Machine learning

Linear regression

Exercise VI

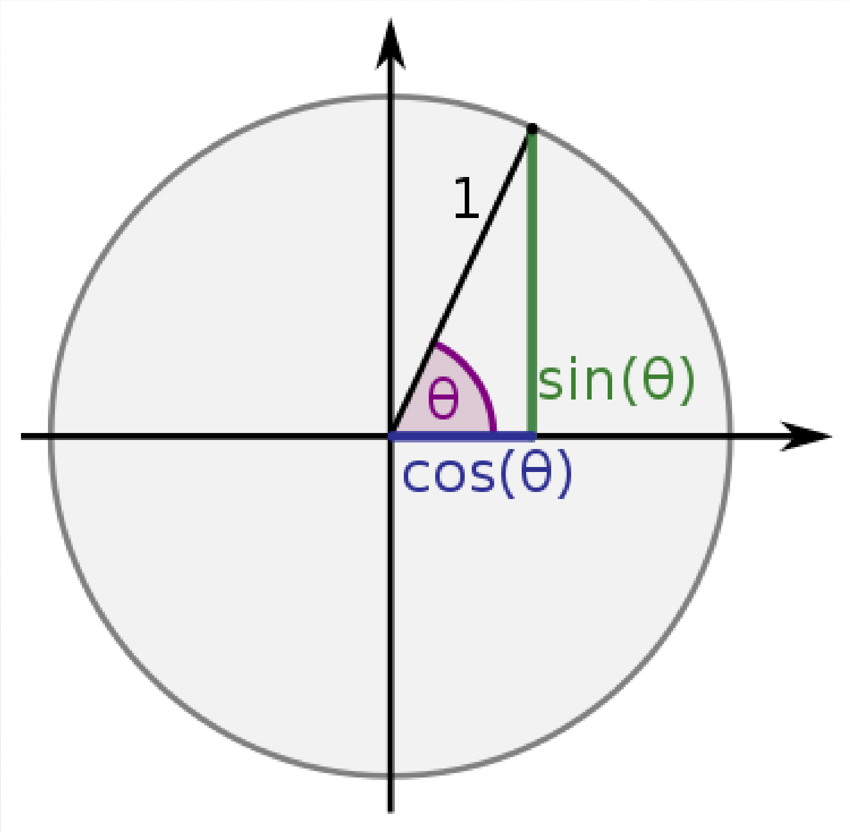
תודות לד"ר יונתן רובין שעזר בהכנת המצגת

קרדיט גם ל- Andrew Ng

פיתוח:
ד"ר יהונתן שלר
משה פרידמן

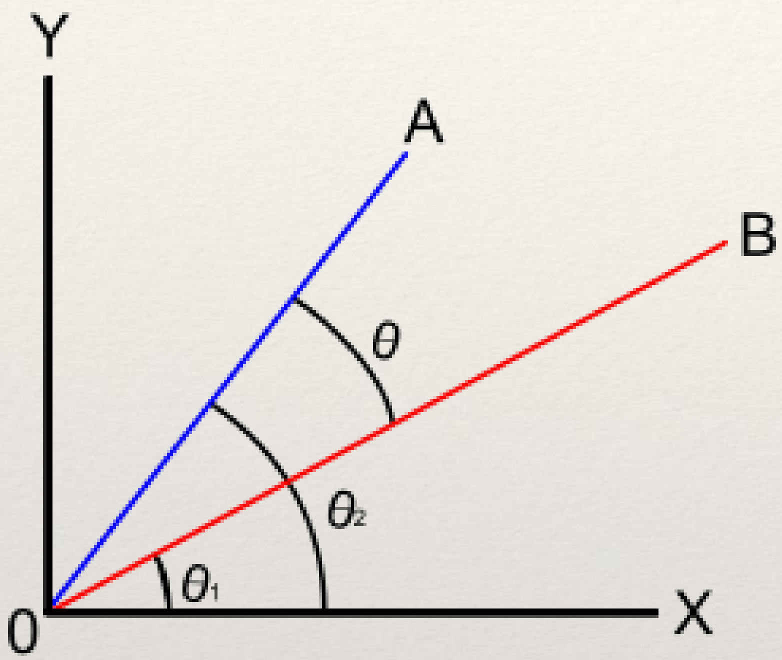
וקטורים ופעולות על וקטורים

מעגל היחידה



- מעגל היחידה – מעגל עם רדיוס r=1
- $\sin\theta$ - ההיטל על ציר ה-y
  - שווה לערך ה-y של נקודה במעגל יחידה, כאשר הזווית בין ציר ה-x לרדיוס הוא θ
- $\cos\theta$ - ההיטל על ציר ה-x
  - שווה לערך ה-x של נקודה במעגל יחידה, כאשר הזווית בין ציר ה-x לרדיוס הוא θ
- לפי פיתגורס - $\sin^2(\theta) + \cos^2(\theta) = 1$

וקטור (פרספקטיבה גאומטרית)



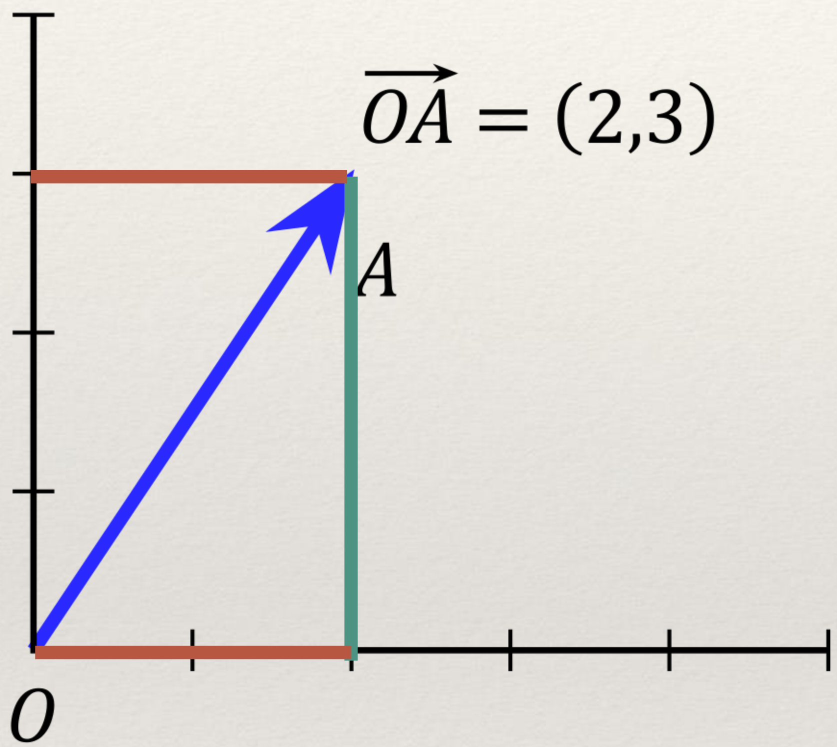
וקטור (vector) - נסמן $\vec{x},\vec{y} \in \mathbb{R}^n$

- הסבר גאומטרי: ישות מתמטית בעלת גודל וכיוון
- נורמה של וקטור - הכללה של מושג ה"אורך" (magnitude) – נסמן $||\vec{x}||$
- כל וקטור $\vec{x}$ מוגדר מבחינה גאומטרית ע"י הנורמה $||\vec{x}||$ והזווית θ_x מציר ה-x

שאלה 1 (סקר) - וקטור ואורך של וקטור

<u>וקטור (vector)</u> - נסמן $\vec{x},\vec{y} \in \mathbb{R}^n$

- <u>הסבר אלגברי</u> (והקשר להסבר הגאומטרי) – ב-$\mathbb{R}^2$ ערכי x, y של הנקודה, זהים לוקטור שתחילתו בראשית הצרים ואורכו כאורך הקו הישר בין הראשית לנקודה. הזוית הנוצרת בין ציר ה-x, לישר שבין הראשית לנקודה זו.



<u>תרגיל (סקר)</u> – חשבו את אורך הוקטור שבתמונה, מבחינת יצוגו הגאומטרי במישור.

<u>תשובות אפשרויות</u>:

א. 4 ב. 2.5 ג. 3.6 ד.14

<u>תשובה</u> – נשתמש בערכי x,y וניצור משולש ההיטלים על ציר ה-x וציר ה-y

- לפי משפט פיתגורס - $\sqrt{2^2+3^2} = \sqrt{13} \approx 3.6$

שאלה 2 - נורמה של וקטור (פרספקטיבה אלגברית)

נורמה של וקטור - הכללה של מושג ה"אורך" (magnitude) – נסמן $\|\vec{x}\|$

- נורמה סטנדרטית (אוקלדית) - נציב p=2

$$L_2 = \|\vec{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}} = \sqrt{x_1 \cdot x_1 + x_2 \cdot x_2 + \cdots + x_n \cdot x_n} =$$

$$\sqrt{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2} = \sqrt{\vec{x}^T \cdot \vec{x}} = \sqrt{\langle \vec{x}, \vec{x} \rangle}$$

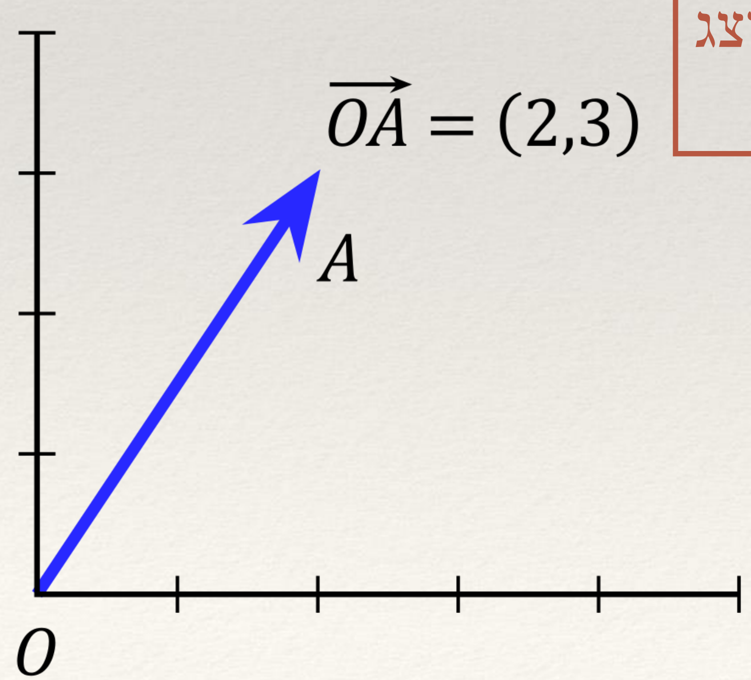
תרגיל - מה ערך הנורמה (הסטנדרטית) של הוקטור שבתמונה?

תשובה –

$$\sqrt{(2,3)^T \cdot (2,3)} = \sqrt{2 \cdot 2 + 3 \cdot 3}$$

$$= \sqrt{13} \approx 3.6$$

שימו לב - חישוב נורמה סטנדרטית, זהה לחישוב המוצג פה של אורך הווקטור.



כפל בסקלר

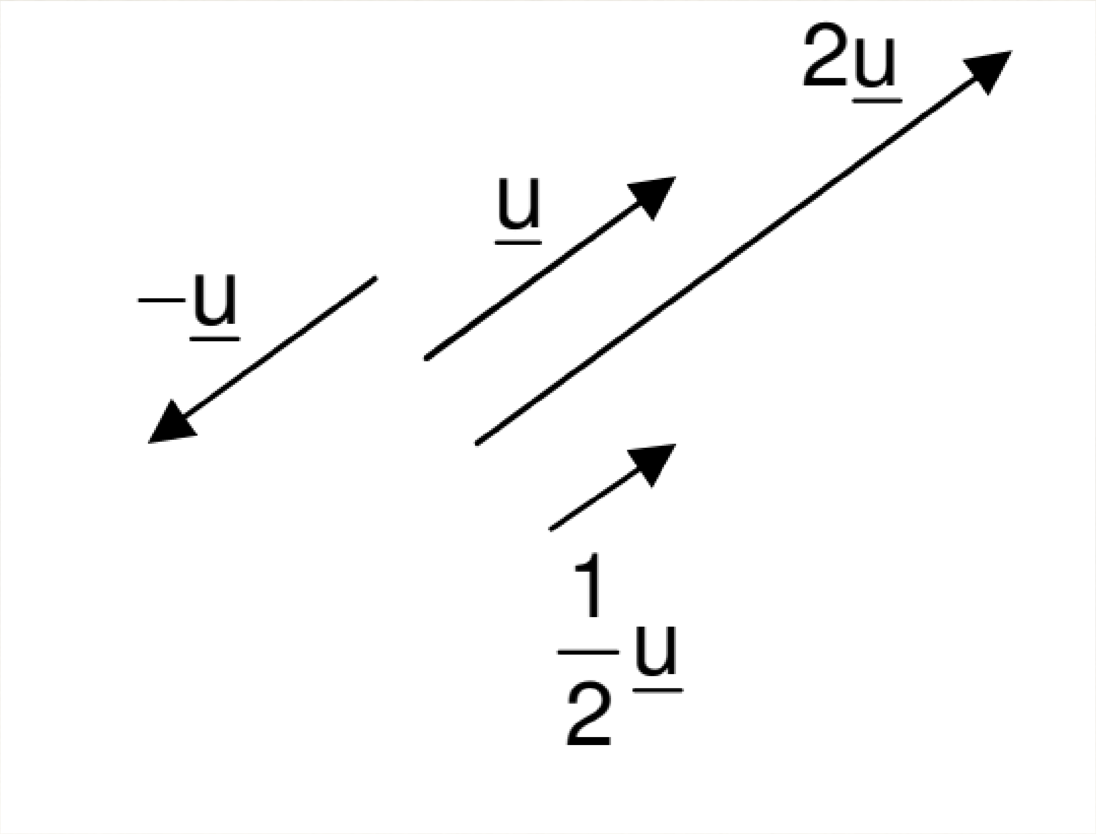
וקטור (vector) - נסמן $\vec{x},\vec{y} \in \mathbb{R}^n$

- $\vec{x} = (x_1, x_2, \dots, x_n)$

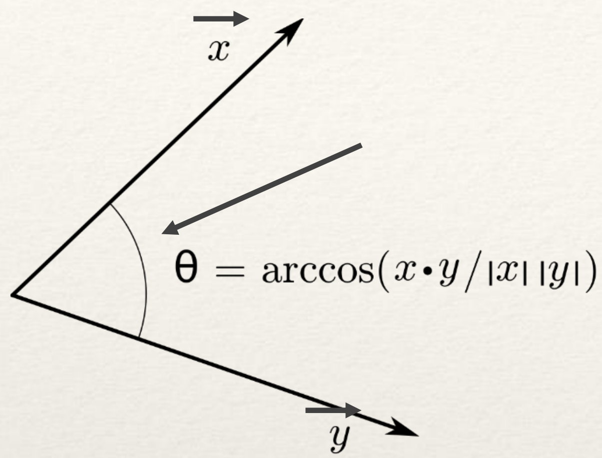
כפל (של וקטור) בסקלר –

$$\vec{x} \cdot \lambda = \lambda \cdot \vec{x} = (\lambda \cdot x_1, \lambda \cdot x_2, \dots, \lambda \cdot x_n)$$

- מבחינה גאומטרית, הזווית לא משתנה, רק "אורך" (הנורמה) הוקטור גדל/קטן (בהתאם לערך של λ)
- כפל של וקטור בסקלר שלילי, הופך גם את כיוונו, ב-$180°$.



שאלה 3 (סקר) מכפלה סקלרית (פרספקטיבה גאומטרית)



θ=הזווית בין הוקטורים $\vec{x},\vec{y}$

<u>מכפלה סקלרית (dot product) של וקטורים</u> – נסמן $\vec{x} \cdot \vec{y}$

מבחינה גאומטרית - $\vec{x} \cdot \vec{y} = \|\vec{x}\| \cdot \|\vec{y}\| \cdot \cos\theta$

- שימו לב שמכאן נובע $\cos\theta = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \cdot \|\vec{y}\|}$

<u>מסקנות נוספות</u>

- אם $\theta = 0^\circ$, המכפלה הסקלרית = $\|\vec{x}\| \cdot \|\vec{y}\|$, כיוון ש-$\cos 0 = 1$
- אם $\theta = 90^\circ$, המכפלה הסקלרית = 0, כיוון ש-$\cos 90 = 0$
- אם $\theta = 180^\circ$, המכפלה הסקלרית = $-1 \cdot \|\vec{x}\| \cdot \|\vec{y}\|$, כיוון ש-$\cos 180 = -1$

<u>תרגיל</u> – נתון $\|\vec{y}\| = 3$ $\|\vec{x}\| = 5$ $\theta = 60^\circ$, כמה שווה $\vec{x} \cdot \vec{y}$?

<u>תשובות אפשרויות</u>:

א. 15 ב. 10 ג. 7.5 ד. 12.5

<u>תשובה</u>: $\vec{x} \cdot \vec{y} = \|\vec{x}\| \cdot \|\vec{y}\| \cdot \cos\theta = 5 \cdot 3 \cdot 0.5 = 7.5$

שאלה 4 - מכפלה סקלרית (פרספקטיבה אלגברית)

<u>וקטור (vector)</u> - נסמן $\vec{x},\vec{y} \in \mathbb{R}^n$

<u>מכפלה סקלרית (dot product) של וקטורים</u> – נסמן $\vec{x} \cdot \vec{y}$

<u>מבחינה אלגברית</u> - $\langle \vec{x},\vec{y} \rangle = \sum_{i=1}^{n} x_i \cdot y_i = x_1 \cdot y_1 + x_2 \cdot y_2 + \cdots + x_n \cdot y_n$

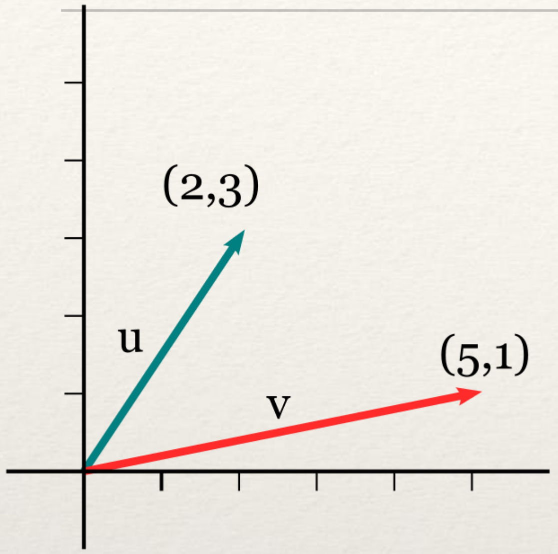
- בצורה מטריציונית - $\langle \vec{x},\vec{y} \rangle = \vec{x}^T \cdot \vec{y} = \sum_{i=1}^{n} x_i \cdot y_i$

<u>תרגיל</u> – נתון $\vec{x} = (4, -7,3)$, $\vec{w} = (3,0, -4)$, למה שווה $\vec{w}^T \cdot \vec{x}$ ומה ניתן להסיק גאומטרית?

<u>תשובה</u>: 3·4+0·-7+-4·3=0

- מסקנה: מבחינה גאומטרית, הוקטורים ניצבים, כיוון ש- $\cos 90 = 0$

שאלה 5 - חיבור וקטורי



<u>חיבור של וקטורים</u> – נסמן $\vec{x}+\vec{y}$

<u>מבחינה אלגברית</u> $=(x_1+y_1,x_2+y_2,\ldots,x_n+y_n)$

<u>תרגיל</u> – חשבו את הווקטור המתקבל, מחיבור הווקטורים $\vec{u}+\vec{v}$ (המופיעים בתמונה).

תשובה – $\vec{u}+\vec{v}=(2,3)+(5,1)=(7,4)$

נסמן $\vec{z}=\vec{u}+\vec{v}$, מתקיים $\vec{z}-\vec{u}=\vec{z}+-\vec{u}=\vec{v}$

משוואה של קו ישר (משוואה לינארית)

שאלה 6 (סקר) - משוואה לינארית

שאלות:

1. איזו משוואה מהמשוואות הבאות הינה משוואה לינארית?

א. x_1+5x_2 ב. $2x_1^3+x_2+2$

תשובות אפשריות:

1. א – לינארית, ב- לינארית
2. א – לא לינארית ב- לינארית
3. א – לינארית ב- לא לינארית
4. א – לא לינארית ב- לא לינארית

תשובה נכונה:

ג. א. לינארית. ב. לא לינארית (פולינומיאלית)

שאלה 7 (סקר) - משוואות קו ישר

איך יראו המשוואות $H_W(x) = w_0 + w_1 \cdot x$ עם הפרמטרים שמופיעים בתחתית?

השיפוע: a. מקביל לציר ה-x, b. מקביל לציר ה-y, c. יורד, d. עולה,

החיתוך: a. נחתך בראשית, b. ללא חיתוך בציר ה-x, c. ללא חיתוך בציר ה-y,

תשובות אפשריות:

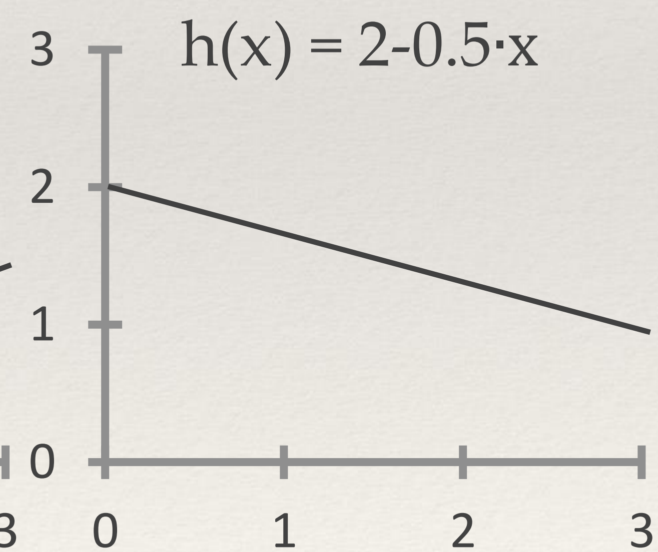
א. ישר i – עולה, ללא חיתוך בציר x; ישר ii – יורד ; ישר iii - מקביל לציר ה-y

ב. ישר i – יורד, ללא חיתוך בציר x; ישר ii – עולה , ללא חיתוך בציר ה-y; ישר iii - מקביל לציר ה-x,

ג. ישר i – יורד ; ישר ii – עולה; חיתוך בראשית ; ישר iii - מקביל לציר ה-x

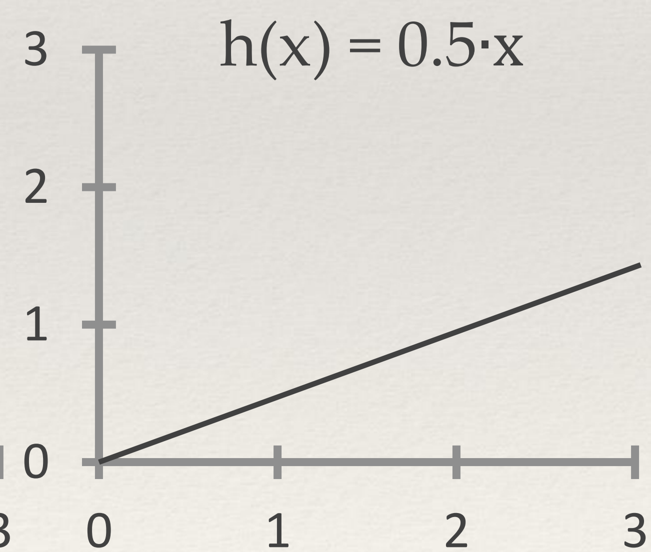
תשובה: ג

ישר i



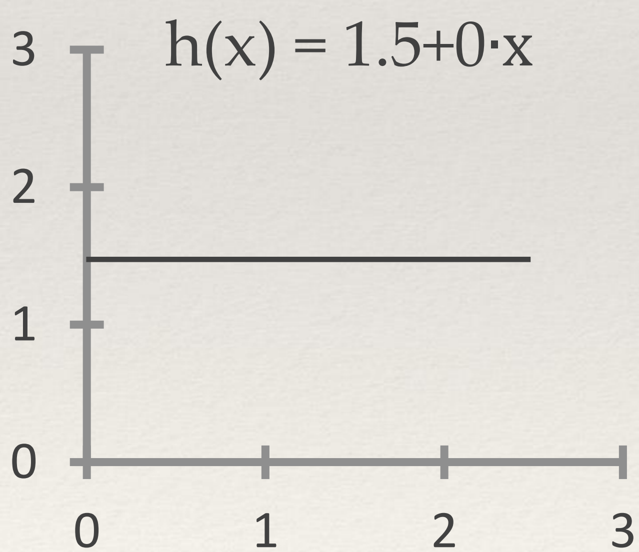
w_0=2
$w_1 = -0.5$

ישר ii



w_0=0
$w_1 = 0.5$

ישר iii



w_0=1.5
$w_1 = 0$

מוטיבציה – גובה המשכורת

<u>שאלה</u>: נתון משכורות, כיצד נקבל החלטה אם העובד/ת ירוויח משכורת גבוהה או נמוכה?

<u>תשובה</u>: לפי המיקום ביחס לקו הישר.

- מעל הקו – גבוהה
- מתחת לקו - נמוכה

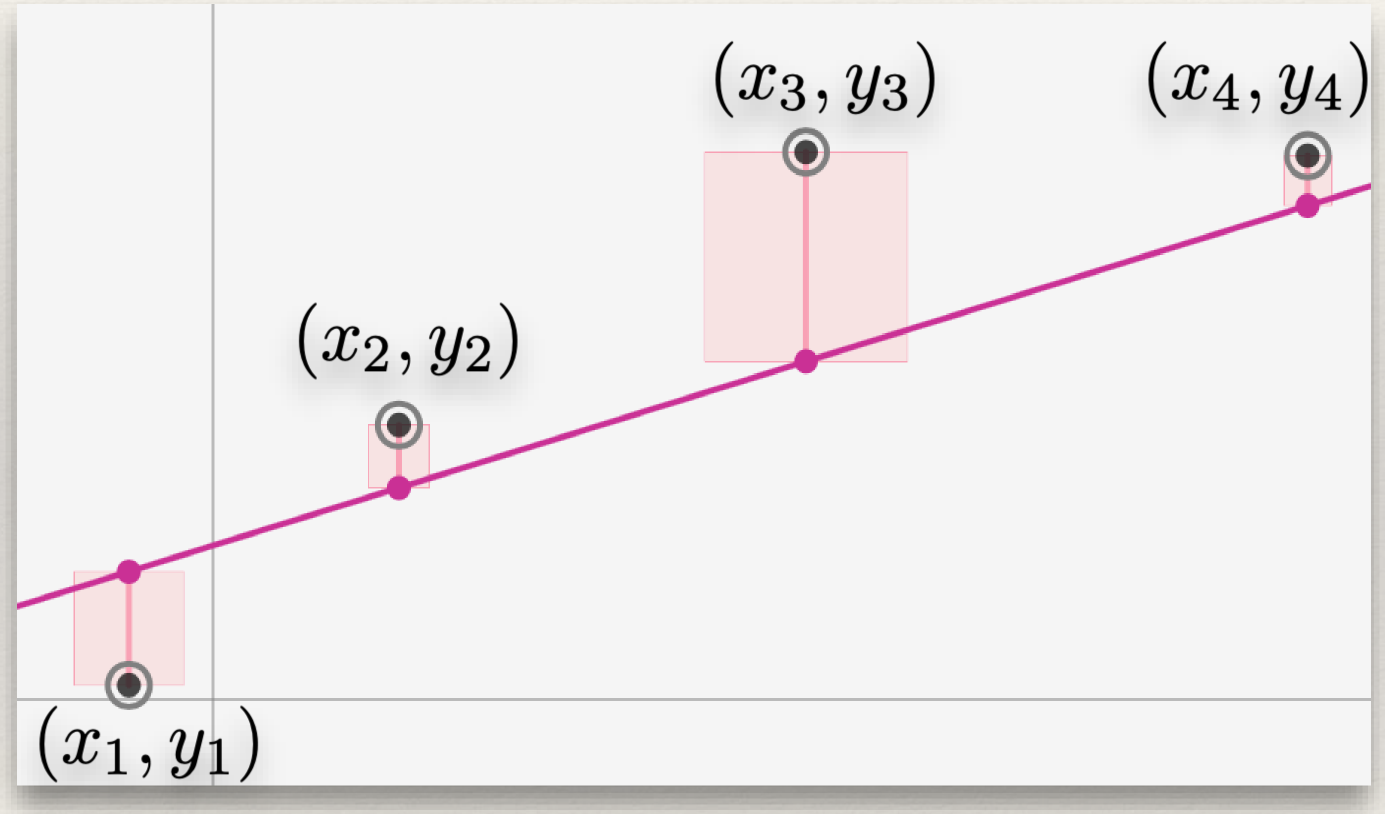
בעית רגרסיה

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

רגרסיה לינארית

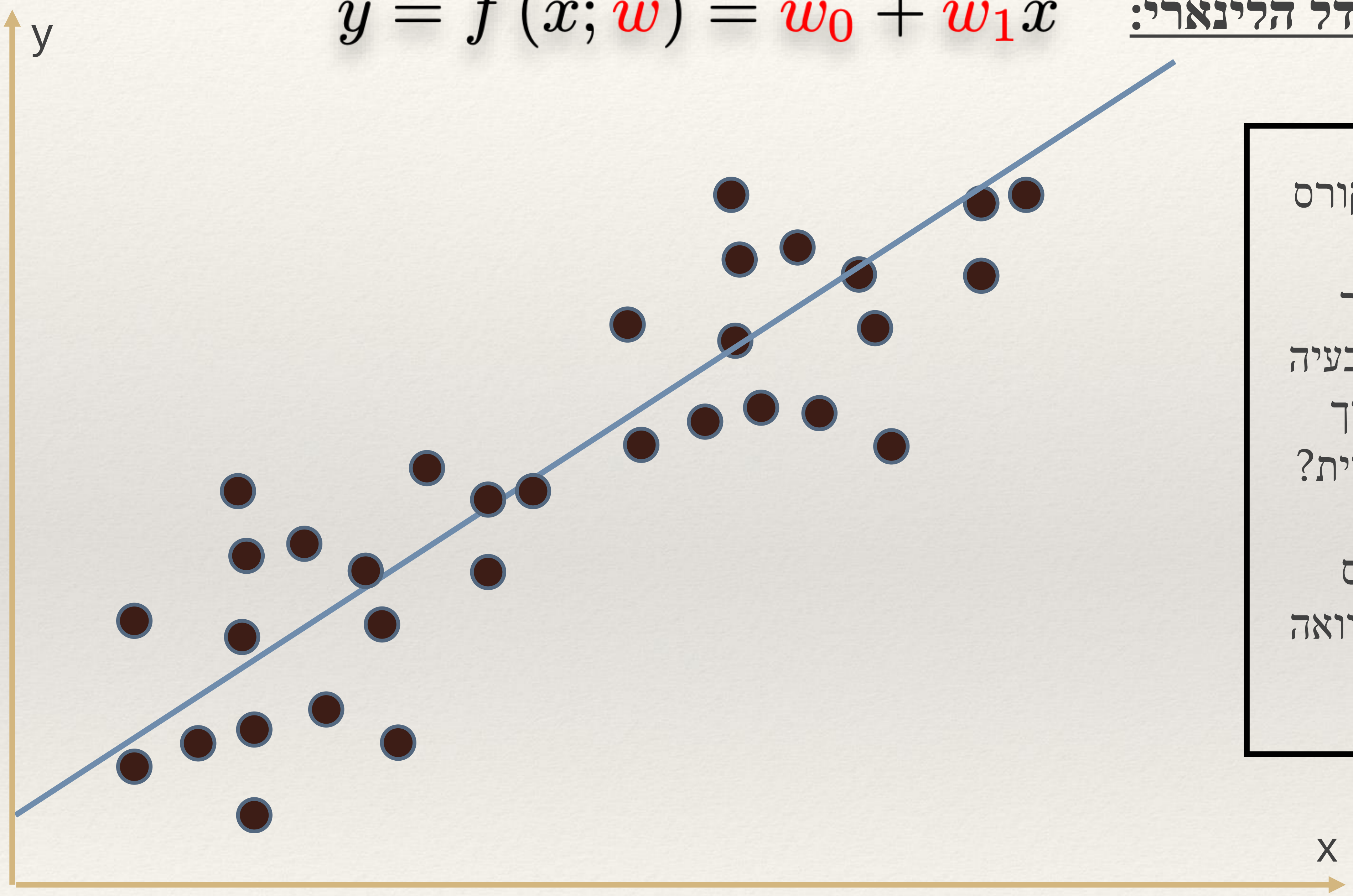
ברגרסיה לינארית נחזה את הקשר בין המאפיינים לבין הערך אותו נרצה לחזות, כקשר לינארי.



מוטיבציה – שכר

המודל הלינארי:

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



נניח ש-x הוא הציון בקורס למידת מכונה
ו-y הוא גובה השכר בעבודה, כיצד תראה הבעיה –כבעיית רגרסיה, תוך שימוש ברגרסיה לינארית?

תשובה: נמדל את היחס בין הציון לשכר, כמשוואה של קו ישר

שיערוך מודל רגרסיה
(regression model evaluation)

שיערוך מודל רגרסיה

SAE: $SAE = \sum_{I=1}^{n} |y_i - \hat{y}_i|$

MAE: $MAE = \frac{1}{n}\sum_{I=1}^{n} |y_i - \hat{y}_i|$

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

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

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

$\bar{y}$: ממוצע הערכים המונחים, כלומר $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

SST (Sum of Squared Total): = $SST = \sum_{I=1}^{n} (y_i - \bar{y})^2$

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

שאלה 8 (סקר)

$SAE = \sum_{I=1}^{n}|y_i - \hat{y}_i|$:**<u>SAE</u>**

$MAE = \frac{1}{n}\sum_{I=1}^{n}|y_i - \hat{y}_i|$:**<u>MAE</u>**

$SSE = \sum_{I=1}^{n}(y_i - \hat{y}_i)^2$:**<u>SSE</u>**

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

$RMSE = \sqrt{\frac{1}{n}\sum_{I=1}^{n}(y_i - \hat{y}_i)^2}$:**<u>RMSE</u>**

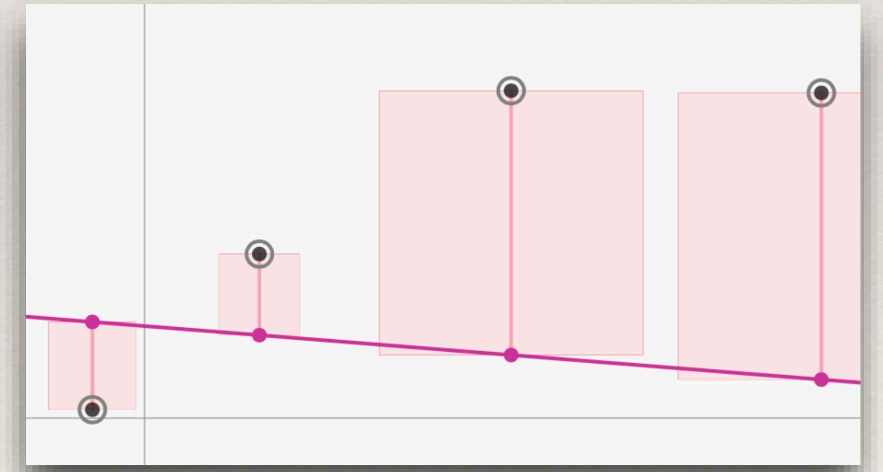
$\bar{y}$: ממוצע הערכים המונחים, כלומר $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

SST (Sum of Squared Total): = $SST = \sum_{I=1}^{n}(y_i - \bar{y})^2$

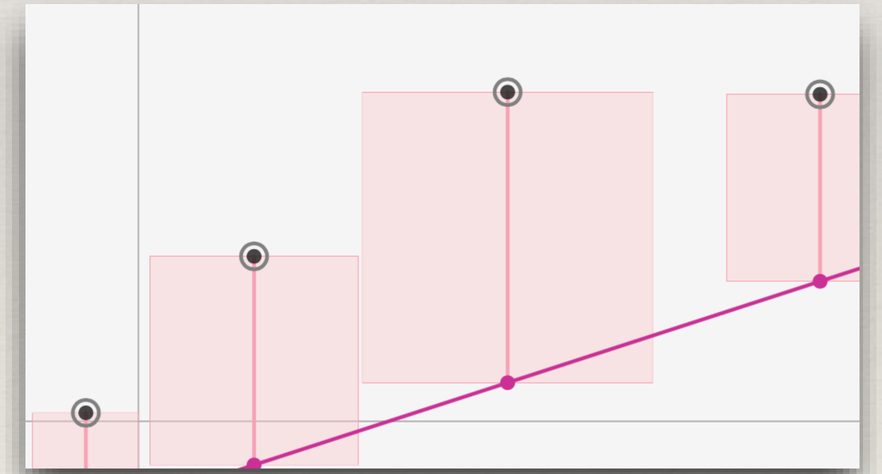
:R-SQARE $\boldsymbol{R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{MSE}{\sigma^2}}$

לאיזו מהפונקציות הלינאריות הבאות הטעות המינימלית?

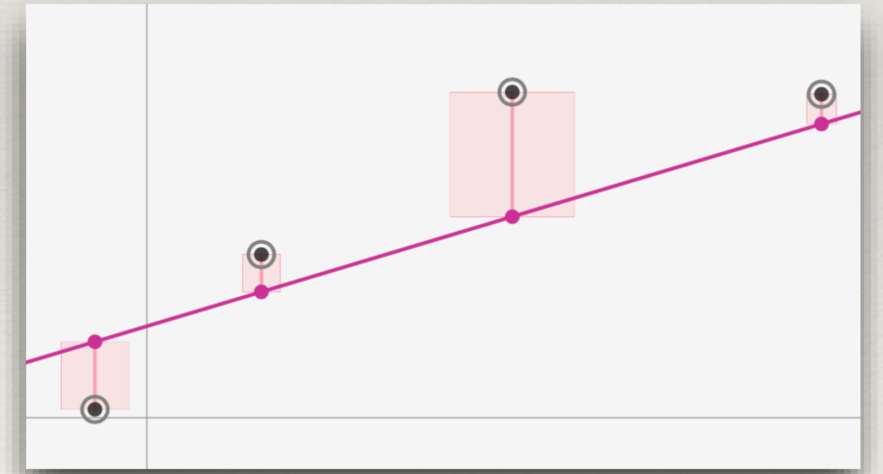
<u>א.</u>



<u>ב.</u>



<u>ג.</u>



[Demo]()

שאלה 9א – שיערוך מודל רגרסיה

<u>תרגיל</u> –
נתונים 2 משתנים – גובה ומשקל. רוצים לשערך את המשקל כתלות בגובה, ומצאו קו רגרסיה ליניארי:
$\hat{y}=9.2+0.8x_1$

<u>חשבון</u> את ה- SAE וה- MAE:

Height (X)	Weight (Y)	Predicted (Ŷ)	Error (Y-Ŷ)	Absolute-Error |(Y-Ŷ)|
43	41	43.6	-2.6	2.6
44	45	44.4	0.6	0.6
45	49	45.2	3.8	3.8
46	47	46	1	1
47	44	46.8	-2.8	2.8
Regression line = y=9.2+0.8x				

<u>**SAE**</u>: $SAE = \sum_{I=1}^{n} |y_i - \hat{y}_i|$

<u>**MAE**</u>: $MAE = \frac{1}{n} \sum_{I=1}^{n} |y_i - \hat{y}_i|$

<u>פתרון</u>
- קודם נשערך את Y
- כעת נחשב את הטעות המוחלטת
- נסכום ונקבל <u>SAE</u> = 10.8
- נחלק בכמות הדו' ונקבל

<u>MAE</u> $= \frac{1}{5} \cdot 10.8 = 2.16$

שאלה 9ב – שיערוך מודל רגרסיה

<u>תרגיל</u> –
נתונים 2 משתנים – גובה ומשקל.
רוצים לשערך את המשקל כתלות בגובה, ומצאו קו רגרסיה (הנ"ל)

<u>חשבו</u> את ה- SSE, ה-MSE וה- RMSE:

Height (X)	Weight (Y)	Predicted (Ŷ)	Error (Y-Ŷ)	squared error $(Y-\hat{Y})^2$
43	41	43.6	-2.6	6.76
44	45	44.4	0.6	0.36
45	49	45.2	3.8	14.44
46	47	46	1	1
47	44	46.8	-2.8	7.84
Regression line = y=9.2+0.8x				

<u>**SSE**</u>: $SSE = \sum_{I=1}^{n}(y_i - \hat{y}_i)^2$

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

<u>**RMSE**</u>: $RMSE = \sqrt{\frac{1}{n}\sum_{I=1}^{n}(y_i - \hat{y}_i)^2}$

<u>פתרון</u>:

- נשתמש בשערוך הקודם של Y (Ŷ) ובחישוב הטעות (Y-Ŷ)
- כעת נחשב את הטעות הריבועית
- נסכום ונקבל <u>SSE</u>=30.4
- נחלק בכמות הדו' ונקבל

<u>MSE</u>=$\frac{1}{5} \cdot 30.4 = 6.08$

<u>RMSE</u> ≈ 2.46

שאלה 9ג – שיערוך מודל רגרסיה

תרגיל –
נתונים 2 משתנים – גובה ומשקל.
רוצים לשערך את המשקל כתלות בגובה, ומצאו קו רגרסיה (הנ"ל)

חשבון את ה- R-Squared:

Height (X)	Weight (Y)	Y−Ȳ	squared dist from avg $(Y-\bar{Y})^2$
43	41	-4.2	17.64
44	45	-0.2	0.04
45	49	3.8	14.44
46	47	1.8	3.24
47	44	-1.2	1.44
Regression line = y=9.2+0.8x			

פתרון:
- נשתמש בחישוב הקודם של SSE:
קיבלנו SSE=30.4
- נחשב את הממוצע $\bar{y}$
$\bar{y}$=45.2
- ונחשב את המרחקים הריבועיים מהממוצע $(Y-\bar{Y})^2$
- נחשב את SST (סכום המרחקים הריבועיים הנ"ל): SST=36.8
- נחשב את R-Squared

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{30.4}{36.8} \approx 0.174$$

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

$\bar{y}$: ממוצע הערכים המונחים, כלומר $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

SST (Sum of Squared Total): $SST = \sum_{I=1}^{n}(y_i - \bar{y})^2$

R-Squared: $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{MSE}{\sigma^2}$

רגרסיה לינארית (linear regression)

רגרסיה לינארית

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

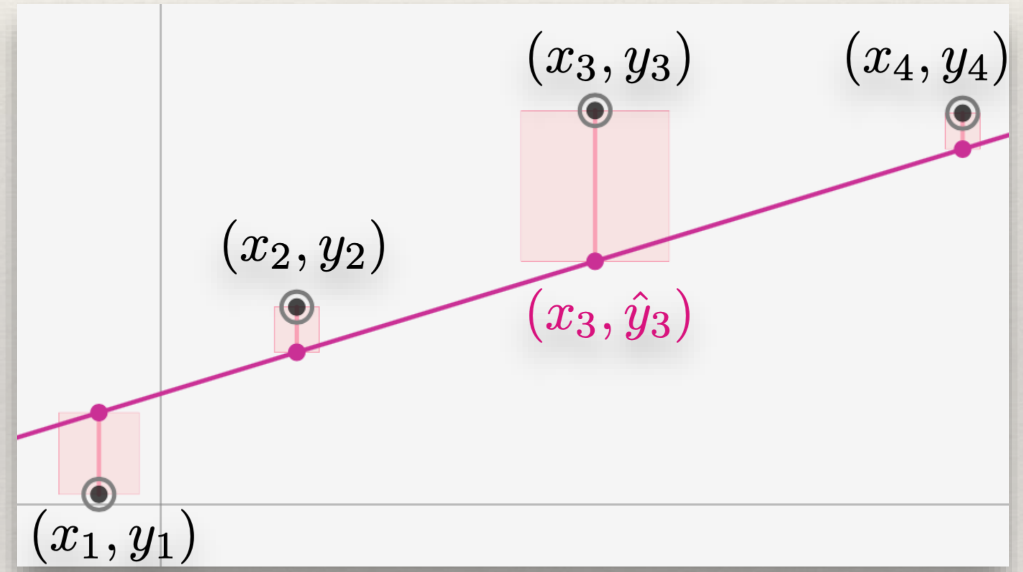
מקרה פשוט (כמו בדוגמה) : יש רק מאפיין אחד בווקטור המאפיינים

שכר יומי בש"ח	סכום הציונים
460	2104
232	1416
315	1534
178	852
…	…

רגרסיה לינארית

המודל הלינארי:

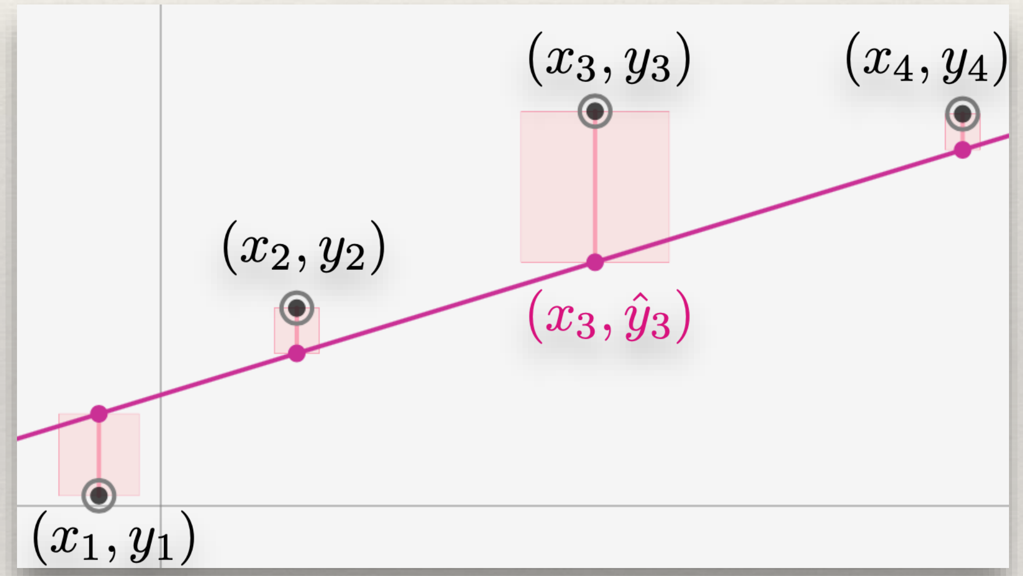
$$\hat{y} = f\left(x; \vec{w}\right) = w_0 + w_1 x$$



פונקצית מחיר (Cost Function)

המודל הליניארי:

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

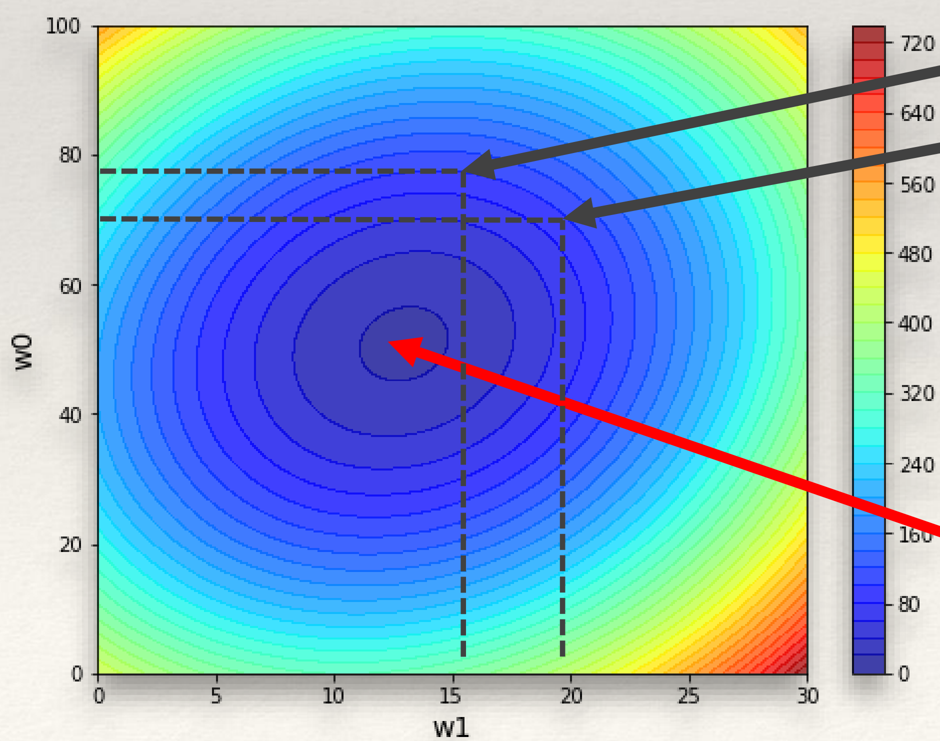


מטרה: למצוא היפותזה עם טעות מינימלית ע"י שימוש בפו' J

פונקצית המחיר – מוגדרת ע"י ממוצע הטעות הריבועית

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

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



ערכי w0, w1 שונים, בעלי אותה הטעות

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

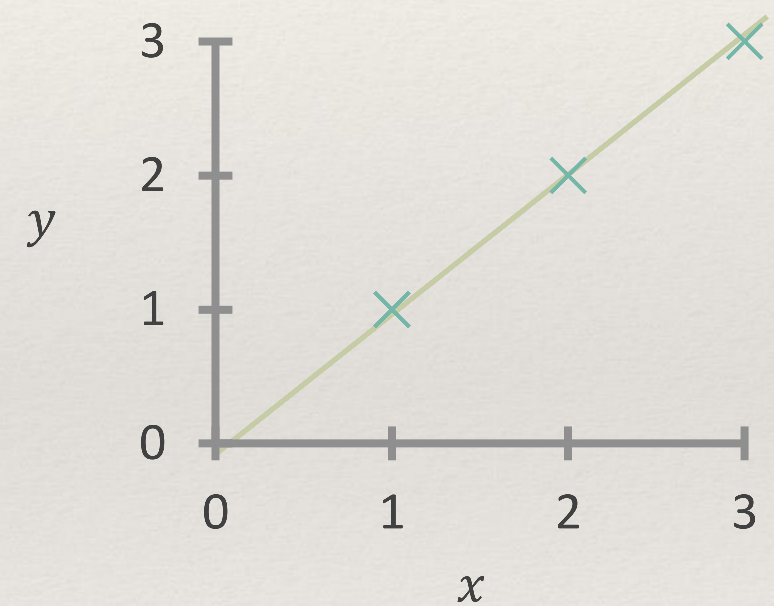
שאלה 11א - פונקציית מחיר

<u>המודל הלינארי:</u>

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

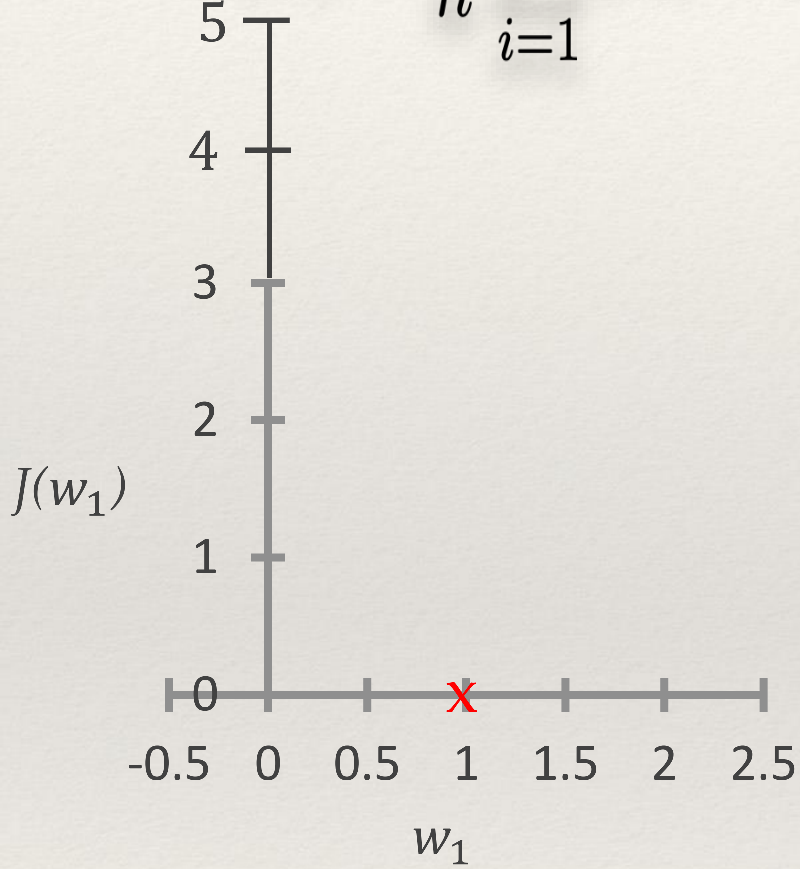
נניח משוואת ישר שיוצאת מראשית הצירים (ישנו רק הפרמטר w_1).

- חשבו מה תהיה הטעות עבור המשוואה - $\hat{y} = x$



<u>פונקציית המחיר:</u>

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



<u>תשובה:</u> הטעות = $\frac{1}{3} \cdot \left((1-1)^2 + (2-2)^2 + (3-3)^2\right) = 0$

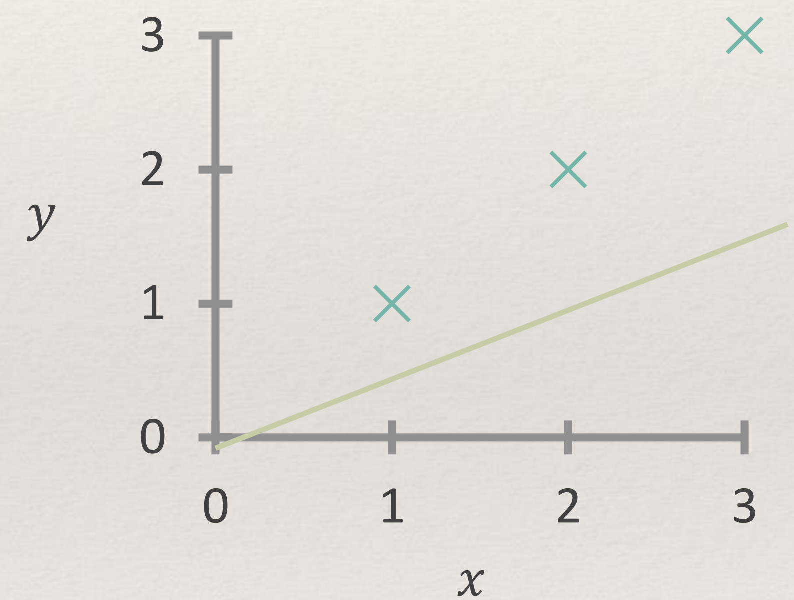
שאלה 11ב - פונקצית מחיר

המודל הלינארי:

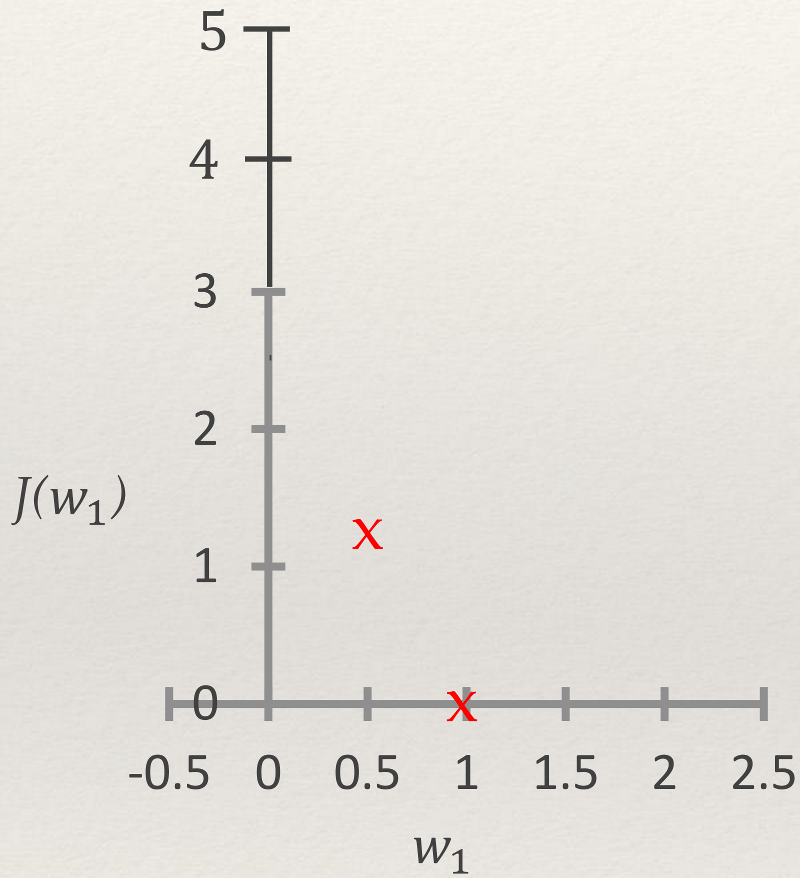
$$\hat{y} = f\left(x; \vec{w}\right) = w_0 + w_1 x$$

נניח משוואת ישר שיוצאת מראשית הצירים (ישנו רק הפרמטר w_1).

- חשבו מה תהיה הטעות עבור המשוואה - $\hat{y} = 0.5x$



פונקצית המחיר:



תשובה: הטעות = $\frac{1}{3}\cdot\left((0.5-1)^2+(1-2)^2+(1.5-3)^2\right) \approx 1.166$

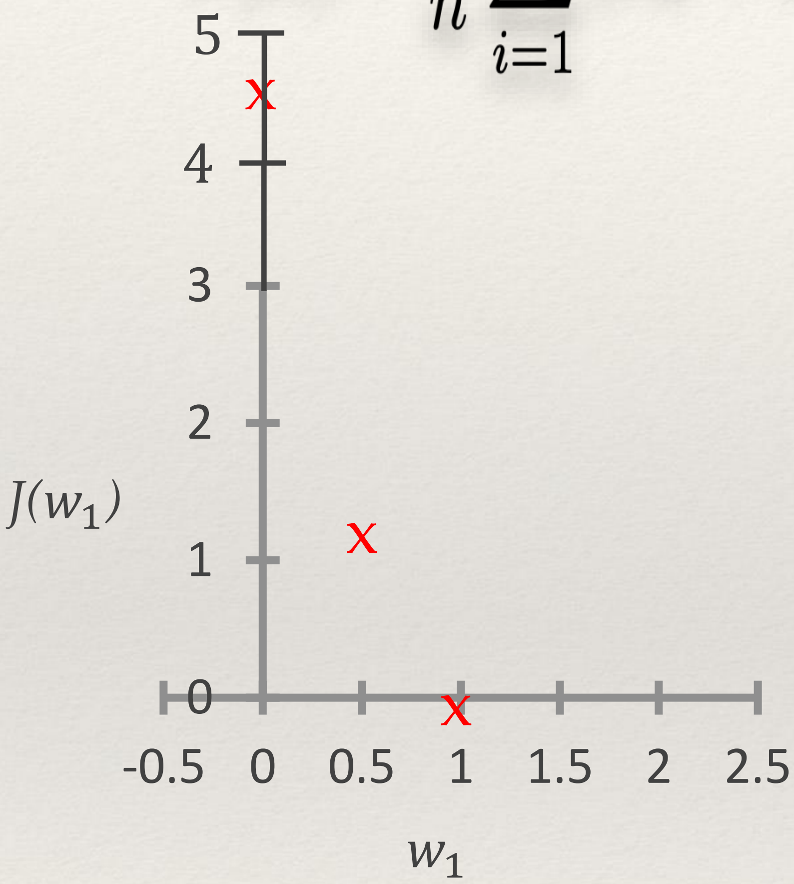
שאלה 11ג - פונקציית מחיר

המודל הלינארי:

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

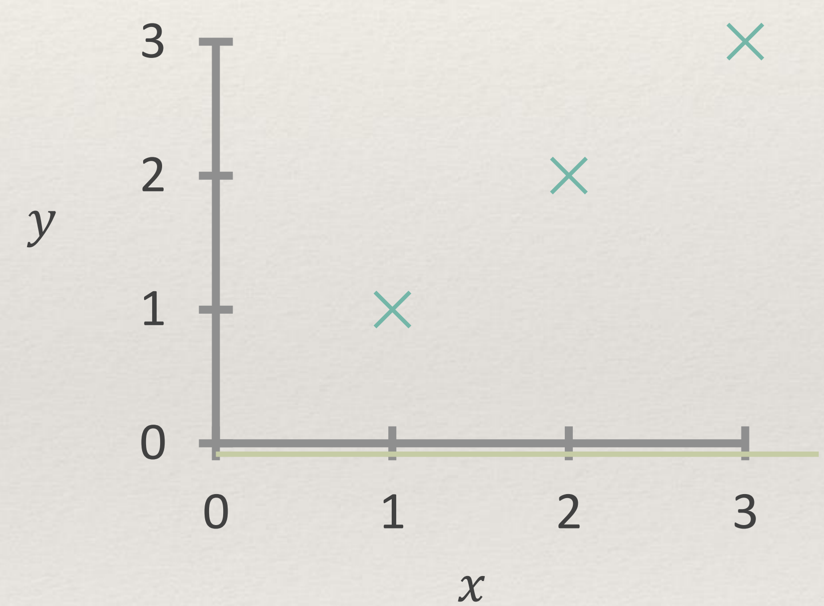
פונקציית המחיר:

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



נניח משוואת ישר שיוצאת מראשית הצירים (ישנו רק הפרמטר w_1).

- חשבו מה תהיה הטעות עבור המשוואה - $\hat{y} = 0x$

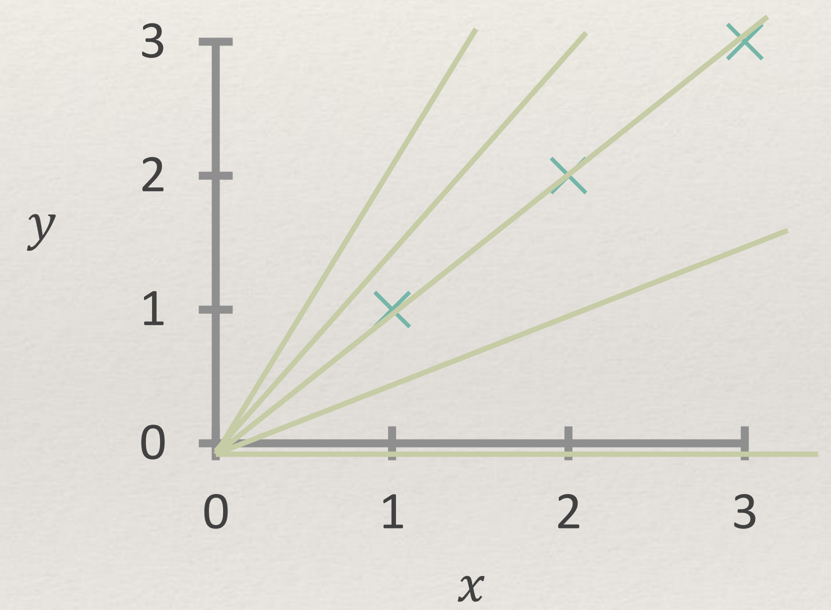


תשובה: הטעות = $\frac{1}{3} \cdot \left((0-1)^2 + (0-2)^2 + (0-3)^2\right) \approx 4.66$

פונקציית מחיר

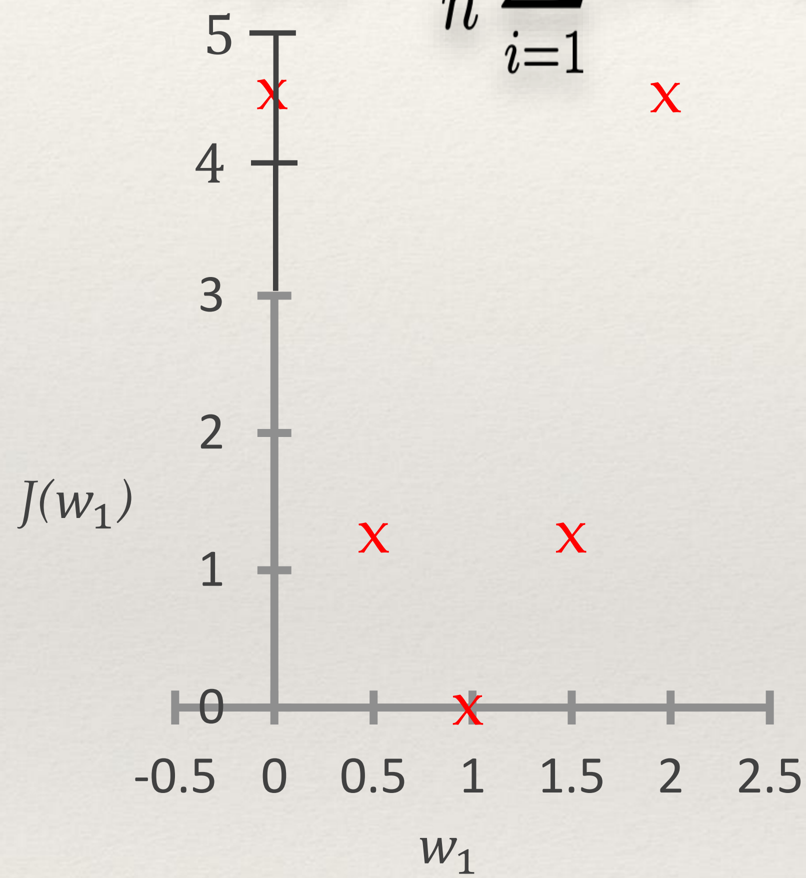
המודל הלינארי:

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



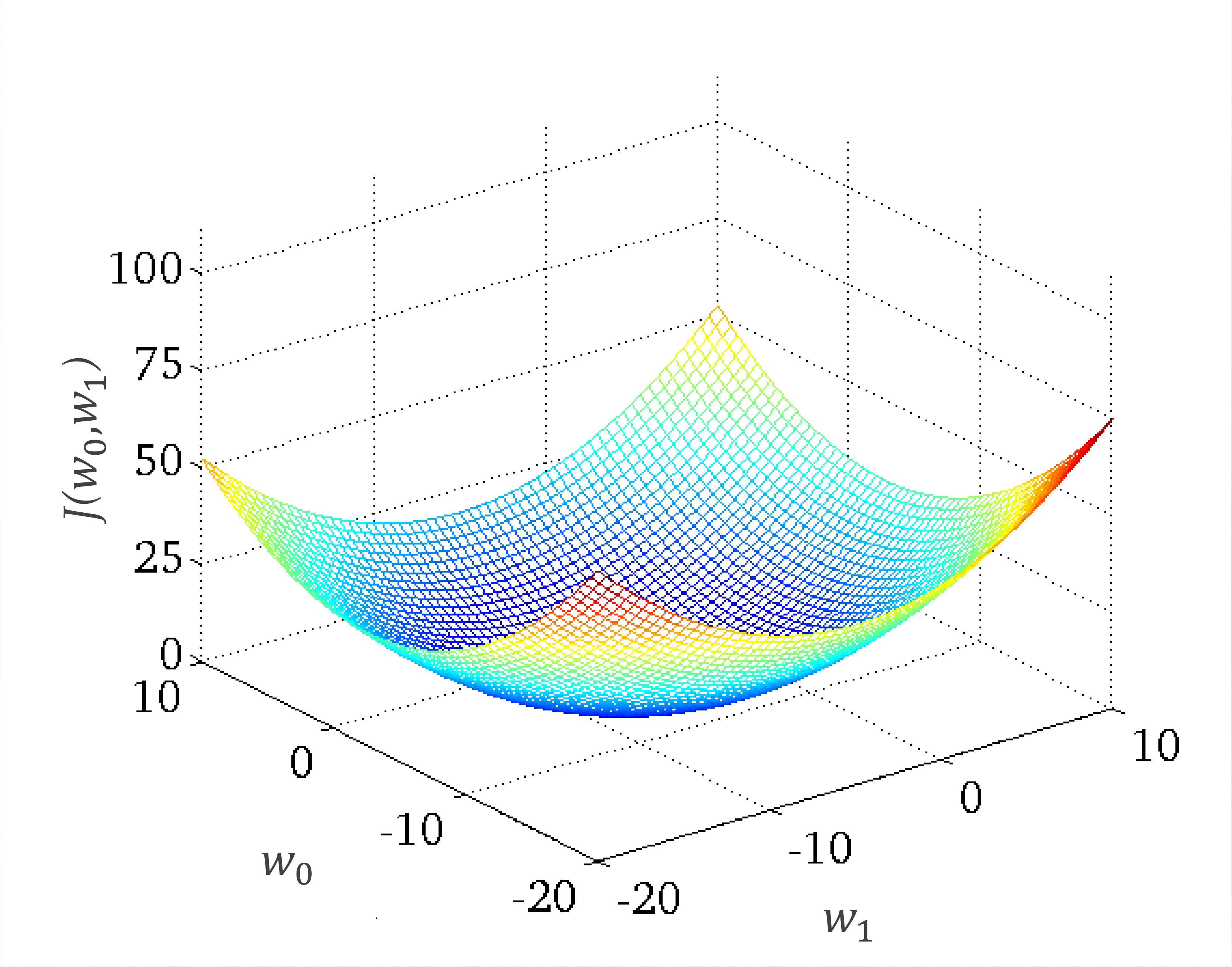
פונקציית המחיר:

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



כמו איזו פונקציה נראית פונקציית המחיר?

פונקצית מחיר עבור 2 פרמטרים w_0,w_1



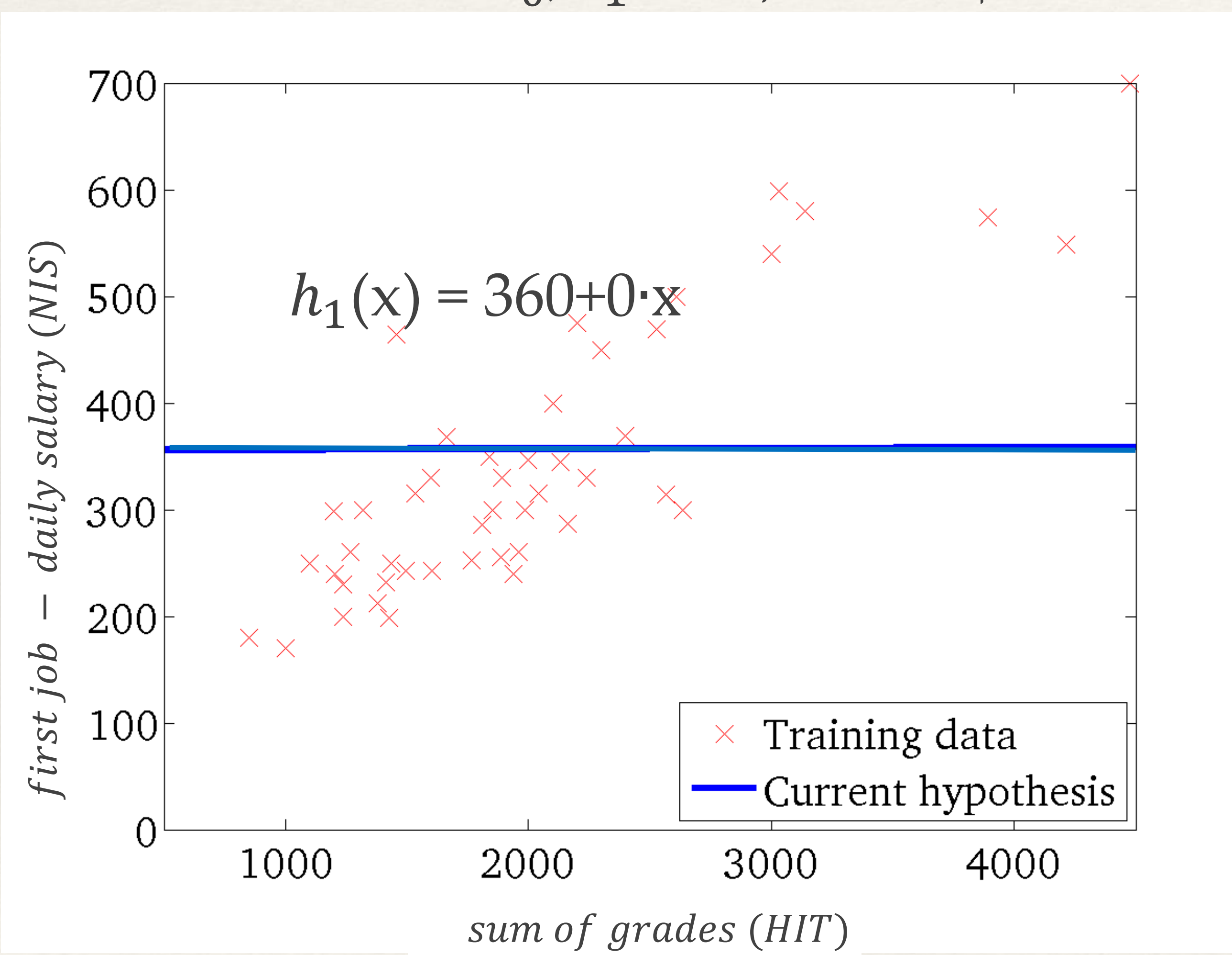
שאלה 12א - פונקצית מחיר עבור 2 פרמטרים w_0,w_1

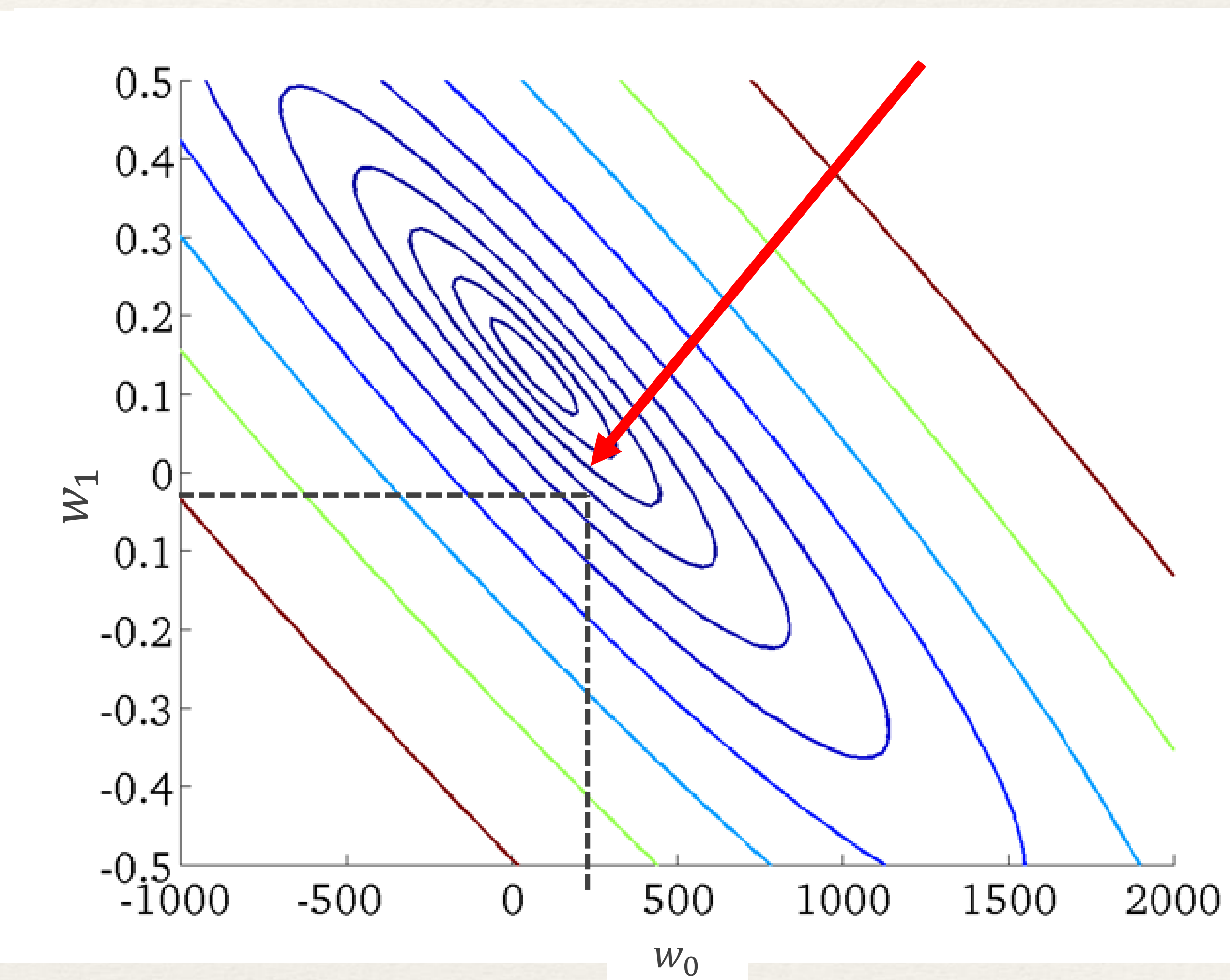
$J(w_0,w_1)$

- פונקצית המחיר (מדמה גרף תלת מימדי)
- פונקצית המחיר - פונקציה של w_0,w_1

$H_W(x)$

- ישר הרגרסיה, מאפיין אחד
- פונקציה של x, עבור w_0,w_1 מסוימים





שאלה: איפה בגרף נמצאת הטעות של ההיפותזה $h_1(x) = 360+0\cdot x$?

תשובה: נסמן את נקודת המפגש של $w_0 = 360, w_1 = 0$

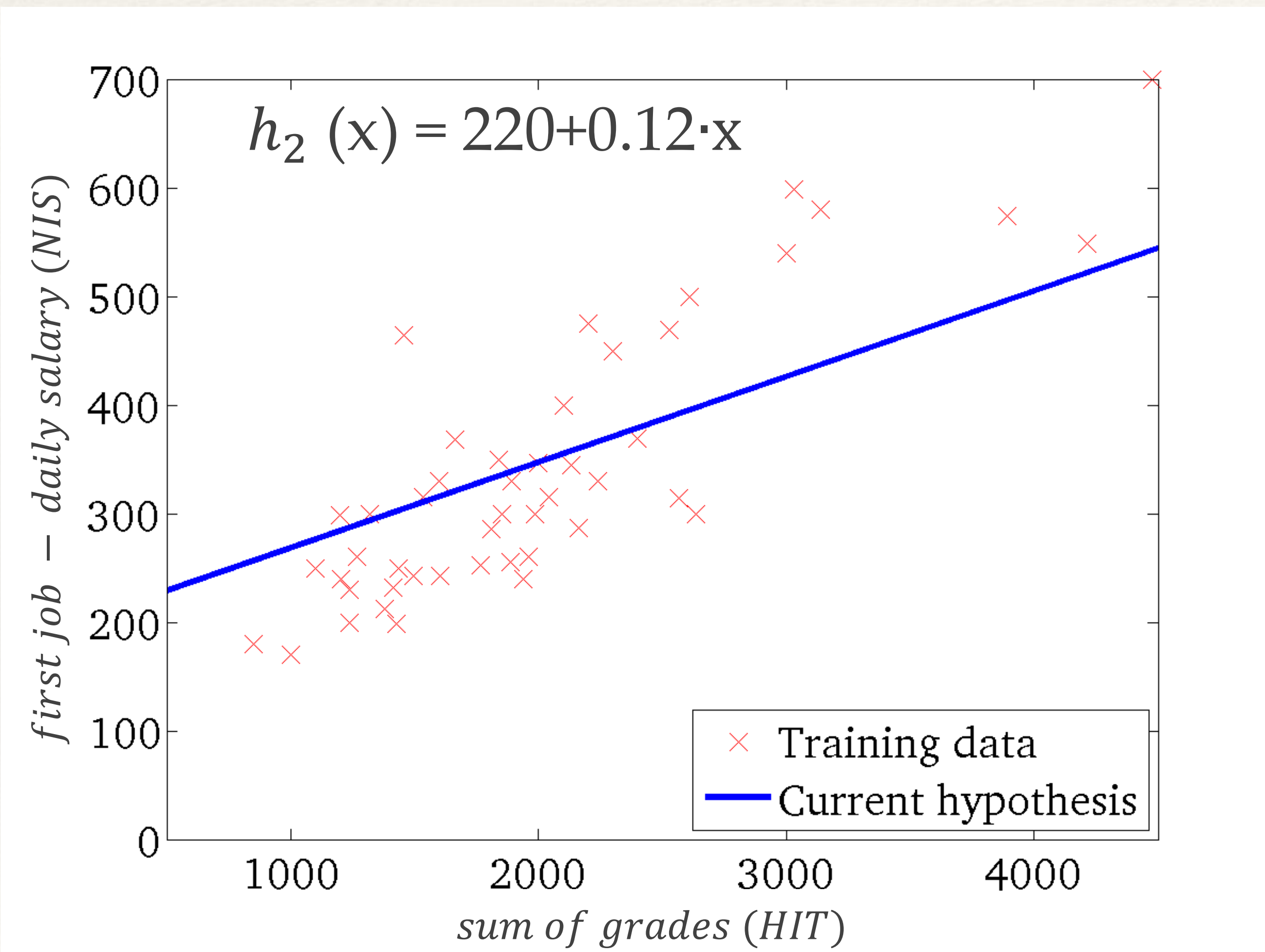
שאלה 12ב - פונקצית מחיר עבור 2 פרמטרים w_0,w_1

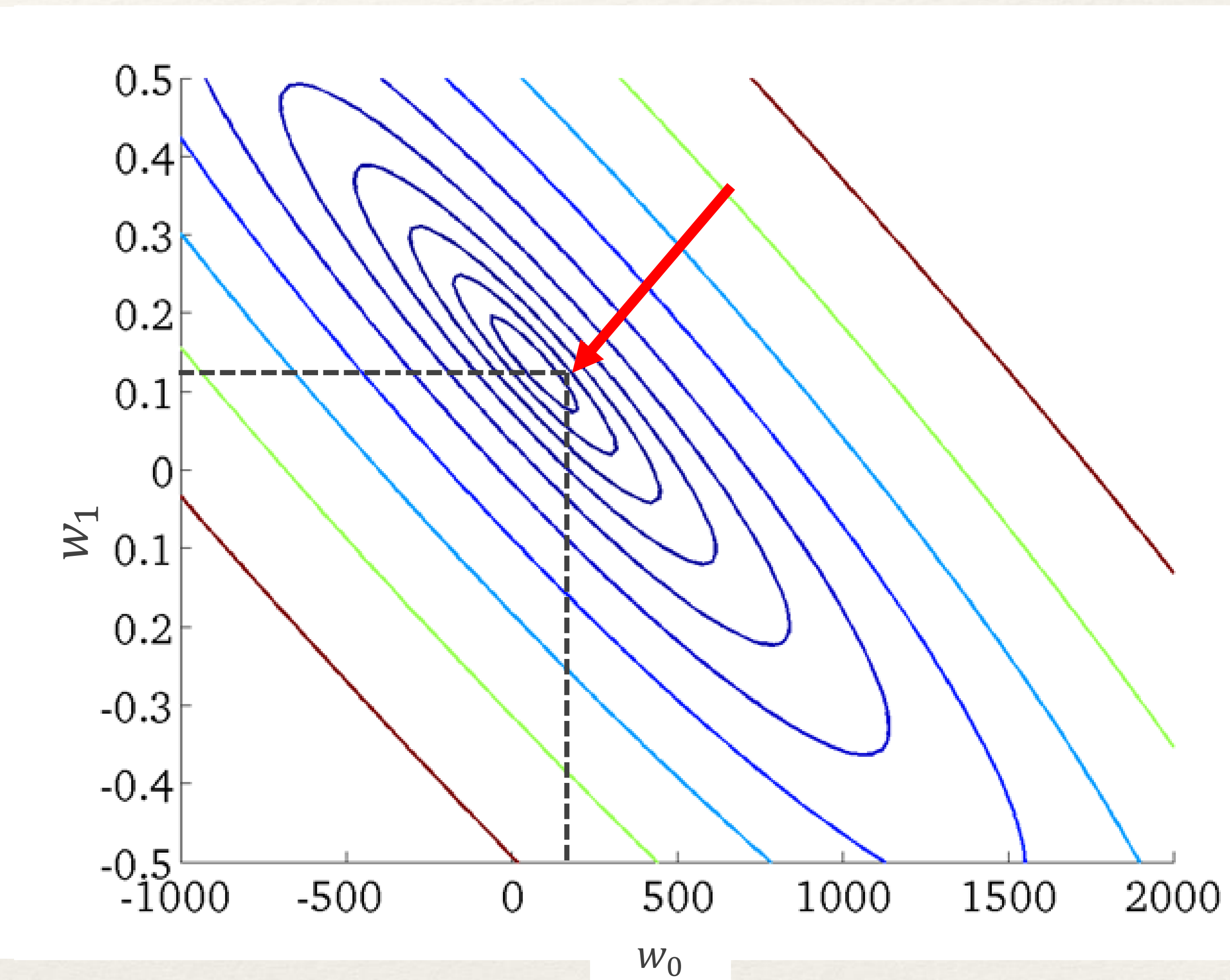
$J(w_0,w_1)$

- פונקצית המחיר (מדמה גרף תלת מימדי)
- פונקצית המחיר - פונקציה של w_0,w_1

$H_W(x)$

- ישר הרגרסיה, מאפיין אחד
- פונקציה של x, עבור w_0,w_1 מסוימים





שאלה: איפה בגרף נמצאת הטעות של ההיפותזה h_2 (x) = 220+0.12·x?

תשובה: נסמן את נקודת המפגש של $w_0 = 220, w_1 = 0.12$

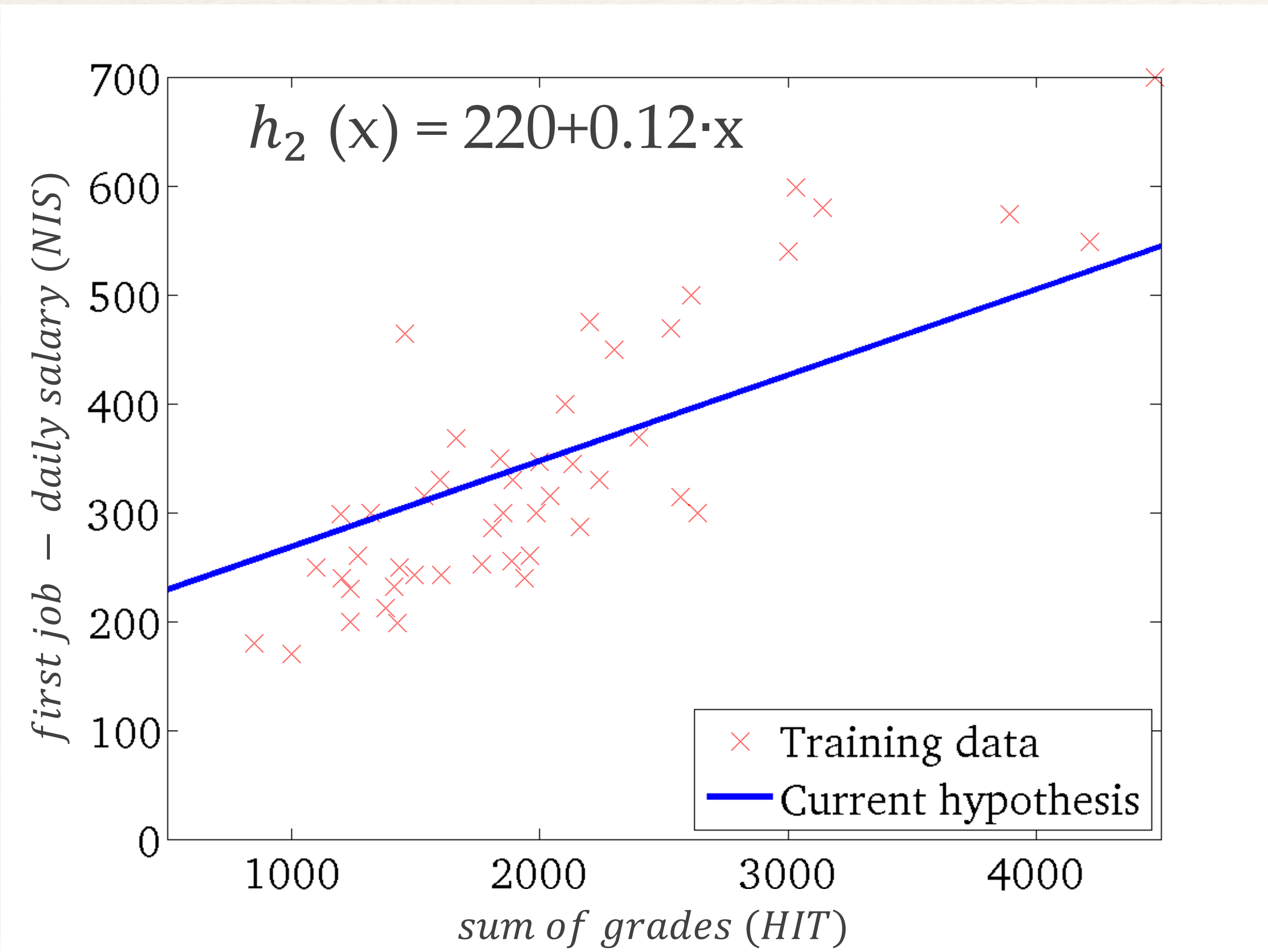
שאלה 12ג - פונקצית מחיר עבור 2 פרמטרים w_0,w_1

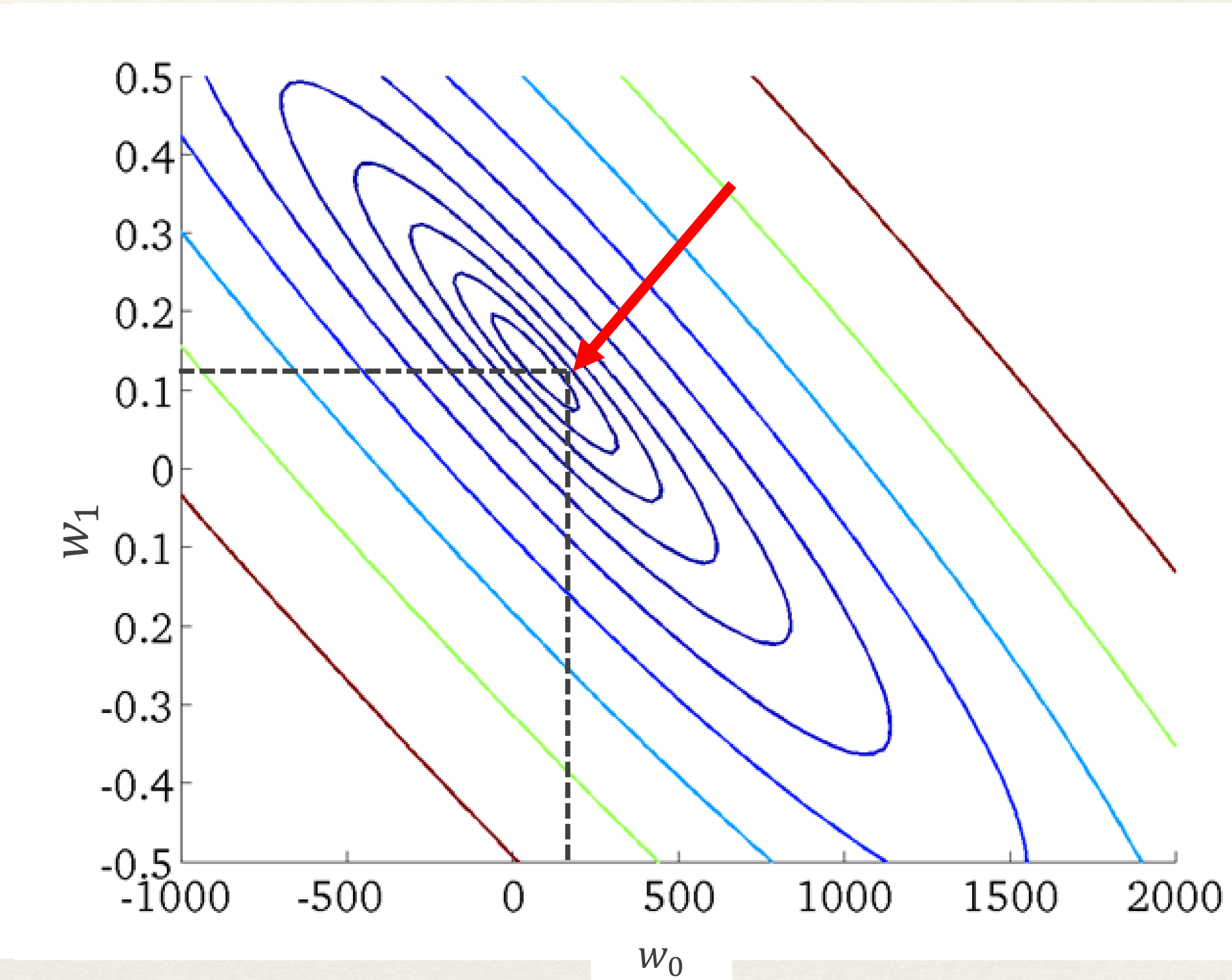
$J(w_0,w_1)$

- פונקצית המחיר (מדמה גרף תלת מימדי)
- פונקצית המחיר - פונקציה של w_0,w_1

$H_W(x)$

- ישר הרגרסיה, מאפיין אחד
- פונקציה של x, עבור w_0,w_1 מסוימים



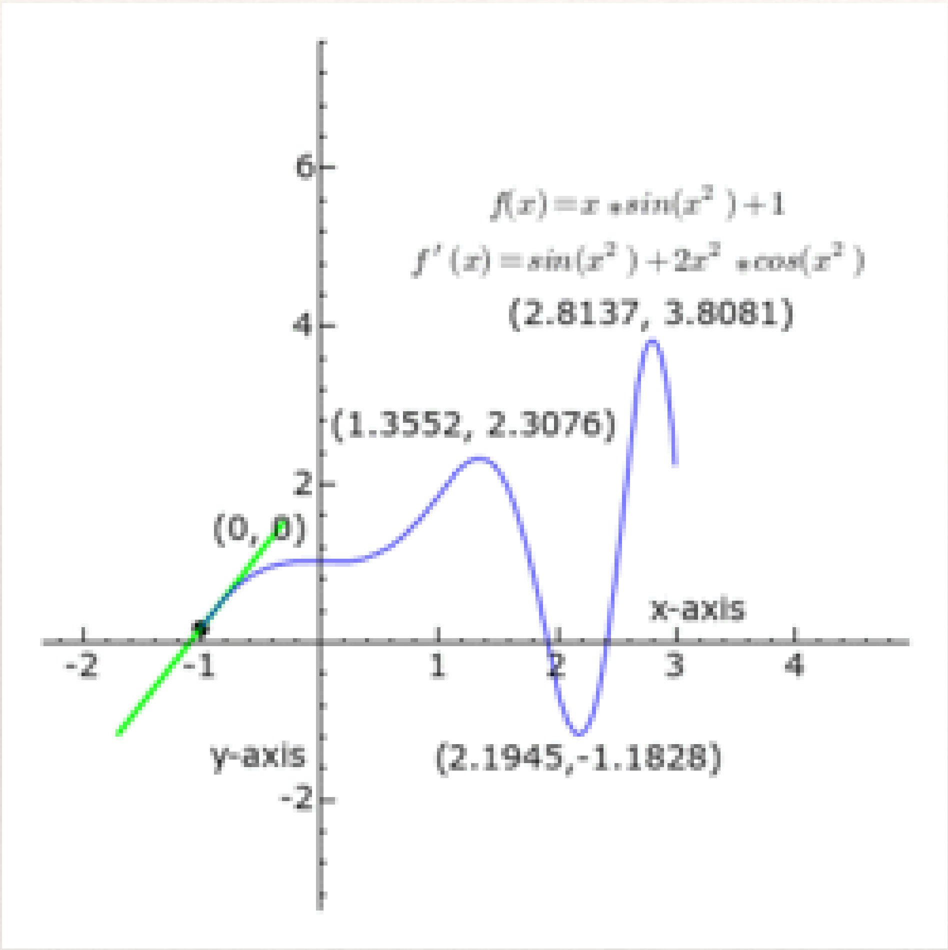


שאלה: לאיזו היפותזה טעות קטנה יותר, h_1 (x) או h_2 (x)? איך ניתן לראות זאת בגרף?

תשובה: h_2 (x), מבחינה גרפית, היא קרובה יותר למעגל הפנימי, שם נמצא המינימום.

נגזרת ונקודות קיצון

מושגים – נגזרת של פונקציה



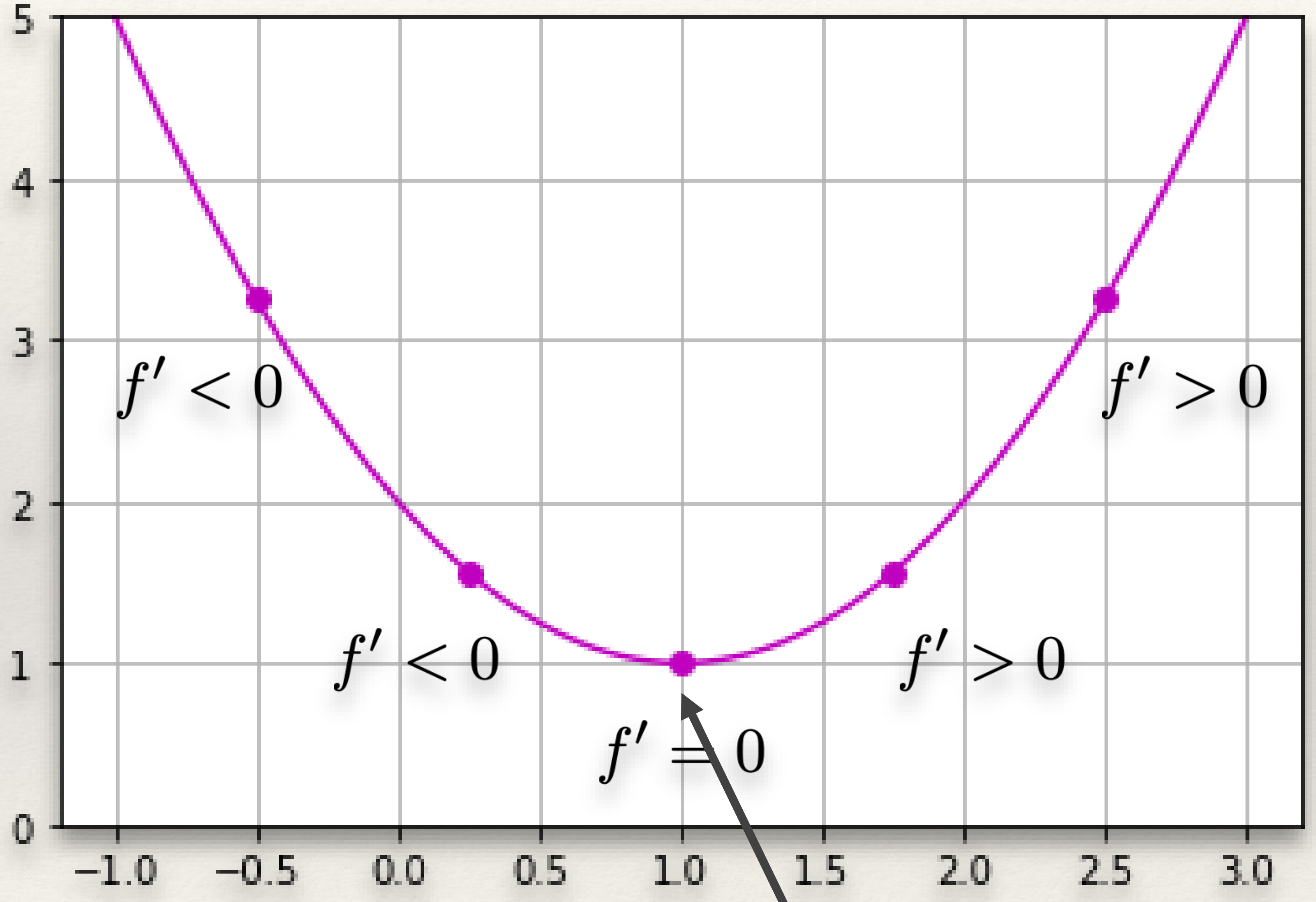
- <u>נגזרת</u> - הנגזרת של פונקציה ממשית מתארת את קצב ההשתנות של הפונקציה

  - <u>מבחינה גאומטרית</u> - הנגזרת של פונקציה בנקודה שווה לשיפוע המשיק באותה נקודה, כלומר, לכיוון של העקומה שהפונקציה מתארת.

מה משמעות ערכי נגזרת $f' > 0$

שאלה: מה המשמעות של ?f ‘(x)>0

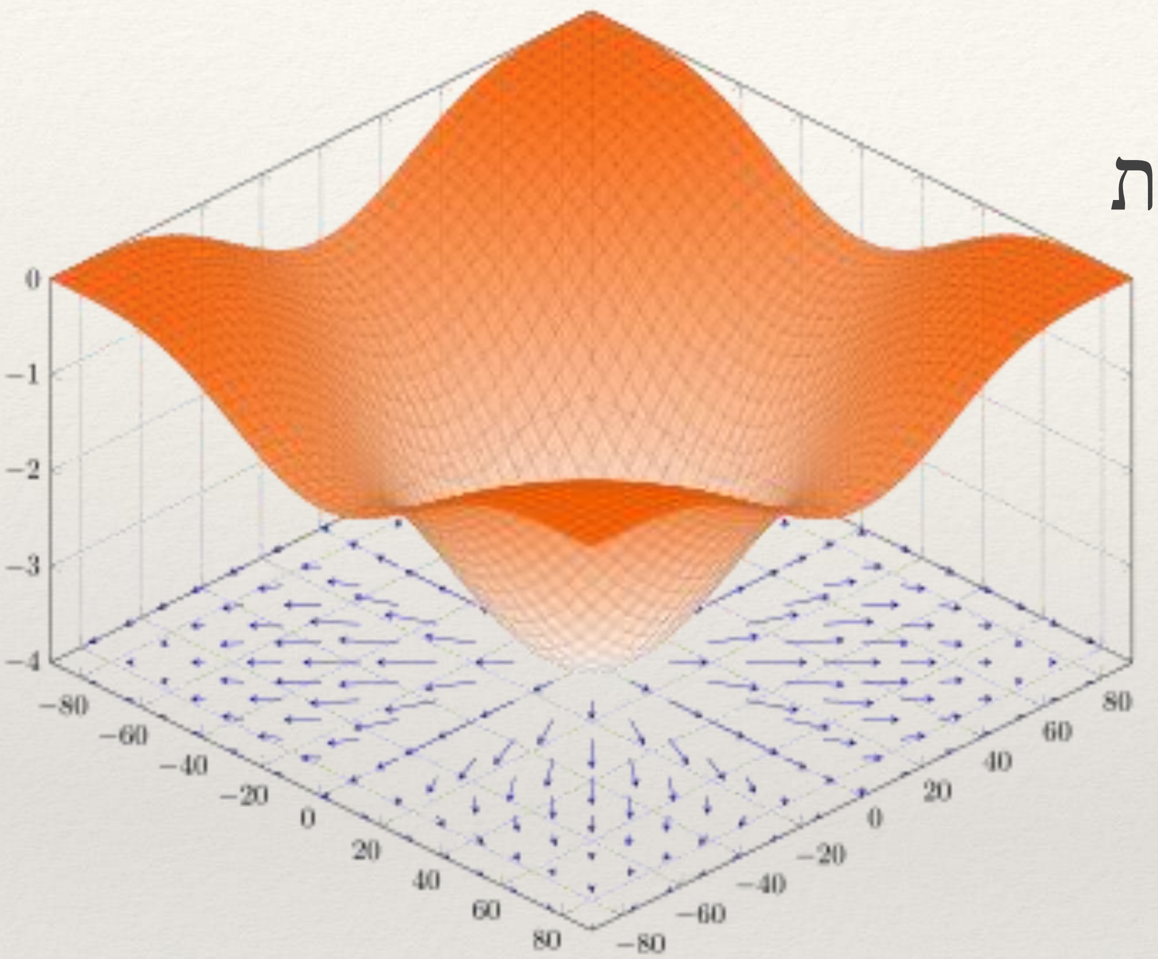
תשובה: המשמעות שכיוון הפונקציה בעליה מנקודות שמתקרבות ל-x מלמטה עד ל-x וכן מ-x לנקודות קרובות לאחר x



מינימום מקומי,
מתקיים f’’(x)>0

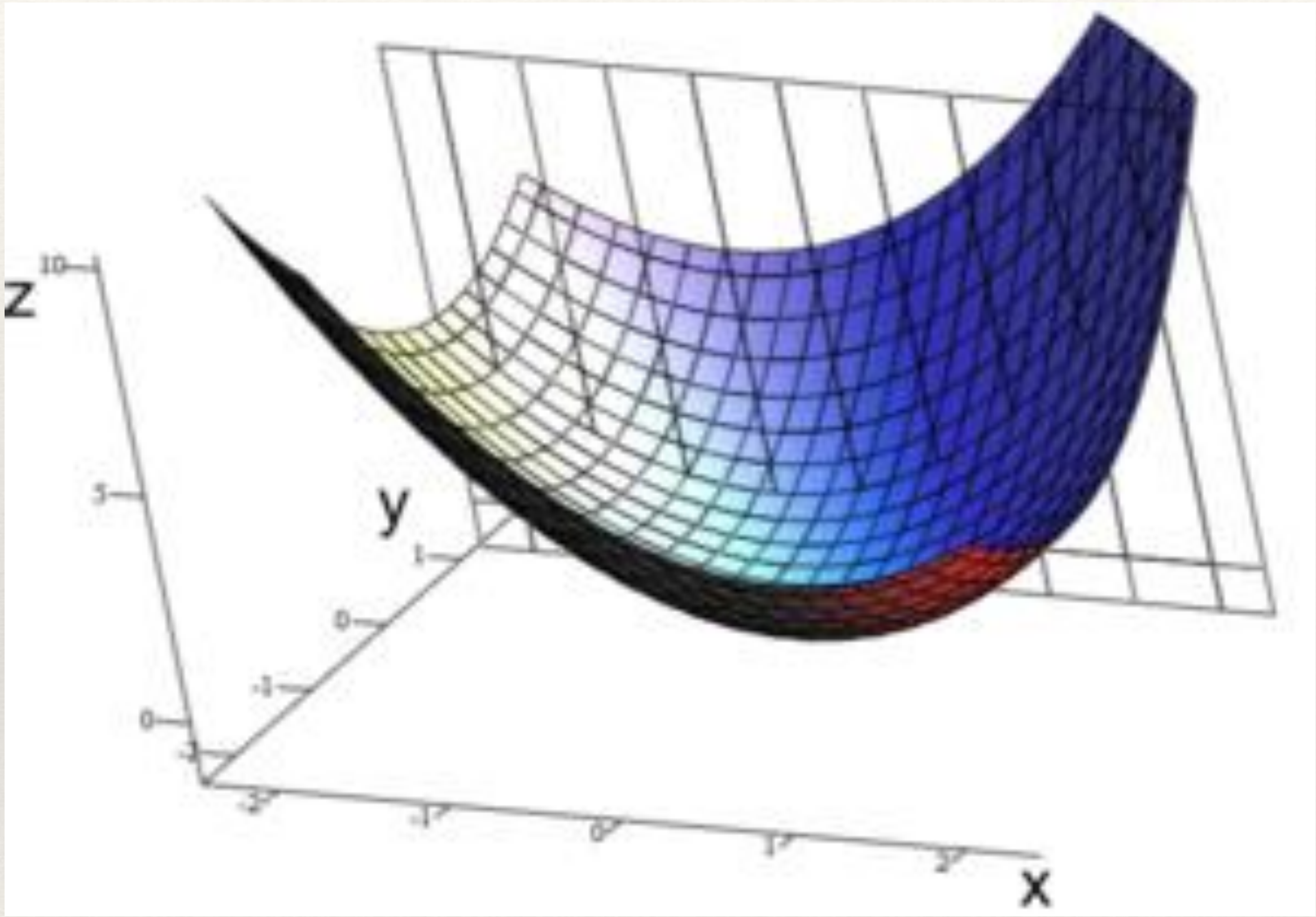
גרדיאנט

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



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

פונקציה קמורה – רב מימדית



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

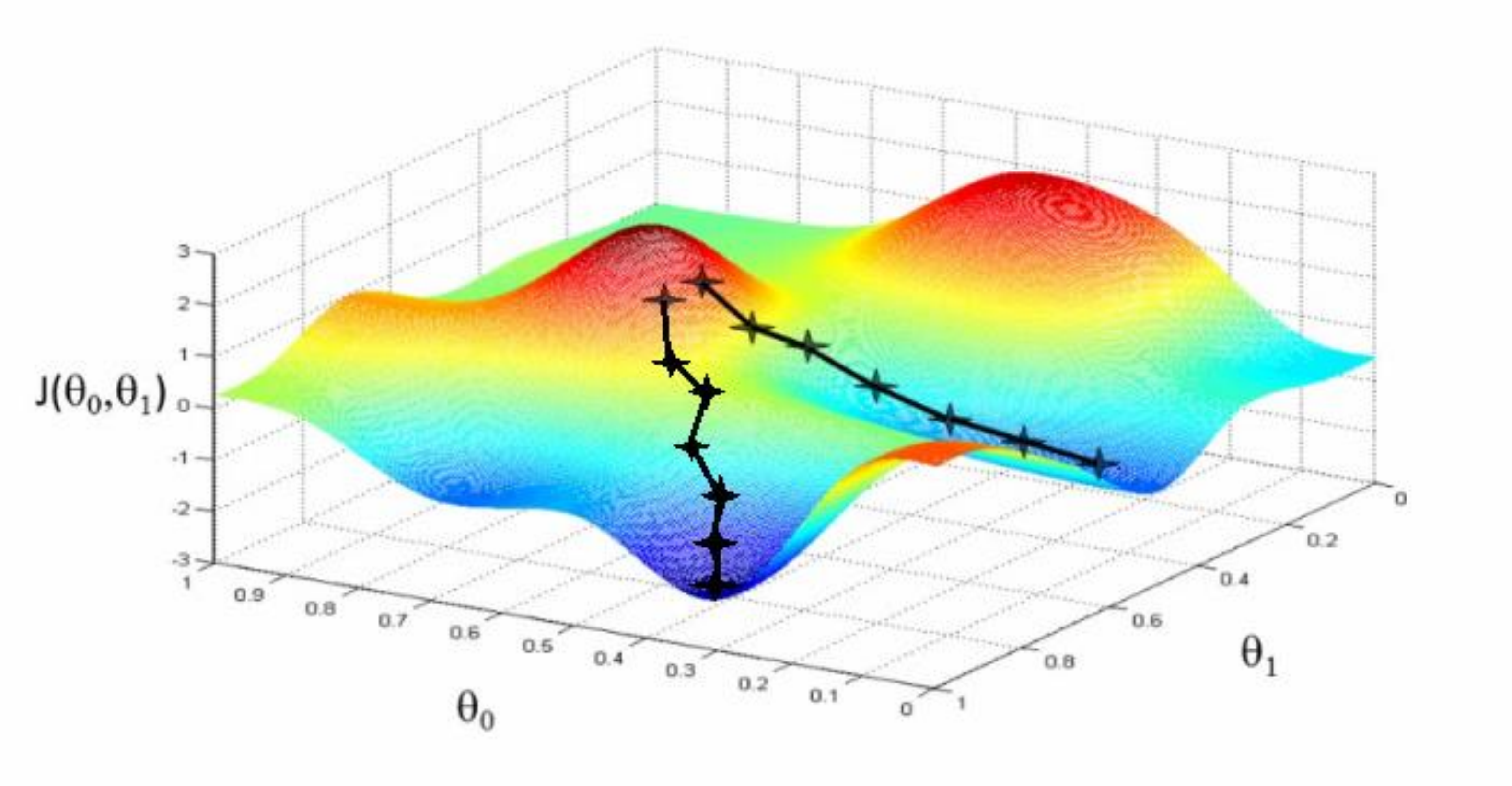
תרגיל – מהם הניגזרות החלקיות לפי x ולפי y?

תשובה –

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

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

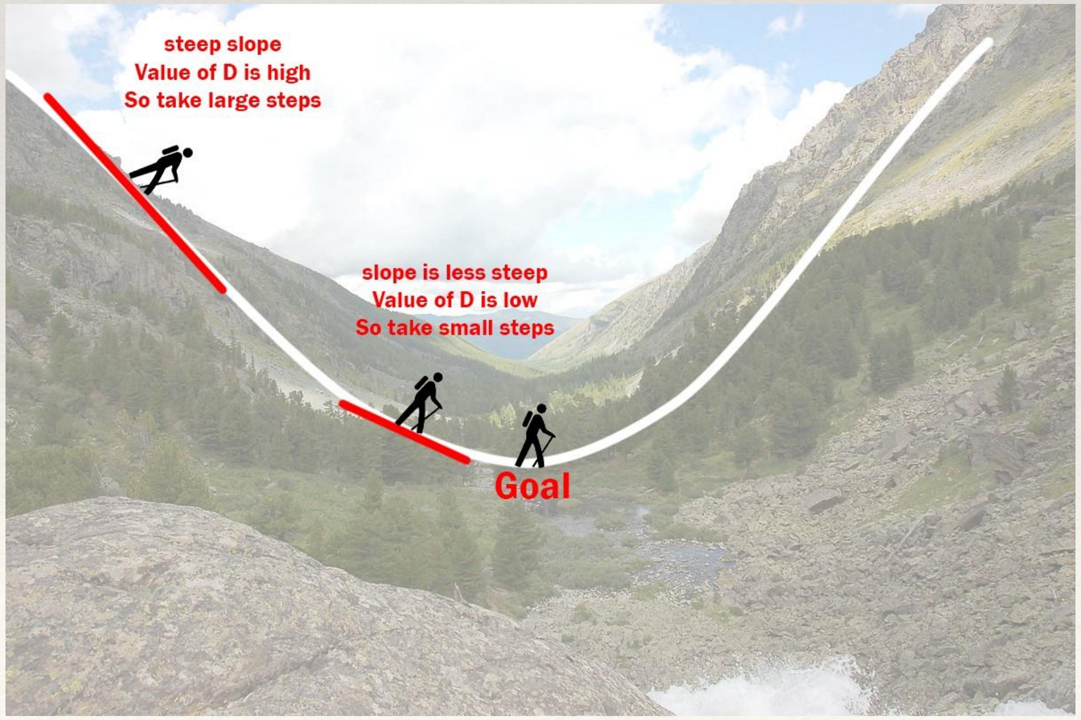
אלגוריתם Gradient Descent (מורד הגרדיאנט)



מבוא ל- Gradient Descent

שאלה: אנחנו נמצאים במקום הררי ורוצים להגיע לתחתית, כיצד נדע לאיזה כיוון ללכת?

תשובה - נניח שיש נחל, הנובע בין ההרים, אשר מגיע לבסוף לים, נעקוב אחרי כיוון הנחל, ומובטח לנו שנגיע לתחתית.



הבעיה שלנו - אנחנו רוצים למצוא את המינימום של הפונקציה, אך אין לנו איך לחשב זאת

הפתרון – נוכל לפתור את הבעיה, כבעיית אופטימיזציה, ע"י שימוש בגרדיאנט.

- הגדיאנט, נותן קירוב לפונקציה
- דומה לכיוון ע"י הנגזרת

אינטואיציה - Gradient Descent

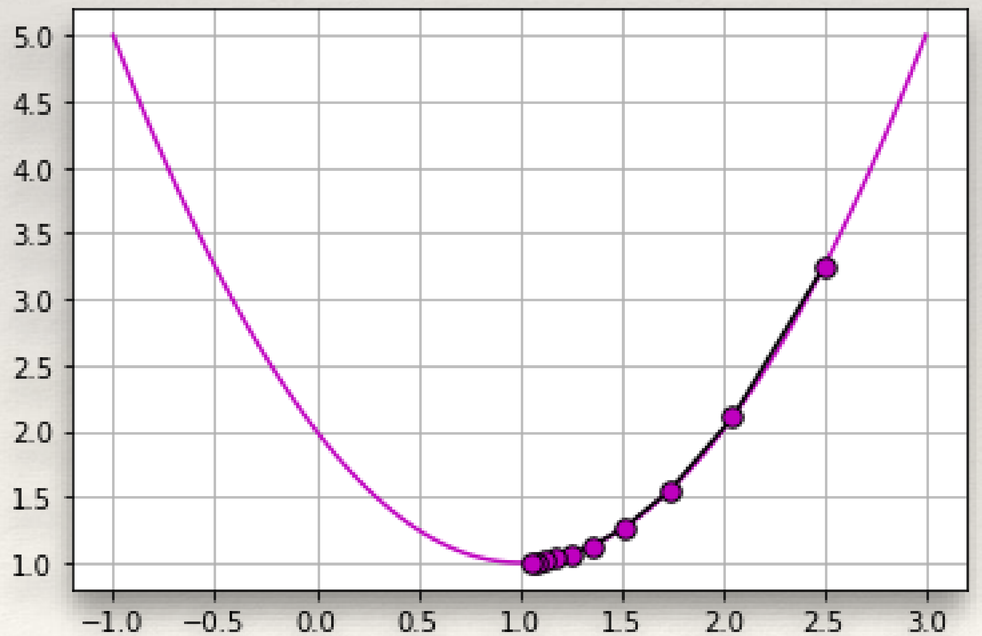
נרצה בכל שלב לעדכן את ערכי המשתנים,
כדי להתקרב עוד ועוד למינימום

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

במקרה של פרמטר יחיד, בכל שלב
באלגוריתם, נשתמש בנגזרת, כדי להתקרב
למינימום

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

קבוע הלמידה (learning rate) –
קובע את קצב ההתכנסות

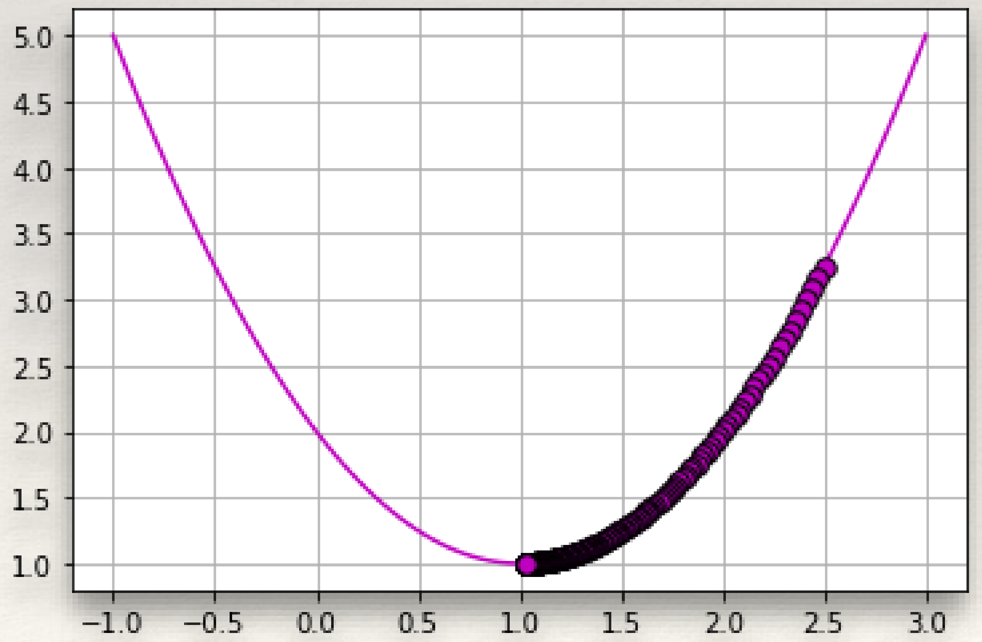


אינטואיציה - Gradient Descent

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

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

קבוע למידה קטן יגרור
התכנסות איטית

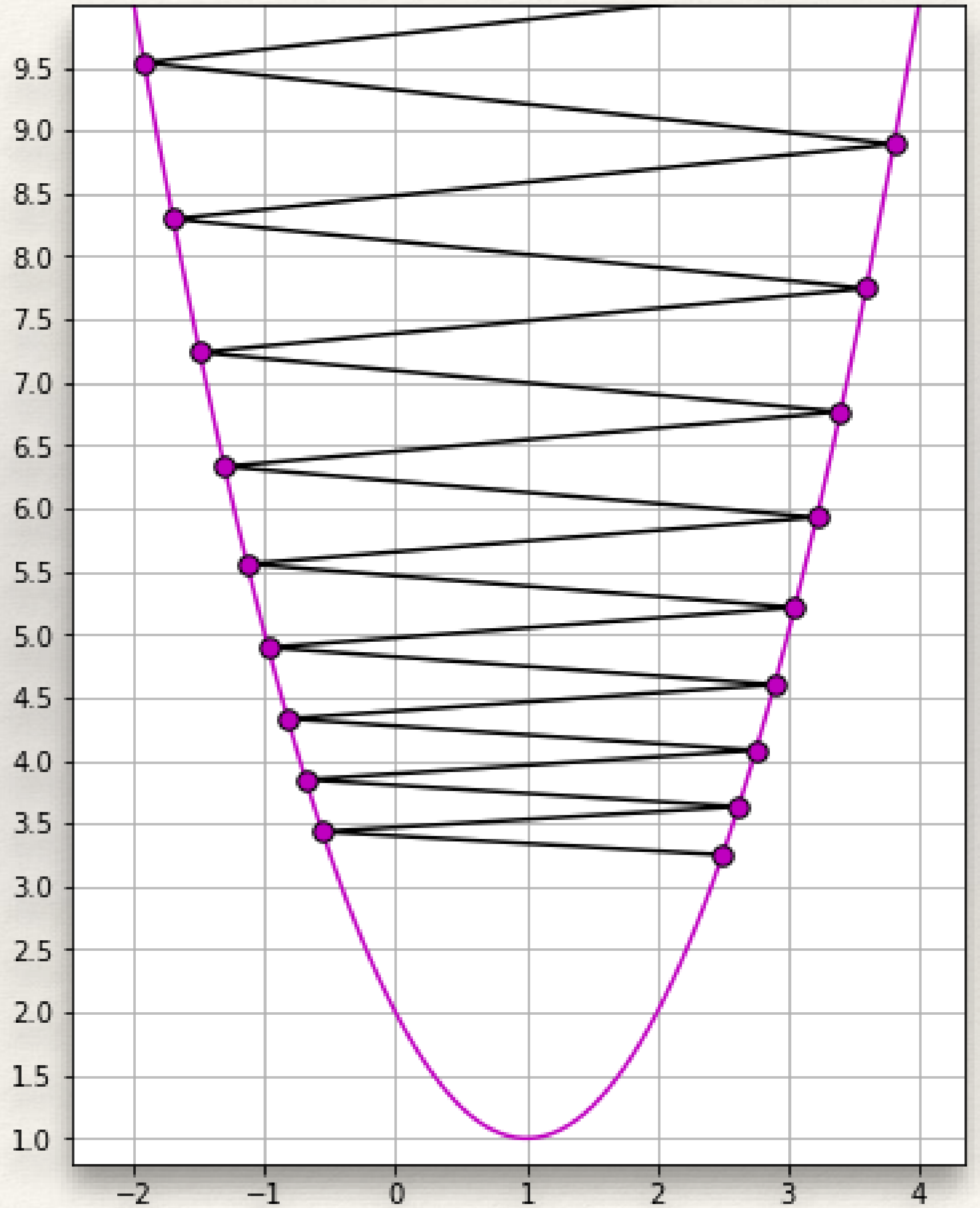


אינטואיציה - Gradient Descent

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

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

קבוע למידה גדול יגרור ש-J
אולי לא יקטן בכלל סבב, ואולי
אף יתבדר



אלגוריתם - Gradient Descent

מתווה האלגוריתם:

- אתחל ערכים התחלתיים לפרמטרים (למשל עבור w_0,w_1)
- מעדכנים את הפרמטרים עד להתכנסות (בתקווה הגעה למינימום):
  - $w_0^{t+1} := w_0^t - \alpha \cdot \frac{\partial f}{\partial w_0^t}$
  - $w_1^{t+1} := w_1^t - \alpha \cdot \frac{\partial f}{\partial w_1^t}$

תרגיל -

- נתונה הפונקציה $g(w_0,w_1) = w_0 + w_1$.
- נתונים הערכים ההתחלתיים הבאים: $w_0^0 = 2, w_1^0 = 6$, כ"כ נתון $\alpha = 0.01$
  - חשבו את w_0^1, w_1^1 (כלומר את ערכי הפרמטרים לאחר סבב אחד)

תשובה – ראשית יש לחשב את הנגזרות החלקיות $\frac{\partial f}{\partial w_0^0} = 1, \frac{\partial f}{\partial w_1^0} = 1$

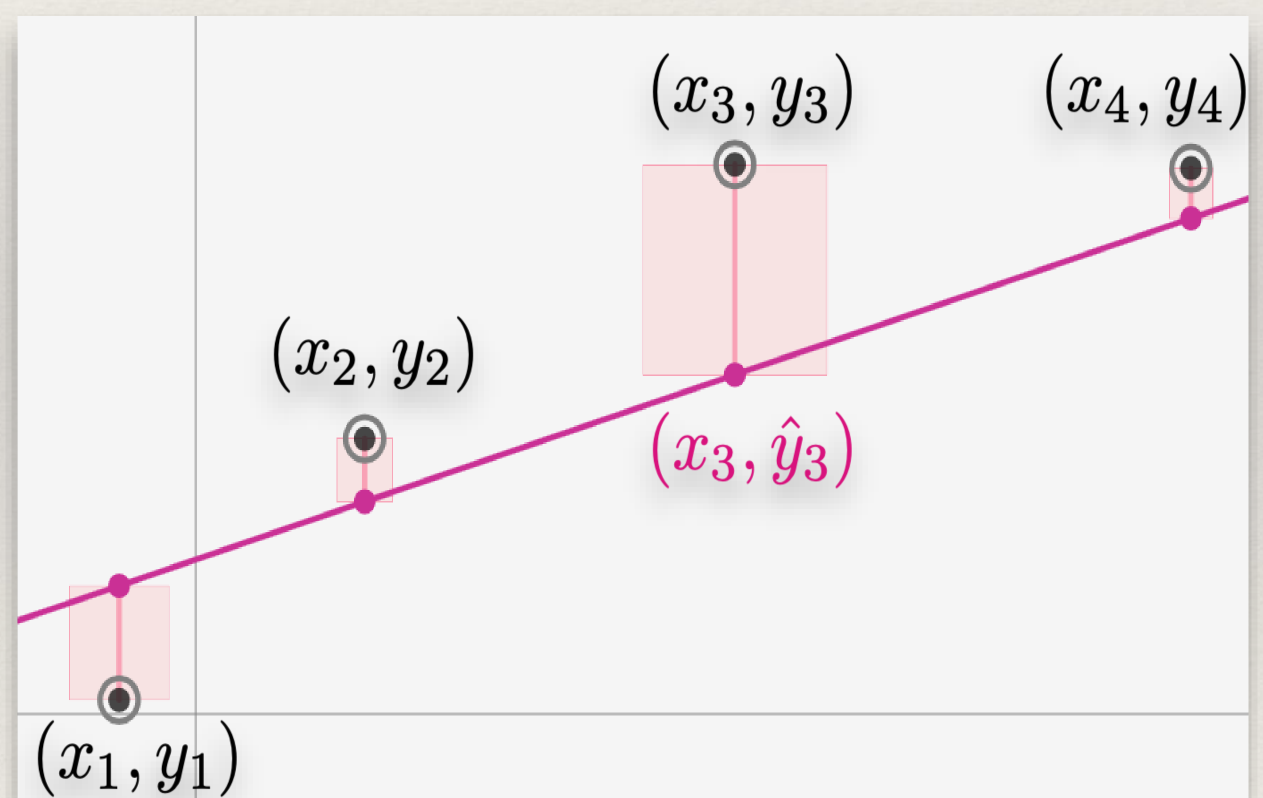
- $w_0^1 \coloneqq 2 - 0.01 \cdot 1 = 1.99$
- $w_1^1 \coloneqq 6 - 0.01 \cdot 1 = 5.99$

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

פונקצית מחיר (Cost Function)

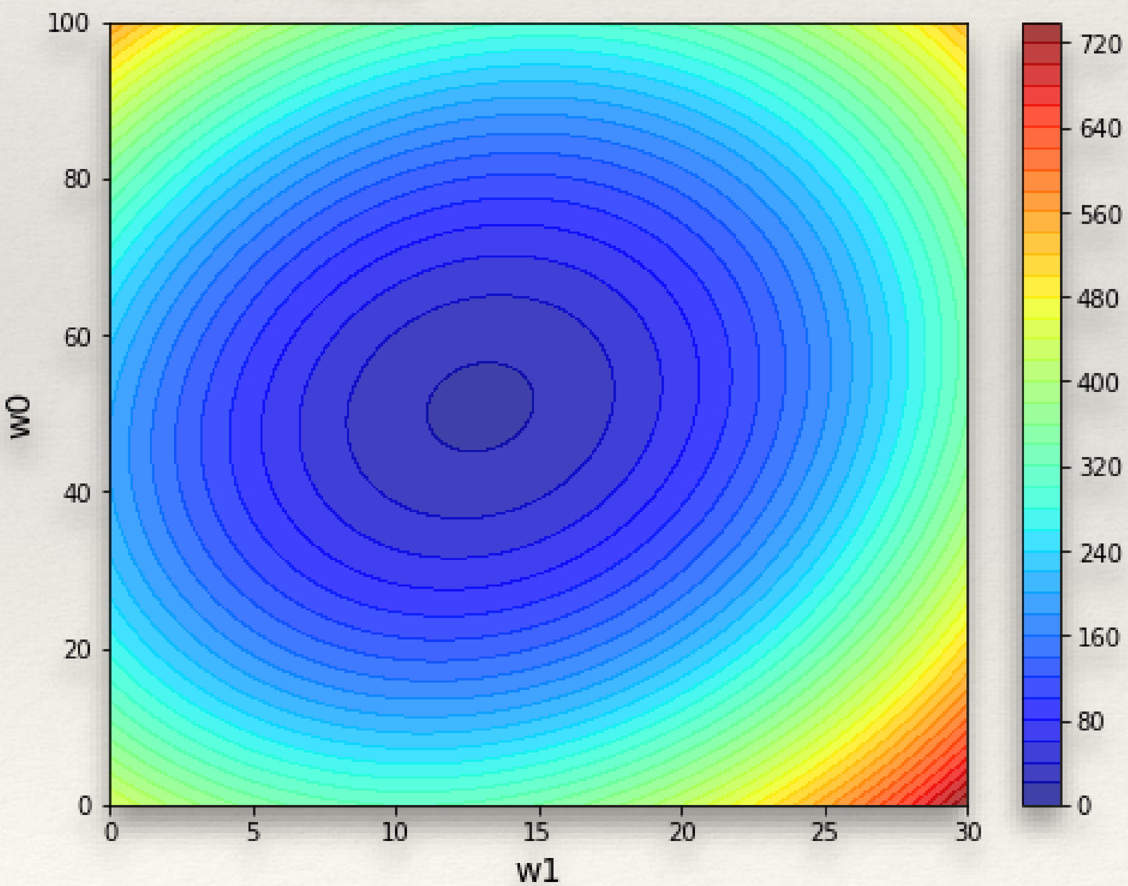
המודל הליניארי:

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



פונקצית המחיר – מוגדרת ע"י ממוצע הטעות הריבועית

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



Linear Regression via Gradient Descent

פונקצית המחיר:

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

הנגזרות החלקיות

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

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

עדכון הפרמטרים

$$w_0 = w_0 - \alpha \cdot \frac{\partial J}{\partial w_0} = w_0 - \alpha \cdot \frac{2}{n} \cdot \sum\nolimits_{i=1}^{n}\left[\left(w_0 + w_1 \cdot x_{i,1} - y_i\right)\cdot 1\right]$$

$$w_1 = w_1 - \alpha \cdot \frac{\partial J}{\partial w_1} = w_1 - \alpha \cdot \frac{2}{n} \cdot \sum\nolimits_{i=1}^{n}\left[\left(w_0 + w_1 \cdot x_{i,1} - y_i\right)\cdot x_{i,1}\right]$$

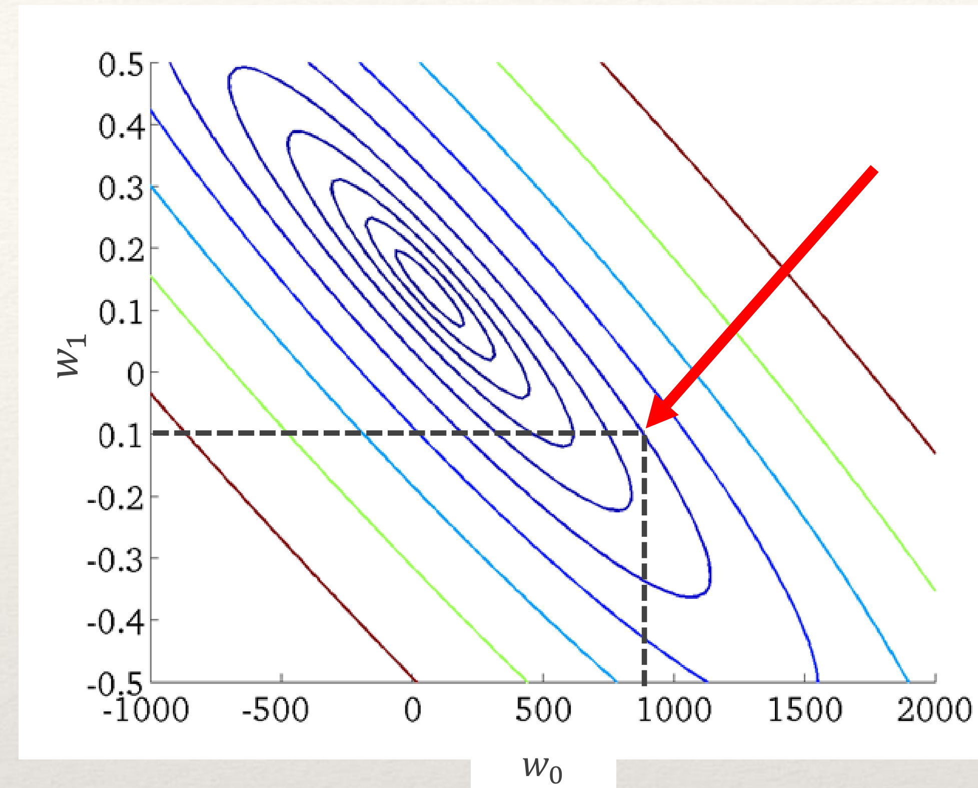
שאלה 13

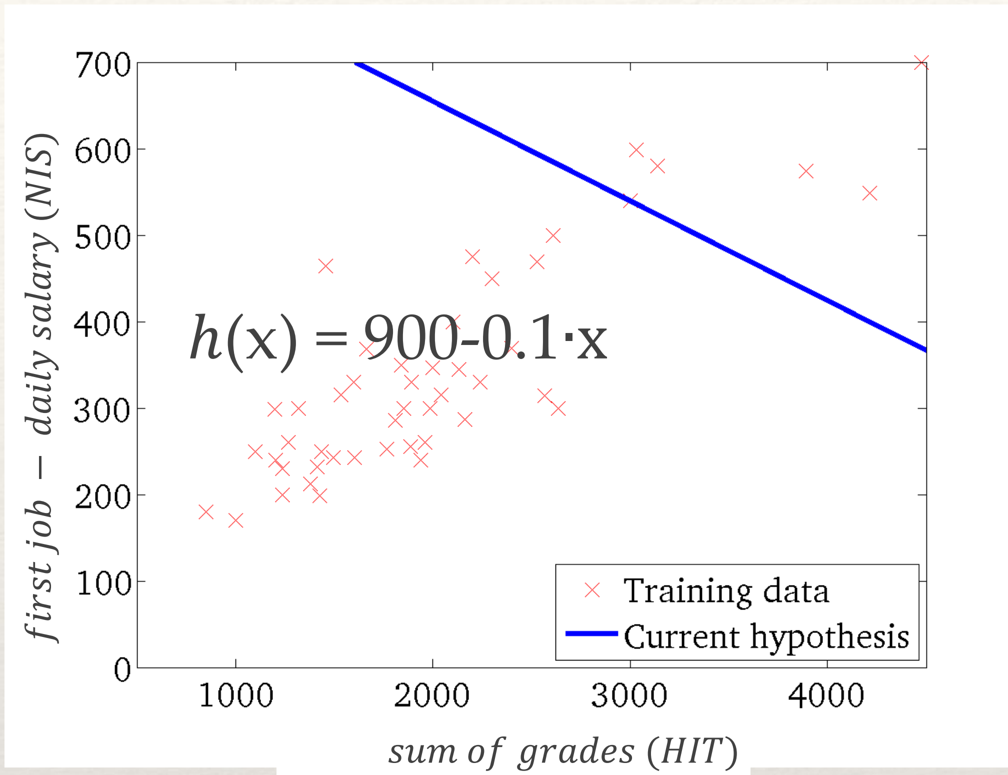
$J(w_0,w_1)$

- פונקצית המחיר (מדמה גרף תלת מימדי)
- פונקצית המחיר - פונקציה של w_0,w_1

$H_W(x)$

- ישר הרגרסיה, מאפיין אחד
- פונקציה של x, עבור w_0,w_1 מסוימים





תרגיל – בתחילת האיטרציה (הסבב) הרביעית בtraining, קיבלנו את המשקולות המתאימות למשוואה הבאה: $h(x) = 900-0.1\cdot x$, נתונים $\alpha= 0.0000001$, ועבור ה- train-set

<u>חשבון</u> את w_0,w_1, לאיטרציה הבאה בעזרת Gradient Descent עבטר הרגרסיה הלינארית

- חשבו סבב נוסף

סכום הציונים	שכר יומי בש"ח
2104	460
1534	315
852	178

[Train set]

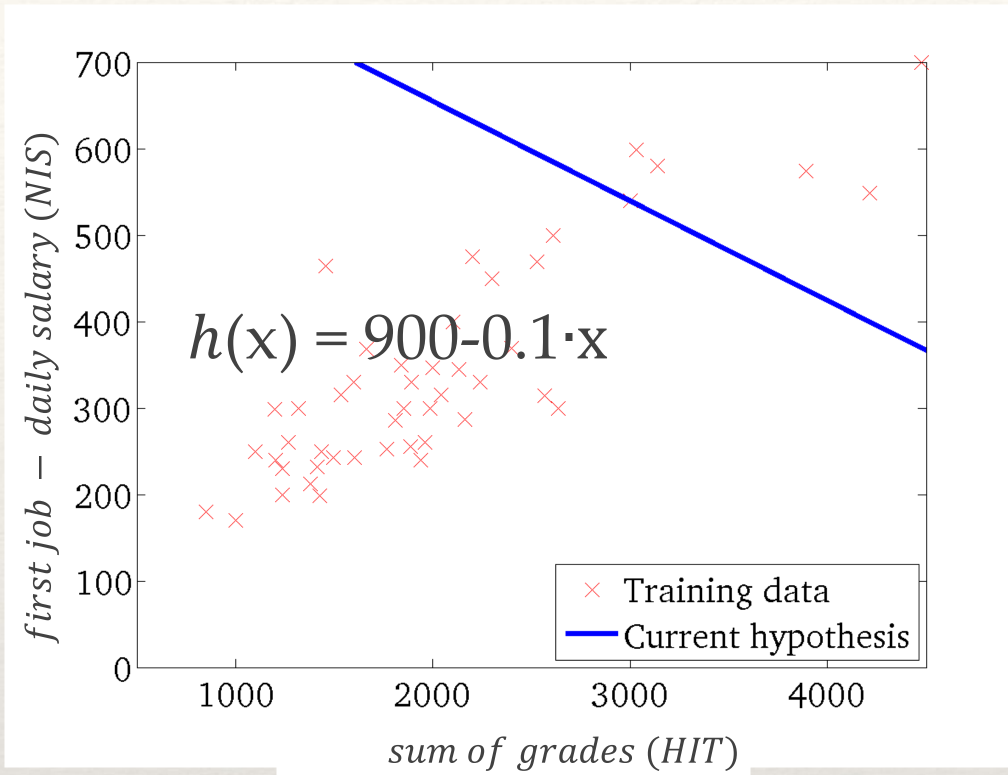
שאלה 13

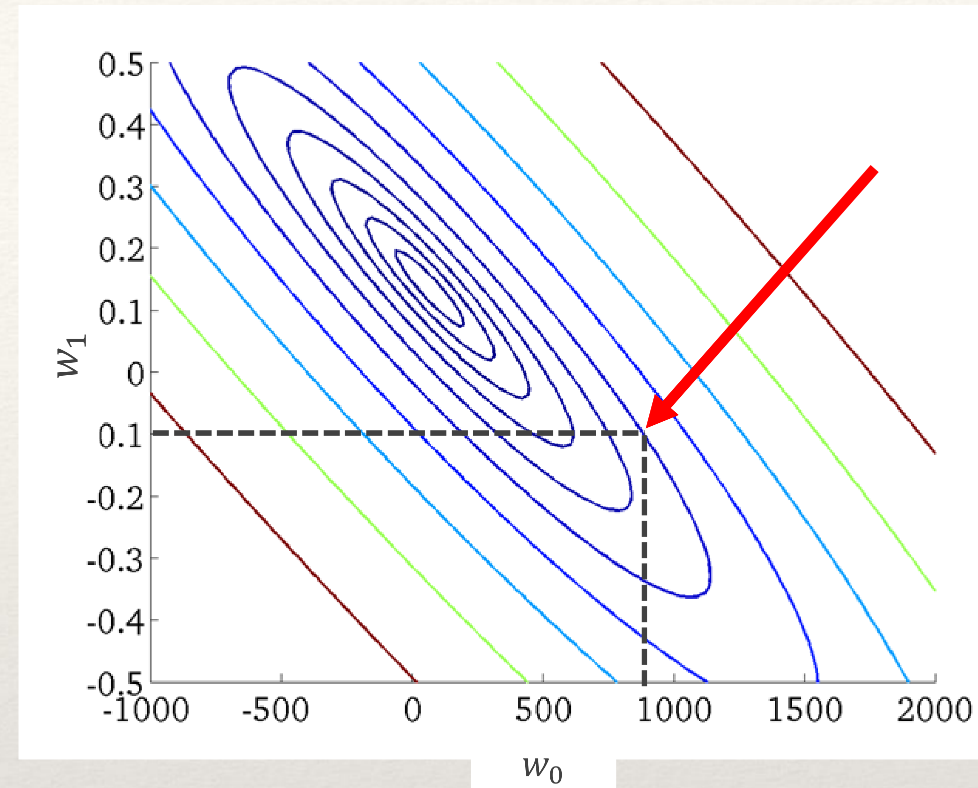
$J(w_0,w_1)$

- פונקציית המחיר (מדמה גרף תלת מימדי)
- פונקציית המחיר - פונקציה של w_0,w_1

$H_W(x)$

- ישר הרגרסיה, מאפיין אחד
- פונקציה של x, עבור w_0,w_1 מסוימים





שכר יומי בש"ח	סכום הציונים
460	2104
315	1534
178	852

Train set

פתרון:

חישוב w_0 חדש

$$\boldsymbol{w_0} = w_0 - \alpha \cdot \frac{\partial J}{\partial w_0}$$

$$= 900 - 0.0000001 \cdot \frac{2}{3}$$

$$\cdot [(900 - 0.1 \cdot 2104 - 460) \cdot 1$$

$$+ (900 - 0.1 \cdot 1534 - 315) \cdot 1$$

$$+ (900 - 0.1 \cdot 852 - 178) \cdot 1]]$$

$$= 900 - 0.0000001 \cdot \frac{2}{3} \cdot 1{,}298 \approx \mathbf{899.99}$$

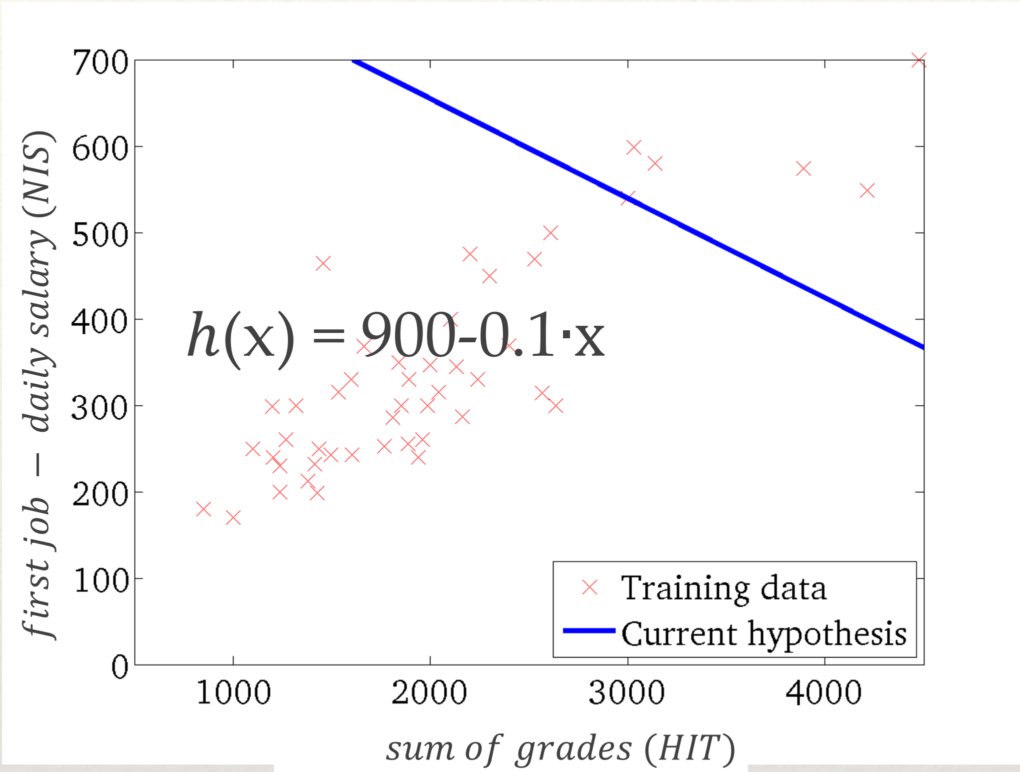
שאלה 13

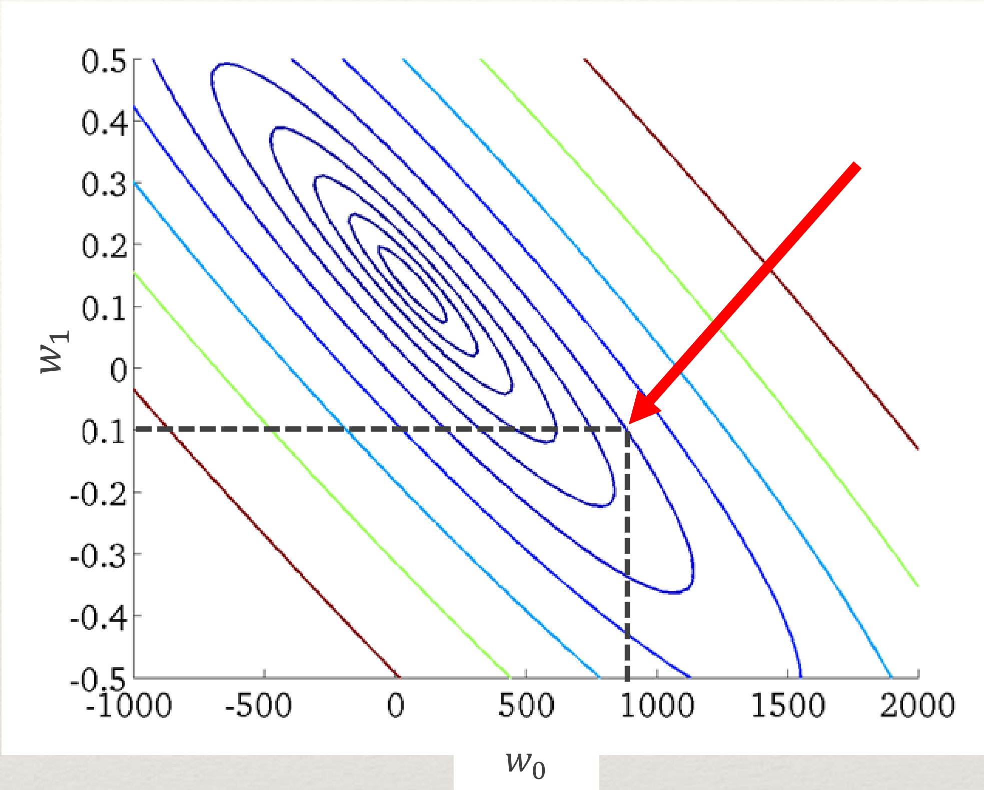
$J(w_0,w_1)$

- פונקצית המחיר (מדמה גרף תלת מימדי)
- פונקצית המחיר - פונקציה של w_0,w_1

$H_W(x)$

- ישר הרגרסיה, מאפיין אחד
- פונקציה של x, עבור w_0,w_1 מסוימים





סכום הציונים	שכר יומי בש"ח
2104	460
1534	315
852	178

[Train set]

המשך פתרון:

חישוב w_1 חדש

$$\boldsymbol{w_1} = w_1 - \alpha \cdot \frac{\partial J}{\partial w_1}$$

$$= -0.1 - 0.0000001 \cdot \frac{2}{3}$$

$$\cdot [(900 - 0.1 \cdot 2104 - 460) \cdot 2104$$

$$+ (900 - 0.1 \cdot 1534 - 315) \cdot 1534$$

$$+ (900 - 0.1 \cdot 852 - 178) \cdot 852]]$$

$$= -0.1 - 0.0000001 \cdot \frac{2}{3} \cdot 2{,}212{,}864.8 \approx -0.147$$

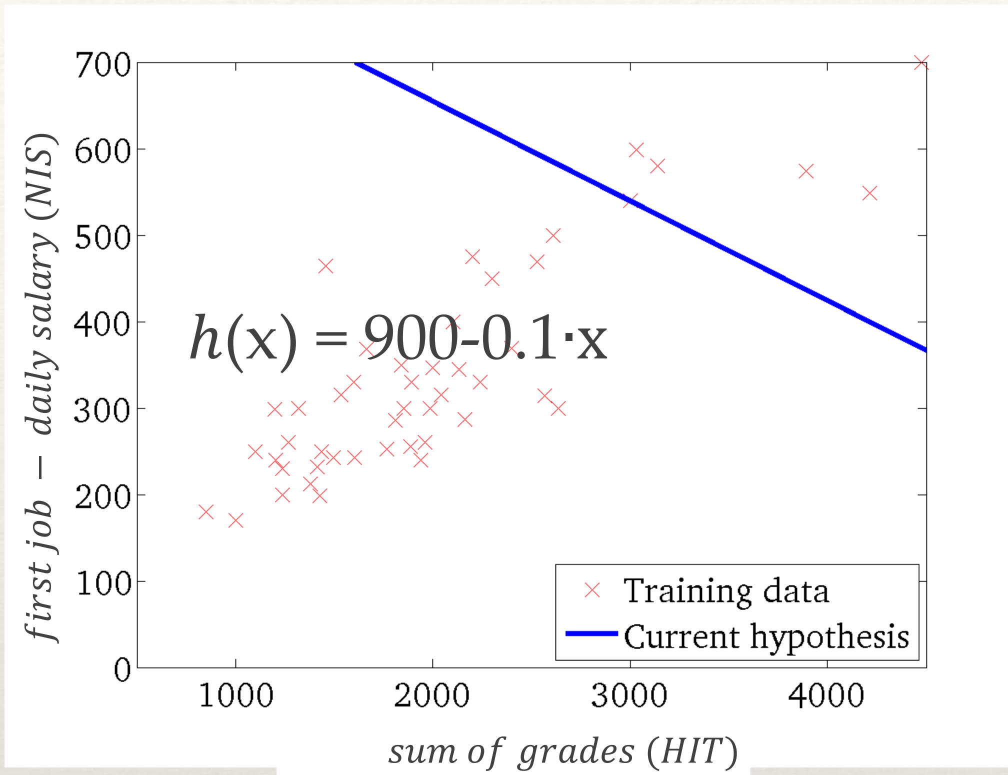
שאלה 13

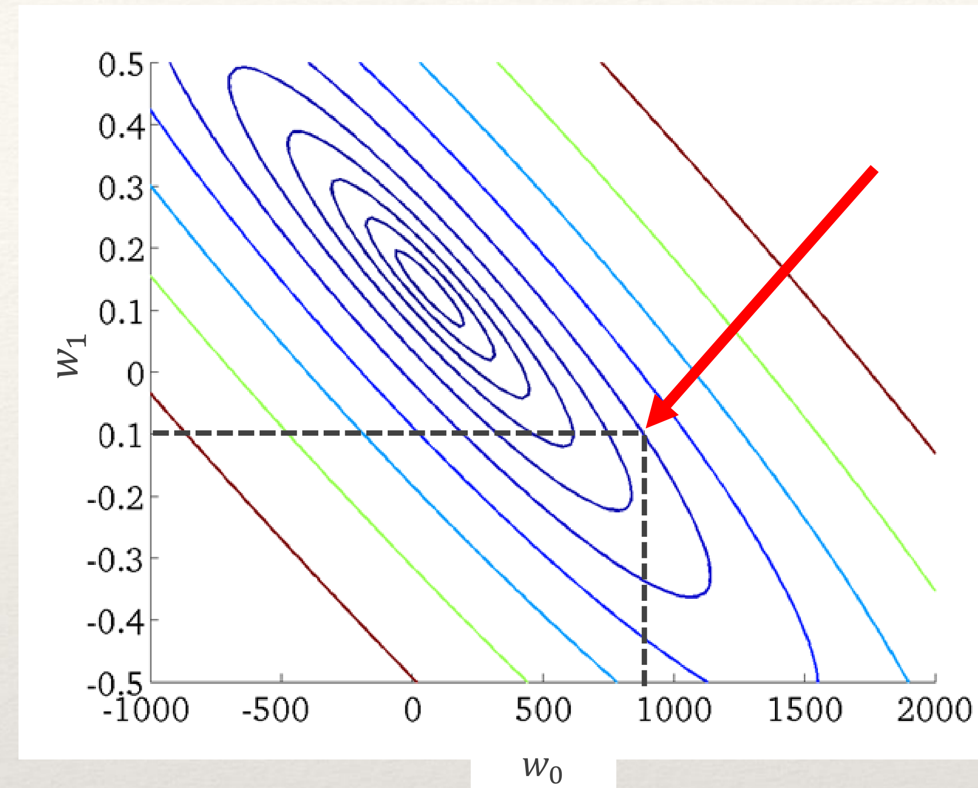
$J(w_0,w_1)$

- פונקצית המחיר (מדמה גרף תלת מימדי)
- פונקצית המחיר - פונקציה של w_0,w_1

$H_W(x)$

- ישר הרגרסיה, מאפיין אחד
- פונקציה של x, עבור w_0,w_1 מסוימים





סכום הציונים	שכר יומי בש"ח
2104	460
1534	315
852	178

Train set

פתרון לאחר איטרציה אחת:

w_0 = **899.99**

w_1 =**-0.147**

$H_W(x)$= 899.99- 0.047x

שאלה 13

$J(w_0,w_1)$

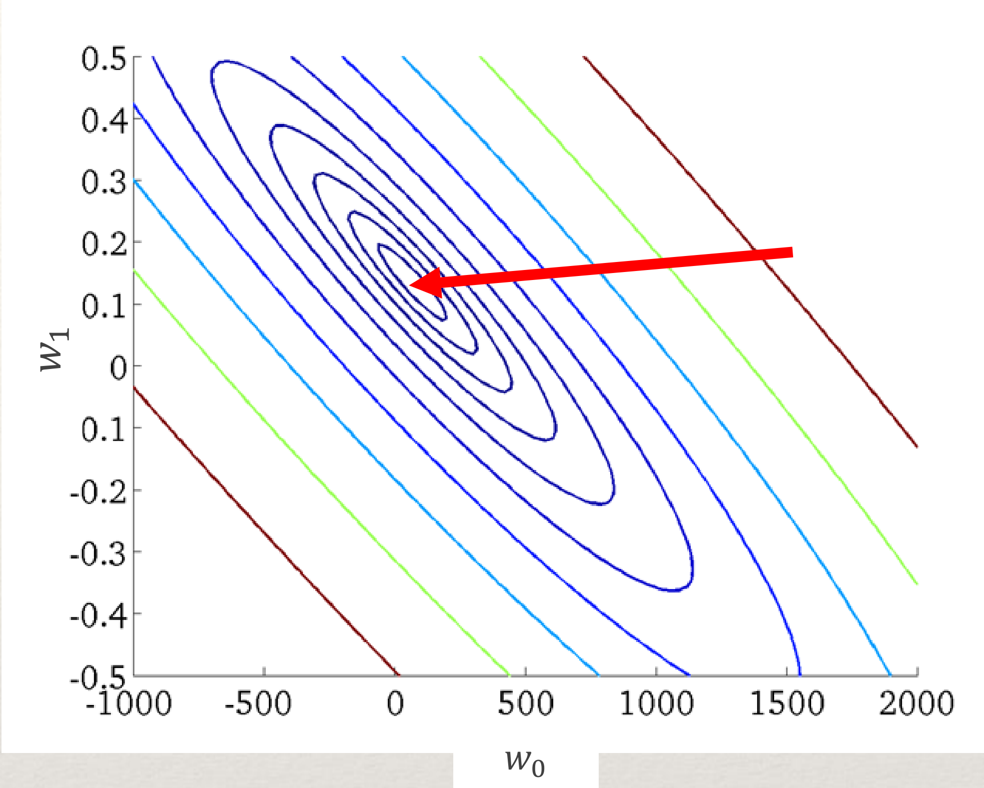
- פונקצית המחיר (מדמה גרף תלת מימדי)

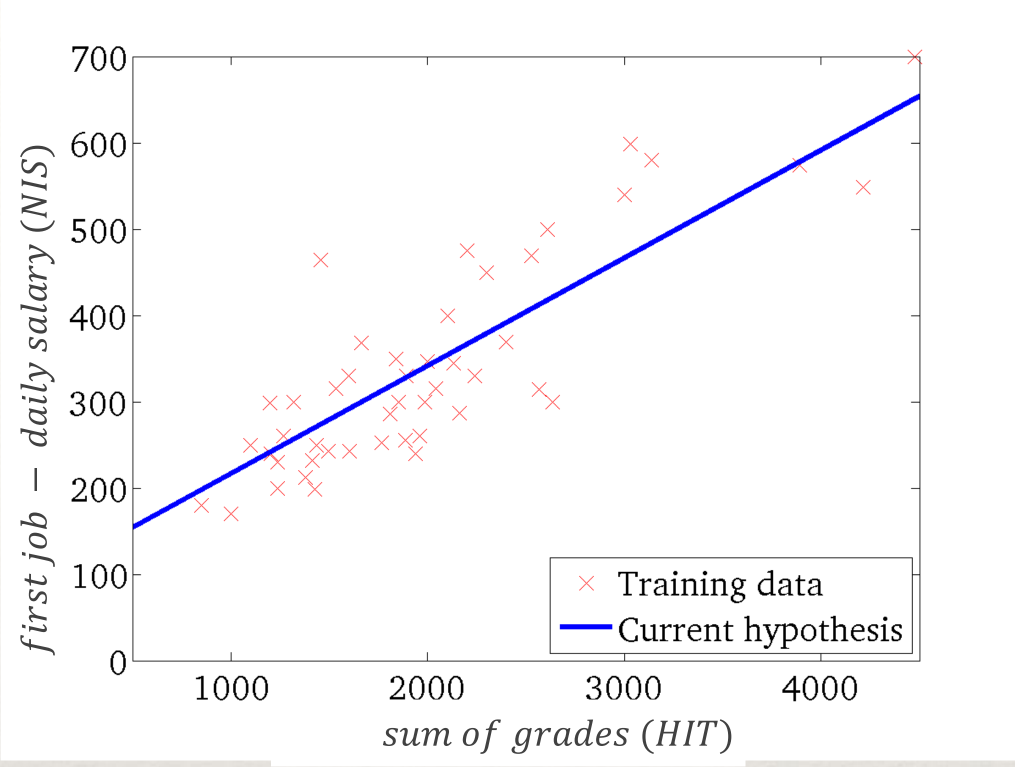
- פונקצית המחיר - פונקציה של w_0,w_1

$H_W(x)$

- ישר הרגרסיה, מאפיין אחד

- פונקציה של x, עבור w_0,w_1 מסוימים





סכום הציונים	שכר יומי בש"ח
2104	460
1534	315
852	178

Train set

... לאחר כמה סבבים

שאלת 14 (סקר)

במהלך שלב האימון ברגרסיה לינארית בעזרת אלגוריתם gradient descent, אשר בו נתון ש-
$\alpha=0.5$, לאחר האיטרציה השלישית התקבלו הערכים: w0=4, w1=2

נתון ה- training set הבא:

מספר הדוגמה (i)	x_i	y_i
1	1	7
2	0	3
3	4	13
4	-4	-4

מה מהתשובות שלהלן מתקיים, לגבי עדכון הפרמטרים
באיטרציה הרביעית (הכפילו את תוצאת הגדיאנט ב- $(-\alpha*(2/n)$:

<u>תשובות אפשרויות</u>:

א. דוגמה מספר 2 לא תשפיע על עדכון w0
ב. דוגמה מספר 1 תתרום לכך שערכו של w1 ירד ב1.
ג. דוגמה מספר 4 תתרום לכך שערכו של w0 יעלה ב1.
ד. דוגמה מספר 3 תתרום לכך שערכו של w1 יעלה ב1.

שאלת 14 (סקר)

במהלך שלב האימון ברגרסיה לינארית בעזרת אלגוריתם gradient descent, אשר בו נתון ש-
$\alpha=0.5$, לאחר האיטרציה השלישית התקבלו הערכים: w0=4, w1=2

נתון ה- training set הבא:

מספר הדוגמה (i)	x_i	y_i
1	1	7
2	0	3
3	4	13
4	-4	-4

מה מהתשובות שלהלן מתקיים, לגבי עדכון הפרמטרים
באיטרציה הרביעית (הכפילו את תוצאת הגדיאנט ב- $(-\alpha*(2/n)$):

<u>תשובות אפשרויות:</u>

א. דוגמה מספר 2 לא תשפיע על עדכון w0
ב. דוגמה מספר 1 תתרום לכך שערכו של w1 ירד ב1.
ג. דוגמה מספר 4 תתרום לכך שערכו של w0 יעלה ב1.
ד. דוגמה מספר 3 תתרום לכך שערכו של w1 יעלה ב1.

<u>התרומה של דוגמה 3 לעדכון ערכו של w1:</u>

$-\alpha*(2/n)*(\hat{y}-y)*x1_{(3)}=-\alpha*(2/n)*(w0+w1*x1_{(3)}-y1_{(3)})*x1_{(3)}=$
$-0.5*(2/4)*(4+2*4-13)*4=-(1/4)*(-1)*4=\mathbf{1}$

ולכן, דוגמה מספר 3 תתרום לכך שערכו של w1 יעלה ב1

שאלת 14 (סקר)

במהלך שלב האימון ברגרסיה לינארית בעזרת אלגוריתם gradient descent, אשר בו נתון ש- $\alpha=0.5$, לאחר האיטרציה השלישית התקבלו הערכים: w0=4, w1=2

נתון ה- training set הבא:

מספר הדוגמה (i)	x_i	y_i
1	1	7
2	0	3
3	4	13
4	-4	-4

מה מהתשובות שלהלן מתקיים, לגבי עדכון הפרמטרים באיטרציה הרביעית (הכפילו את תוצאת הגדיאנט ב- $(-\alpha*(2/n)$):

<u>תשובות אפשרויות:</u>

א. דוגמה מספר 2 לא תשפיע על עדכון w0
ב. דוגמה מספר 1 תתרום לכך שערכו של w1 ירד ב1.
ג. דוגמה מספר 4 תתרום לכך שערכו של w0 יעלה ב1.
ד. דוגמה מספר 3 תתרום לכך שערכו של w1 יעלה ב1.

<u>התרומה של דוגמה 1 לעדכון ערכו של w1:</u>

$\alpha*(2/n)*(\hat{y}-y)*x1_{(1)}=\alpha*(2/n)*(w0+w1*x1_{(1)}-x1_{(1)})*x1_{(1)}=$
$-0.5*(2/4)*(4+2*1-7)*4=-(1/4)*(-1)*4=\mathbf{1}$

דוגמה מספר 1 תתרום לכך שערכו של w1 יעלה ב1
תשובה ב' <u>לא נכונה</u>

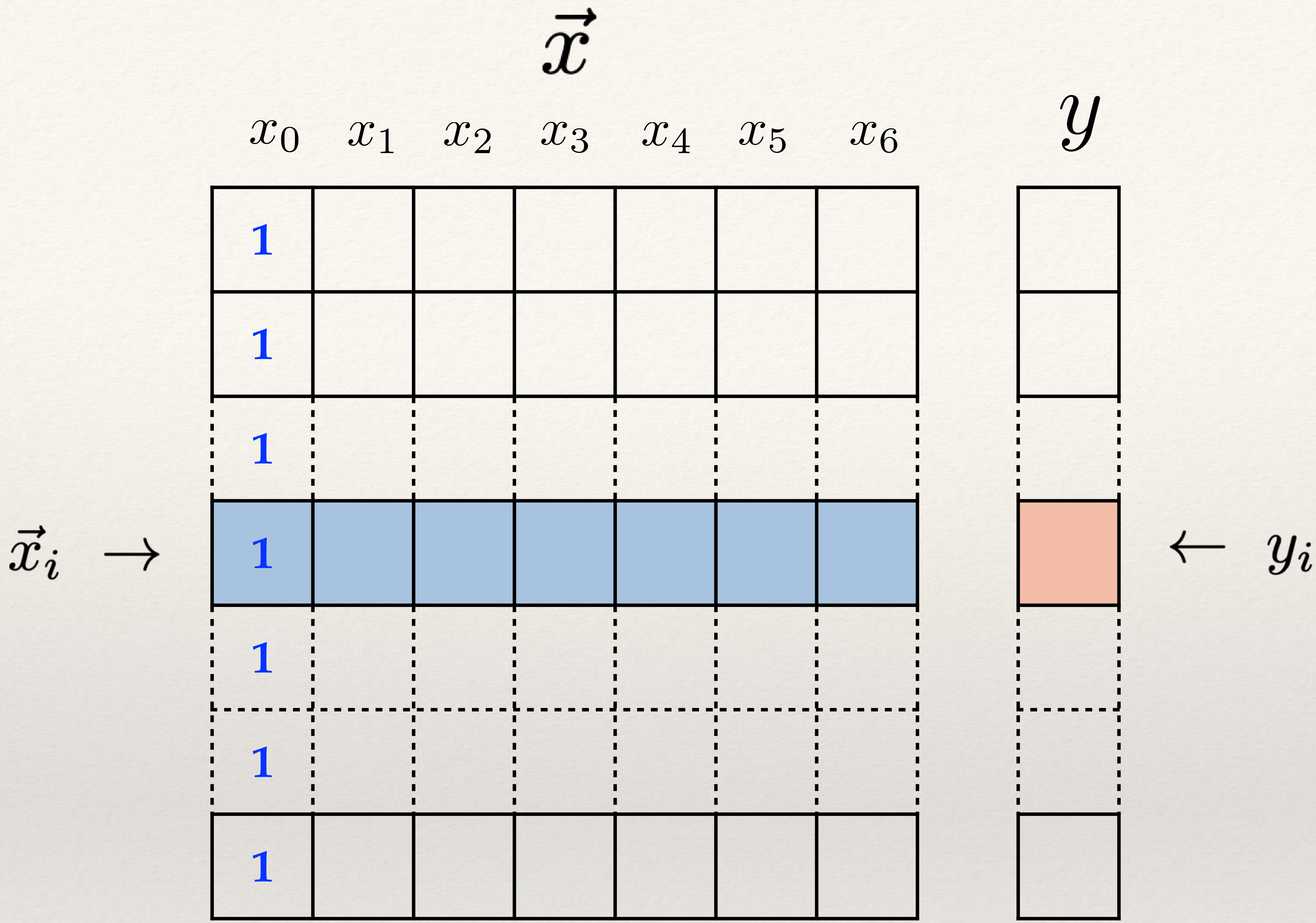
רגרסיה לינארית (linear regression)
ריבוי משתנים (multivariate) - מוטיבציה

multivariate linear regression

במקרה זה, לכל וקטור באימון יש יותר ממאפיין אחד (למשל: גודל הדירה, קומה, כיווני-אוויר, וכו')

Price ($K) (y)	Size ($meter^2$) (x4)	Number of bedrooms **(x3)**	Number of floors **(x2)**	Age of home (years) **(x1)**
460	2104	5	1	45
232	1416	3	2	40
315	1534	3	2	30
178	852	2	1	36

$$H_W(x) = \vec{w}^T \cdot \vec{x} = w_0 \cdot x_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_d \cdot x_d$$



המודל הלינארי -עבור רגרסיה מרובת משתנים:

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

קומבינציה לינארית של המאפיינים

פונקצית מחיר (Cost Function)

המודל הליניארי:

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

פונקצית המחיר

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

$(n \geq 1)$

Multivariate Gradient Descent Algorithm - for linear regression:

Repeat until done:

We want w_0 to be partialy derived as the rest of $\vec{w}$, so if j=0, $x_{i,0} = 1$

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

(simultaneously update w_j for j=0,…,d)

Gradient Descent - 15 שאלה

Multivariate Gradient Descent Algorithm - for linear regression:

Repeat until done: $w_j = w_j - \alpha \cdot \frac{\partial J}{\partial w_j} = w_j - \alpha \cdot \frac{2}{n} \cdot \sum_{i=1}^{n} [(\vec{w} \cdot \vec{x}_i - y_i) \cdot x_{i,j}]$

תרגיל –

נתונים α=0. 1, w_j=0, לכל j

ונתון ה-train set

בצעו סבב אחד של Gradient Descent

תשובה:

נחשב את ערכי הנגזרות החלקיות ונכפיל ב-α

נחשב את ערכי וקטור המשקולות ($\vec{w}$) החדשים ...

שאלה: האם ערכי $\vec{w}$ המעודכן נראים סבירים? מדוע זה קרה?

Train set

Size (meter²)	Number of bedrooms	Age of home (years)	Price ($K)
2104	5	45	460
1416	3	40	232
852	2	36	178

Predicted (Ŷ)	(Ŷ-Y)xi,0	(Ŷ-Y)xi,1	(Ŷ-Y)xi,2	(Ŷ-Y)xi,3
0	-460	-967840	-2300	-20700
0	-232	-328512	-696	-9280
0	-178	-151656	-356	-6408
Sum	-870	-1448008	-3352	-36388
α*2*1/n*Sum	-58	-96533.9	-223.4667	-2425.86667
new w vals	58	96533.87	223.46667	2425.866667

דגשים - Gradient Descent

- את העדכון של w_j בכל שלב, יש לבצע במקביל
- למשל עבור מקרה של מאפיין אחד בוקטור המאפיינים:

Correct: Simultaneous update

$$temp0 := w_0 - \alpha \frac{\partial J}{\partial w_0} J(w_0, w_1)$$
$$temp1 := w_1 - \alpha \frac{\partial J}{\partial w_1} J(w_0, w_1)$$
$$w_0 := temp_0$$
$$w_1 := temp_1$$

Incorrect

$$temp0 := w_0 - \alpha \frac{\partial J}{\partial w_0} J(w_0, w_1)$$
$$w_0 := temp_0$$
$$temp1 := w_1 - \alpha \frac{\partial J}{\partial w_1} J(w_0, w_1)$$
$$w_1 := temp_1$$

דגשים - Gradient Descent

- הוספת עוד עמודה $x_0^{(i)}$ - כדי להקל בתהליך, מומלץ להוסיף עוד עמודה עבור $x_0^{(i)}$ במטריצת הוקטורים של ה-train set, כך ולקבוע $x_0^{(i)} = 1$, עבור כל תא ראשון בוקטור (נקרא לו contant_val)
- כאשר יש יותר ממאפיין אחד, צריך לבצע צעד מקדים של scaling
  - נשתמש ב- scaling של t-score או של סילום בטווח [-1,1]
  - שימו לב שאי ביצוע הסילום גרם לחישובי ערכים גבוהים מאוד בתרגילים, ועלול היה לא להתכנס ומאוד משפיע על רגרסיה לינארית
- כאשר רוצים לשערך מודל רגרסיה, צריך להשתמש במטריקות שונות מאשר במודל סיווג
  - נשתמש למשל ב- RMSE

דגשים - Gradient Descent

- תנאי עצירה –
  - יש להפסיק את האלגוריתם כאשר השינוי ב- J (פו' המחיר) קטן מ ε (למשל $\varepsilon < 10^{-3}$)
- קביעת ערכים ראשוניים –
  - יכול להשפיע על התוצאות
    - ניתן להציב $\theta_j = 0$
  - לגבי α, ניתן לקבוע ערכים של $10^{-3}, 3 \cdot 10^{-3}, 10^{-2}, 3 \cdot 10^{-2}$ (Andrew Ng)

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