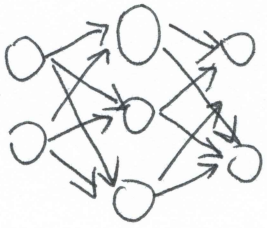
e.g. 2-layer Fully-connected NN (MLP):



input: $\underline{x} \in \mathbb{R}^D$

output: $\hat{\underline{y}} \in \mathbb{R}^C$

hidden layer: $\underline{h} \in \mathbb{R}^H$

activation function: $g(\,)$, $f(\,)$

parameters: $\theta = \{ \underline{W}_I, \underline{b}_I, \underline{W}_o, \underline{b}_o \}$

feed-forward:

$$\underline{h} = g(\underline{W}_I \underline{x} + \underline{b}_I)$$

$$\hat{\underline{y}} = f(\underline{W}_o \underline{h} + \underline{b}_o)$$

note: I dropped transpose according to the advice I got from class to avoid confusion

parameter estimation: $\theta = \min_{\theta} L(\hat{\underline{y}}(\underline{x}; \theta), \underline{y})$

gradient descent: $\theta_{new} = \theta_{old} - \alpha \nabla_\theta L$

where α is the learning rate

$\nabla_\theta L$ is the gradient of the loss w.r.t. θ

backpropagation:

let $\underline{\delta}_0 = \frac{\partial L}{\partial \underline{z}_0}$ for $\underline{z}_0 = \underline{W}_0 \underline{h} + \underline{b}_0$ (pre-activation output)

$\underline{\delta}_0 = \frac{\partial L}{\partial \hat{\underline{y}}} \odot f'(\underline{z}_0)$ ← chain rule

$\therefore \nabla_{\underline{W}_0} L_{i,c} = \frac{\partial L}{\partial (\underline{W}_0)_{i,c}} = \frac{\partial L}{\partial (\underline{z}_0)_c} \cdot \frac{\partial (\underline{z}_0)_c}{\partial (\underline{W}_0)_{i,c}} = (\underline{\delta}_0)_c \cdot \underline{h}_i$

$\Downarrow$

$\nabla_{\underline{W}_0} L = \underline{h}\, \underline{\delta}_0^T$

$\nabla_{\underline{b}_0} L = \underline{\delta}_0$

let $\underline{\delta}_I = \frac{\partial L}{\partial \underline{z}_I}$ where $\underline{z}_I = \underline{W}_I \underline{x} + \underline{b}_I$

$\underline{\delta}_I = \frac{\partial L}{\partial \underline{h}} \odot g'(\underline{z}_I) = \left(\frac{\partial \underline{z}_0}{\partial \underline{h}}^T \cdot \frac{\partial L}{\partial \underline{z}_0}\right) \odot g'(\underline{z}_I)$

$= \underline{W}_0 \underline{\delta}_0 \odot g'(\underline{z}_I)$

$\therefore \nabla_{\underline{W}_I} L = \underline{x}\, \underline{\delta}_I^T$, $\nabla_{\underline{b}_I} L = \underline{\delta}_I$