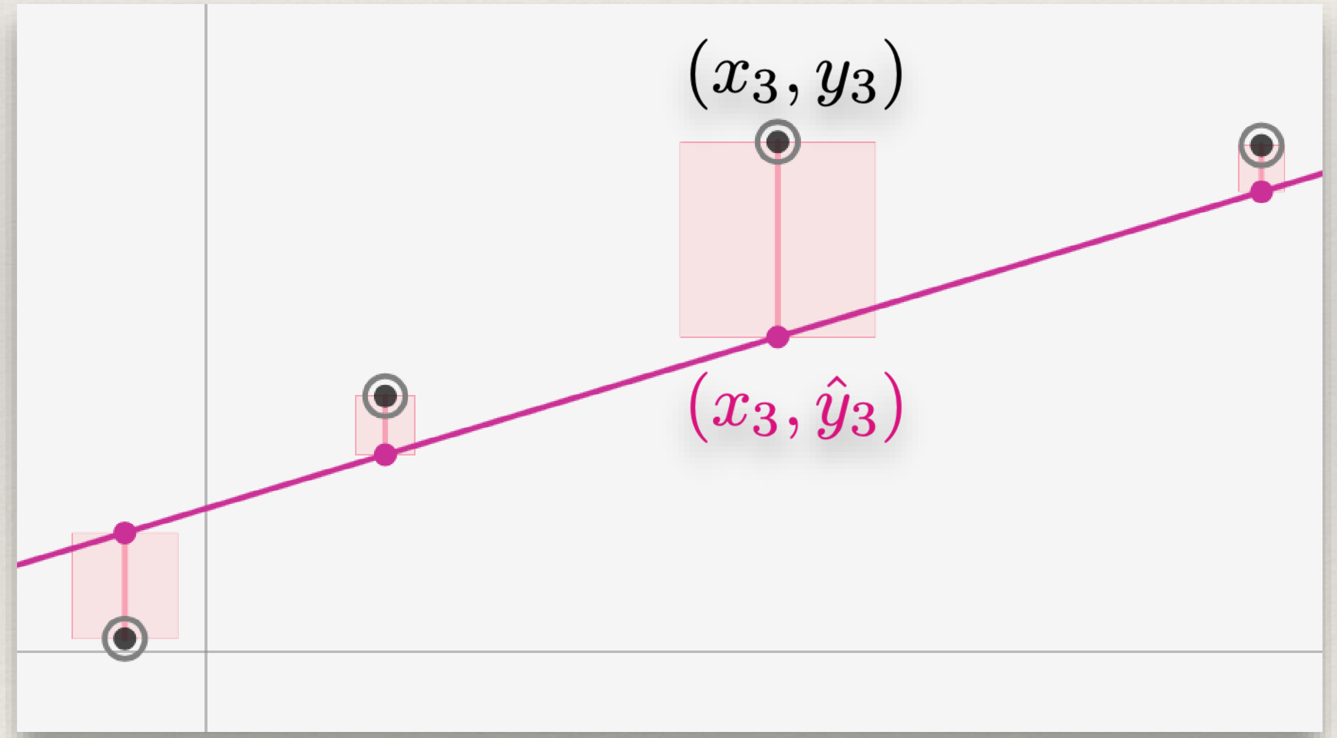
**תשובה: <u>ג</u>**

# Linear Regression (1-D)

cost function: $J\left(\vec{w}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2$

**הטעות:** $MSE = \frac{1}{n}\sum_{I=1}^{n}(\hat{y}_i - y_i)^2$

optimization problem: $\min_{w}\left[J\left(\vec{w}\right)\right]$




מה נחשיב מודל טוב ביחס לטעות?

- נשאף ל-MSE, קטן ככל הניתן
- איך מקבלים את התוצאה הרצויה?

תשובה בהמשך ...

- אבל קודם כל – נמחיש את המטרה

Demo

# Cost Function

<u>cost function</u> - $J(\vec{w})$ - is the loss function

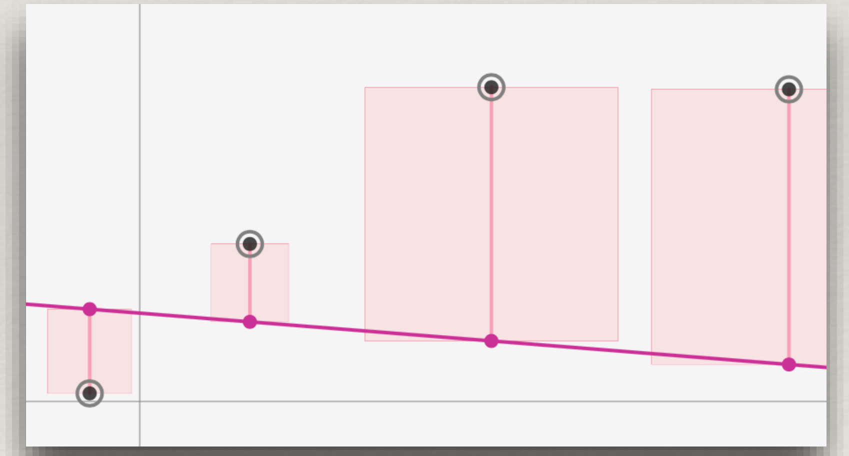
- פונקציית ה"מחיר" היא פונקציית ה"הפסד"

**<u>שאלה</u> (סקר): לאיזו מהפונקציות הלינאריות הבאות הטעות המינימלית?**

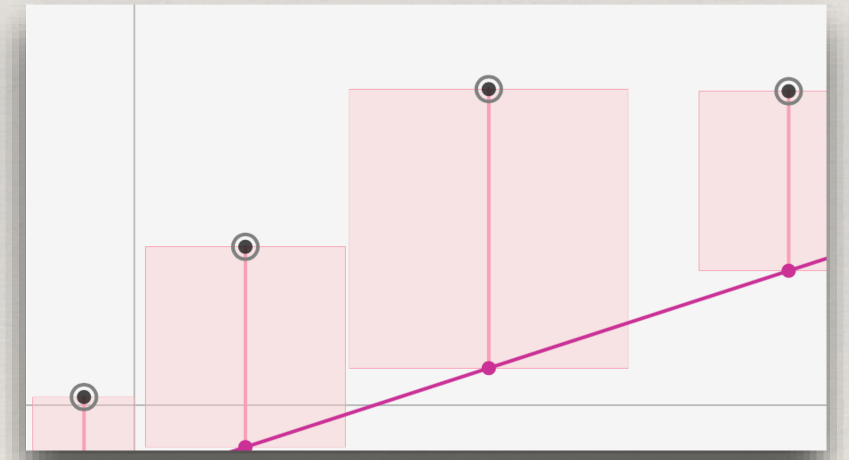
$$\vec{w}^{*}=\arg\min_{w} J(\vec{w})$$

- נרצה למזער את ההפסד

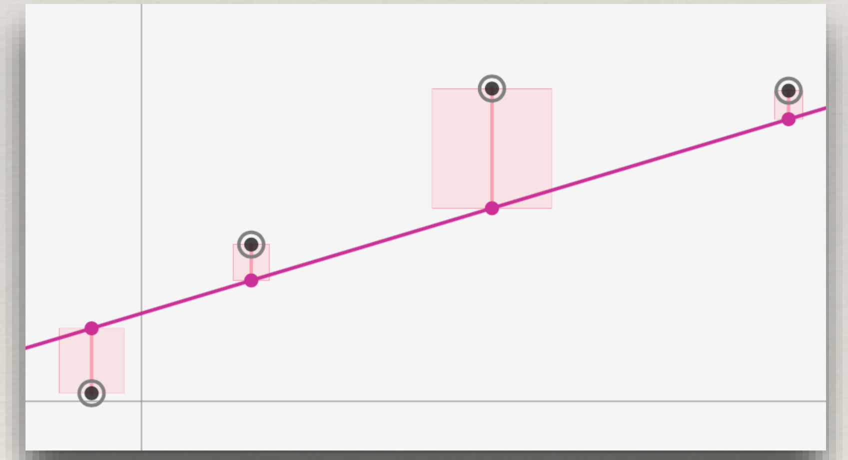
<u>א.</u>



<u>ב.</u>



<u>ג.</u>



[Demo]

**<u>תשובה: ג</u>**

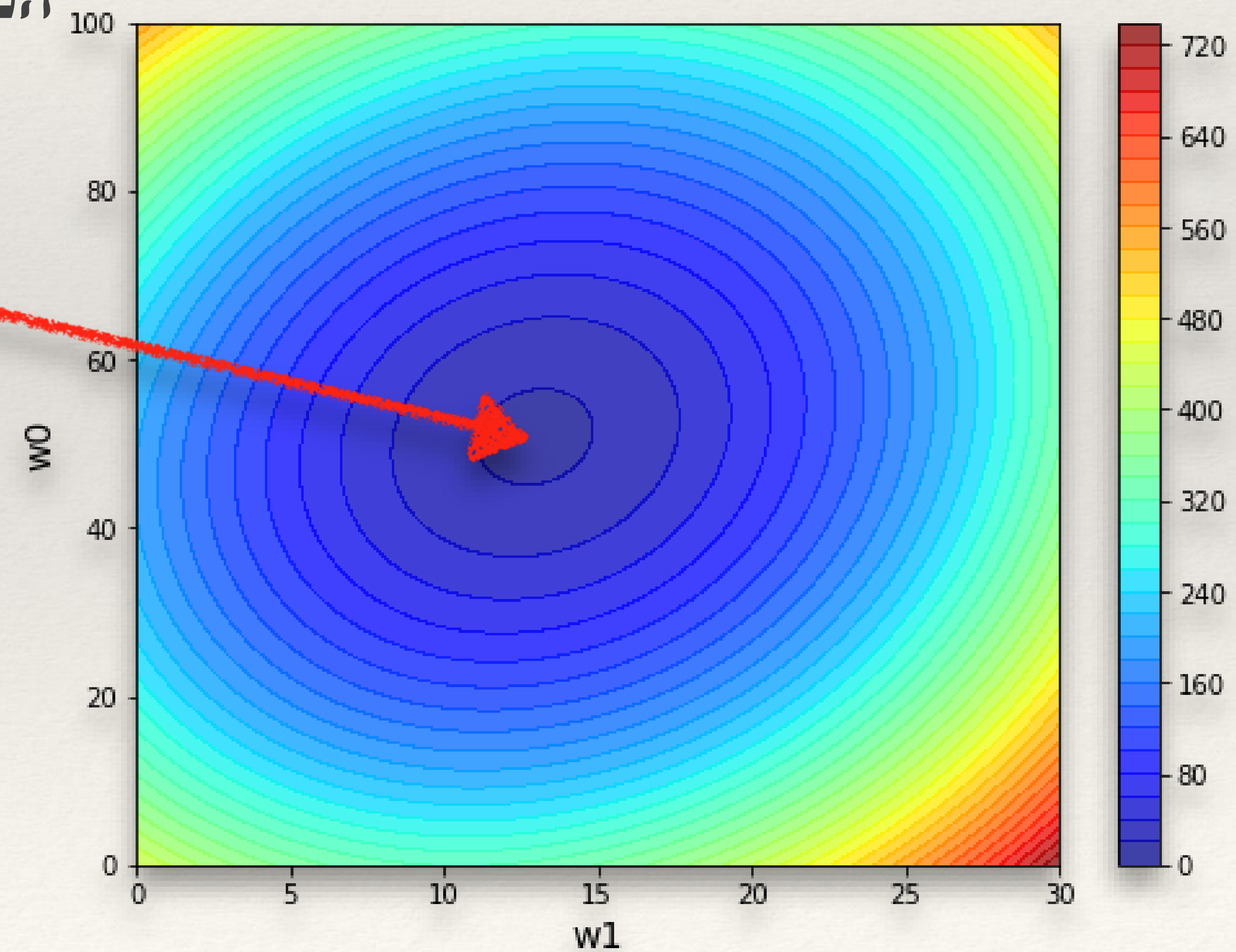
# Cost Function

$$J\left(\vec{w}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(w_0 + w_1 x_i - y_i\right)^2$$

הטעות: $MSE = \frac{1}{n}\sum_{I=1}^{n}(\hat{y}_i - y_i)^2$

$$\min_{w}\left[J\left(\vec{w}\right)\right]$$

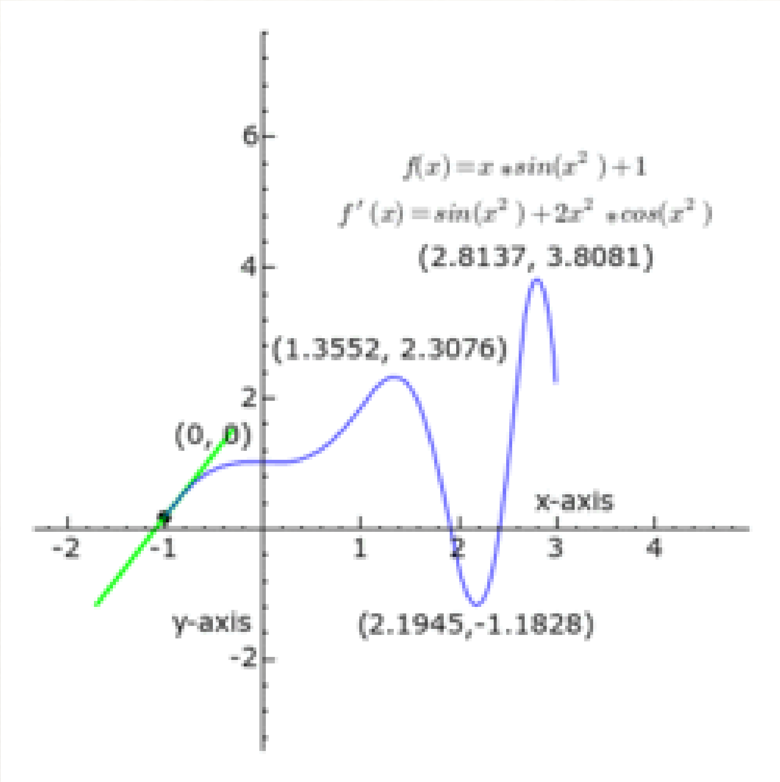
$J\left(\vec{w}\right)$



# פונקציה עם משתנה אחד - נגזרת ונקודות קיצון

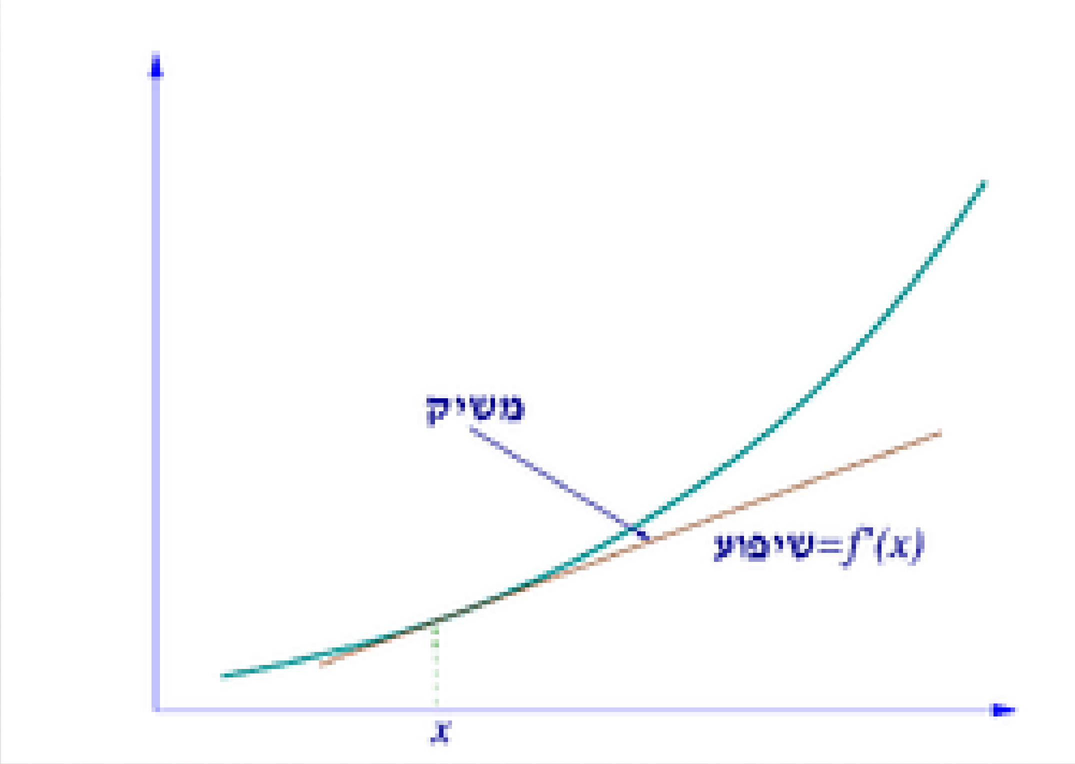
# מושגים – נגזרת של פונקציה

- <u>נגזרת</u> - הנגזרת של פונקציה ממשית מתארת את קצב ההשתנות של הפונקציה
- לדוגמה, הנגזרת לפי משתנה הזמן של פונקציית המיקום (העתק) של מכונית נוסעת, היא המהירות של המכונית

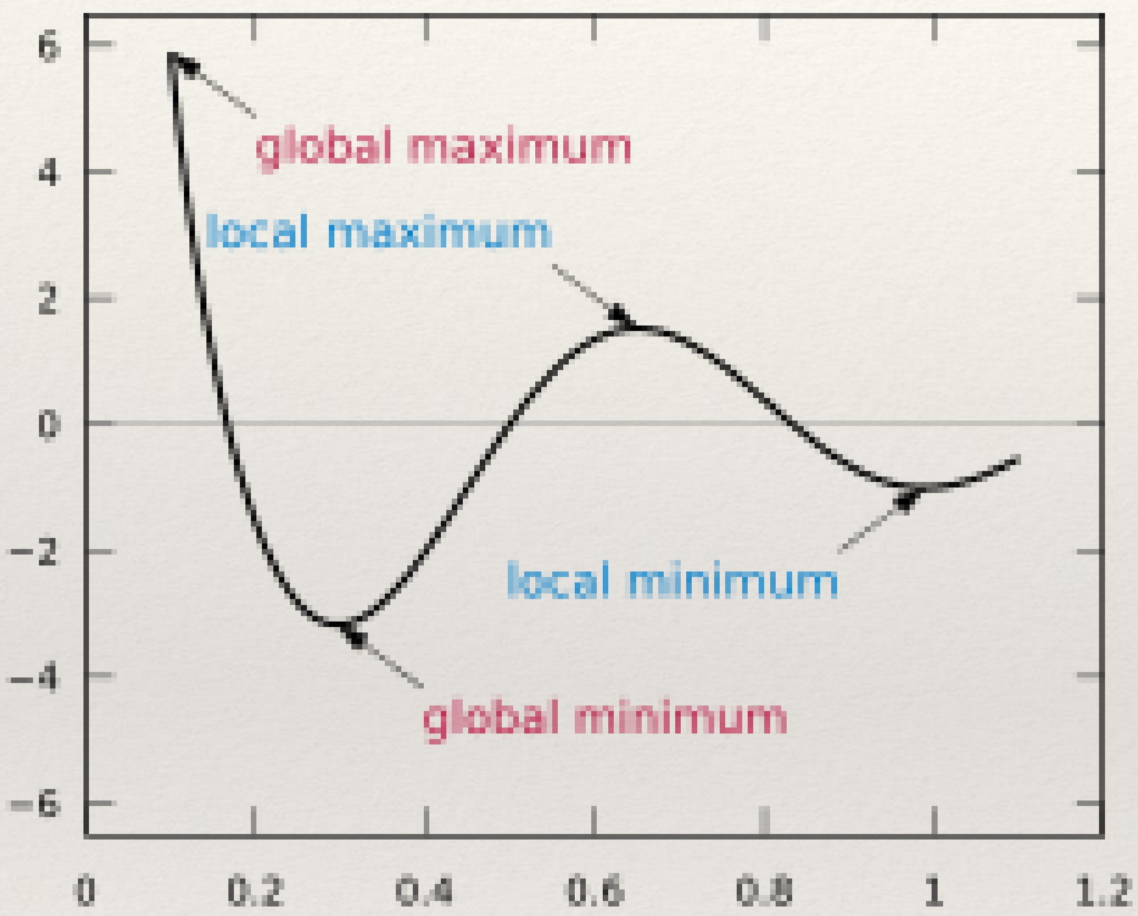


# מושגים – נגזרת של פונקציה

* נגזרת - הנגזרת של פונקציה ממשית מתארת את קצב ההשתנות של הפונקציה
    * מבחינה גאומטרית - הנגזרת של פונקציה בנקודה שווה לשיפוע המשיק באותה נקודה, כלומר, לכיוון של העקומה שהפונקציה מתארת.
* גזירות – תכונה של פונקציה, שניתנת לגזירה.



# מושגים – נקודות קיצון של פונקציה




נקודת קיצון - נקודה שבה ערכה של הפונקציה הוא גבוה ביותר או נמוך ביותר.

- יש להבדיל בין נקודות קיצון מקומיות ובין נקודות קיצון מוחלטות (גלובליות).

נקודת מינימום מקומית – בפו' עם משתנה 1, אם הפונ' f גזירה פעמיים, מתקיים $f'(x)=0$, $f''(x)>0$, אז זוהי נקודות מינימום מקומית

נקודת נקסימום מקומית – בפו' עם משתנה 1, אם הפונ' f גזירה פעמיים, מתקיים $f'(x)=0$, $f''(x)<0$, אז זוהי נקודות מקסימום מקומית

# שאלות סקר

1. בפונקציה רציפה וגזירה פעמיים, כיצד נמצא את נקודת המינימום?

<u>תשובות אפשרויות:</u>

א. אם הפונקציה גזירה ורציפה, אין לה נקודת מינימום.

ב. אם הפונקציה, בעל נגזרת ראשונה >0, לפני הנקודה, ונגזרת שנייה שווה לאפס, זוהי נקודת מינימום.

ג. אם הפונקציה בעלת ערך=0 בנגזרת הראשונה בנקודה ונגזרת שניה>0, זוהי נקודת מינימום.

<u>תשובה</u> – <u>**ג.**</u>

רגרסיה לינארית (linear regression) - בעזרת אלגוריתם Gradient Descent (מורד הגרדיאנט) - הרעיון

# Cost Function

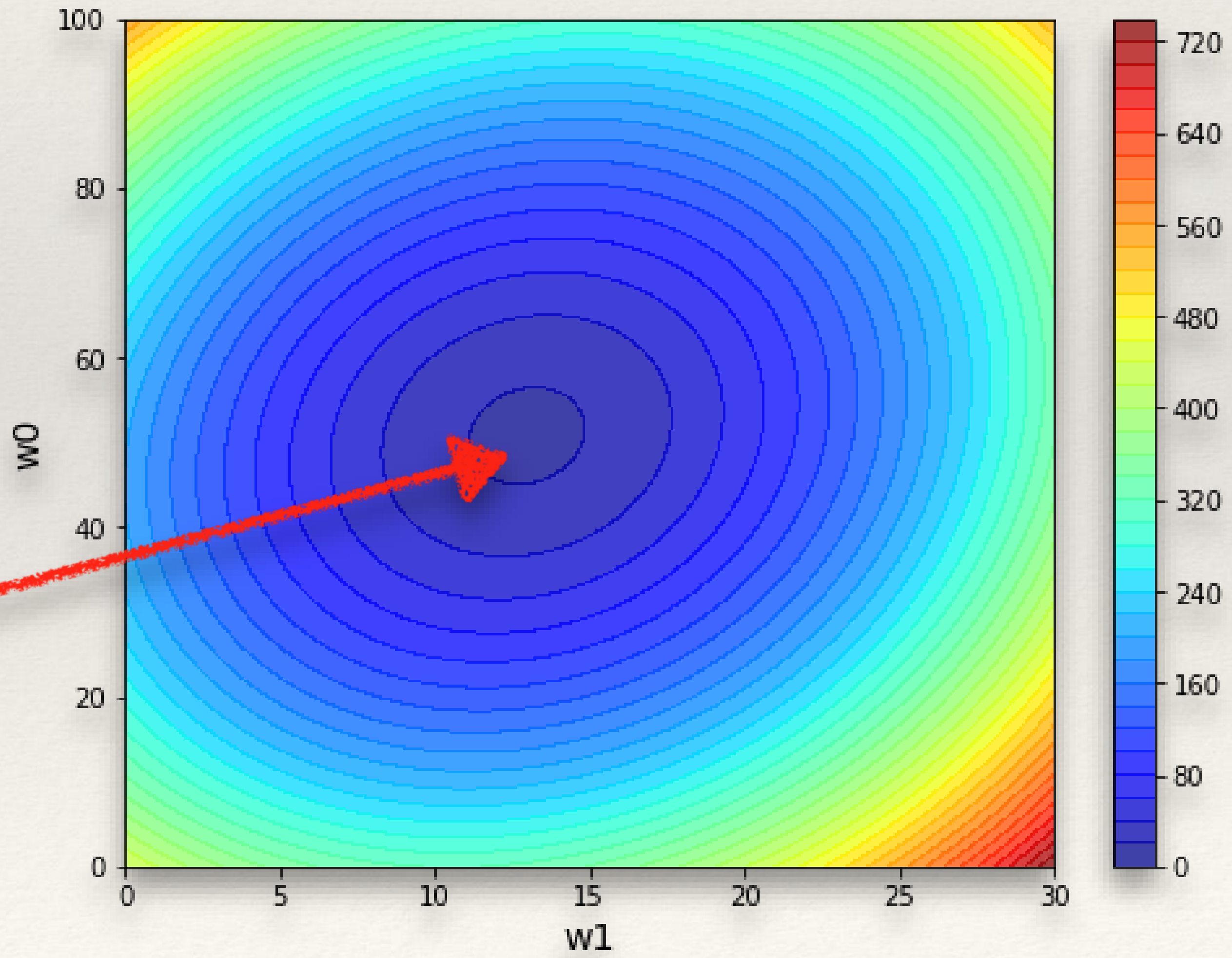
המודל הליניארי: $\hat{y} = f\left(x; \vec{w}\right) = w_0 + w_1 x$

$$J\left(\vec{w}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(w_0 + w_1 x_i - y_i\right)^2$$

הטעות: $MSE = \frac{1}{n}\sum_{I=1}^{n}(\hat{y}_i - y_i)^2$

$$\min_{w}\left[J\left(\vec{w}\right)\right]$$

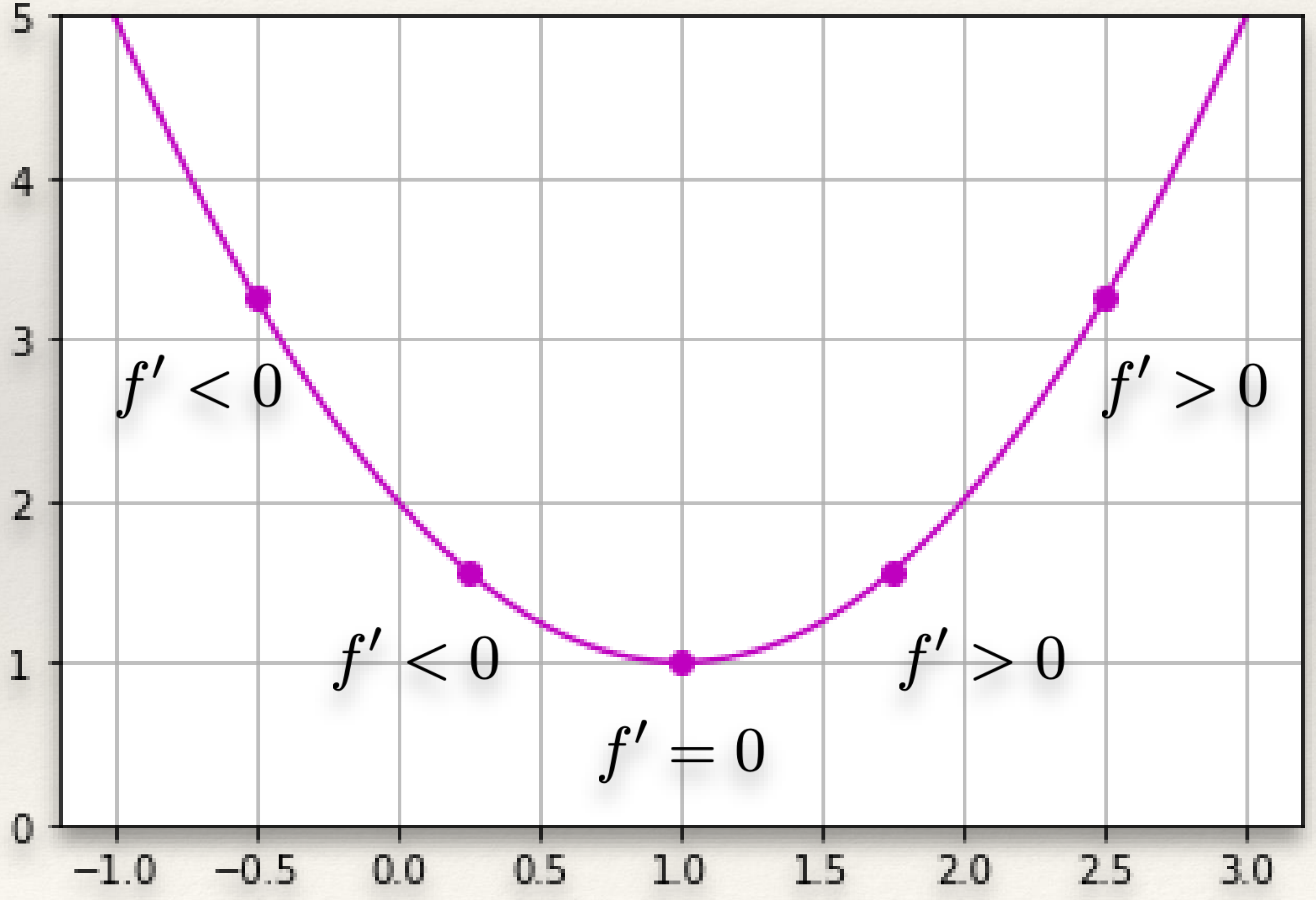
$J\left(\vec{w}\right)$



# The Gradient

$$f(x) : \mathbb{R} \to \mathbb{R}$$

$$f' = \frac{df}{dx}$$



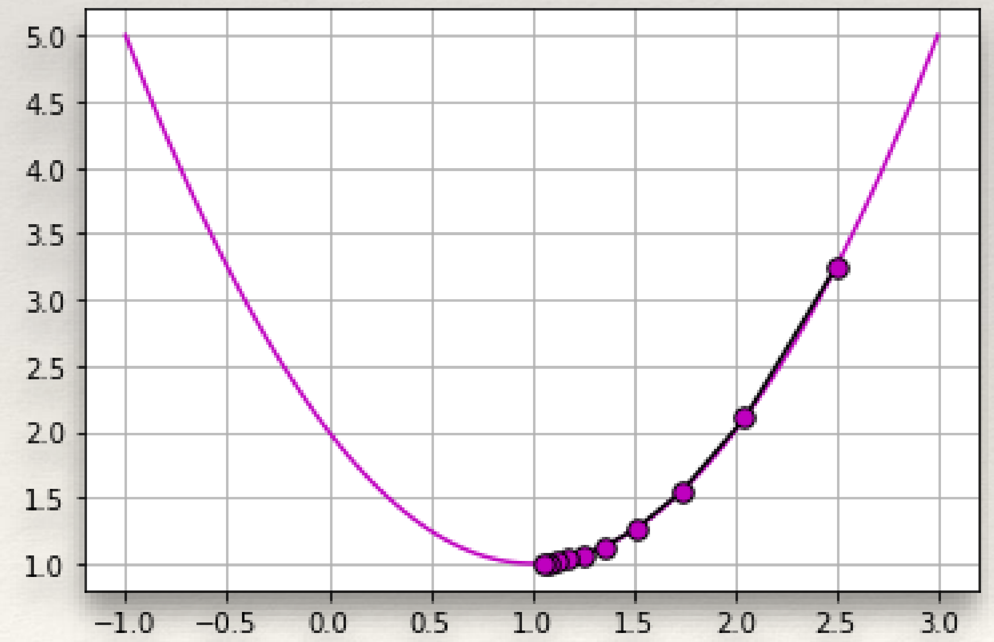
# Gradient Descent

$$x^{(0)} \to x^{(1)} \to x^{(2)} \to \dots \to x^{(T)}$$

$$x^{(t+1)} = x^{(t)} - \alpha f'\left(x^{(t)}\right)$$

$$x := x - \alpha f'(x)$$

*learning rate*

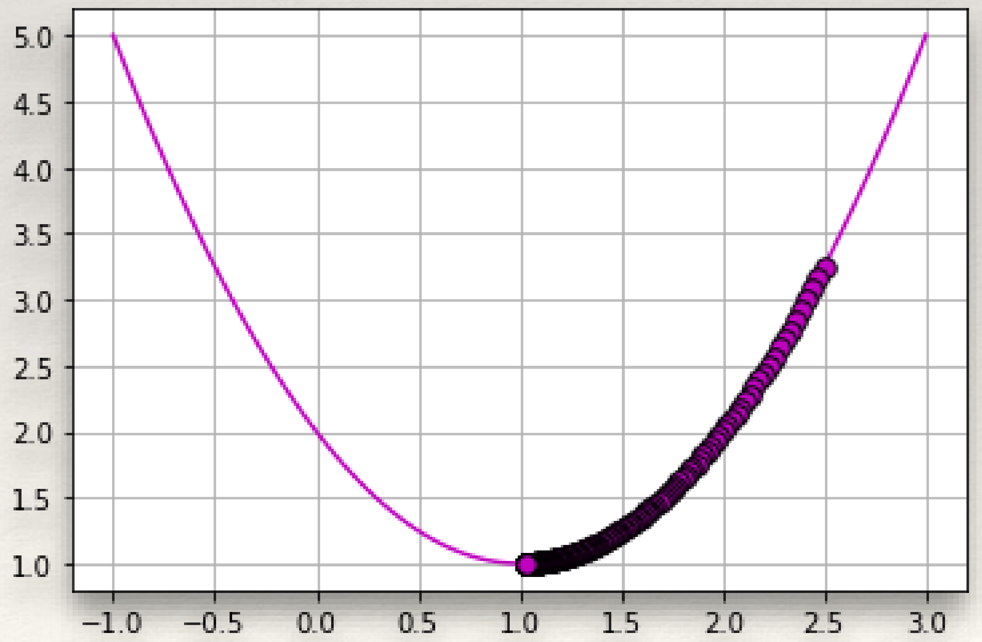


# Gradient Descent

$$x^{(0)} \to x^{(1)} \to x^{(2)} \to \ldots \to x^{(T)}$$

$$x := x - \alpha f'(x)$$

*learning rate*

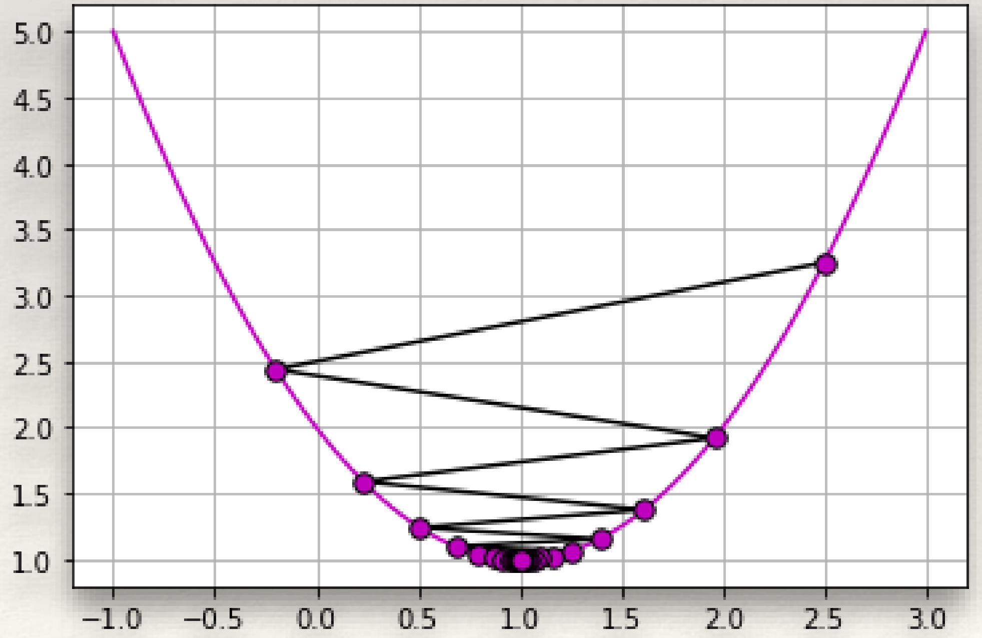


# Gradient Descent

$$x^{(0)} \to x^{(1)} \to x^{(2)} \to \ldots \to x^{(T)}$$

$$x := x - \alpha f'(x)$$

*learning rate*

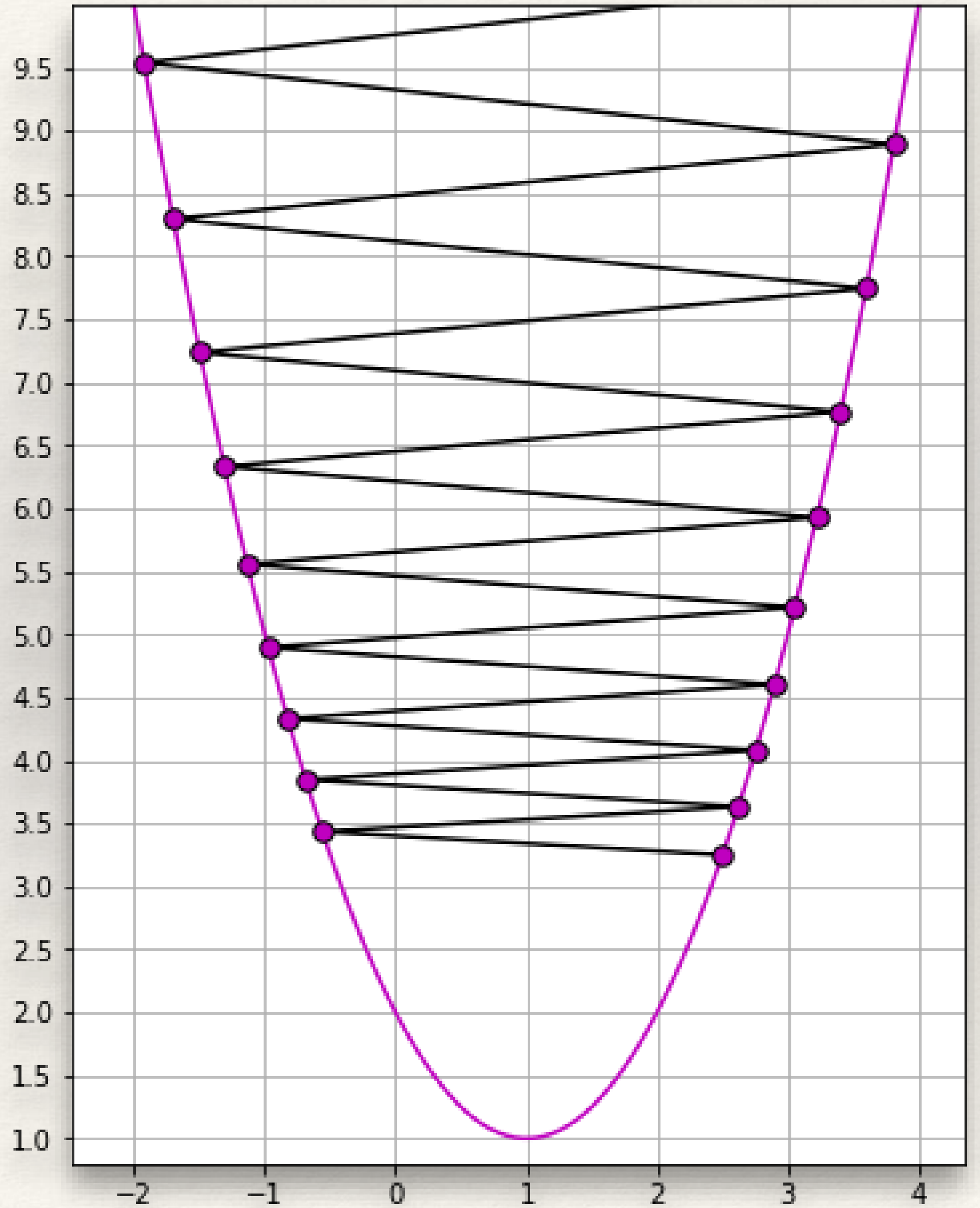


# Gradient Descent

$$x^{(0)} \rightarrow x^{(1)} \rightarrow x^{(2)} \rightarrow \ldots \rightarrow x^{(T)}$$

$$x := x - \alpha f'(x)$$

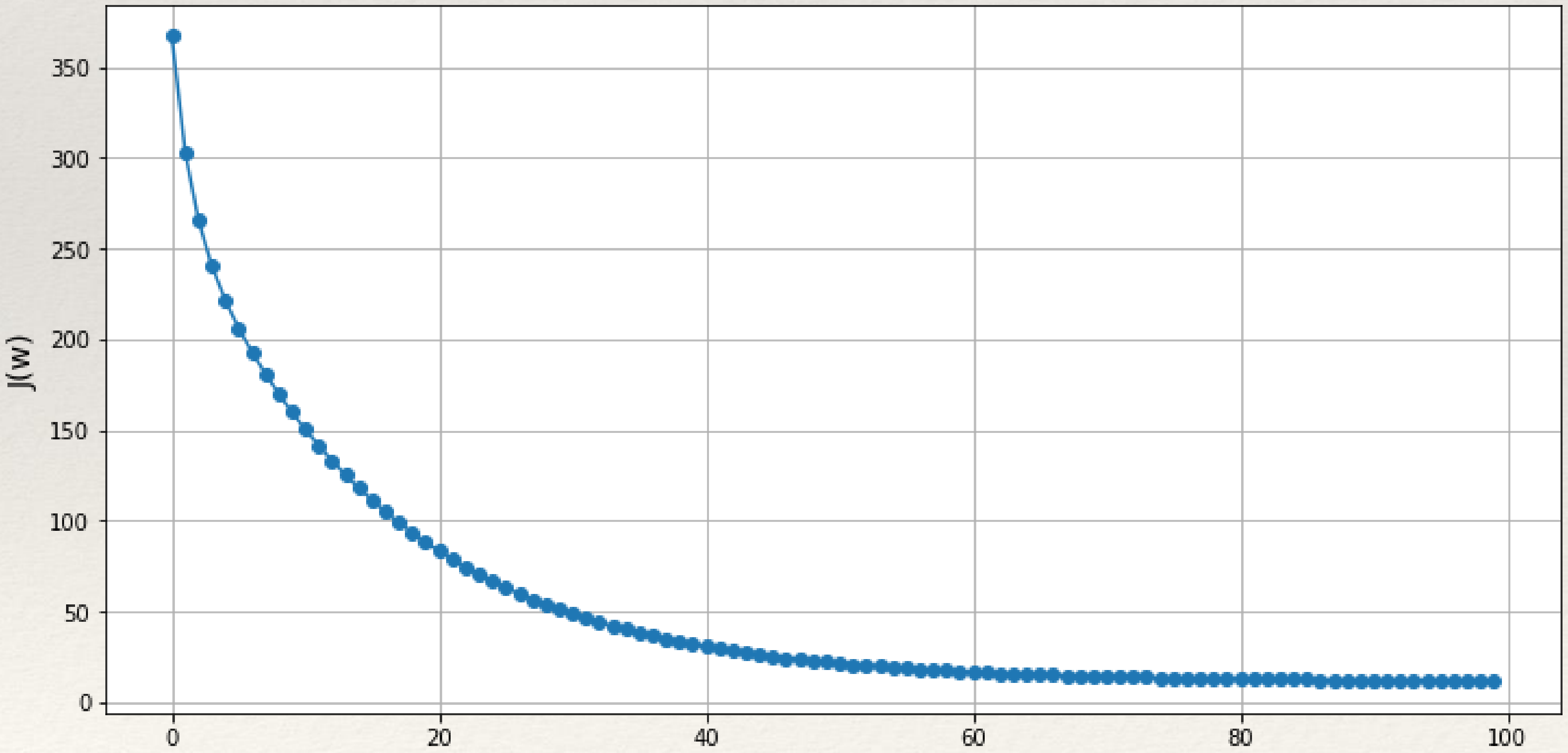
*learning rate*

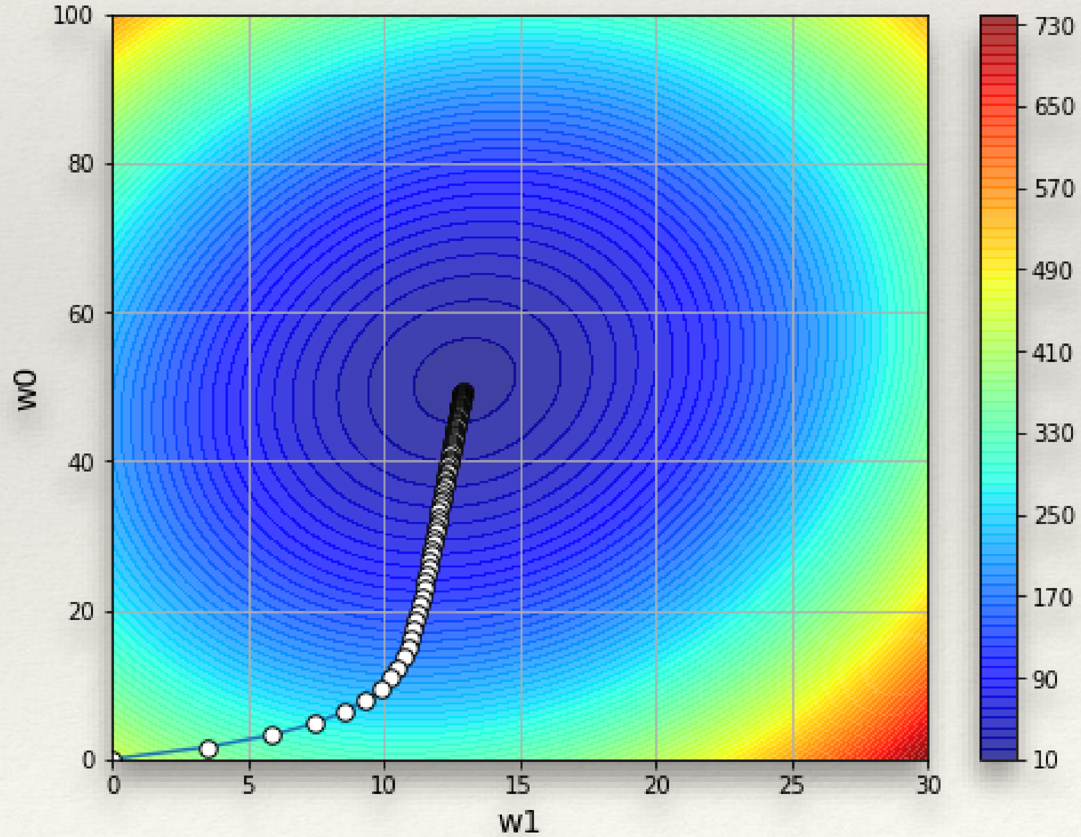


# Linear Regression via Gradient Descent

$$w_0 := w_0 - \frac{2\alpha}{n}\sum_{i=1}^{n}(w_0 + w_1 x_i - y_i)$$

$$w_1 := w_1 - \frac{2\alpha}{n}\sum_{i=1}^{n}(w_0 + w_1 x_i - y_i)\cdot x_i$$



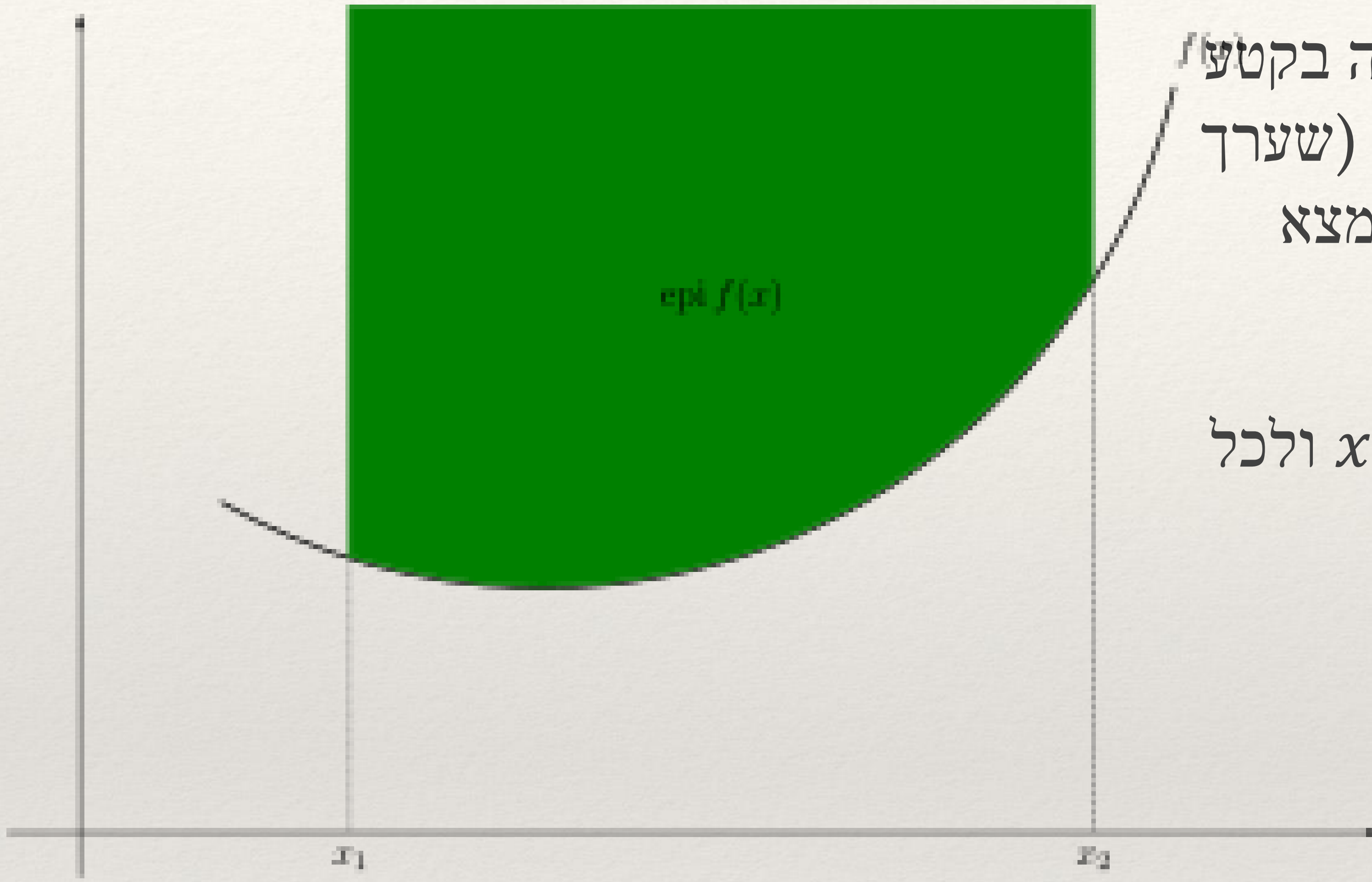

איך מגיעים לזה?

## פונקציה עם משתנה אחד - גרדיאנט ופונקציה קמורה

# מושגים – גרדיאנט של פונקציה עם משתנה אחד

- פונקציה דיפרנציאבילית - פונקציה ממשית בעלת כמה משתנים, שיש לה קירוב ליניארי (דיפרנציאל).
    - עבור פונקציות סקלריות במשתנה יחיד, מושג הדיפרנציאל קשור קשר הדוק למושג הנגזרת
- נגזרת חלקית (partial derivative) - נגזרת חלקית של פונקציה בכמה משתנים היא נגזרת של הפונקציה באחד ממשתניה.
    - עבור פונקציות סקלריות במשתנה יחיד, מדובר בעצם בנגזרת
- גראדיאנט (gradiant) – גרדאינט של פונקציה וקטורית, הוא הוקטור של הנגזרות החלקיות.
- עבור פונקציות סקלריות במשתנה יחיד, מדובר בעצם בוקטור עם ערך יחיד - הנגזרת
    - כיוון וקטור הגרדיאנט מצביע אל הכיוון בו השינוי בשדה הסקלרי מקסימלי (חיובי). $\text{grad } f(a) = \vec{\nabla} f(a)$

# מושגים – פונקציה קמורה (Convex)




פונקציה קמורה (Convex) - פונקציה קמורה בקטע מסוים, אם לכל שתי נקודות על גרף הפונקציה (שערך ה- x שלהן נמצא בקטע), הקו המחבר ביניהן נמצא מעל לגרף הפונקציה (או עליו).

עבור הקטע $I$, הפו' קמורה, אם לכל $x_1, x_2 \in I$ ולכל סקלר $0 \le \lambda \le 1$, מתקיים:

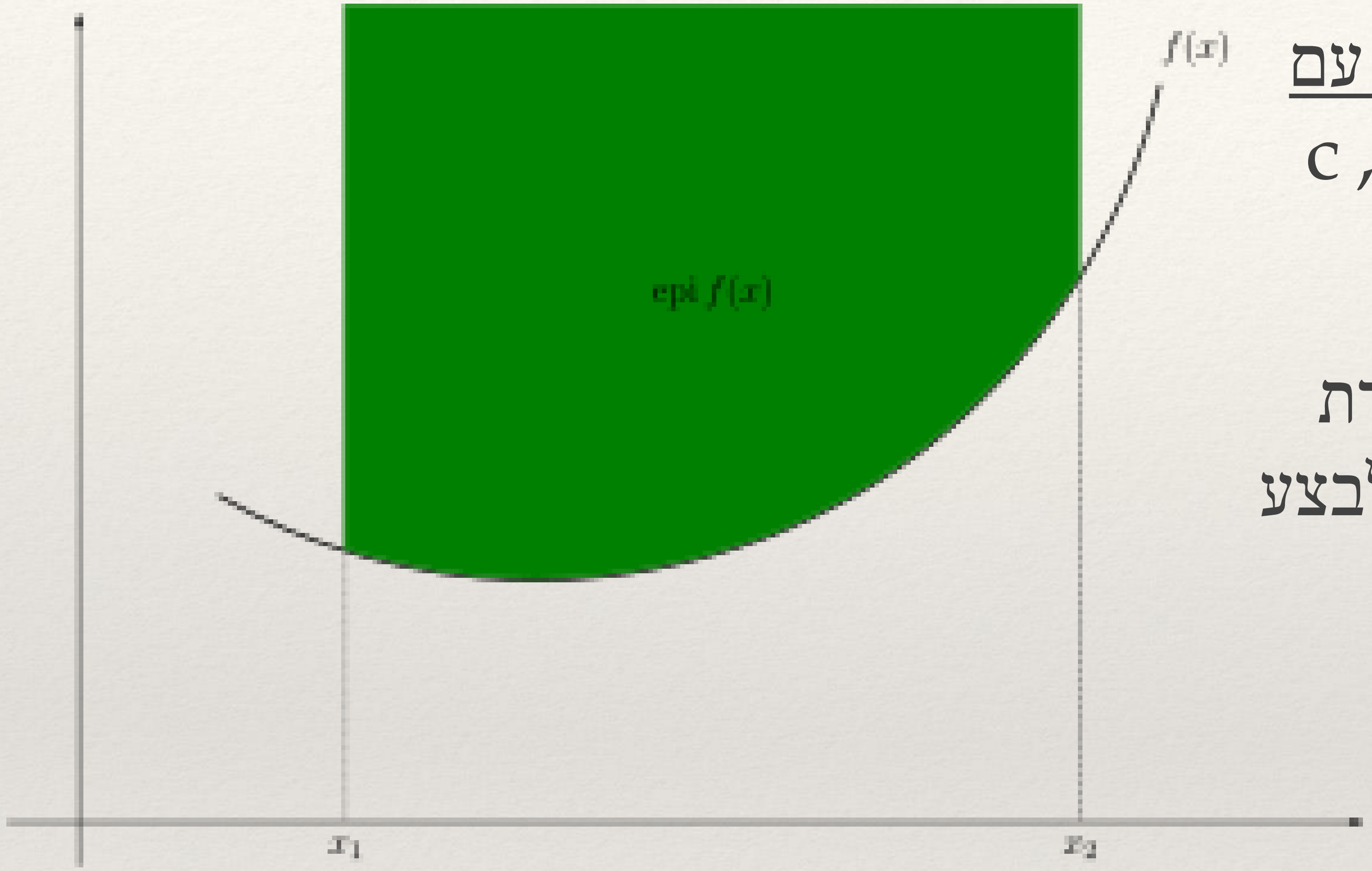
קמירות חלשה:

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$$

קמירות חזקה:

$$f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda) f(x_2)$$

# מושגים – פונקציה קמורה ונקודות קיצון




<u>מינימום גלובלי עבור פונקציה קמורה עם משתנה אחד</u> - בנקודה c, אם f′(c)=0, c היא מינימום גלובלי.

- כאשר יש יותר ממשתנה אחד, נגזרת אינה אפשרית, ולכן נהיה חייבים לבצע אופטימיזציה.
- כיצד עושים זאת? בהמשך ...

<u>קמירות חלשה</u>:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

<u>קמירות חזקה</u>:

$$f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda)f(x_2)$$

# שאלת סקר

1. לאיזה כיוון נצטרך להתקדם, בשביל להתקרב למינימום בפונקציה?

תשובות אפשרויות:

א. נתקדם ע"י ערך הפכי מהערך בו נמצאים כרגע

ב. נתקדם בכיוון הפוך מכיוון הנמדד על ידי הנגזרת בנקודה.

ג. נתקדם בכיוון הנמדד על ידי הנגזרת בנקודה

.

תשובה – **ב.**

# שאלת סקר

2. מהי פונקציה קמורה (פונקציית convex)?

<u>תשובות אפשרויות:</u>

א. פונקציה גזירה ורציפה

ב. פונקציה בה יש נקודת מקסימום

ג. פונקציה בצורת קערה, אשר בה נקודת המינימום היא מינימום גלובלי

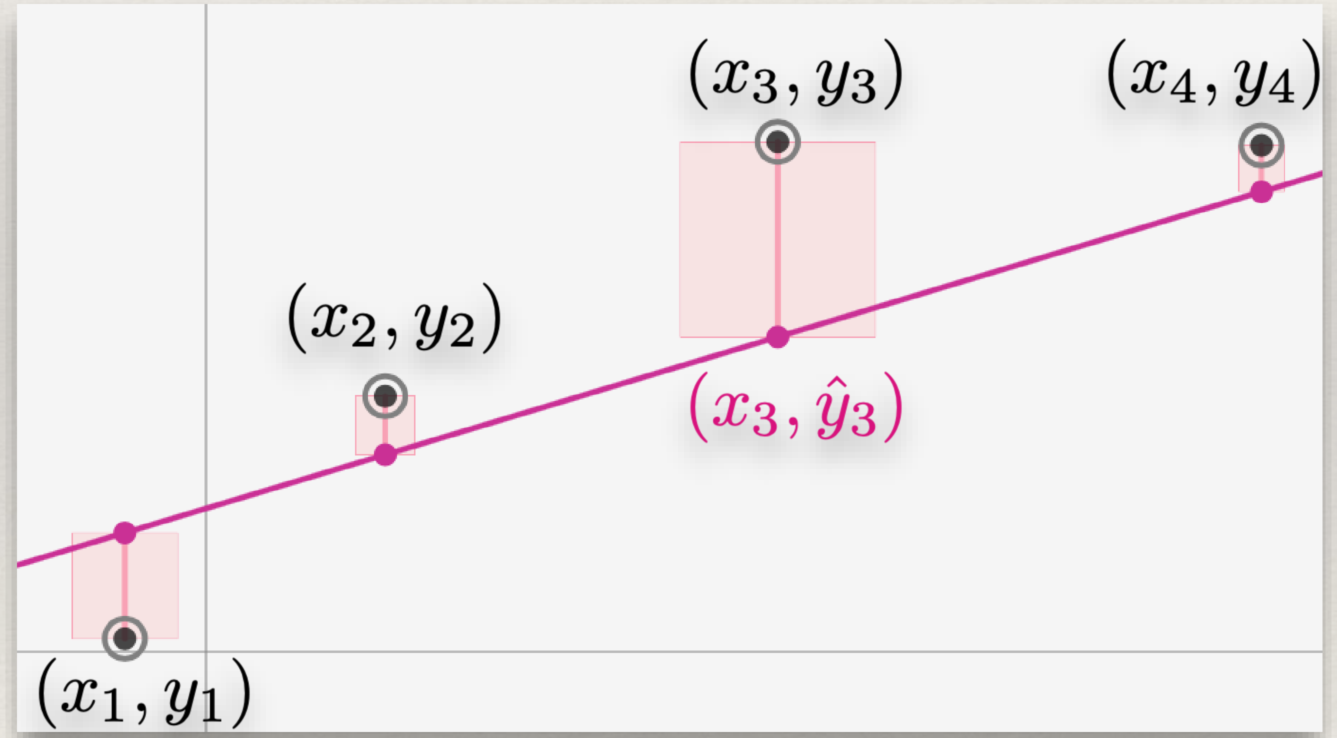
.

<u>תשובה</u> – <u>**ג.**</u>

# חזרה ל- Gradient Descent

# Reminder: Linear Regression (1-D)

*data-set* $\left\{(x_i, y_i)\right\}_{i=1}^{n}$




**linear model:**

$$\hat{y} = f\left(x; \vec{w}\right) = w_0 + w_1 x$$

**cost function:**

$$J\left(\vec{w}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2$$

**optimization problem:**

$$\min_{w}\left[J\left(\vec{w}\right)\right]$$

# Cost Function

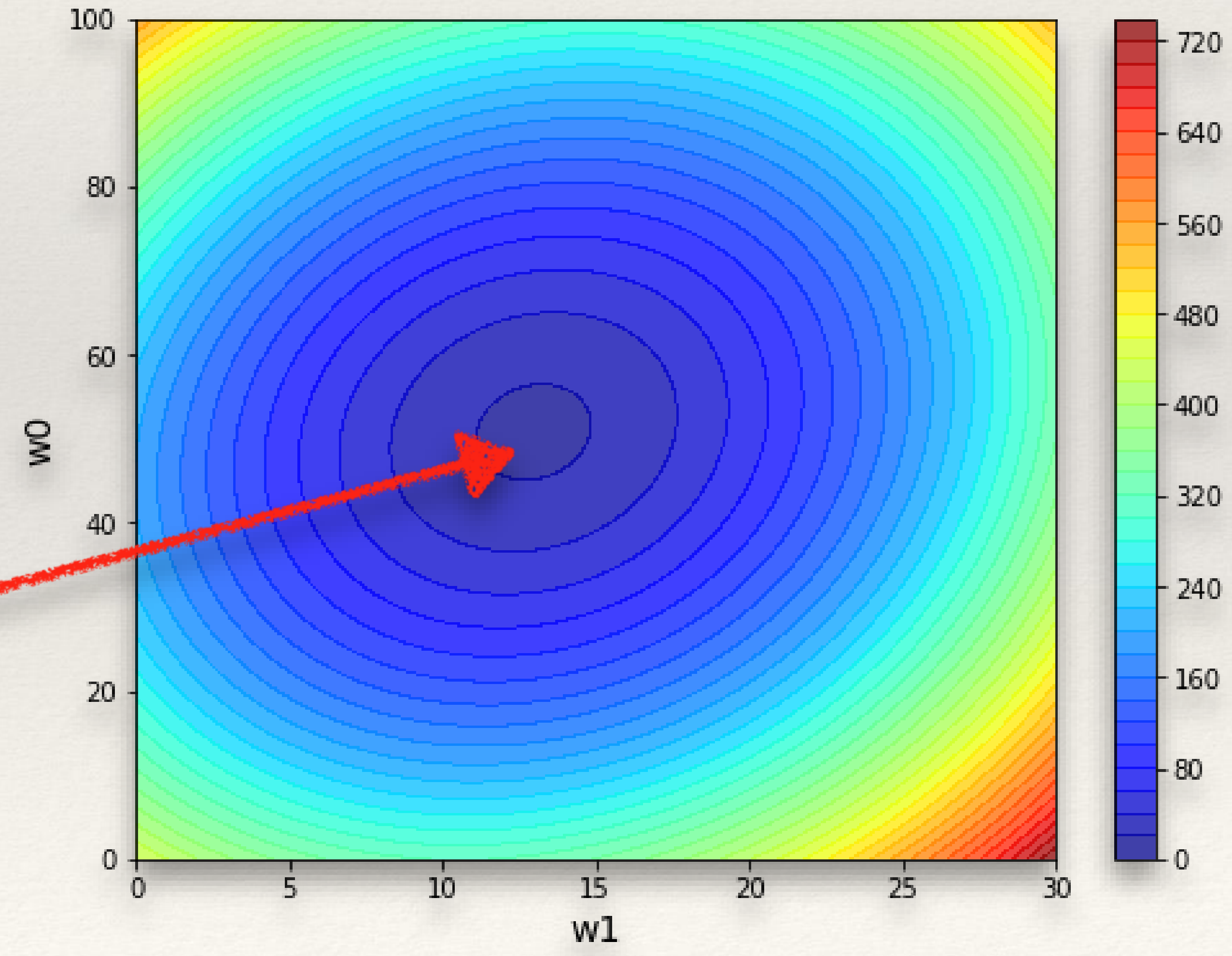
**המודל הליניארי:** $\hat{y} = f\left(x; \vec{w}\right) = w_0 + w_1 x$

$$J\left(\vec{w}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(w_0 + w_1 x_i - y_i\right)^2$$

**הטעות:** $MSE = \frac{1}{n}\sum_{I=1}^{n}(\hat{y}_i - y_i)^2$

$$\min_{w}\left[J\left(\vec{w}\right)\right]$$

$J\left(\vec{w}\right)$



# Gradient Descent

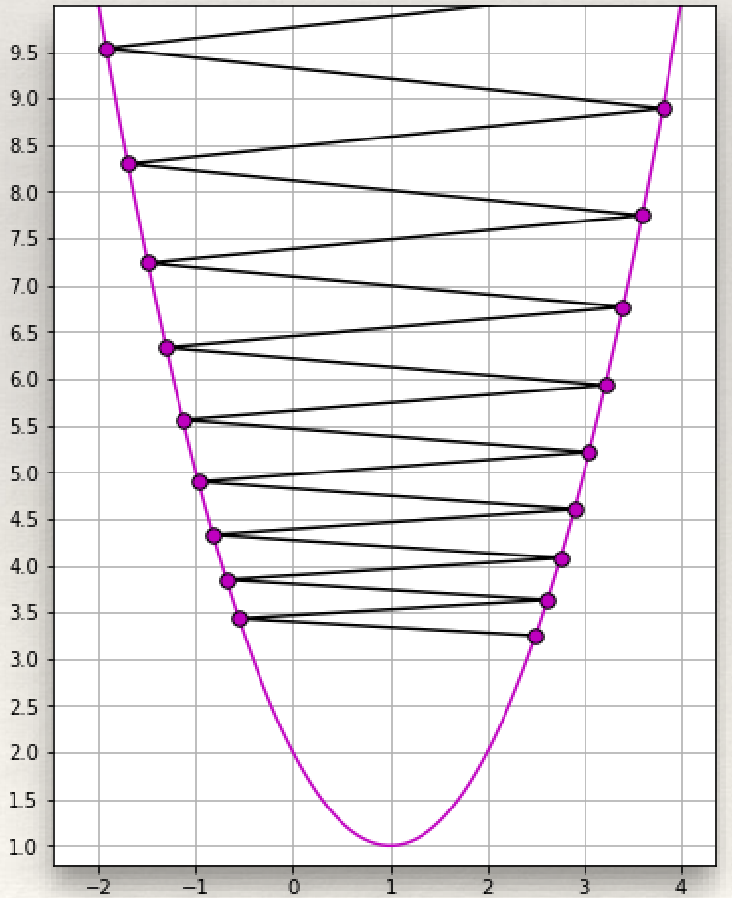
$\vec{x} = (x_1) \longleftarrow$ וקטור עם משתנה אחד

$Gradient = \vec{\nabla}\, \mathrm{f}(x) = \left(\frac{\partial f}{\partial x_1}\right) = (f'(x)) \longleftarrow$ וקטור של הנגזרת

$\vec{x} := \vec{x} - \alpha \nabla f\,(\vec{x}) \longleftarrow$ עדכון של המשתנה

$\uparrow$

*קבוע הלמידה*



# Linear Regression via Gradient Descent

**cost function:** $J\left(\vec{w}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(w_0 + w_1 x_i - y_i\right)^2$

$\vec{x} = (x_1)$

$Gradient = \vec{\nabla}\,\mathrm{f}(x) = \left(\frac{\partial f}{\partial x_1}\right) = (f'(x))$

$$\frac{\partial J}{\partial w_0} = \frac{2}{n}\sum_{i=1}^{n}\left(w_0 + w_1 x_i - y_i\right)\cdot 1$$

יוסבר בהמשך

$$\frac{\partial J}{\partial w_1} = \frac{2}{n}\sum_{i=1}^{n}\left(w_0 + w_1 x_i - y_i\right)\cdot x_i$$

הנגזרות החלקיות

# Linear Regression via Gradient Descent

$$\frac{\partial J}{\partial w_0} = \frac{2}{n}\sum_{i=1}^{n}(w_0 + w_1 x_i - y_i)\cdot 1$$

יוסבר בהמשך

$$\frac{\partial J}{\partial w_1} = \frac{2}{n}\sum_{i=1}^{n}(w_0 + w_1 x_i - y_i)\cdot x_i$$

הנגזרות החלקיות

$$w_0 = w_0 - \alpha\cdot\frac{\partial J}{\partial w_0} = w_0 - \alpha\cdot\frac{2}{n}\cdot\sum_{i=1}^{n}[(w_0 + w_1\cdot x_i - y_i)\cdot 1]$$

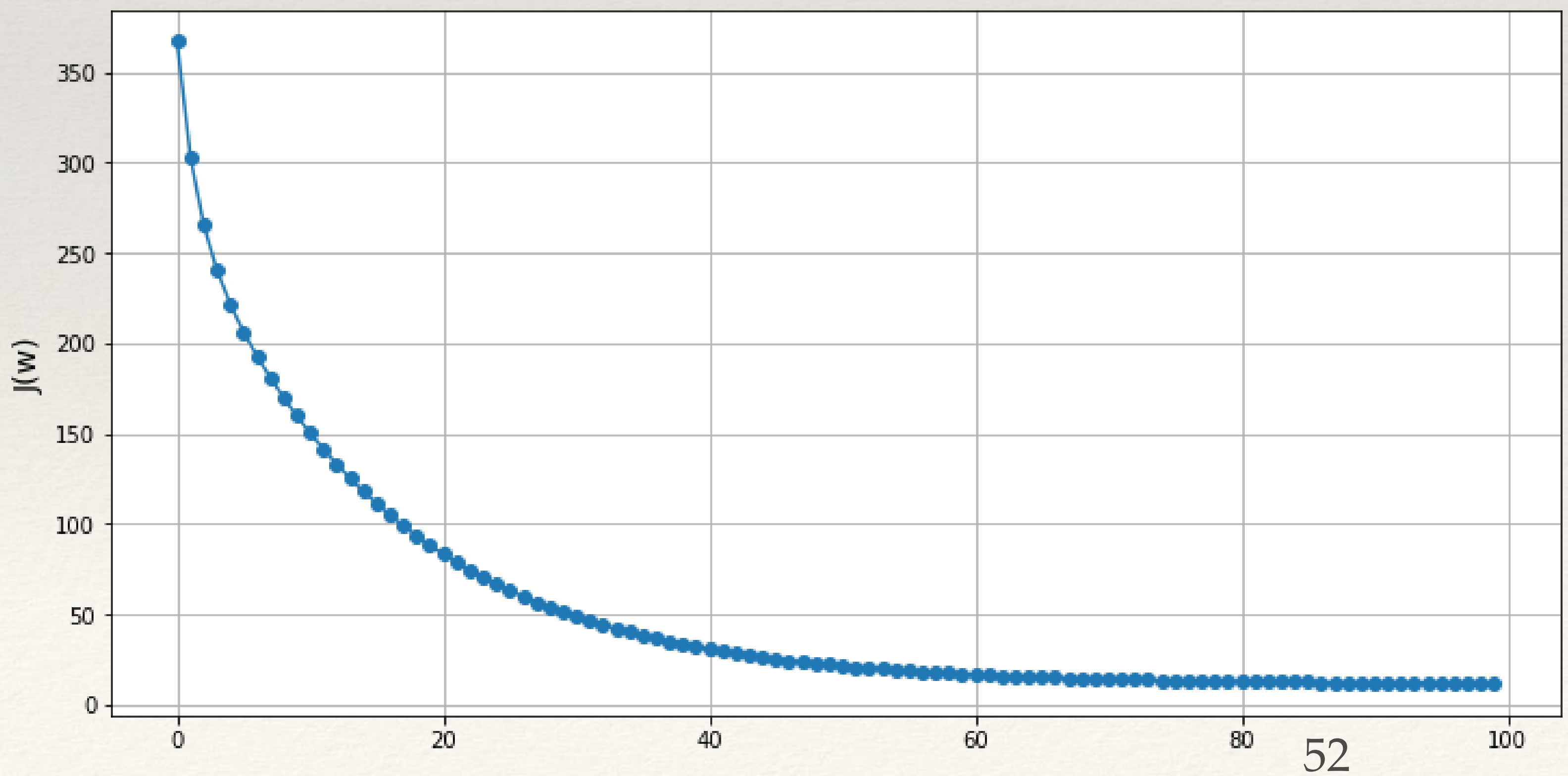
$$w_1 = w_1 - \alpha\cdot\frac{\partial J}{\partial w_1} = w_1 - \alpha\cdot\frac{2}{n}\cdot\sum_{i=1}^{n}[(w_0 + w_1\cdot x_i - y_i)\cdot x_i]$$

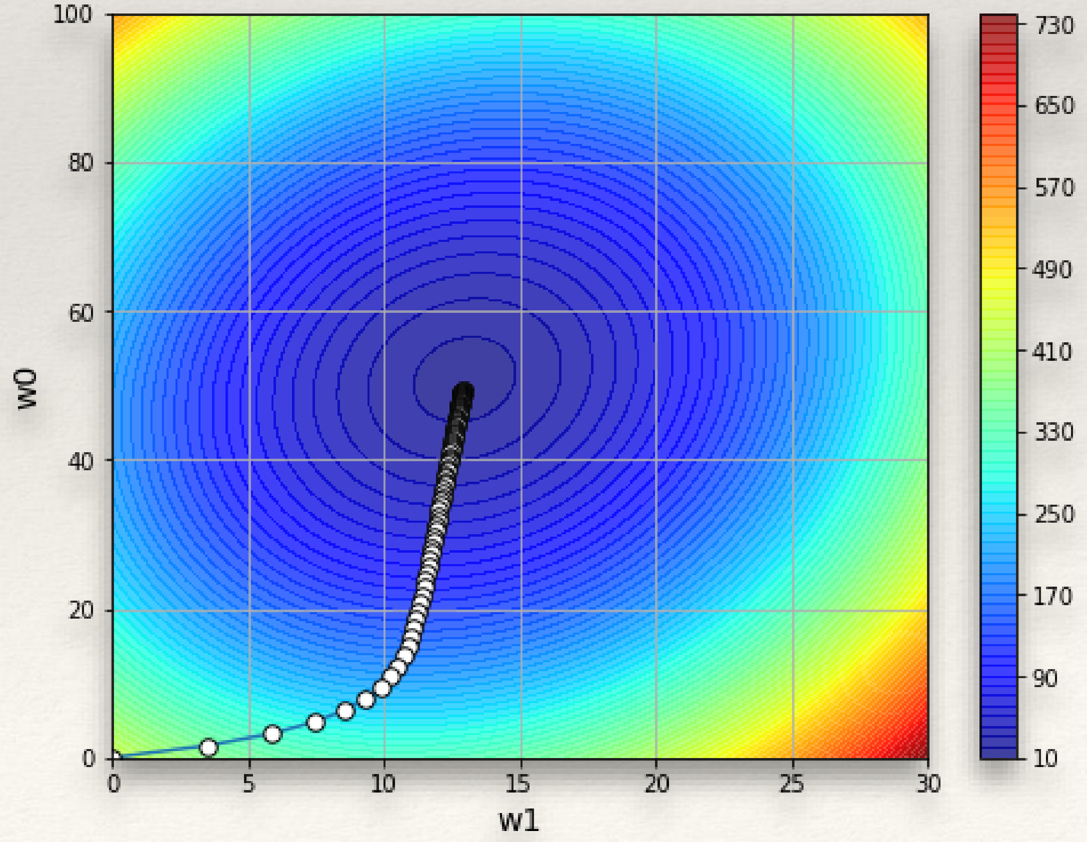
עדכון הפרמטרים

# Linear Regression via Gradient Descent

$$w_0 := w_0 - \frac{2\alpha}{n}\sum_{i=1}^{n}(w_0 + w_1 x_i - y_i)$$

$$w_1 := w_1 - \frac{2\alpha}{n}\sum_{i=1}^{n}(w_0 + w_1 x_i - y_i)\cdot x_i$$



# רגרסיה לינארית (linear regression)
# ריבוי משתנים (multivariate) - מוטיבציה

# Multivariate Linear Regression

- קודם, לכל דוגמא היה רק **מאפיין יחיד** (למשל: גודל הדירה)
- עכשיו, לכל דוגמא יש **כמה מאפיינים** (למשל: גודל הדירה, קומה, כיווני-אוויר וכו')

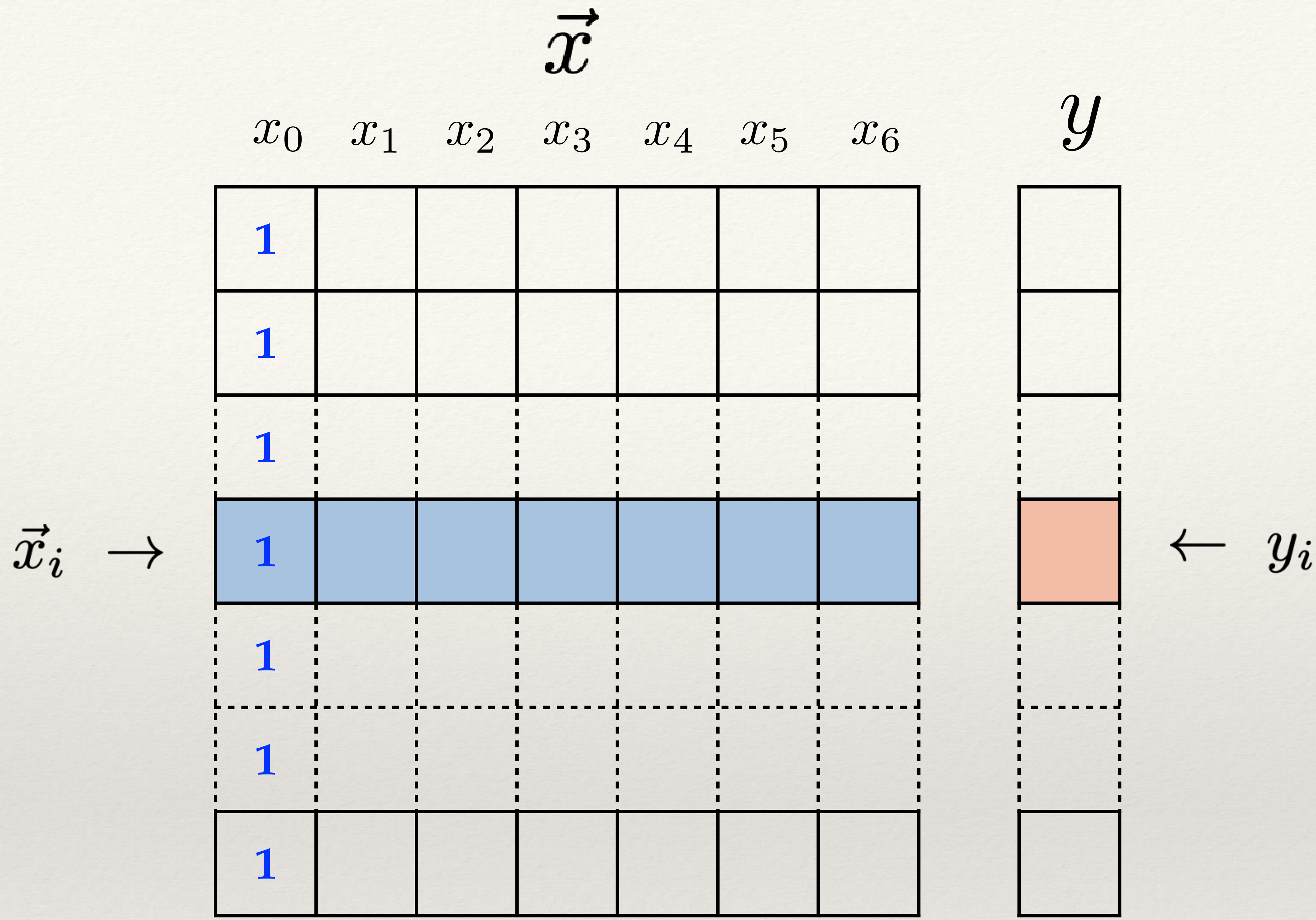
$$\vec{x} = (x_1, x_2, ..., x_d) \in \mathbb{R}^d$$

- נתונות $n$ דוגמאות, ולכל אחת $d$ מאפיינים:

$$\left\{ (\vec{x}_i, y_i) \right\}_{i=1}^{n}$$

$$\left[ \vec{x}_i \right]_j = x_{i,j}$$

הרכיב ה $j$ של הדוגמא ה $i$




המודל הלינארי -עבור רגרסיה מרובה משתנים:

$$\hat{y} = w_0 x_0 + w_1 x_1 + w_2 x_2 + \ldots + w_6 x_6$$

קומבינציה לינארית של המאפיינים

# Multivariate Linear Regression

❖ התחזית של המודל עבור הדוגמא ה-$i$:

$$\hat{y}_i = \vec{w} \cdot \vec{x}_i \qquad\qquad i = 1, \ldots, n$$

❖ נרצה שהתחזיות יהיו קרובות לנתונים:

$$\hat{y}_i \approx y_i$$

❖ נשתמש (שוב) בשגיאה ריבועית (MSE):

$$J\left(\vec{w}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\vec{w} \cdot \vec{x}_i - y_i\right)^2$$

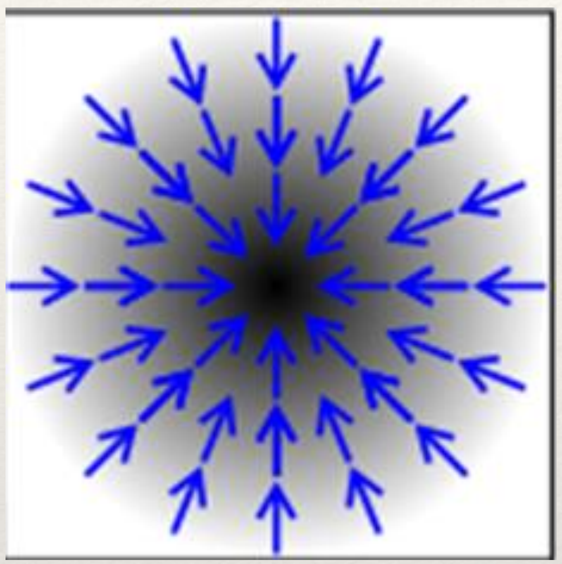
## מושגים מתמטיים – עבור פונקציה רבת משתנים

# מושגים – נגזרת של פונקציה עם כמה משתנים

- פונקציה דיפרנציאבילית - פונקציה ממשית בעלת כמה משתנים, שיש לה קירוב ליניארי (דיפרנציאל).
- נגזרת חלקית (partial derivative) - נגזרת חלקית של פונקציה בכמה משתנים היא נגזרת של הפונקציה באחד ממשתניה.
- עבור הפונקציה $f = a_1 \cdot x_1 + a_2 \cdot {x_2}^2 + \cdots + a_n \cdot x_n + b$
- עבור $x_2$, נסמן את הנגזרת החלקית כך: $\frac{\partial}{\partial x_2} f$ (∂ - d מסולסלת)
- למשל עבור $f(x,y) = x^y$
- $\frac{\partial f}{\partial_y} = \ln(x) \cdot x^y$ $\qquad$ $\frac{\partial f}{\partial_x} = y \cdot x^{y-1}$

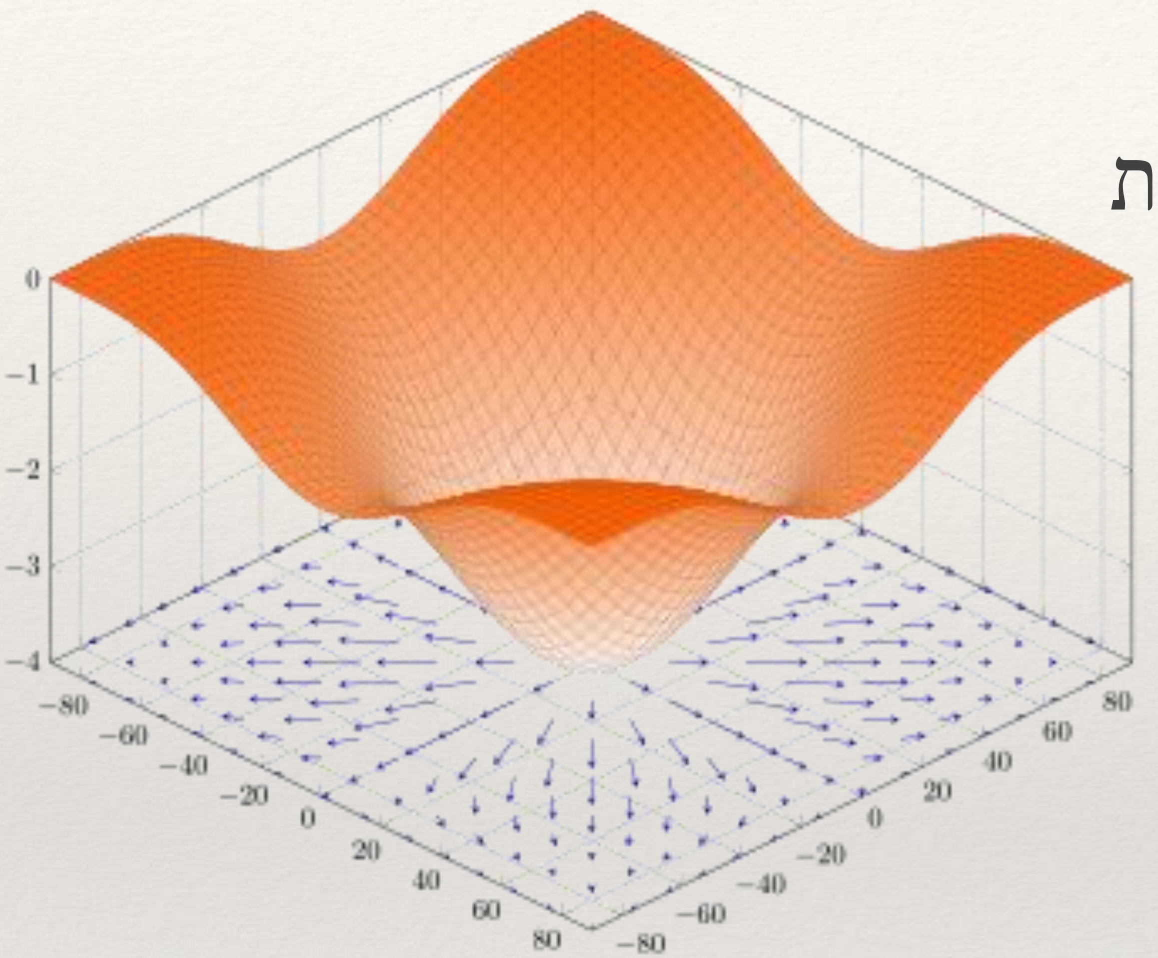
# מושגים – גרדיאנט

- פונקציה דיפרנציאבילית - פונקציה ממשית בעלת כמה משתנים, שיש לה קירוב ליניארי (דיפרנציאל).
  - פונקציה עם כמה משתנים, נקראת גם פונקציה וקטורית
- נגזרת חלקית - נגזרת חלקית של פונקציה בכמה משתנים היא נגזרת של הפונקציה באחד ממשתניה.
- גראדיאנט (gradiant) – גרדאינט של פונקציה וקטורית, הוא הוקטור של הנגזרות החלקיות.
- $\text{grad } \mathrm{f}(\mathrm{a}) = \vec{\nabla}\, \mathrm{f}(a) = (\frac{\partial}{\partial_{x_1}} f(a), \frac{\partial}{\partial_{x_2}} f(a), \dots, \frac{\partial}{\partial_{x_n}} f(a))$



האזורים הקהים, בעלי ערכים גבוהים יותר

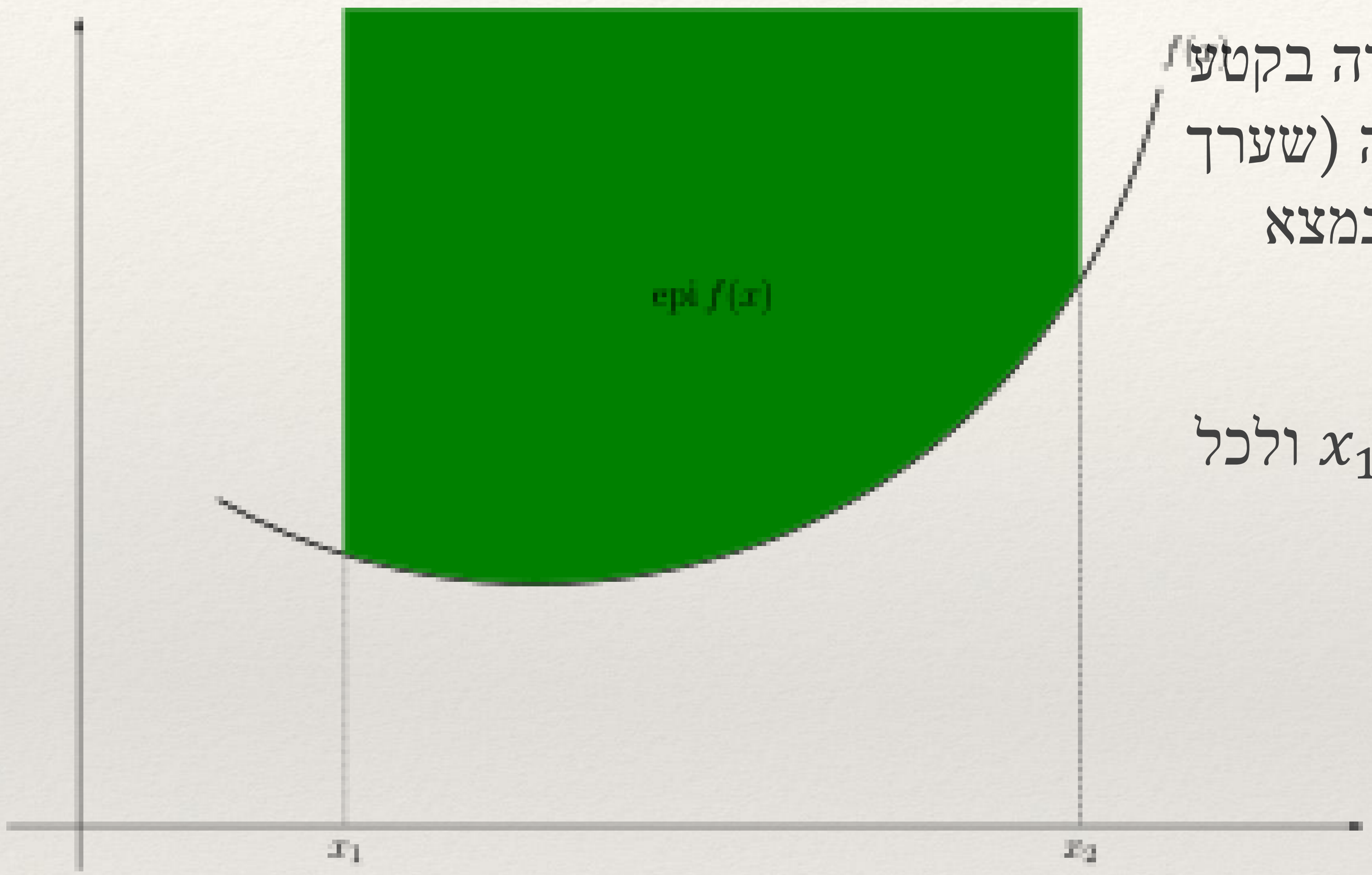
# מושגים – גרדיאנט




The gradient of the function
$f(x,y) = -(\cos^2 x + \cos^2 y)^2$

- גראדיאנט (gradiant) – גרדאינט של פונקציה וקטורית, הוא הוקטור של הנגזרות החלקיות.
- $\text{grad } f(a) = \vec{\nabla} f(a)$ <u>כיוון וקטור</u> <u>הגרדיאנט</u> מצביע אל הכיוון בו השינוי בשדה הסקלרי מקסימלי (חיובי). גודל וקטור הגרדיאנט כשיעור השינוי המקסימלי

# תזכורת – פונקציה קמורה (Convex)




פונקציה קמורה (Convex) - פונקציה קמורה בקטע מסוים, אם לכל שתי נקודות על גרף הפונקציה (שערך ה- x שלהן נמצא בקטע), הקו המחבר ביניהן נמצא מעל לגרף הפונקציה (או עליו).

עבור הקטע $I$, הפו' קמורה, אם לכל $x_1, x_2 \in \mathrm{I}$ ולכל סקלר $0 \leq \lambda \leq 1$, מתקיים:

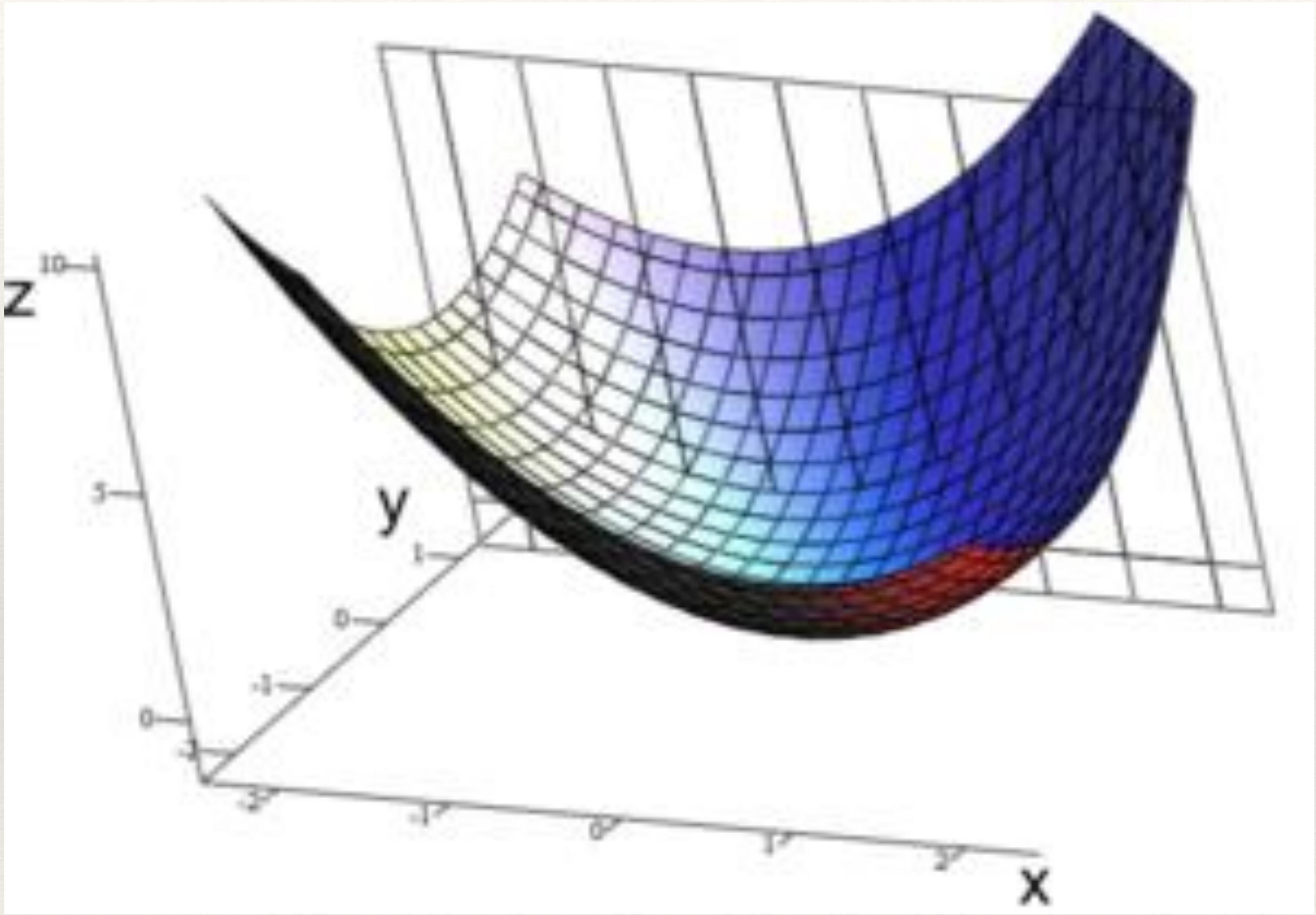
<u>קמירות חלשה</u>:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

<u>קמירות חזקה</u>:

$$f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda) f(x_2)$$

# מושגים – פונקציה קמורה – רב מימדית




A graph of the bivariate convex function $x^2 + xy + y^2$

# מושגים – מטריצת הסיאן (Hessian)

$$\mathbf{H} = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n} \\[2.2ex] \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n} \\[2.2ex] \vdots & \vdots & \ddots & \vdots \\[2.2ex] \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

<u>מטריצת הסיאן</u> - היא מטריצה ריבועית, שאיבריה הם הנגזרות החלקיות מסדר שני של פונקציה.

עבור הנקודה a: $\vec{a} = (a_1, \dots, a_n)$

$[H(f)]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$ - הערך של האיבר ij הוא הערך של הנגזרת השניה של f בנקודה a, כאשר קודם גוזרים על פי המשתנה $x_j$ ואח"כ לפי המשתנה $x_i$

# מושגים – מינימום בפונקציה רבת משתנים

$$\mathbf{H} = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

<u>מטריצת הסיאן</u> - היא מטריצה ריבועית, שאיבריה הם הנגזרות החלקיות מסדר שני של פונקציה.

עבור הנקודה a: $\vec{a} = (a_1, \dots, a_n)$

$[H(f)]_{ij} = \dfrac{\partial^2 f}{\partial x_i \partial x_j}$ - הערך של האיבר ij הוא הערך של הנגזרת השניה של f בנקודה a, כאשר קודם גוזרים על פי המשתנה $x_j$ ואח"כ לפי המשתנה $x_i$

<u>מינימום בפונקציה רבת משתנים:</u>

עבור פונקציה f, אם מתקיים שבנקודה a הגרדיאנט $\vec{\nabla}$ f$(a)$ (וקטור הנגזרות החלקיות) הינו וקטור ה-0 ומתקיים שמטריצת ההסיאן בנקודה a, חיובית עבור כל ערכיה (ערכי הנגזרות השניה), הנקודה היא נקודת מינימום

# שאלת סקר

1. מהו גראדיאנט?

<u>תשובות אפשרויות</u>:
א. שיטה לחישוב המשקולות ברגרסיה לינארית
ב. נקודת המינימום בפונקציית convex.
ג. וקטור הנגזרות החלקיות של פונקציה רב מימדית
.

<u>תשובה</u> – <u>**ג.**</u>

# שאלת סקר

2. מהי מטריצת הסיאן ומדוע לא נשתמש בה לאימון מודל רגרסיה לינארית?

<u>תשובות אפשרויות:</u>
א. מטריצת הסיאן היא מטריצת הנגזרות החלקיות השניות של פונקציה רב מימדית, לא משתמשים בה, בגלל הקושי לחשב את ערכיה.
ב. מטריצת הסיאן היא מטריצת הנגזרות החלקיות השניות של פונקציה רב מימדית, כן משתמשים בה, בגלל הוודאות שבמציאת נקודת המינימום
ג. מטריצת הסיאן היא מטריצת של הגראדיאנטים של ההיפותזות השונות, ולא ניתן לדעת אם זו פונקציית convex.

<u>תשובה</u> – <u>**א.**</u>
הערה – אכן מטריצת הסיאן, אם יש לנו, יכולה לחשב את נקודת המינימום, אולם קשה לחשב אותה.

# חזרה ל- multivariate linear regression

# פונקצית מחיר

$$J\left(\vec{w}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\vec{w}\cdot\vec{x}_i - y_i\right)^2$$

$$\min_{w}\left[J\left(\vec{w}\right)\right] \xrightarrow{\text{Gradient Descent}} \vec{w} := \vec{w} - \alpha\nabla J\left(\vec{w}\right)$$

$$\nabla J = \left(\frac{\partial J}{\partial w_0}, \frac{\partial J}{\partial w_1}, \frac{\partial J}{\partial w_2}, \ldots, \frac{\partial J}{\partial w_d}\right)$$

נחשב את הגרדיאנט (כלומר את כל הנגזרות החלקיות ... )

# חישוב הגרדיאנט – ע"י הנגזרות החלקיות

$$J\left(\vec{w}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\vec{w}\cdot\vec{x}_i - y_i\right)^2$$

- נחשב לדוגמא את הנגזרת ביחס ל : $w_2$

$$\frac{\partial J}{\partial w_2} = \frac{2}{n}\sum_{i=1}^{n}\left(\vec{w}\cdot\vec{x}_i - y_i\right)\cdot\frac{\partial\left(\vec{w}\cdot\vec{x}_i\right)}{\partial w_2}$$

# חישוב הגרדיאנט – ע"י הנגזרות החלקיות

$$J(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\vec{w}\cdot\vec{x}_i - y_i)^2$$

- נחשב לדוגמא את הנגזרת ביחס ל : $w_2$

$$\frac{\partial J}{\partial w_2} = \frac{2}{n}\sum_{i=1}^{n}(\vec{w}\cdot\vec{x}_i - y_i)\cdot\frac{\partial(\vec{w}\cdot\vec{x}_i)}{\partial w_2}$$

- תזכורת: $\vec{w}\cdot\vec{x}_i = w_0 x_{i,0} + w_1 x_{i,1} + w_2 x_{i,2} + \ldots + w_d x_{i,d}$

$$\frac{\partial J}{\partial w_2} = \frac{2}{n}\sum_{i=1}^{n}(\vec{w}\cdot\vec{x}_i - y_i)\cdot x_{i,2}$$

# חישוב הגרדיאנט – ע"י הנגזרות החלקיות

$$J\left(\vec{w}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\vec{w}\cdot\vec{x}_i - y_i\right)^2$$

- תזכורת:

$$\vec{w}\cdot\vec{x}_i = w_0 x_{i,0} + w_1 x_{i,1} + w_2 x_{i,2} + \ldots + w_d x_{i,d}$$

- ובאופן דומה:

$$\frac{\partial J}{\partial w_0} = \frac{2}{n}\sum_{i=1}^{n}\left(\vec{w}\cdot\vec{x}_i - y_i\right)\cdot x_{i,0}$$

$$\frac{\partial J}{\partial w_1} = \frac{2}{n}\sum_{i=1}^{n}\left(\vec{w}\cdot\vec{x}_i - y_i\right)\cdot x_{i,1}$$

$$\vdots$$

$$\frac{\partial J}{\partial w_\mathrm{d}} = \frac{2}{n}\sum_{i=1}^{n}\left(\vec{w}\cdot\vec{x}_i - y_i\right)\cdot x_{i,d}$$

# חישוב הגרדיאנט ע"י הנגזרות החלקיות – הסבר

$$\nabla J = \frac{2}{n}\sum_{i=1}^{n}(\vec{w}\cdot\vec{x}_i - y_i)\cdot\vec{x}_i$$

$$\frac{\partial J}{\partial w_0} = \frac{2}{n}\sum_{i=1}^{n}(\vec{w}\cdot\vec{x}_i - y_i)\cdot x_{i,0}$$

$$\frac{\partial J}{\partial w_1} = \frac{2}{n}\sum_{i=1}^{n}(\vec{w}\cdot\vec{x}_i - y_i)\cdot x_{i,1}$$

$$\vdots$$

$$\frac{\partial J}{\partial w_d} = \frac{2}{n}\sum_{i=1}^{n}(\vec{w}\cdot\vec{x}_i - y_i)\cdot x_{i,d}$$

# אלגוריתם ה- Gradient Descent

$$\min_{w}\left[J\left(\vec{w}\right)\right] \xrightarrow{\text{Gradient Descent}} \boxed{\vec{w} := \vec{w} - \alpha \nabla J\left(\vec{w}\right)}$$

והגרדיאנט שחישבנו:

$$\nabla J = \frac{2}{n}\sum_{i=1}^{n}\left(\vec{w}\cdot\vec{x}_i - y_i\right)\cdot\vec{x}_i$$

**Gradient Descent:**

$$\boxed{\vec{w} := \vec{w} - \frac{2\alpha}{n}\sum_{i=1}^{n}\left(\vec{w}\cdot\vec{x}_i - y_i\right)\cdot\vec{x}_i}$$

# סיכום – Gradient Descent (לרגרסיה לינארית)

**univariate linear regression**

$$\left\{(x_i, y_i)\right\}_{i=1}^{n}$$

$$\hat{y}_i = w_0 + w_1 x_i$$

$$J\left(\vec{w}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2$$

$$\min_{w}\left[J\left(\vec{w}\right)\right]$$

$$w_0 := w_0 - \frac{2\alpha}{n}\sum_{i=1}^{n}\left(w_0 + w_1 x_i - y_i\right)$$

$$w_1 := w_1 - \frac{2\alpha}{n}\sum_{i=1}^{n}\left(w_0 + w_1 x_i - y_i\right)\cdot x_i$$

**multivariate linear regression**

$$\left\{(\vec{x}_i, y_i)\right\}_{i=1}^{n}$$

$$\hat{y}_i = \vec{w}\cdot\vec{x}_i$$

$$J\left(\vec{w}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2$$

$$\min_{w}\left[J\left(\vec{w}\right)\right]$$

$$\vec{w} := \vec{w} - \frac{2\alpha}{n}\sum_{i=1}^{n}\left(\vec{w}\cdot\vec{x}_i - y_i\right)\cdot\vec{x}_i$$

# רגרסיה לינארית – יתרונות וחסרונות

יתרונות:
- קל למימוש, להבנה והסבר
- ניתן להסיק על חשיבות המאפיינים
- עובד די טוב גם עם train-set קטן
- זמן אימון מהיר
- סיבוכיות מקום נמוכה
- רוב החסרונות ברות טיפול (למשל טיפול ב-overfitting  בעזרת regularization)

חסרונות:
- לא מתאים כשאין קשר לינארי ומתקשה שההיפותזה מורכבת
- נטיה ל-overfitting – במיוחד בריבוי מאפיינים
- מתקשה לטפל במאפיינים לא רלוונטים וברעש
- לא עובד טוב ללא סילום (scaling)
- צריך לוודא חוסר תלות בין המאפיינים
- הטעות צריכה להתפלג נורמלית