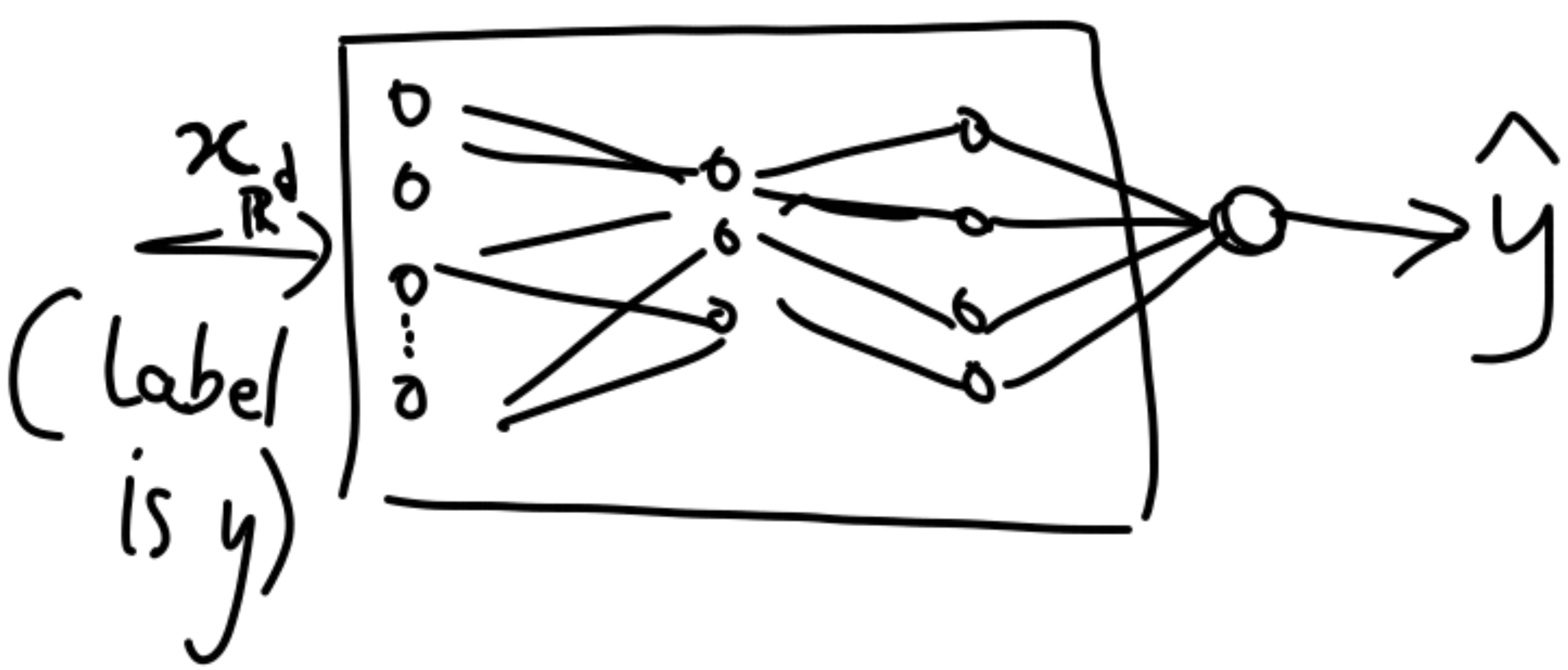
CS725

Loss functions typically used in training neural networks.

Ⓐ REGRESSION

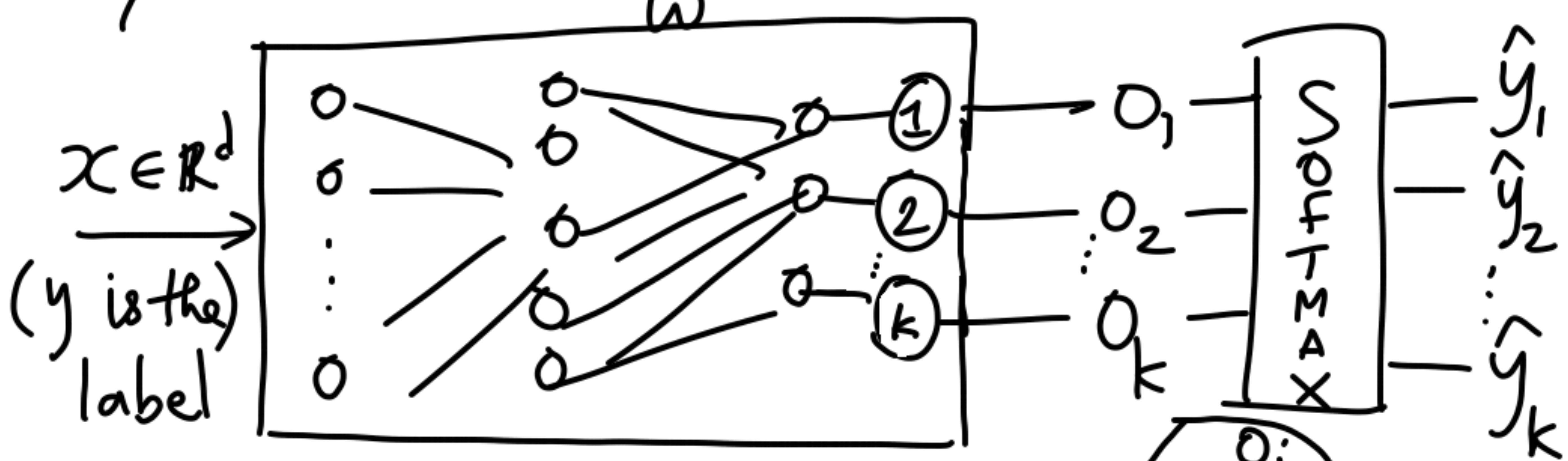


$f_{NN}(x; w)$

$$L(\hat{y}, y; w) = \frac{1}{2}(\hat{y} - y)^2$$

[Convenience when computing the gradient of L]

B) CLASSIFICATION



for a K-class Classification task

$$\frac{e^{o_j}}{\sum_j e^{o_j}}$$

Predicted Probabity Distribution

CROSS-ENTROPY LOSS

$$\mathcal{L}(\hat{y}, y; w) = \sum_{k=1}^{K} -y_k \log \hat{y}_k$$

$$[\hat{y}_1, \dots, \hat{y}_k]$$

y_k is either 0 or 1 & $\hat{y}_k$ is a probability

Training a neural network

Neural networks: Loss function is <u>non-convex</u>

Obvious choice to optimize the neural network loss function: $\boxed{\text{GRADIENT DESCENT}}$

$\Theta = \{w_0, w_1, \ldots, w_N\}$ is set of all parameters of a neural network

$\Theta \leftarrow$ initialize the weight vector

while (not stopping criterion) do

Pick a single example $(x, y) \in \mathcal{D}$ // SGD

$L \leftarrow L(f_\Theta(x), y)$

$\Theta \leftarrow \Theta - \boxed{\eta} \nabla_\Theta L$

end while

CRITICAL ASPECTS OF NN TRAINING:

(1) wt initialization

(2) Stopping criteria

(3) Learning rate

For (S)GD training, we need to compute $\nabla_\theta L$

which is $\frac{\partial L}{\partial w_0}, \frac{\partial L}{\partial w_1}, \ldots, \frac{\partial L}{\partial w_N}$

This gradient is computed efficiently using the $\boxed{\text{BACKPROPAGATION ALGORITHM}}$

Preliminaries

Univariate chain rule ; Given a function $f(y)$, $y = g(x)$

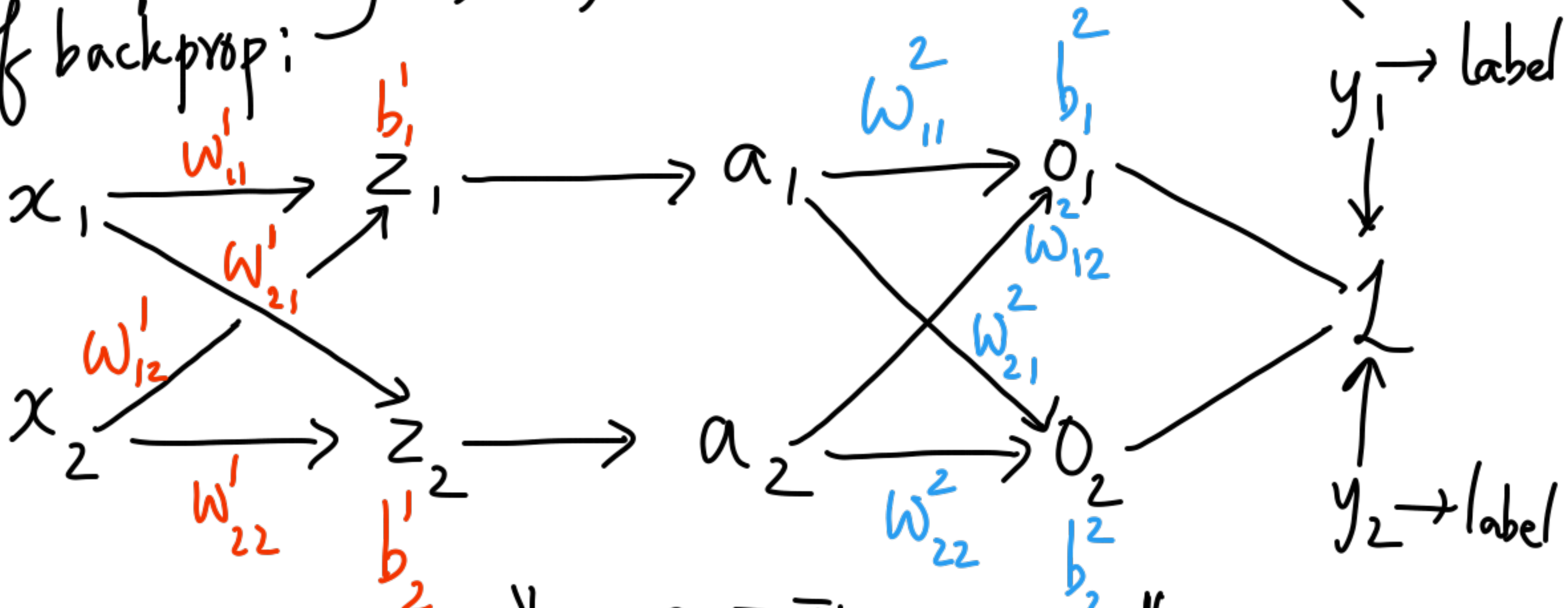
$$\frac{df}{dx} = \frac{df}{dy} \cdot \frac{dy}{dx}$$

Multivariate chain rule ; Given $f(y_1, \dots, y_n)$, $y_i = g_i(x)$

$$\frac{\partial f}{\partial x} = \sum_i \frac{\partial f}{\partial y_i} \cdot \frac{\partial y_i}{\partial x}$$

Backpropagation: Two-pass algorithm

Example using a feedforward neural network (multi-layer perceptron, MLP) of backprop:



"COMPUTATION GRAPH"

① forward pass

$$z_i = \sum_j w^1_{ij} x_j + b^1_i$$

$$a_i = \sigma(z_i)$$

$$O_k = \sum_i W^2_{ki} a_i + b^2_k$$

$$L = \frac{1}{2} \sum_k (O_k - Y_k)^2$$

② BACKWARD PASS

$$\frac{\partial L}{\partial L} = 1 \quad (\text{Base case})$$

$$\frac{\partial L}{\partial o_k} = \frac{\partial L}{\partial L} \cdot \frac{\partial L}{\partial o_k} = o_k - y_k$$

$$\frac{\partial L}{\partial w^2_{ki}} = \frac{\partial L}{\partial o_k} \cdot \frac{\partial o_k}{\partial w^2_{ki}} = \frac{\partial L}{\partial o_k} \cdot a_i$$

$$\frac{\partial L}{\partial b^2_k} = \frac{\partial L}{\partial o_k} \cdot \frac{\partial o_k}{\partial b^2_k} = \frac{\partial L}{\partial o_k}$$

$$\frac{\partial L}{\partial a_i} = \sum_k \frac{\partial L}{\partial O_k} \cdot \frac{\partial O_k}{\partial a_i} = \sum_k \frac{\partial L}{\partial o_k} \cdot w_{ki}^2$$

$$\frac{\partial L}{\partial Z_i} = \frac{\partial L}{\partial a_i} \cdot \frac{\partial a_i}{\partial z_i} = \frac{\partial L}{\partial a_i} \cdot \sigma'(z_i)$$

$$\frac{\partial L}{\partial w_{ij}^1} = \frac{\partial L}{\partial Z_i} \cdot \frac{\partial z_i}{\partial W'_{ij}} = \frac{\partial L}{\partial z_i} \cdot x_j$$

$$\frac{\partial L}{\partial b_i^1} = \frac{\partial L}{\partial z_i} \cdot \frac{\partial z_i}{\partial b_i^1} = \frac{\partial L}{\partial Z_i}$$