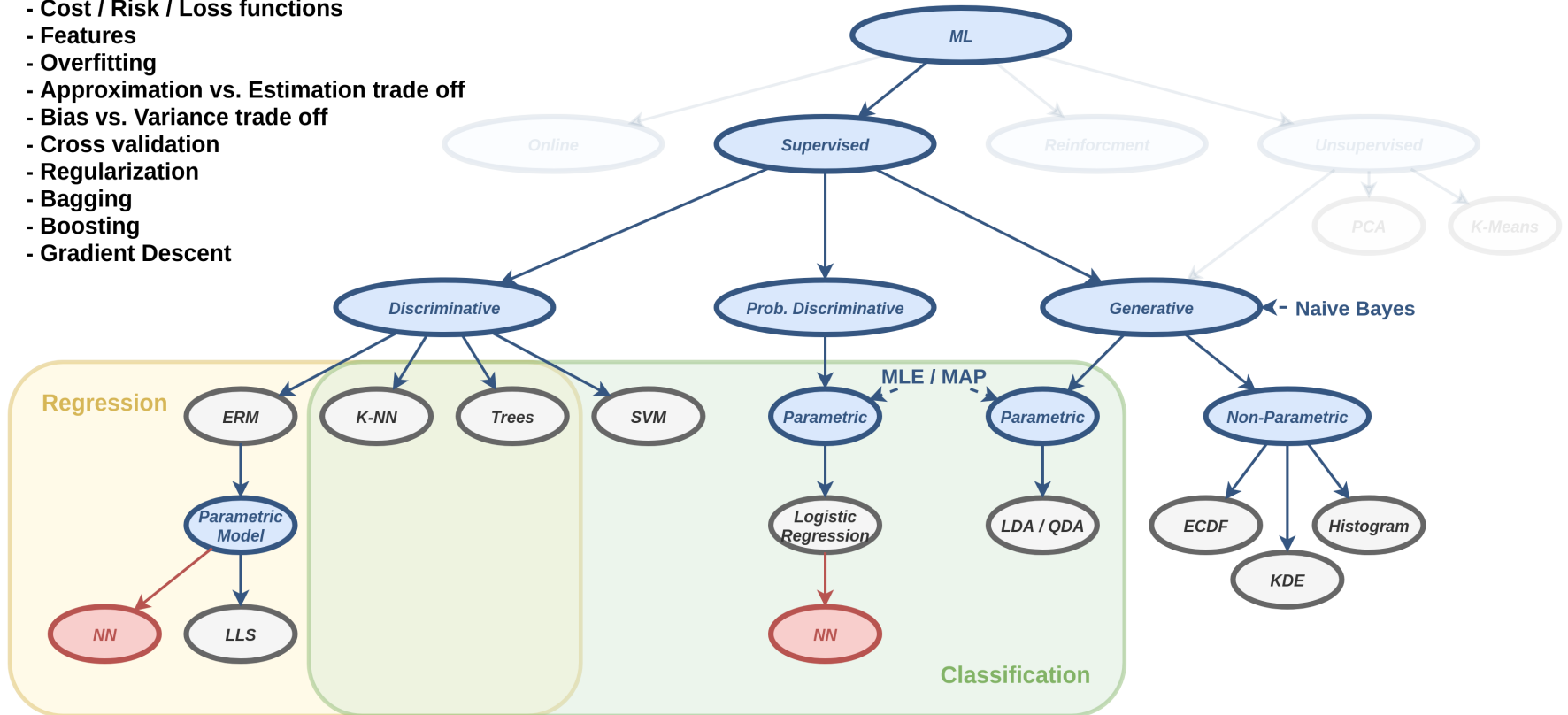


הרצאה 10 - Neural Networks

Subjects Covered in this Course

General concepts:

- Cost / Risk / Loss functions
- Features
- Overfitting
- Approximation vs. Estimation trade off
- Bias vs. Variance trade off
- Cross validation
- Regularization
- Bagging
- Boosting
- Gradient Descent



רשת נוירונים מלאכותית כמודל פרמטרי

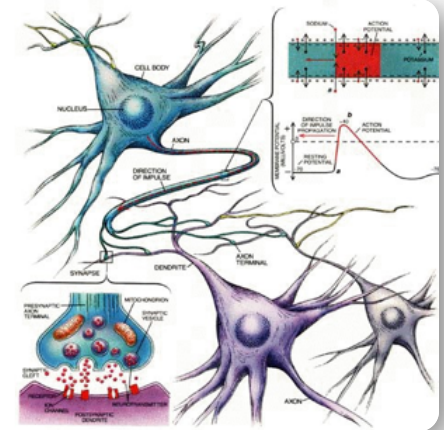
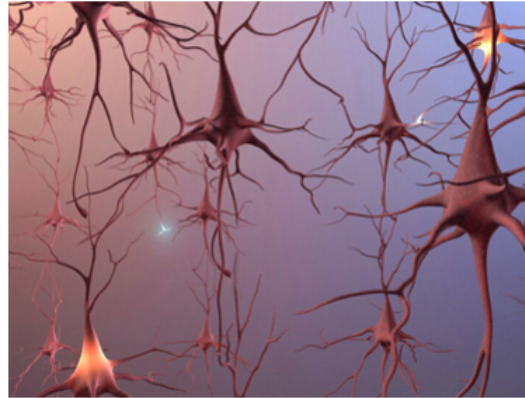
- נתקלנו במספר מקרים בהם ניסינו למצוא פונקציה שתבצע פעולה או תתאר תופעה כלשהי (מציאת חזאי או פונקציית פילוג).
- דרך נוחה לעשות זאת היא בעזרת מודל פרמטרי ומציאת הפרמטרים האופטימאליים.
- עד כה עבדנו עם מודלים לינאריים בפרמטרים.
- ניתן לקרב הרבה מאד פונקציות בעזרת פולינום מסדר מספיק גבוה.
- מודלים אלו הם לא מאד מוצלחים ובעייתיים וכאשר x הוא ממימד גבוה.
- האם ישנם מודלים מתאימים יותר?

רשתות נוירונים מלאכותיות

Artificial Neural Networks - ANN

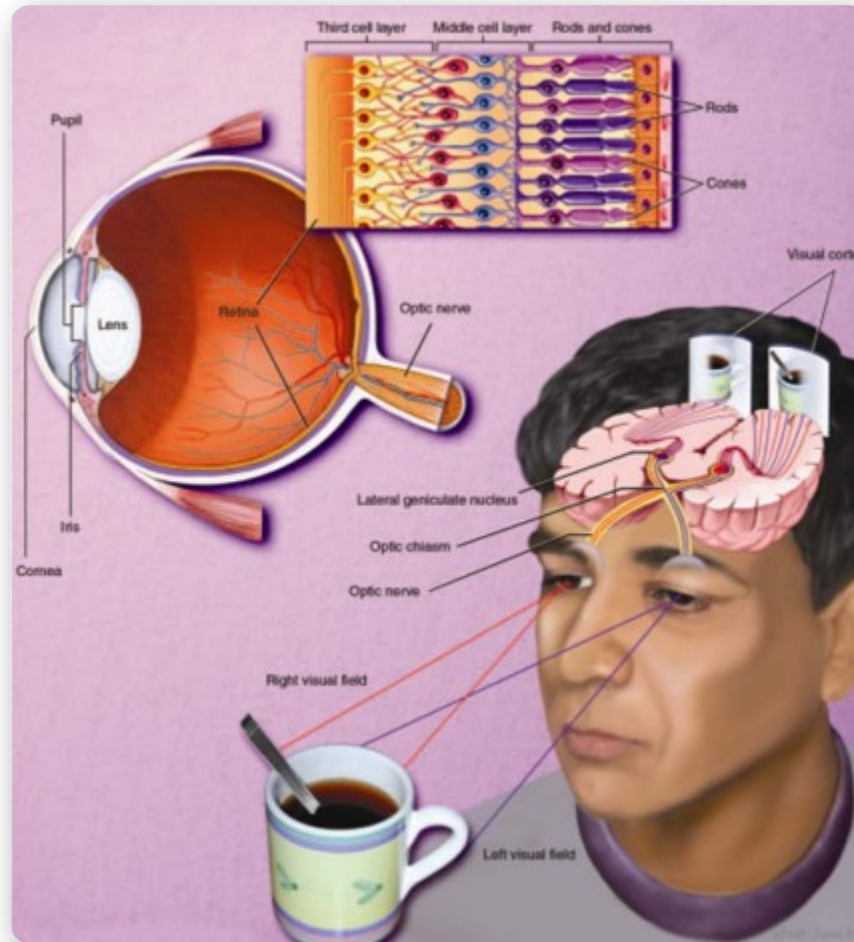
- בשנים האחרונות מודלים אלו הוכיחו את עצמם כמודלים פרמטריים מאד יעילים לפתרון מגוון רחב של בעיות.
- ההשראה לצורה שבה הם בנויים מגיעה מרשתות עצביות ביולוגיות.

רשתות עצביות ביולוגיות

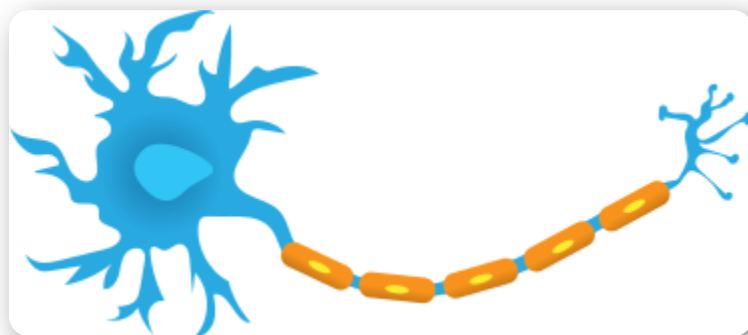


תקשורת בין תאי עצב

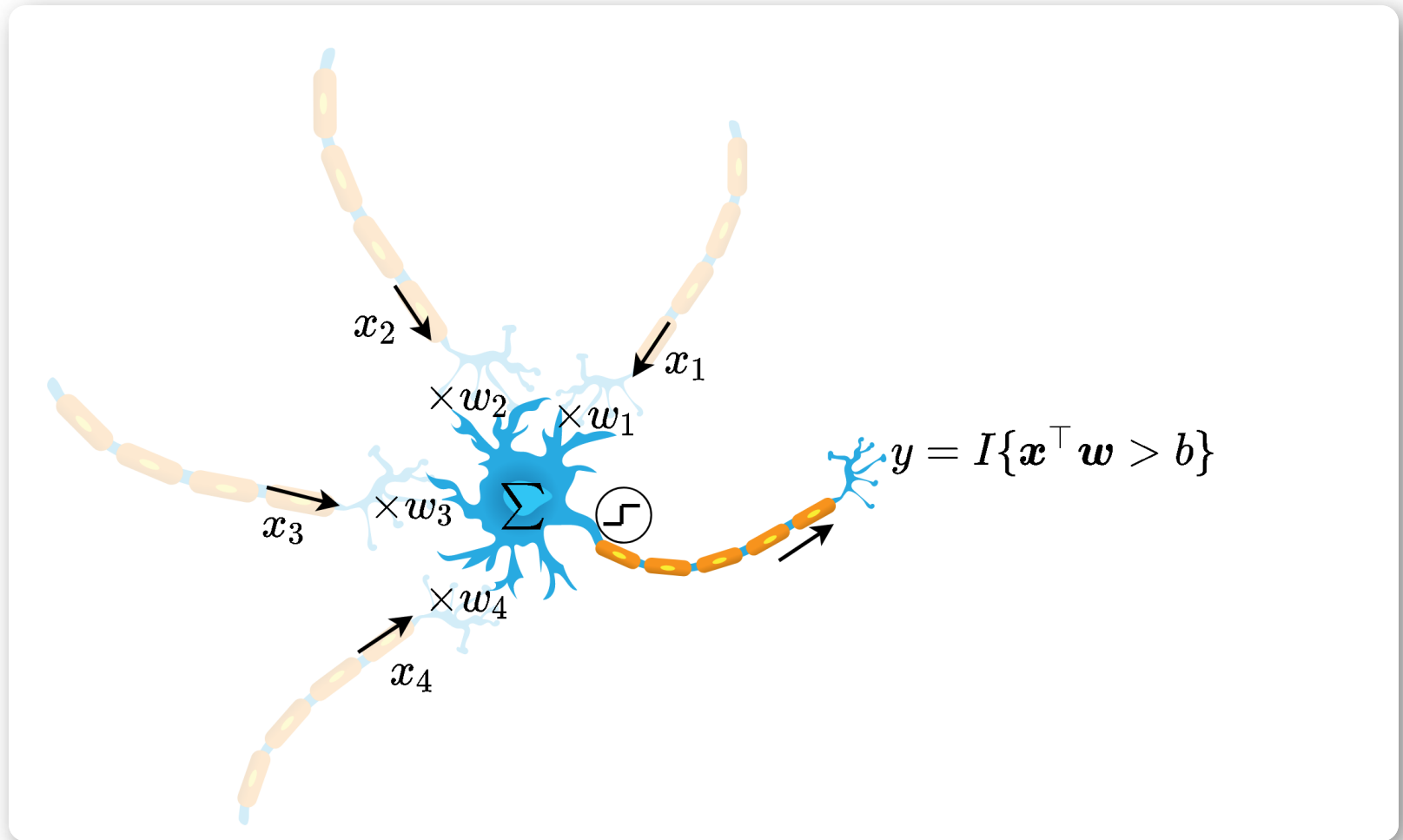
רשתות עצביות ביולוגיות



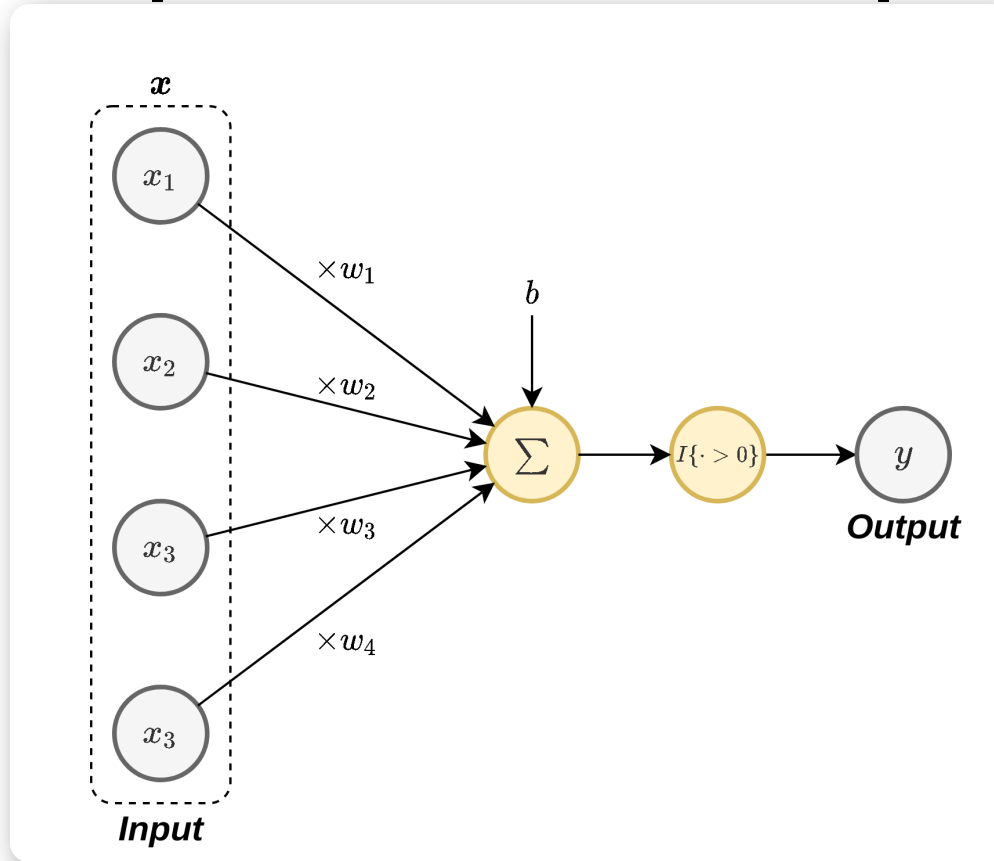
מפוטונים לזיהוי אובייקטים במרחב



בצורה פשטנית ניתן לתאר את האופן בו נירון ביולוגי פועל
כך:



באופן סכימתי ניתן למדל את פעולת הנוירון באופן הבא:



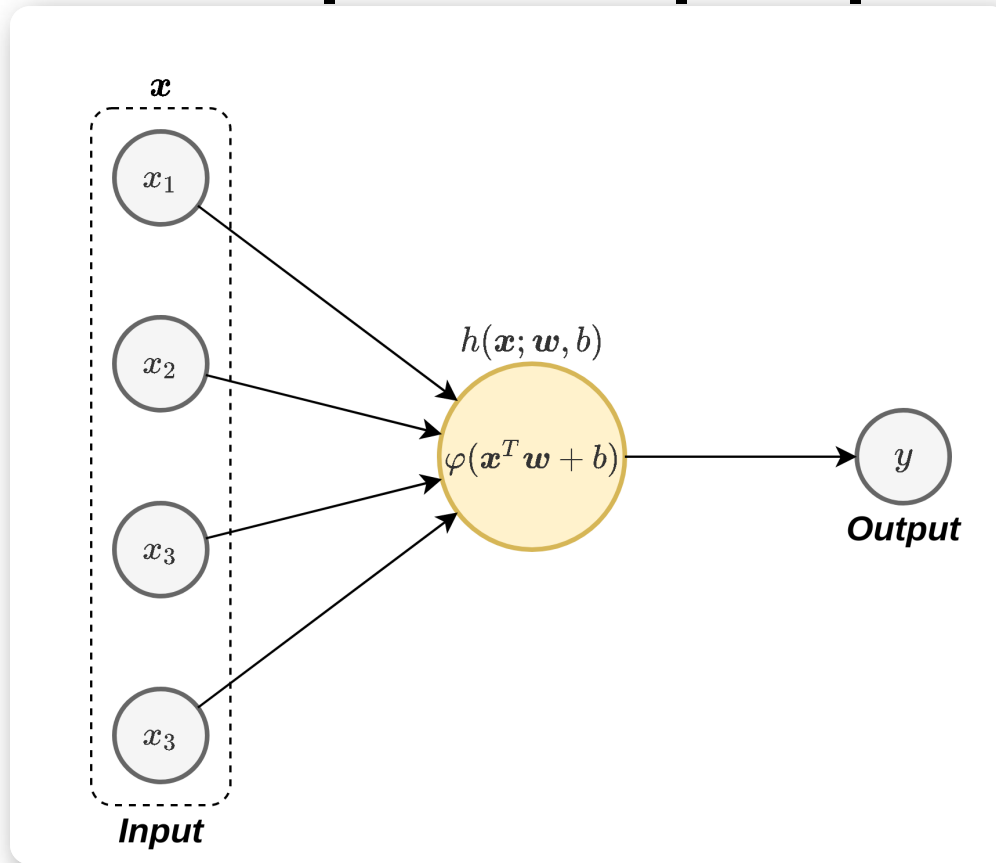
$$y = I\{\mathbf{x}^\top \mathbf{w} + b > 0\}$$

נוירונים ברשת נוירונים מלאכותית

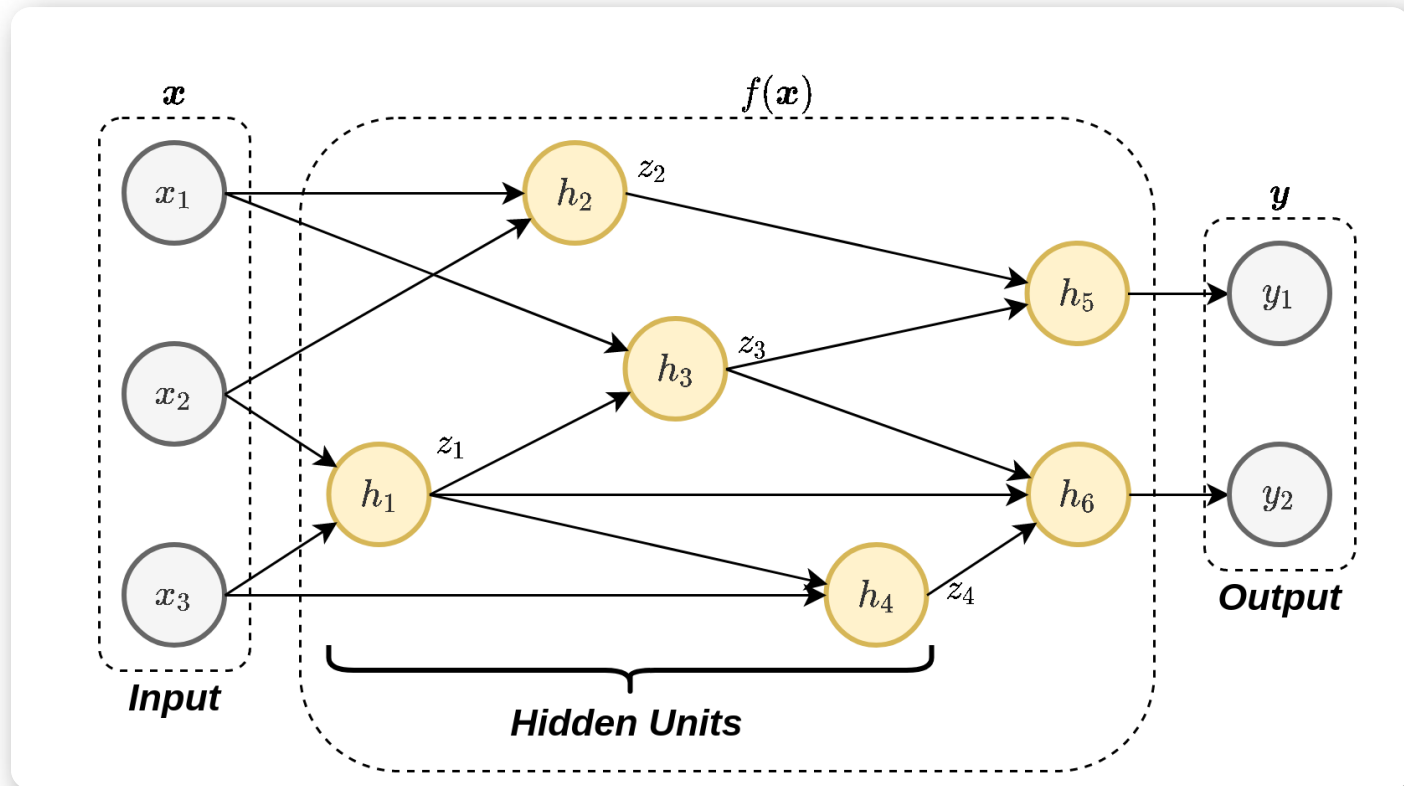
- פונקציית המדרגה היא בפועל מאד בעייתית.
- לשם כך נחליף את פונקציית המדרגה בפונקציה אחרת כל שהיא $\varphi(\cdot)$.
- פונקציה זו מכונה **פונקציית ההפעלה (activation function)**.
- בחירות נפוצות של פונקציית ההפעלה כוללות את:
 - הפונקציה הלוגיסטית (סיגמואיד): $\varphi(x) = \sigma(x) = \frac{1}{1+e^{-x}}$
 - טנגנס היפרבולי: $\varphi(x) = \tanh(x/2)$
 - **ReLU (Rectified Linear Unit)**: $\varphi(x) = \max(x, 0)$.
- פונקציות נוספות שנמצאות הן כל מיני וריאציות של **ReLU**.

נוירונים ברשת נוירונים מלאכותית

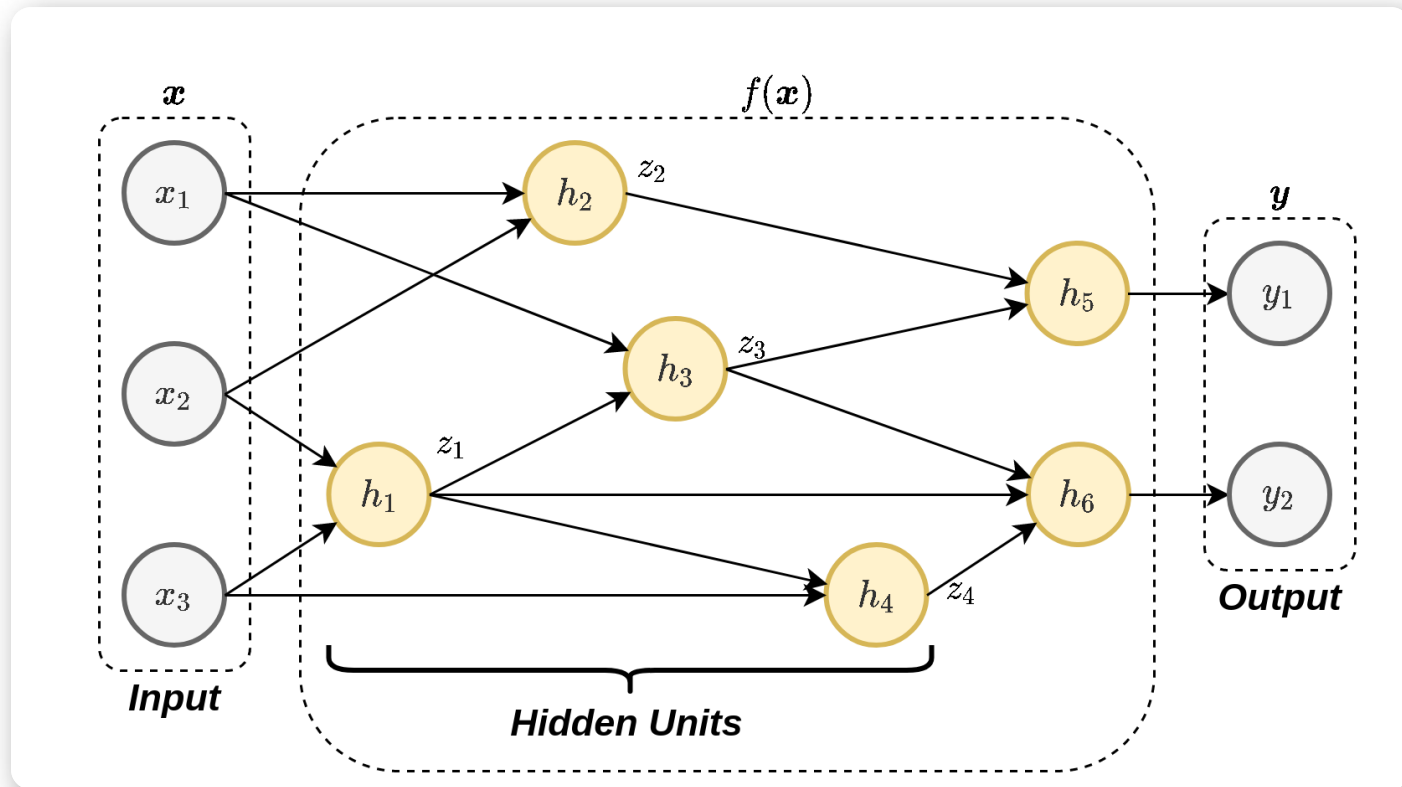
באופן סכימתי נסמן נוירון בודד באופן הבא:



נשלב מספר נוירונים יחד על מנת לבנות רשת נוירונים:



רשת שכזו יכולה לקרב מגוון מאד רחב של פונקציות. הפרמטרים של המודל יהיו הפרמטרים של כל הנוירונים.



לרוב הנוירונים יהיו מהצורה של:

$$h_j(\mathbf{x}; \mathbf{w}_j, b_j) = \varphi(\mathbf{x}^\top \mathbf{w}_j + b_j)$$

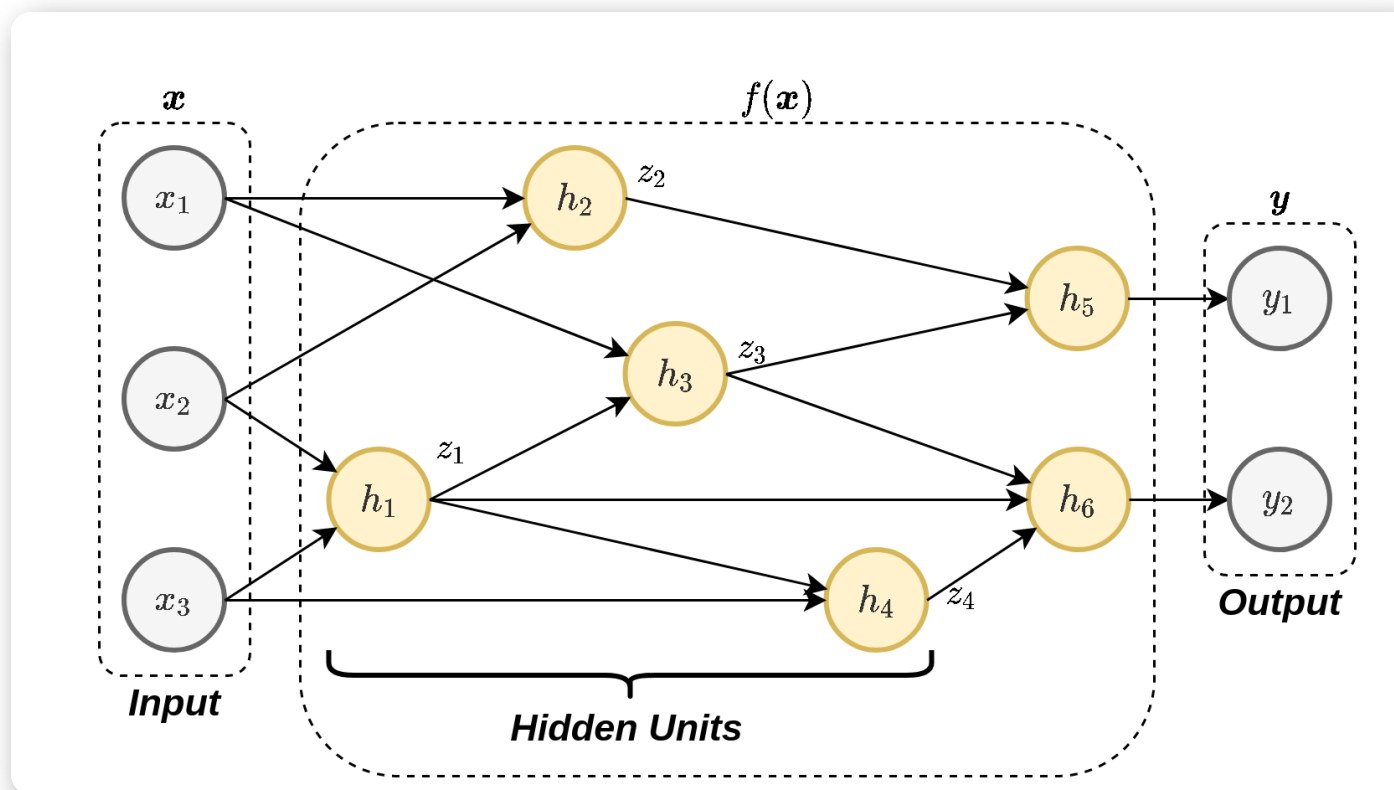
אך ניתן גם לבחור פונקציות אחרות. בקורס זה, אלא אם נאמר אחרת, אנו נניח כי הנוירונים הם מהצורה הזו.

הארכיטקטורה של הרשת

המבנה של הרשת כולל את מספר הנוירונים שהיא מכילה ואת הדרך שבה הם מחוברים אחד לשני.

- **בחירת הארכיטקטורה היא קריטית לביצועים.**
- **לשימושים שונים מתאימות ארכיטקטורות שונות.**
- **חלק גדול מאד מהמחקר שנעשה כיום הוא סביב הנושא של חיפוש ארכיטקטורות.**
- **התהליך של מציאת הארכיטקטורה דורש לא מעט ניסיון, אינטואיציה והרבה ניסוי וטעיה.**
- **לרוב נמצא בעיה דומה ונשתמש בארכיטקטורה שעבדה טוב במקרה זה (נרפרנס).**

הארכיטקטורה של הרשת



- **יחידות נסתרות (hidden units):** הנוירונים אשר אינם מחוברים למוצא הרשת.
- **רשת עמוקה (deep network):** רשת אשר מכילה מסלולים אשר עוברים דרך יותר מיחידה נסתרת אחת.

Feed-forward vs. Recurrent

אנו מבדילים בין שני סוגי ארכיטקטורות:

- **רשת הזנה קדמית (feed-forward network):** ארכיטקטורות אשר אינן מכילות מסלולים מעגליים.
- **רשתות נשנות (recurrent neural network - RNN):** בקורס זה לא נעסוק ברשתות מסוג זה. אלו ארכיטקטורות אשר כן מכילות מסלולים מעגליים.

על החשיבות של פונקציות ההפעלה

- ללא פונקציית ההפעלה הנורונים היו לינאריים ולכן כל הרשת תהיה פשוט מודל לינארי.

Regression + ERM

- הרשת תמדל חזאי אשר אמור להוציא סקלר שמקבל ערכים רציפים בתחום לא מוגבל.
- אנו נרצה שהמוצא של הרשת יתנקז לנוירון בודד ללא פונקציית אקטיבציה כדי לקבל ערכים ממשיים ללא הגבלה.

המוצא של הרשת

סיווג בינארי דיסקרימינטיבי הסתברותי

- הרשת תמדל את $p_{y|x}(1|x)$.
- אנו נרצה שהרשת תוציא ערך סקלרי רציף בתחום בין 0 ל-1.
- שהמוצא של הרשת יתנקז לנוירון בודד עם פונקציית הפעלה שמוציאה ערכים בתחום $[0, 1]$ כדוגמת הפונקציה הלוגיסטית.

סיווג לא בינארי דיסקרימינטיבי הסתברותי

- הרשת תמדל את כל ההסתברויות $p_{y|x}(y|x)$.
- נרצה שהרשת תוציא וקטור באורך C שעליו נפעיל את פונקציית ה `softmax`.

מציאת הפרמטרים של המודל

בעיית האופטימיזציה:

- ב ERM אנו ננסה למזער את ה $risk$ האמפירי.
- בגישה הדיסקרימינטיבית ההסתברותית נשתמש ב MLE או MAP. במקרה של רגרסיה לוגיסטית נוכל להשתמש בפונקציה מהרצאה 9
- כדי לפתור את בעיית האופטימיזציה נשתמש ב $gradient$ descent.

מציאת הפרמטרים של המודל

נסמן את מוצא הרשת $f(x; W) \in \mathbb{R}$.
רגרסיה: לדוגמה, פונקציית ההפסד של least squares היא

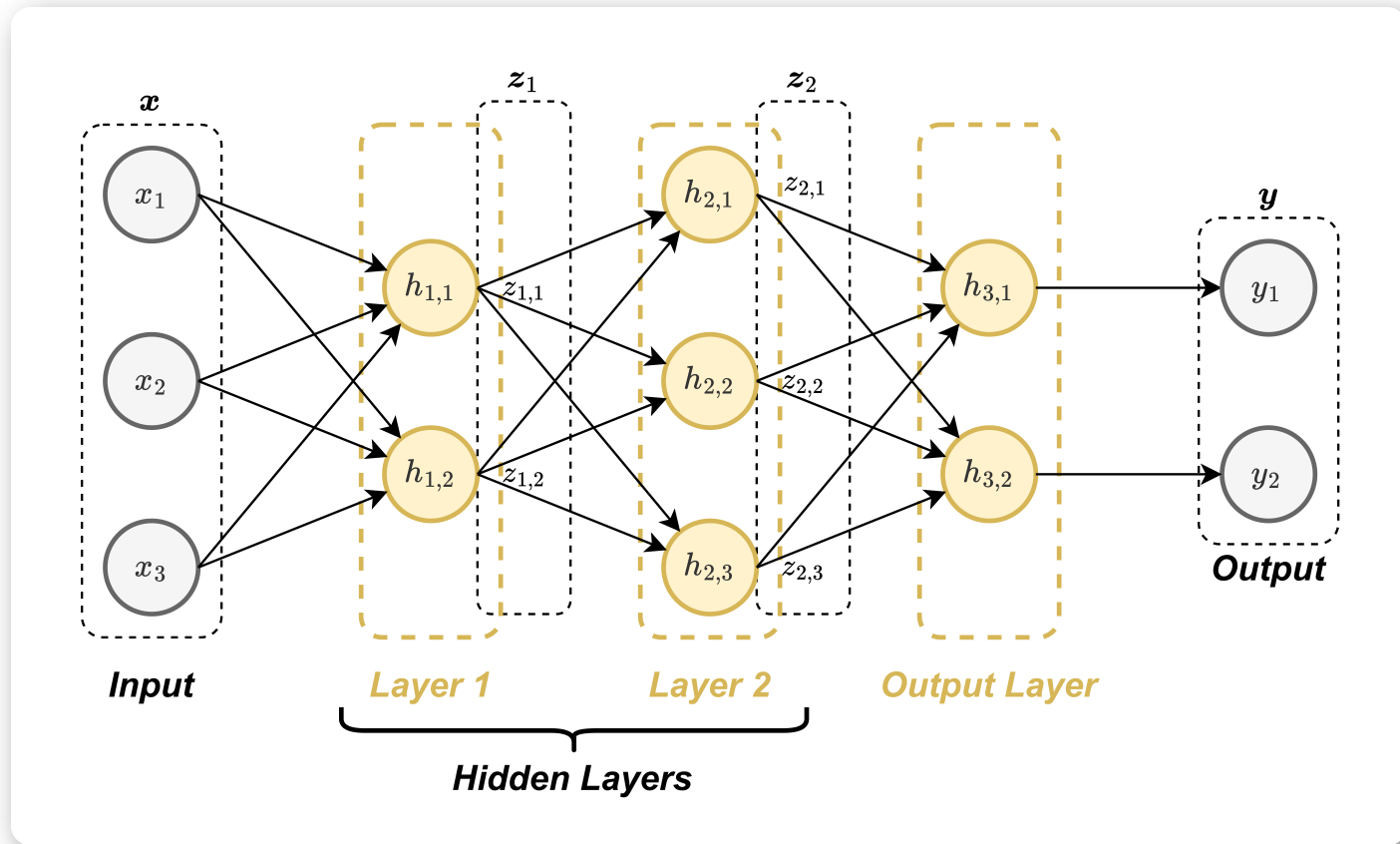
$$\mathcal{L}(W) = \sum_{i=1}^n \left(y^{(i)} - f(x^{(i)}; W) \right)^2$$

סיווג בינארי: במקרה של רגרסיה לוגיסטית ניתן להשתמש בפונקציה מהרצאה 9:

$$\mathcal{L}(W) = - \sum_{i=1}^N \left[y^{(i)} \log \left(\sigma \left(f(x^{(i)}; W) \right) \right) + \left(1 - y^{(i)} \right) \log \left(1 - \sigma \left(f(x^{(i)}; W) \right) \right) \right]$$

עם פונקציית הסיגמואיד $\sigma(z) = 1 / (1 + \exp(-z))$.
במקרה של סיווג רב מחלקתי $(f_1(x; W), \dots, f_c(x; W)) \in \mathbb{R}^c$, ניתן להשתמש בפונקציית softmax ופונקציית ההפסד משקף 28 בהרצאה 9.

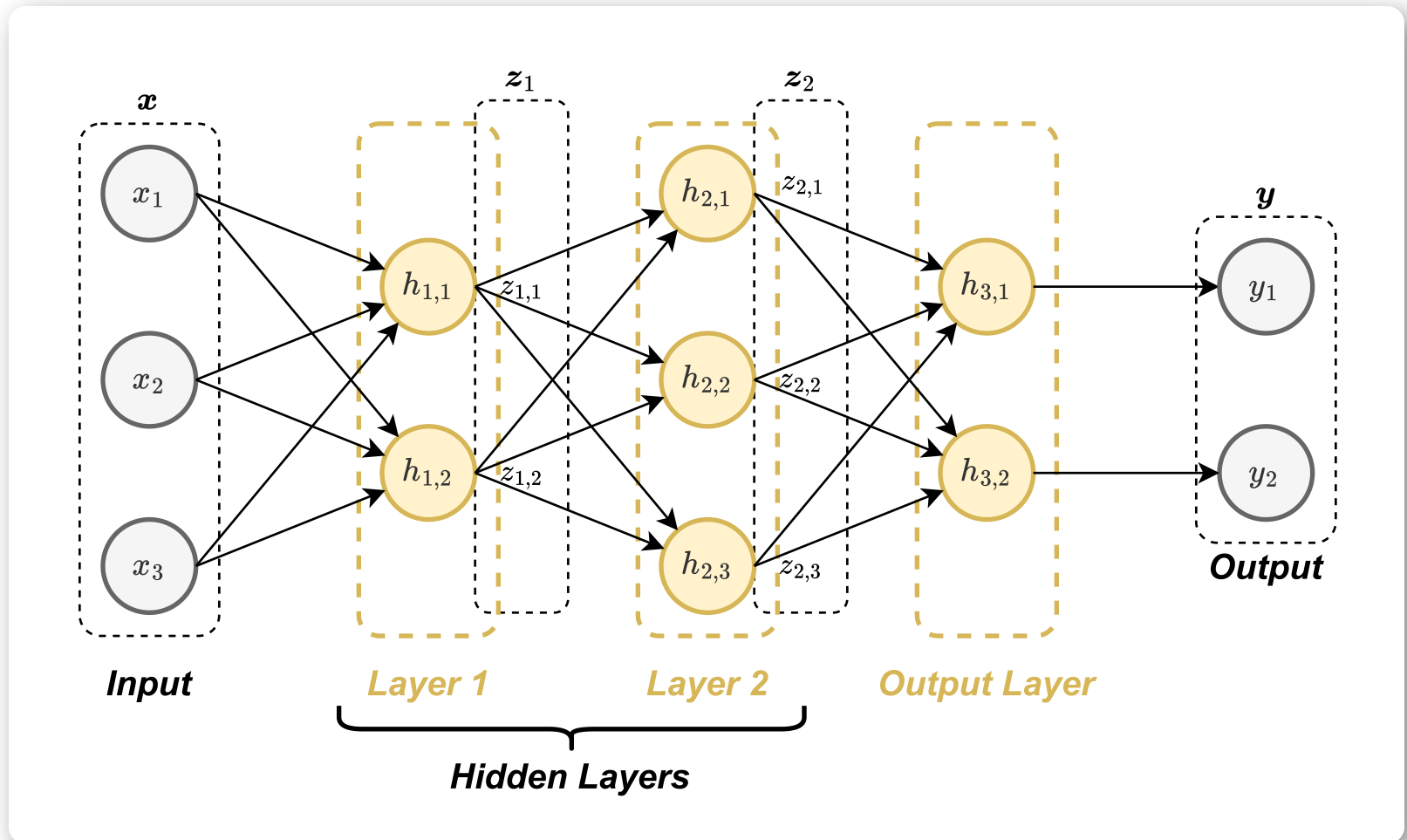
(MultiLayer Perceptron (MLP



- הנירונים מסודרים בשכבות (layers), כשתיים או יותר

- השכבות הן **Fully Connected (FC)** (כל נירון מוזן מכל הנירונים שבשכבה שלפניו).

(MultiLayer Perceptron (MLP



מה שמגדיר את הארכיטקטורה במקרה של MLP הוא מספר השכבות הנסתרות וכמות הנוירונים בכל שכבה (**רוחב השכבה**). בדוגמה הזו, יש ברשת 3 שכבות ברוחב 2, 3 ו 2.

$$W_i = \begin{bmatrix} - & w_{i,1} & - \\ - & w_{i,2} & - \\ & \vdots & \end{bmatrix} \bullet$$

$$b_i = [b_{i,1}, b_{i,2}, \dots]^T \bullet$$

כאשר W_i, b_i הם סט המשקלים וההסטים המתאימים לשכבה i .
בשכבה i -ישנם d_i נוירונים וכן d_{in}, d_{out} הם ממדי המטריצות
בהתאם.

הפונקציה אותה מממשת השכבה כולה הינה:

$$z_i = \varphi(W_i z_{i-1} + b_i)$$

עבור MLP כללי עם L שכבות ניתן לכתוב

$$z_L = \varphi_L (W_L z_{L-1} + b_L) = \varphi_L (W_L \varphi_{L-1} (W_{L-1} z_{L-2} + b_{L-1})) = h_L \circ h_{L-1} \circ$$

כאשר

$$h_\ell (z_{\ell-1}) = \varphi_\ell (W_\ell z_{\ell-1} + b_\ell)$$

שימו לב, φ_ℓ יכולה להיות תלויה בשכבה.
ניתן לכתוב זאת בצורה רקורסיבית

$$\mathbf{z}_0 = \mathbf{x}$$

$$\mathbf{u}_\ell = W_\ell \mathbf{z}_{\ell-1} + \mathbf{b}_\ell \quad \text{for } \ell = 1 \text{ to } L$$

$$\mathbf{z}_\ell = \varphi_\ell(\mathbf{u}_\ell) \quad \text{for } \ell = 1 \text{ to } L$$

כאשר פעולת האקטיבציה φ_ℓ מתבצעת איבר-איבר ו- $\mathbf{y}_L = \mathbf{z}_L$.

"משפט הקירוב האוניברסלי"

בהינתן:

• פונקציית הפעלה רציפה כלשהיא φ שאינה פולינומיאלית.

• ופונקציה רציפה כלשהיא על קוביית היחידה $f : [0, 1]^{D_{\text{in}}} \rightarrow [0, 1]^{D_{\text{out}}}$.

אזי ניתן למצוא פונקציה $f_\varepsilon : [0, 1]^{D_{\text{in}}} \rightarrow [0, 1]^{D_{\text{out}}}$ מהצורה:

$$f_\varepsilon(\mathbf{x}) = W_2 \varphi(W_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2$$

כך ש:

$$\sup_{\mathbf{x} \in [0, 1]^{D_{\text{in}}}} \|f(\mathbf{x}) - f_\varepsilon(\mathbf{x})\| < \varepsilon$$

הערה לגבי נגזרות וקטוריות

זכרו כי עבור פונקציה סקלרית $f(\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^n$

$$\nabla f(\boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left[\frac{\partial f(\boldsymbol{\theta})}{\partial \theta_1}, \dots, \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_n} \right] \in \mathbb{R}^{1 \times n}$$

תהי $g(\boldsymbol{\theta})$ פונקציה וקטורית של וקטור $\boldsymbol{\theta}, g : \boldsymbol{\theta} \mapsto \mathbb{R}^m, g(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \dots, g_m(\boldsymbol{\theta}))$
אזי

$$\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left[\frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_j} \right]_{ij} \in \mathbb{R}^{m \times n}$$

ובמקרה הפשוט בו $g(\boldsymbol{\theta}) = (g_1(\theta_1), \dots, g_m(\theta_m))$ מתקיים כי

$$\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \text{diag}(g'_1(\theta_1), \dots, g'_m(\theta_m)) = \text{diag}(\mathbf{g}'(\boldsymbol{\theta}))$$

Back-Propagation

כדי להשתמש בשיטות גרדיאנט נרצה לחשב נגזרות לפי פרמטרי הרשת. באופן כללי אנו צריכים לחשב את הנגזרות של פונקציית ההפסד ביחס לכל פרמטרי הרשת (משקולות ואיברי הטיה), כלומר

$$\frac{\partial \mathcal{L}(W)}{\partial W_\ell}$$

כאשר W_ℓ הם המשקולות של השכבה ה- ℓ . שימו לב כי

$$\frac{\partial \mathcal{L}(W)}{\partial W_\ell} = \frac{\partial \mathcal{L}(W)}{\partial z_\ell} \frac{\partial z_\ell}{\partial W_\ell} = \frac{\partial \mathcal{L}(W)}{\partial z_\ell} \frac{\partial z_\ell}{\partial u_\ell} \frac{\partial u_\ell}{\partial W_\ell}$$

$$\frac{\partial u_\ell}{\partial W_\ell} = \mathbf{z}_{\ell-1}$$

$$\frac{\partial \mathbf{z}_\ell}{\partial u_\ell} = \text{diag}(\varphi'_\ell(\mathbf{u}_\ell))$$

כאשר הנגזרת המתגרת היחידה לחישוב היא הראשונה. 28

Back-Propagation

שיטה המקלה על חישוב הנגזרות על ידי שימוש בכלל השרשרת.

כלל השרשרת - תזכורת

במקרה הסקלרי:

$$(f(g(x)))' = f'(g(x)) \cdot g'(x)$$

במקרה של מספר משתנים:

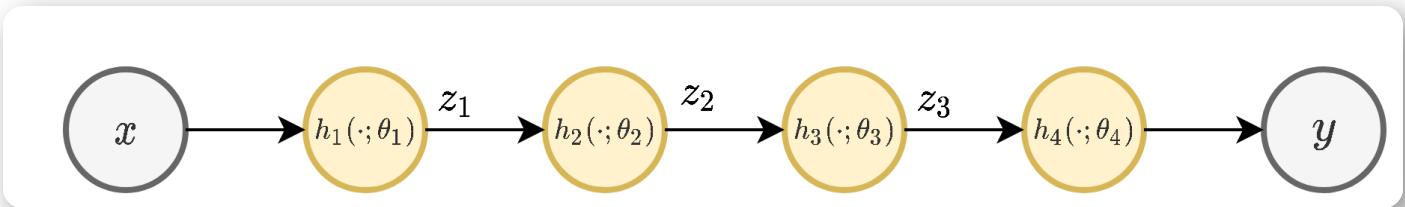
$$\begin{aligned} \frac{d}{dx} f(z_1(x), z_2(x), z_3(x)) = & \left(\frac{\partial}{\partial z_1} f(z_1(x), z_2(x), z_3(x)) \right) \frac{d}{dx} z_1(x) \\ & + \left(\frac{\partial}{\partial z_2} f(z_1(x), z_2(x), z_3(x)) \right) \frac{d}{dx} z_2(x) \\ & + \left(\frac{\partial}{\partial z_3} f(z_1(x), z_2(x), z_3(x)) \right) \frac{d}{dx} z_3(x) \end{aligned}$$

Back-Propagation

לאלגוריתם 2 שלבים:

- **Forward pass**: העברה של הדגימות דרך הרשת ושמירה של כל ערכי הביניים.
- **Backward pass**: חישוב של הנגזרות של הנוירונים מהמוצא של הרשת לכיוון הכניסה.

Back-Propagation: דוגמא פשוטה

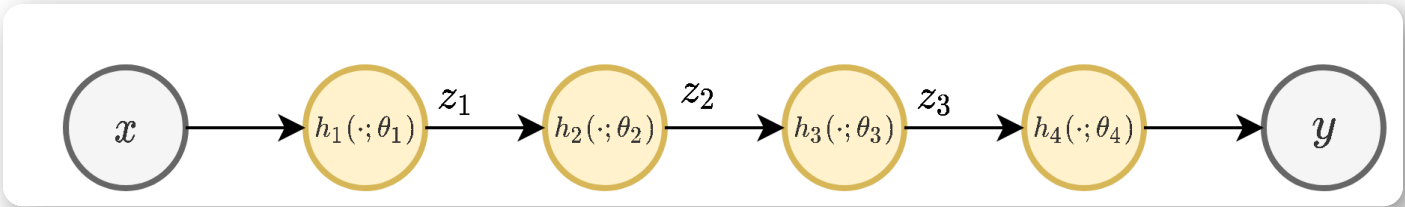


נרצה לחשב את $\partial \mathcal{L} / \partial \theta_i$ עבור פרמטר θ_i כלשהו. למשל, עבור פונקציית ההפסד הריבועית, כאשר $L = (y - t)^2$, הוא הערך האמיתי

$$\frac{\partial L}{\partial \theta_i} = 2(y - t) \frac{\partial y}{\partial \theta_i}$$

ובאופן דומה עבור שאר פונקציות ההפסד. כך, עלינו להתמקד בנגזרת של המוצא ביחס לפרמטר.

Back-Propagation: דוגמא פשוטה



נרשום את הנגזרת של y לפי θ_2 :

$$\frac{\partial y}{\partial \theta_2} = \frac{\partial y}{\partial z_2} \frac{\partial z_2}{\partial \theta_2} = \frac{\partial y}{\partial z_2} \frac{\partial}{\partial \theta_2} h_2(z_1; \theta_2)$$

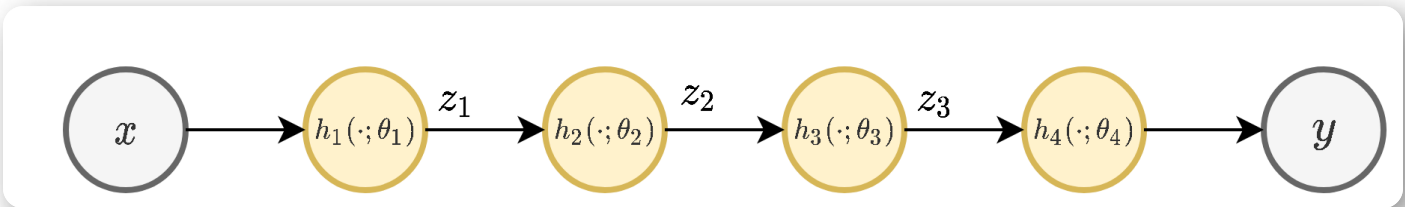
נוכל לפרק גם את הנגזרת של $\frac{dy}{dz_2}$:

$$\frac{\partial y}{\partial z_2} = \frac{\partial y}{\partial z_3} \frac{\partial z_3}{\partial z_2} = \frac{\partial}{\partial z_3} h_4(z_3; \theta_4) \frac{\partial}{\partial z_2} h_3(z_2; \theta_3)$$

לכן:

$$\frac{\partial y}{\partial \theta_2} = \frac{\partial y}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial \theta_2} = \frac{\partial}{\partial z_3} h_4(z_3; \theta_4) \frac{\partial}{\partial z_2} h_3(z_2; \theta_3) \frac{\partial}{\partial \theta_2} h_2(z_1; \theta_2)$$

Back-Propagation: דוגמא פשוטה



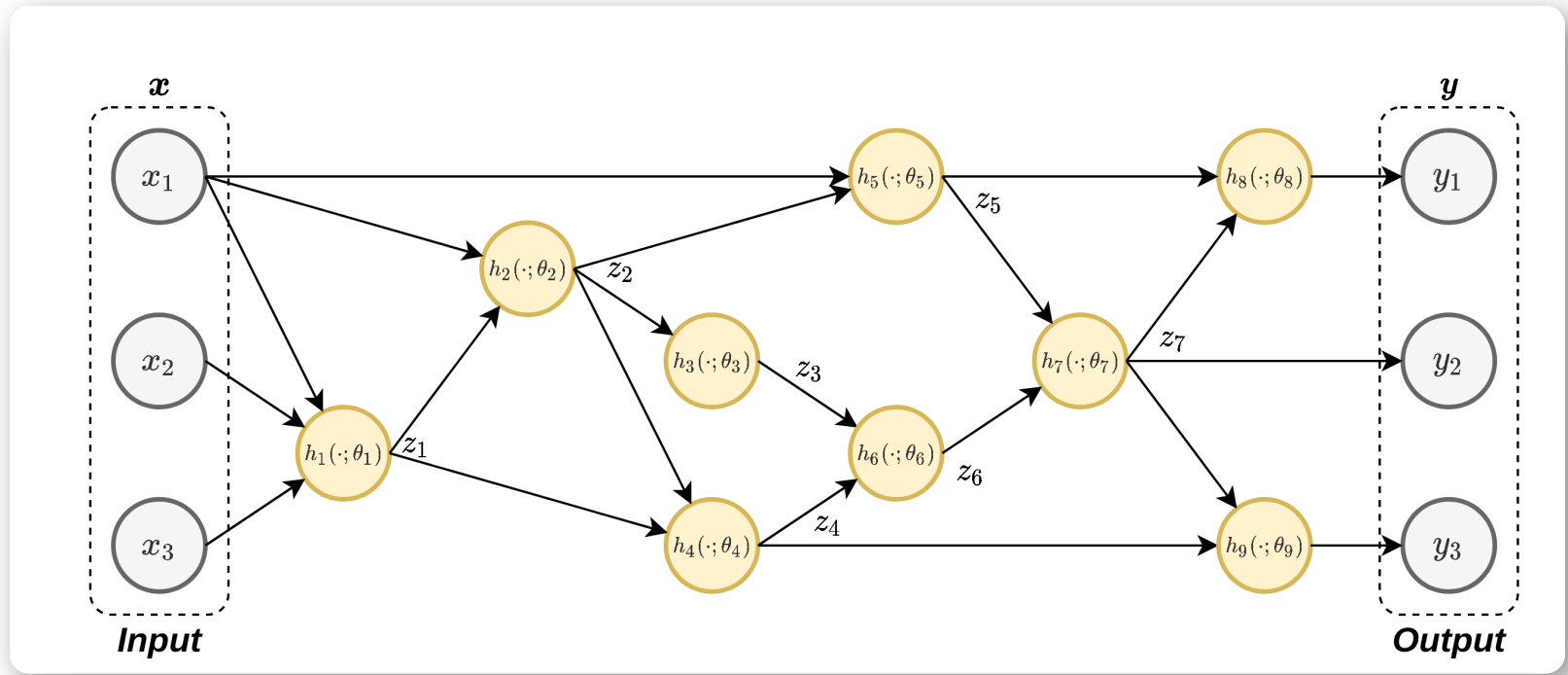
$$\frac{\partial y}{\partial \theta_2} = \frac{\partial y}{\partial z_3} \frac{\partial z_3}{\partial z_2} \frac{\partial z_2}{\partial \theta_2} = \frac{\partial}{\partial z_3} h_4(z_3; \theta_4) \frac{\partial}{\partial z_2} h_3(z_2; \theta_3) \frac{\partial}{\partial \theta_2} h_2(z_1; \theta_2)$$

כדי לחשב את הביטוי שקיבלנו עלינו לבצע את שני השלבים הבאים:

• לחשב את כל ה z_i לאורך הרשת (forward pass).

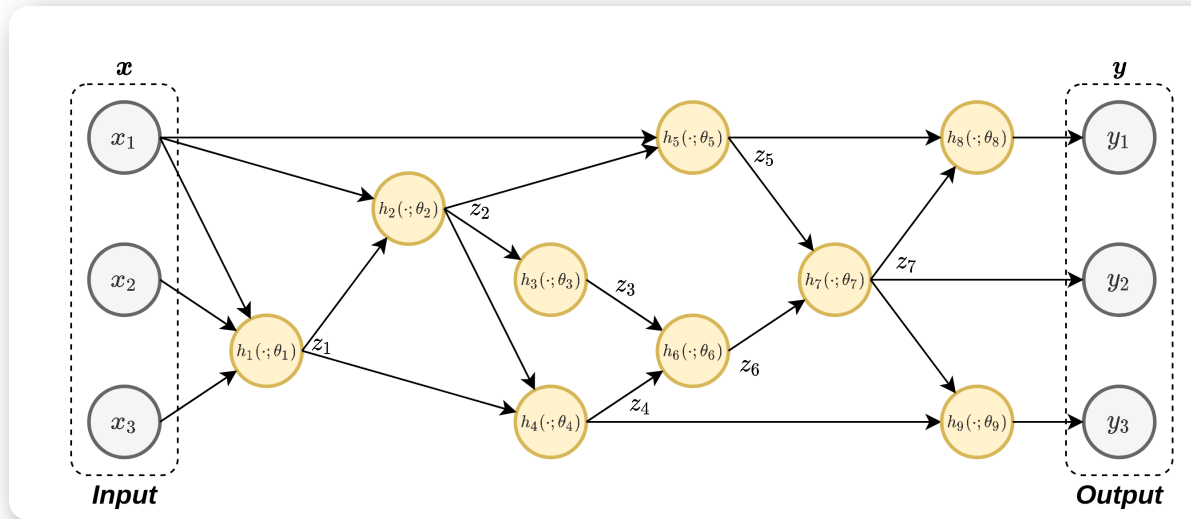
• לחשב את כל הנגזרות מהמוצא של הרשת ועד לנקודה בה נמצא הפרמטר שלפיו רוצים לגזור (backward-pass).

Back-Propagation: דוגמא מעט יותר מורכבת



נחשב את הנגזרת של y_1 לפי θ_3 .

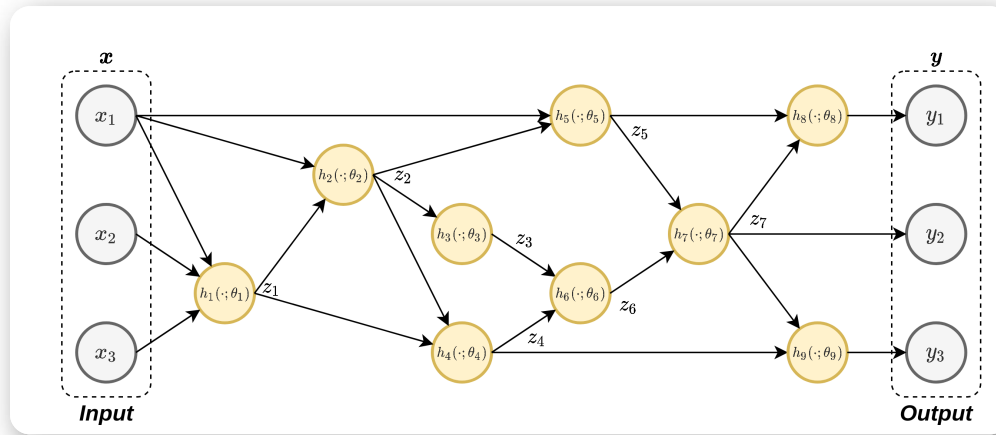
מורכבת Back-Propagation: דוגמא מעט יותר



נפרק את הנגזרת של $\frac{\partial y_1}{\partial \theta_3}$ בדומה למה שחישבנו קודם:

$$\begin{aligned} \frac{\partial y_1}{\partial \theta_3} &= \frac{\partial y_1}{\partial z_7} \frac{\partial z_7}{\partial z_6} \frac{\partial z_6}{\partial z_3} \frac{\partial z_3}{\partial \theta_3} \\ &= \frac{\partial}{\partial z_7} h_8(z_7; \theta_8) \frac{\partial}{\partial z_6} h_7(z_6; \theta_7) \frac{\partial}{\partial z_3} h_6(z_5; \theta_6) \frac{\partial}{\partial \theta_3} h_3(z_2; \theta_3) \end{aligned}$$

Back-Propagation: דוגמא מעט יותר מורכבת

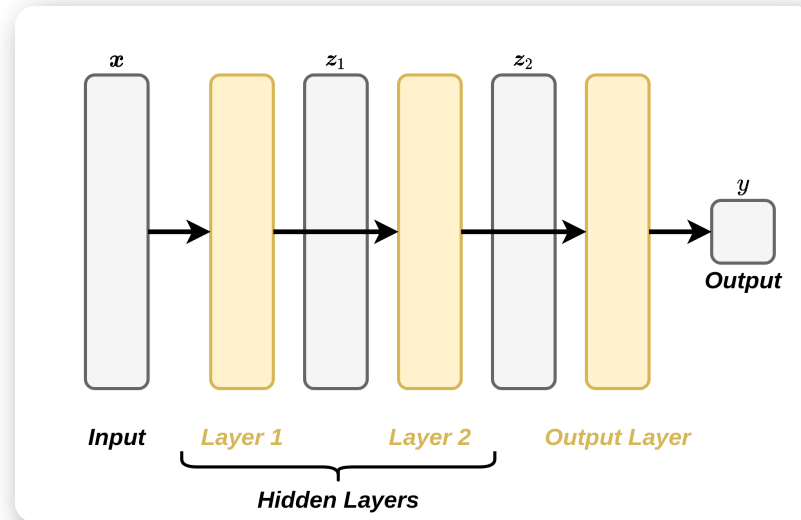


$$\begin{aligned}\frac{\partial y_1}{\partial \theta_3} &= \frac{\partial y_1}{\partial z_7} \frac{\partial z_7}{\partial z_6} \frac{\partial z_6}{\partial z_3} \frac{\partial z_3}{\partial \theta_3} \\ &= \frac{\partial}{\partial z_7} h_8(z_7; \theta_8) \frac{\partial}{\partial z_6} h_7(z_6; \theta_7) \frac{\partial}{\partial z_3} h_6(z_5; \theta_6) \frac{\partial}{\partial \theta_3} h_3(z_2; \theta_3)\end{aligned}$$

• נריץ את ה forward-pass בשביל לחשב את ערכי ה z_i .

• נריץ את ה backward-pass בו נחשב את הנגזרות מהמוצא של הרשת עד לנגזרת של h_3 .

Back-Propagation - MLP



משוואות ה forward-pass:

$$z_1 = \varphi(\underbrace{W_1 x + b_1}_{u_1}), \quad W_1 \sim d_1 \times d_{in}$$

$$z_2 = \varphi(\underbrace{W_2 z_1 + b_2}_{u_2}), \quad W_2 \sim d_2 \times d_1$$

$$y = w_3^T z_2 + b_3, \quad w_3 \sim d_2 \times 1$$

Back-Propagation and MLPs

Forward pass

$$\mathbf{z}_0 = \mathbf{x}$$

$$\mathbf{u}_l = W_l \mathbf{z}_{l-1} + \mathbf{b}_l \quad \text{for } l = 1 \text{ to } L$$

$$\mathbf{z}_l = \varphi_l(\mathbf{u}_l) \quad \text{for } l = 1 \text{ to } L$$

Backward pass

$$\delta_L = \text{diag}(\varphi'_L(u_L)) \frac{\partial \mathcal{L}}{\partial z_L}$$

$$\delta_l = \text{diag}(\varphi'_l(u_l)) W_{l+1}^\top \delta_{l+1} \quad \text{for } l = L - 1 \text{ to } 1$$

Back-Propagation and MLPs

חישוב הגרדיאנט יתבצע באמצעות

$$\nabla_{W_\ell} \mathcal{L} = \delta_\ell z_{\ell-1}^\top \quad ; \quad \nabla_{b_\ell} \mathcal{L} = \delta_\ell$$

ואלגוריתם הגרדיאנט יהיה

$$W_\ell^{(t+1)} = W_\ell^{(t)} - \eta \delta_\ell^{(t)} z_{\ell-1}^{(t)\top} \quad ; \quad b_\ell^{(t+1)} = b_\ell^{(t)} - \eta \delta_\ell^{(t)}$$