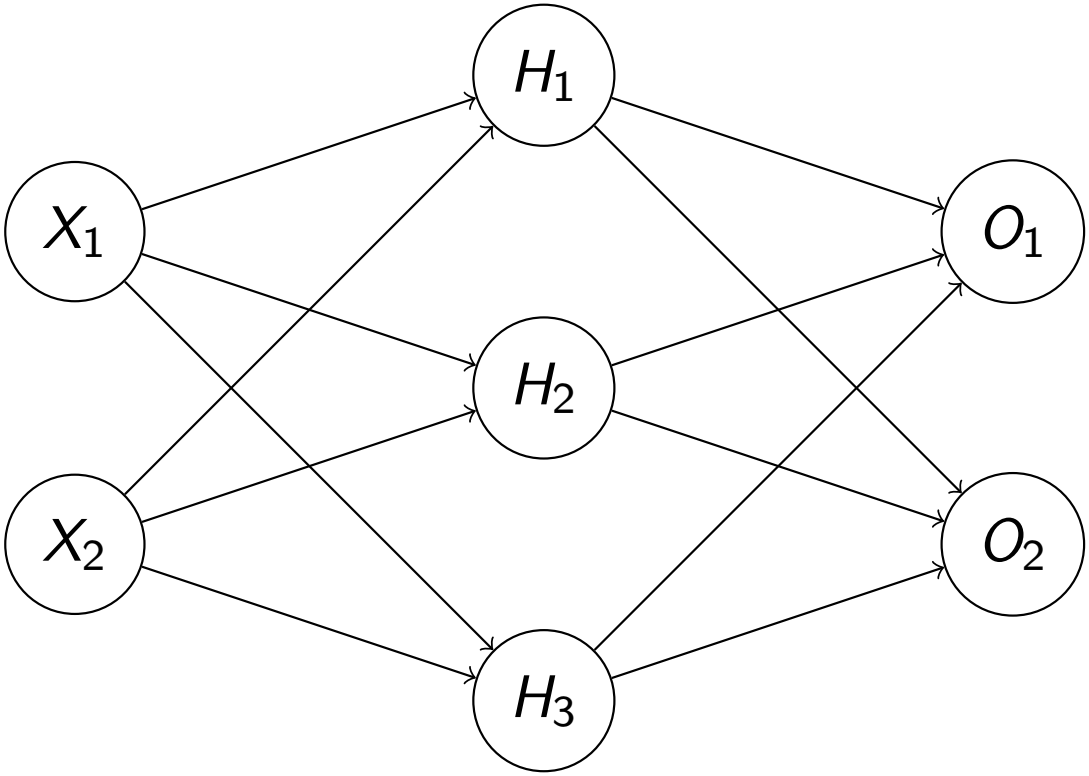
Wskazówki do ćwiczenia 5.

WSI - ćwiczenia 2022Z

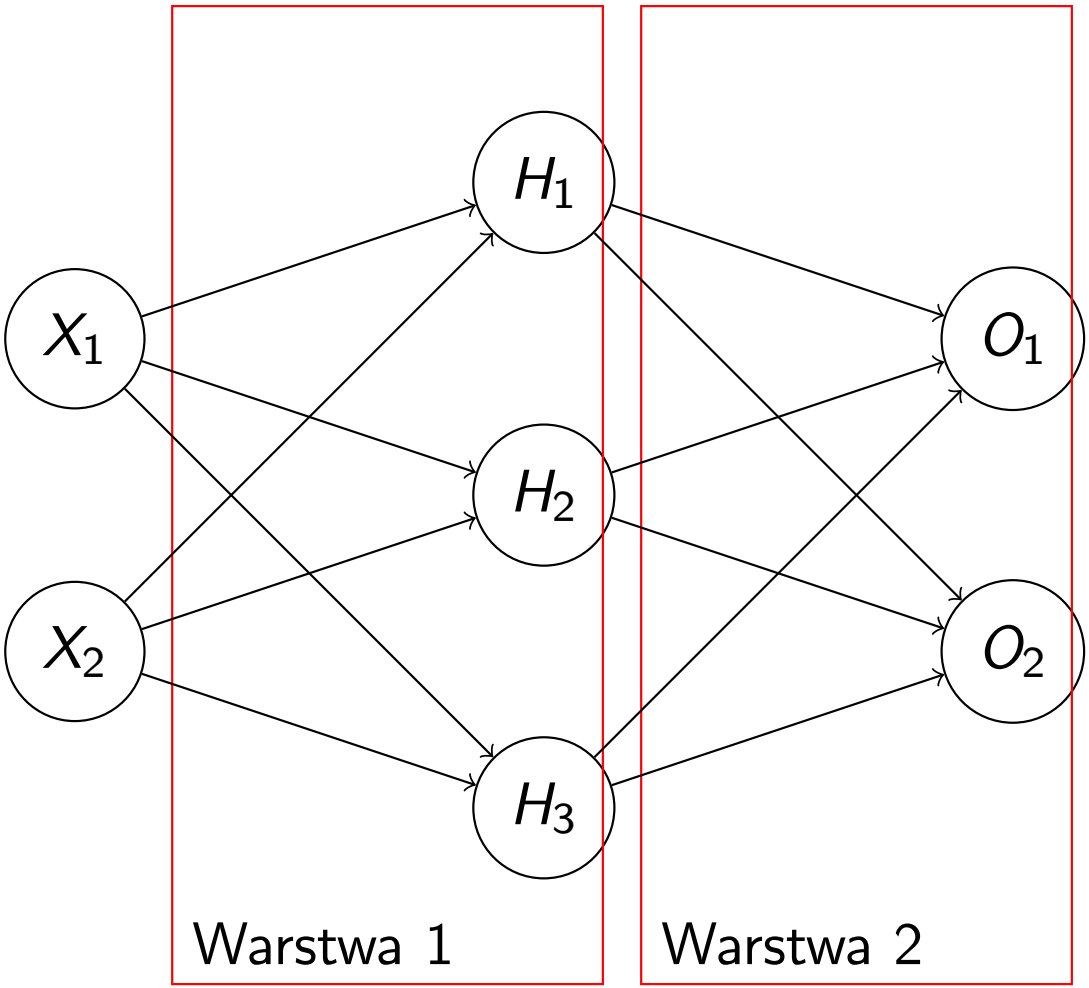
Jakub Łyskawa, Stanisław Pawlak

December 8, 2022

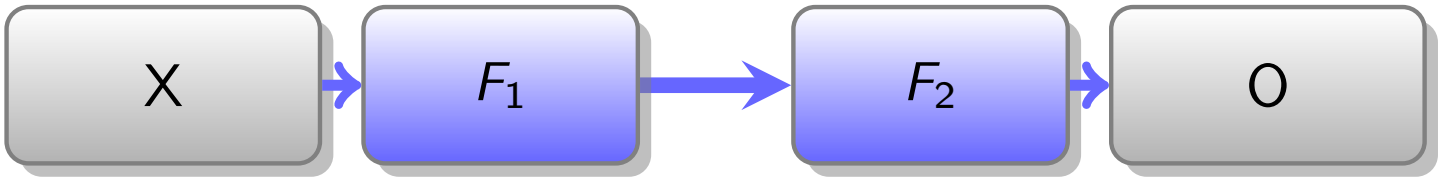
Perceptron wielowarstwowy



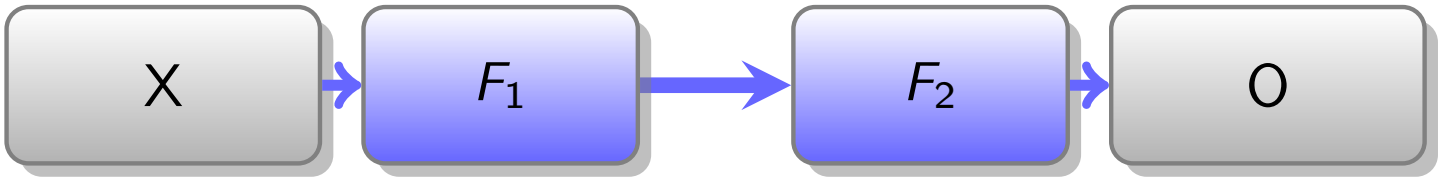
Perceptron wielowarstwowy



Perceptron wielowarstwowy

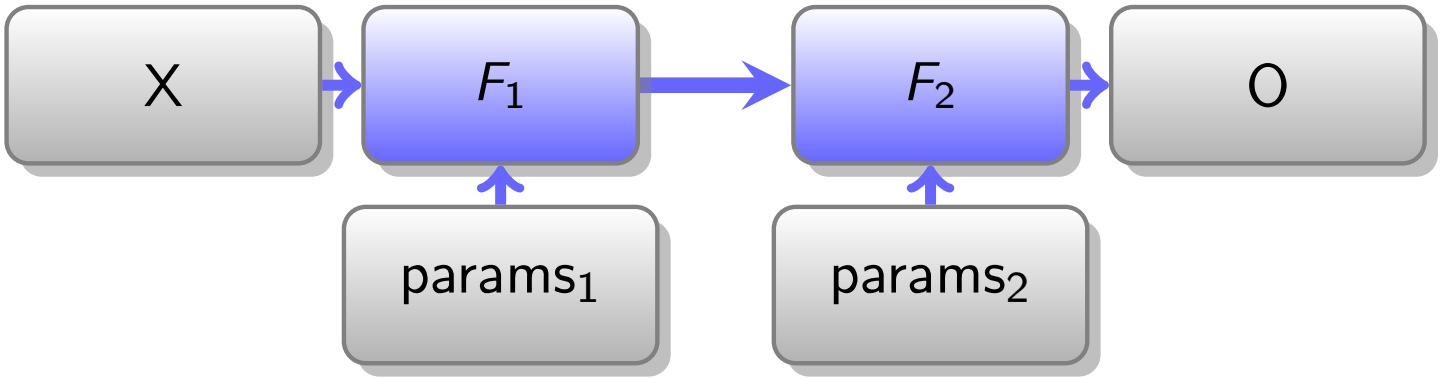


Perceptron wielowarstwowy



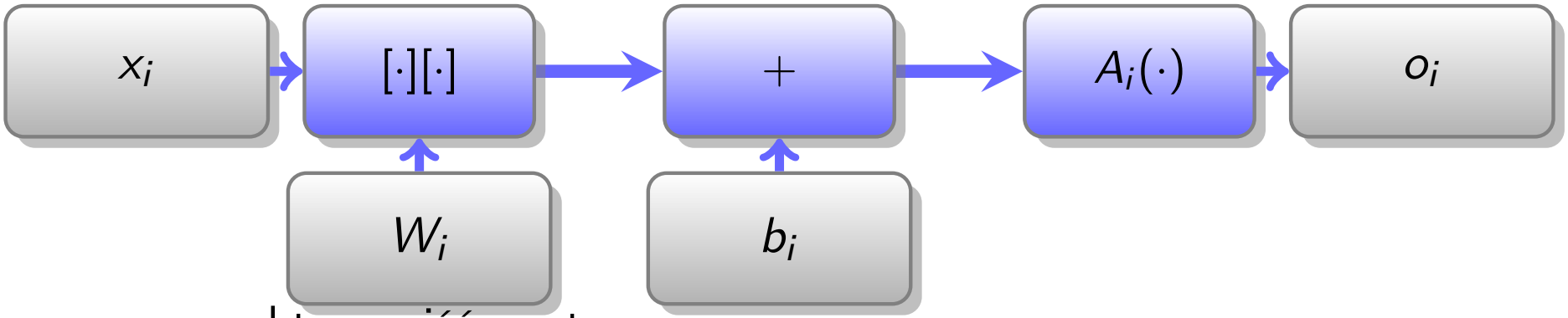
$$O = F_2(F_1(X))$$

Perceptron wielowarstwowy



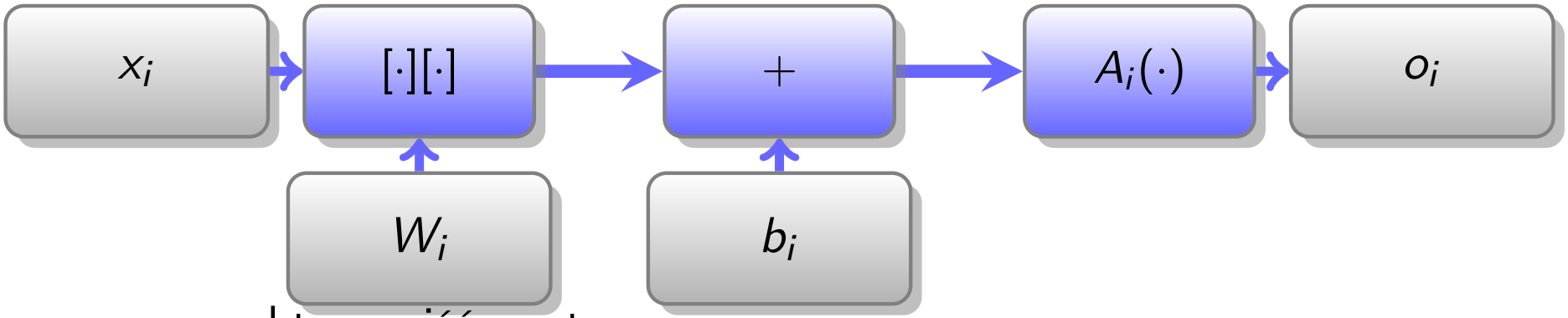
$$O = F_2(F_1(X, \text{params}_1), \text{params}_2)$$

Warstwa



- x_i - wektor wejść wartswy
- W_i - macierz wag
- b_i - wektor progu („bias”)
- A_i - funkcja aktywacji (np. ReLU, sigmoid, liniowa)
- o_i - wyjście warstwy

Warstwa



- x_i - wektor wejść wartswy
- W_i - macierz wag
- b_i - wektor progu („bias”)
- A_i - funkcja aktywacji (np. ReLU, sigmoid, liniowa)
- o_i - wyjście warstwy

$$o_i = A_i(W_i x_i + b_i)$$

Błąd sieci

$O_i = F_{...}(F_2(F_1(X_i)))$
$\frac{1}{N}\sum_{i=1}^{N} L(O_i, Y_i)$

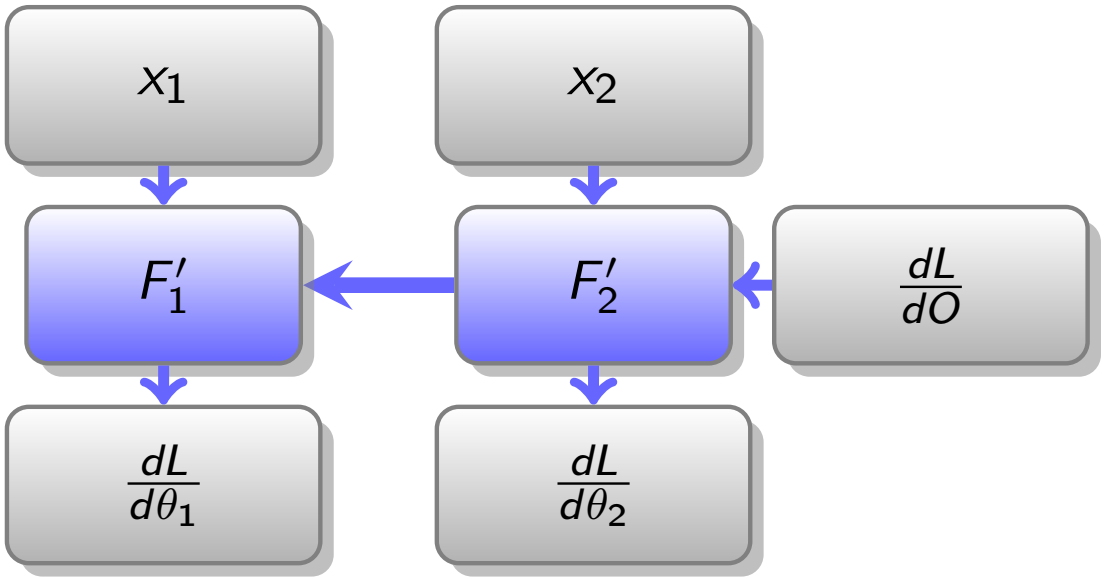
Na przykład strata średniokwadratowa dla
$L(O, Y) = \frac{1}{2}|O - Y|^2$

Uczenie sieci

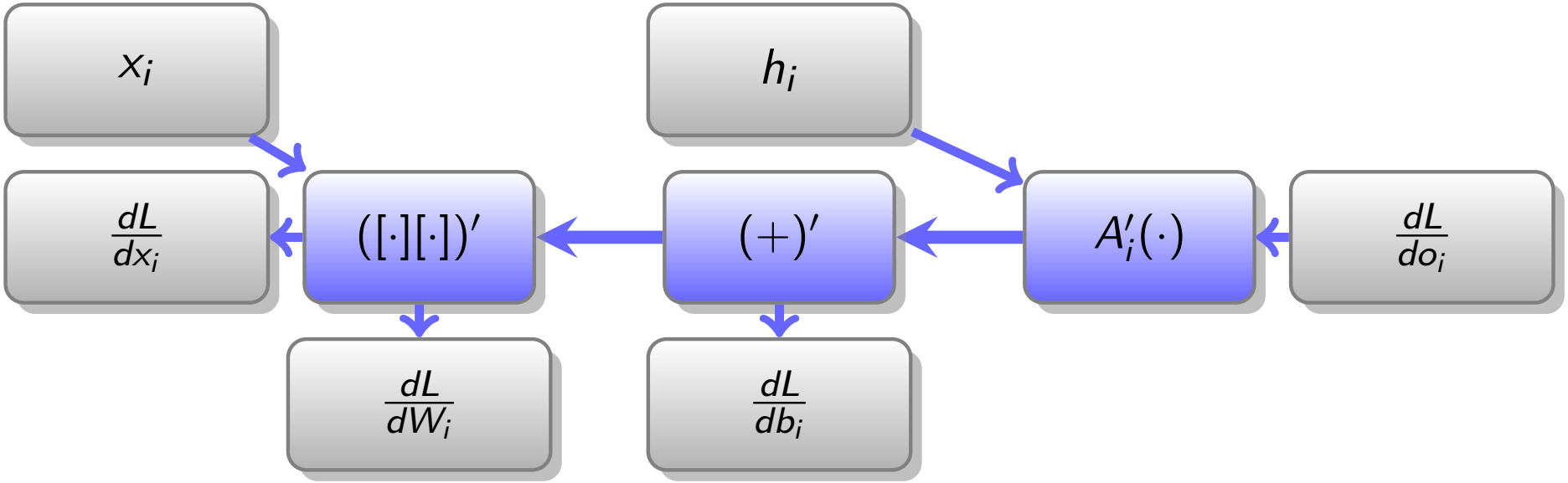
Minimalizacja funkcji straty metodami gradientowymi (np. najszybszego spadku)

Gradient wyjścia O_i (dla wejścia X_i): $\frac{dL}{dO_i} = O_i - Y_i$

Propagacja wsteczna

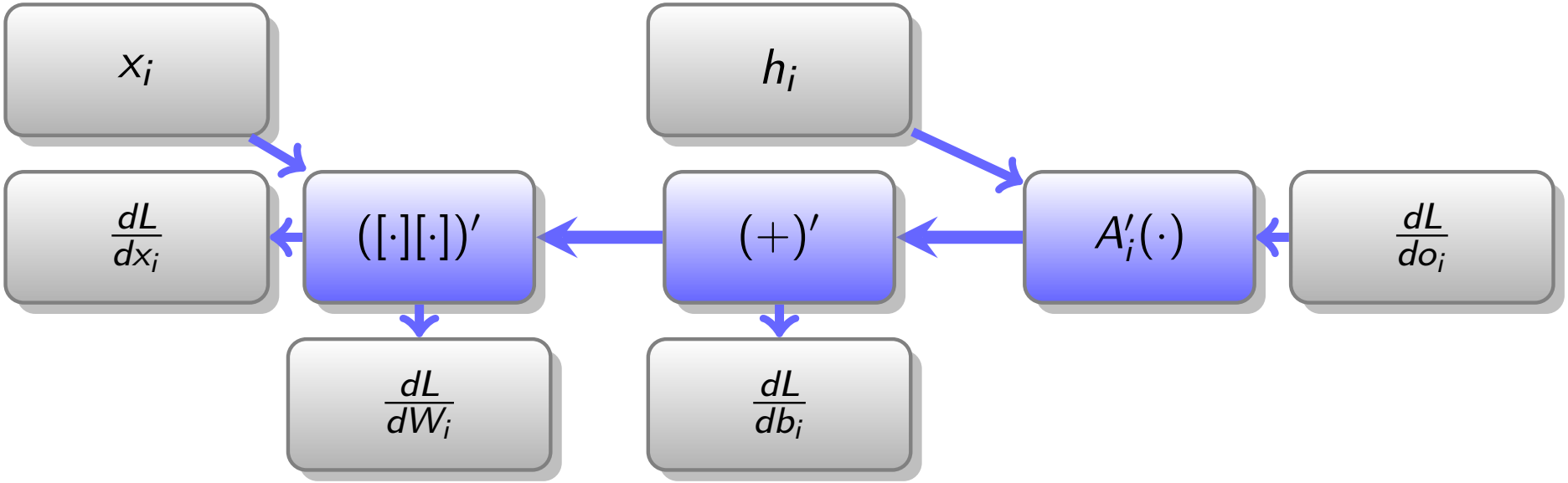


Propagacja wsteczna - warstwa



h_i - Wyjście warstwy przed aktywacją

Propagacja wsteczna - warstwa



h_i - Wyjście warstwy przed aktywacją

$$\frac{dL}{dh_i} = \frac{dL}{do_i} \circ A_i'(h_i) \qquad \frac{dL}{db_i} = \frac{dL}{dh_i} \qquad \frac{dL}{dx_i} = W_i^T \frac{dL}{dh_i} \qquad \frac{dL}{dW_i} = \frac{dL}{dh_i} x_i^T$$

$\circ$ - wektor iloczynów
(dla f. aktywacji działających niezależnie na kolejnych elementach)

Końcowe uwagi

- Może być więcej warstw niż 2

Końcowe uwagi

- Może być więcej warstw niż 2
- Uczy się zazwyczaj na "paczkach" N próbek - wtedy wartości f. straty i gradienty uśrednia się

Końcowe uwagi

- Może być więcej warstw niż 2
- Uczy się zazwyczaj na "paczkach" N próbek - wtedy wartości f. straty i gradienty uśrednia się
- Zamiast obliczać wyniki dla pojedynczych próbek, można dla całych list - wtedy wejścia, wyjścia i wyniki pośrednie są macierzami (typowe w zewnętrznych bibliotekach)

Końcowe uwagi

- Może być więcej warstw niż 2
- Uczy się zazwyczaj na "paczkach" N próbek - wtedy wartości f. straty i gradienty uśrednia się
- Zamiast obliczać wyniki dla pojedynczych próbek, można dla całych list - wtedy wejścia, wyjścia i wyniki pośrednie są macierzami (typowe w zewnętrznych bibliotekach)
- Dla klasyfikacji dla więcej niż 2 klas zwykle koduje się wyjścia kodowaniem 1 z n (one-hot encoding).

Końcowe uwagi

- Może być więcej warstw niż 2
- Uczy się zazwyczaj na "paczkach" N próbek - wtedy wartości f. straty i gradienty uśrednia się
- Zamiast obliczać wyniki dla pojedynczych próbek, można dla całych list - wtedy wejścia, wyjścia i wyniki pośrednie są macierzami (typowe w zewnętrznych bibliotekach)
- Dla klasyfikacji dla więcej niż 2 klas zwykle koduje się wyjścia kodowaniem 1 z n (one-hot encoding).
- Funkcja aktywacji ostatniej warstwy zwykle jest liniowa albo dedykowana do zadania.

Końcowe uwagi

- Może być więcej warstw niż 2
- Uczy się zazwyczaj na "paczkach" N próbek - wtedy wartości f. straty i gradienty uśrednia się
- Zamiast obliczać wyniki dla pojedynczych próbek, można dla całych list - wtedy wejścia, wyjścia i wyniki pośrednie są macierzami (typowe w zewnętrznych bibliotekach)
- Dla klasyfikacji dla więcej niż 2 klas zwykle koduje się wyjścia kodowaniem 1 z n (one-hot encoding).
- Funkcja aktywacji ostatniej warstwy zwykle jest liniowa albo dedykowana do zadania.
- Sieci neuronowe to *de facto* złożone funkcje - metody gradientowe mogą wymagać wielu iteracji z małym (0.01, 0.001 itp) parametrem kroku.