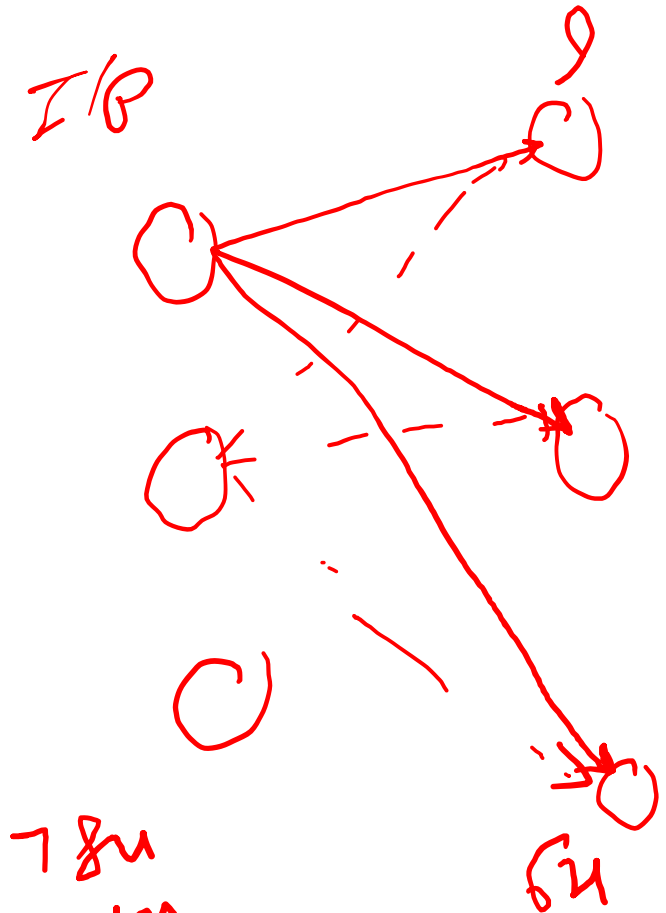
Lecture 17

*) Back propagation.

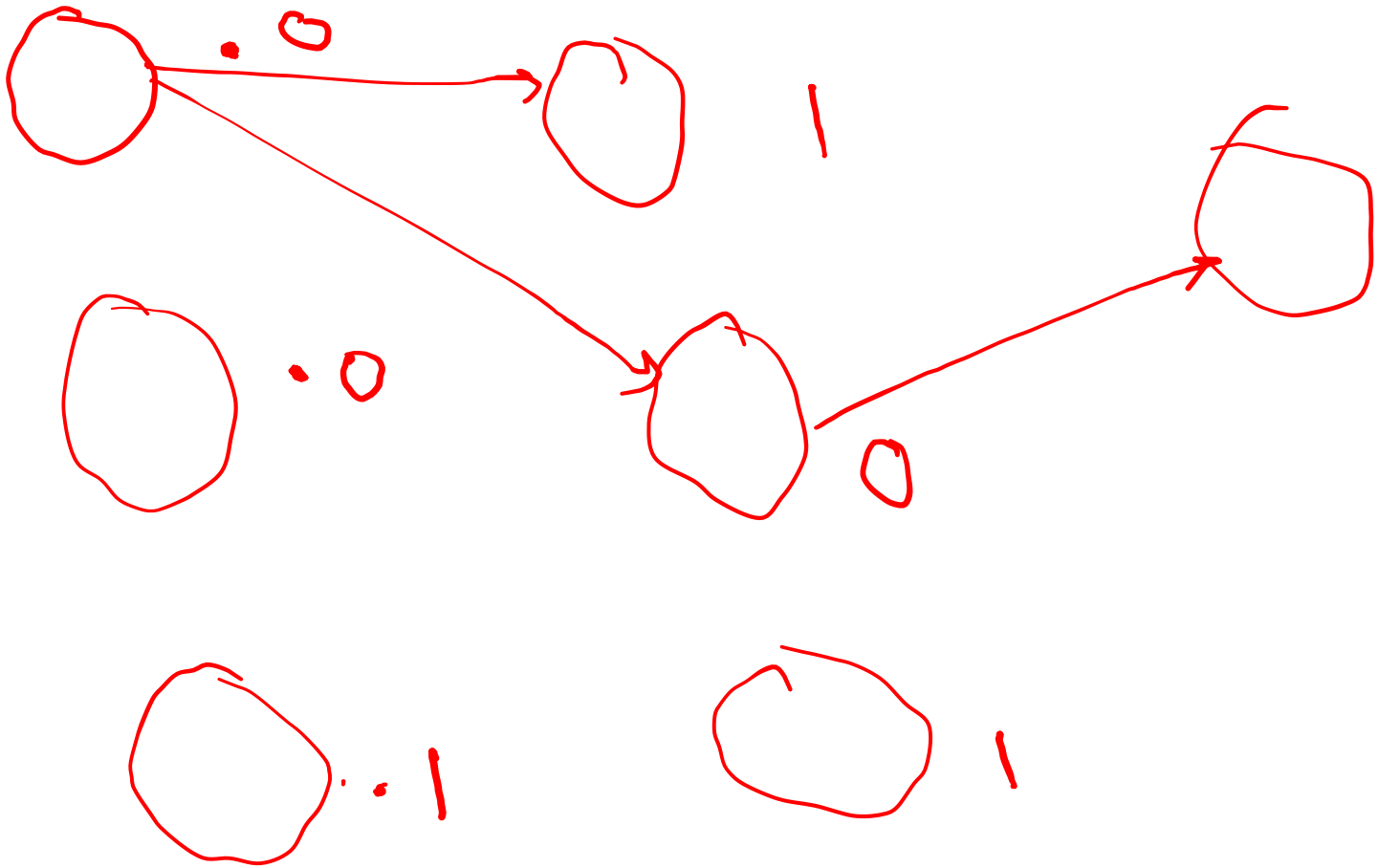


$$\delta_i = \sum_j \delta_j U_{ji} \sigma'(a_i)$$

$$\delta_{kj} : \delta_k U_{kj} \sigma'(a_j)$$

$$\delta_k : \frac{\partial E}{\partial a_k}$$

$$\hat{y} : \sigma(a_k)$$


0
1
0
0
1
1

Momentum

SGD with momentum remembers the update Δw at each iteration.

Each update is as a (convex) combination of the gradient and the previous update.

$$\Delta w := \eta \nabla Q_i(w) + \alpha \Delta w$$

$$w := w - \Delta w$$

Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. (8 October 1986). "Learning representations by back-propagating errors". Nature 323 (6088): 533536.

at $\theta = 0$, $w^1 \leftarrow w^0 - \eta \left[\nabla_w J(w)\right]^{\theta=0}$

$$w \leftarrow w^0 - \eta \begin{bmatrix} 2(w_1 + w_2 x \cdot y) \\ 2x(w_1 + w_2 x - y) \end{bmatrix}_{w_1 = w_1^0,\ w_2 = w_2^0}$$

at $\theta = 1$, $w^2 \leftarrow w^1 - \eta \left[\nabla_w J(w)\right]^{\theta=1}_{w_1 = w_1^1,\ w_2 = w_2^1} - \eta\alpha \left[\nabla_w J(w)\right]^{\theta=0}$

$W = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ are the weights.

W at $t = 0$ step $\quad W^0 = \begin{bmatrix} w_1^0 \\ w_2^0 \end{bmatrix}$

$W^0 \leftarrow$ initialized.

W at $t = 1$ step $\quad W^1 = \begin{bmatrix} w_1^1 \\ w_2^1 \end{bmatrix}$

*) Regression: $(y-\hat{y})^2 = E$

*) Classification: Binary classification.

Class labels $y \in \{0, 1\}$

$$E = -\left[y \log \hat{y} + (1-y) \log (1-\hat{y})\right] \leftarrow \text{Binary cross-entropy}$$

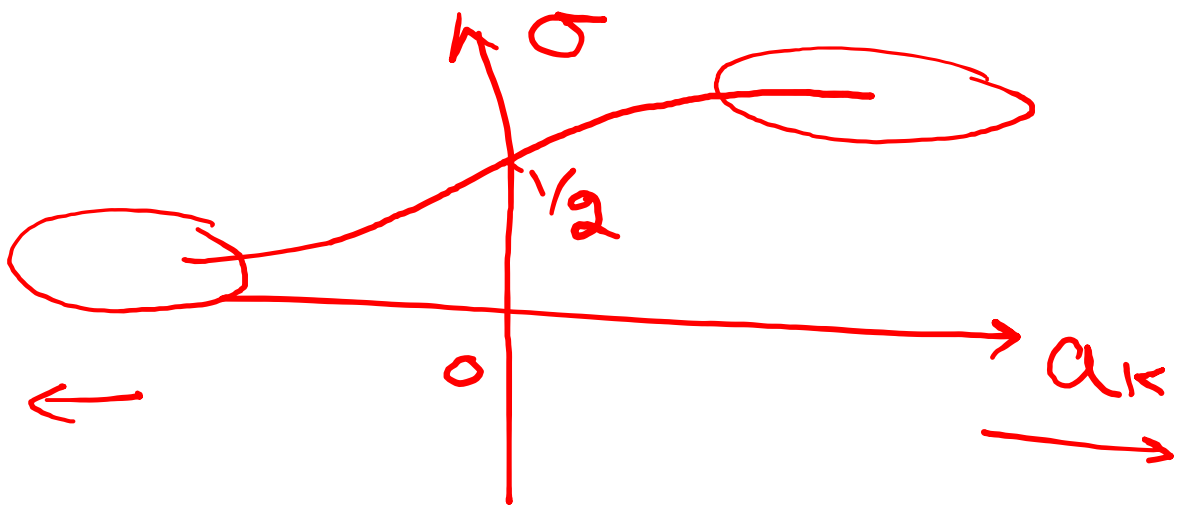
$\hat{y} \in (0, 1)$

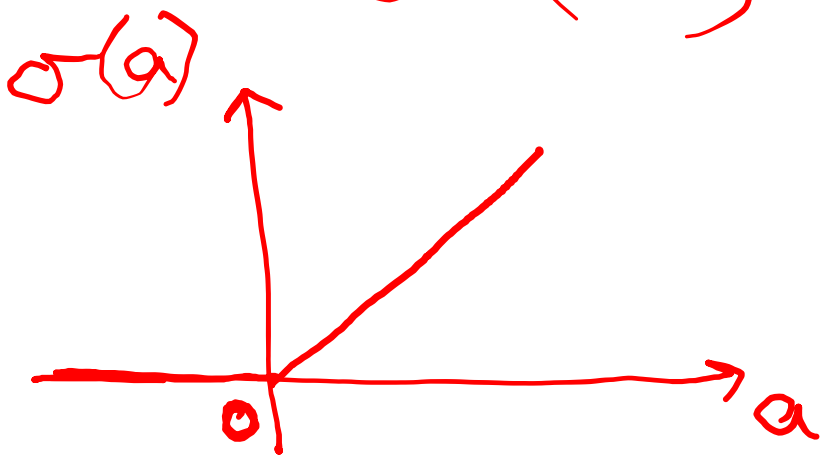
$$\delta_k = \frac{\partial E}{\partial a_k} = -\frac{\partial}{\partial a_k}\left[y \log \sigma(a_k) + (1-y) \log (1-\sigma(a_k))\right]$$

$$= -\frac{y\, \sigma(a_k)(1-\sigma(a_k))}{\sigma(a_k)} - \frac{(1-y)(-1)\sigma'(a_k)}{1-\sigma(a_k)}$$

$-y(1-\sigma(a_k))$
$+(1-y)\sigma(a_k)$

If $y = 0$, $\delta_k = \sigma(a_k)$ $\sigma'(a_k)$





Vanishing gradient

$\sigma(a) = \max(0, a)$

ReLU → Rectified linear unit.

$$E = -y \log \hat{y} - (1-y) \log (1-\hat{y})$$

$y = 0, \quad \hat{y} = 0.99$

$$E = -\log (1 - .99) = 2$$

$y = 1, \quad \hat{y} = 0.99 \rightarrow$

$$E = -1 \log (.99) \approx 0$$

$$E = e^{-y\hat{y}}$$

error

$$E = e^{-y\hat{y}} \qquad y = 1$$

Weight vector W

$$\| W^{t+1} - W^t \| < \text{threshold.}$$

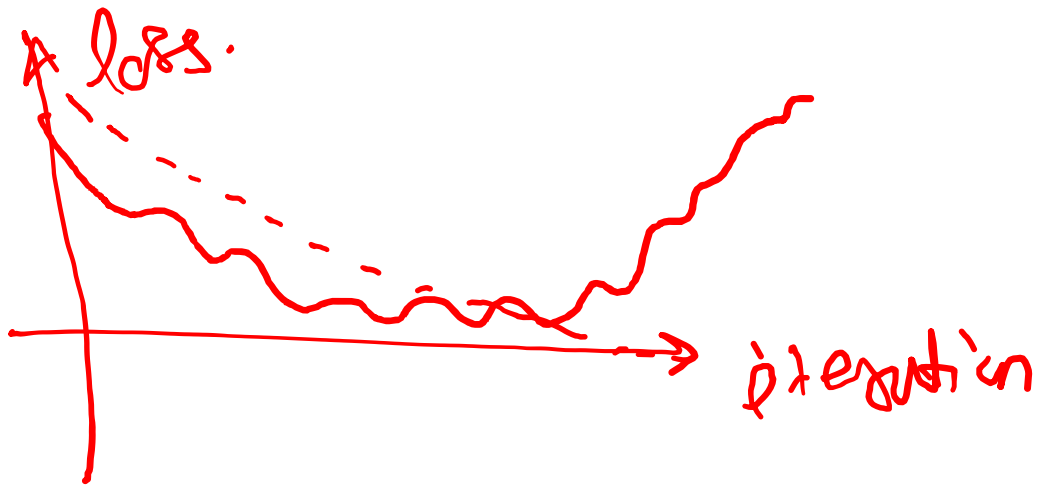
$$< \alpha \| \nabla_w E(w) \|$$

threshold $= 10^{-3}$

$t+1^{th}$ iteration.

t^{th} iteration.

If weights do not change much b/w subsequent iterations, exit the loop.



*) multiclass cross-entropy.

Classes → $y \in \{0, 1, 2, \ldots, C-1\}$

no. of classes, C

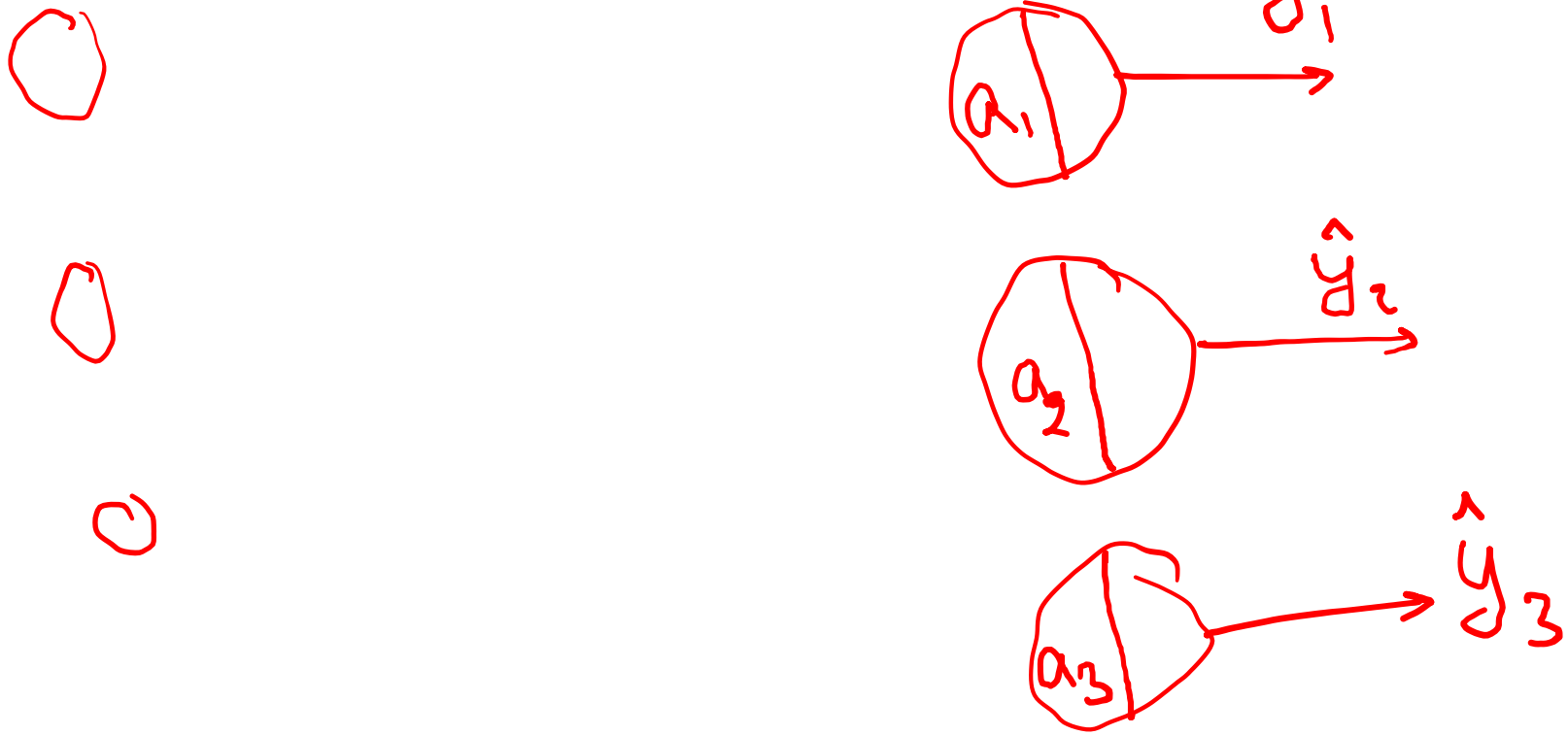
repr. of k^{th} class →

$t = [0, 0, \ldots 1, \ldots 0]$ → one hot encoding.

(k^{th} position)

$E = -t^T \log \hat{y}$

i/p l i j K



o/p K^{th} layer will have C nodes for C-classes.

let C = 3

activation → Softmax.

$$\hat{y}_1 = \frac{e^{a_1}}{\sum_{p=1}^{C} e^{a_p}} \qquad \hat{y}_2 = \frac{e^{a_2}}{\sum_{p=1}^{C} e^{a_p}}$$

$$\hat{y}_k = e^{a_k} \Big/ \sum_{p=1}^{C} e^{a_p}$$

$\hat{y}_k$ → Probability