Week 5: Neural Networks

Matthew Caldwell

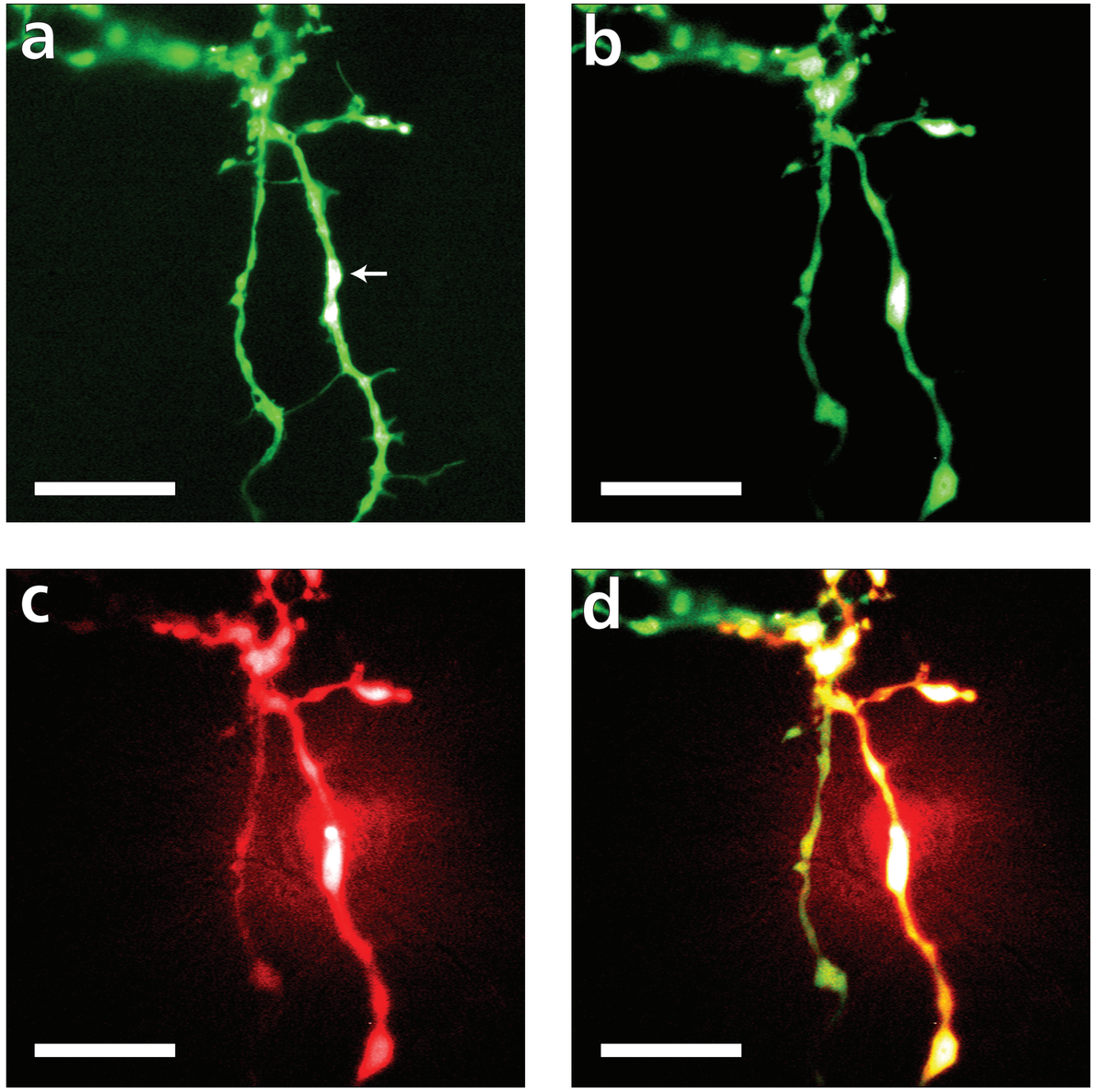
COMP0088 Introduction to Machine Learning • UCL Computer Science • Autumn 2024

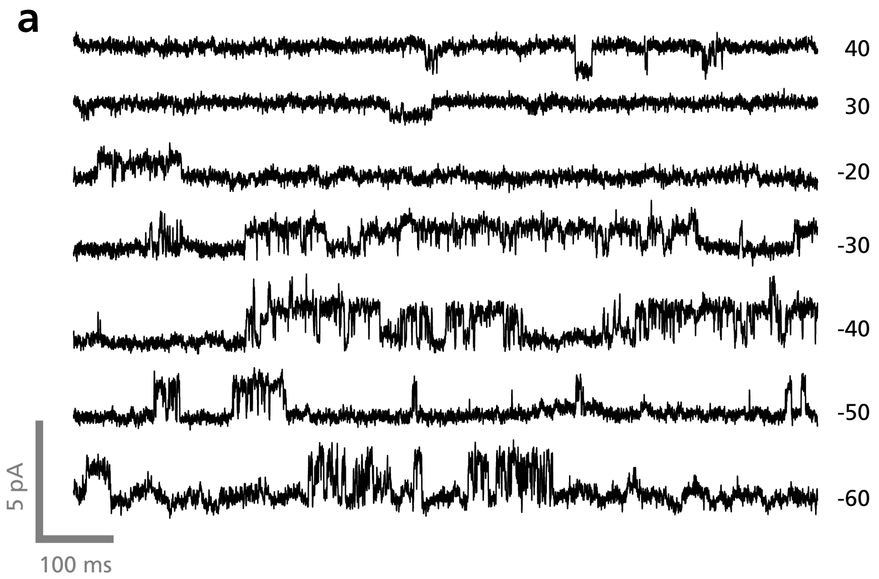
Admin

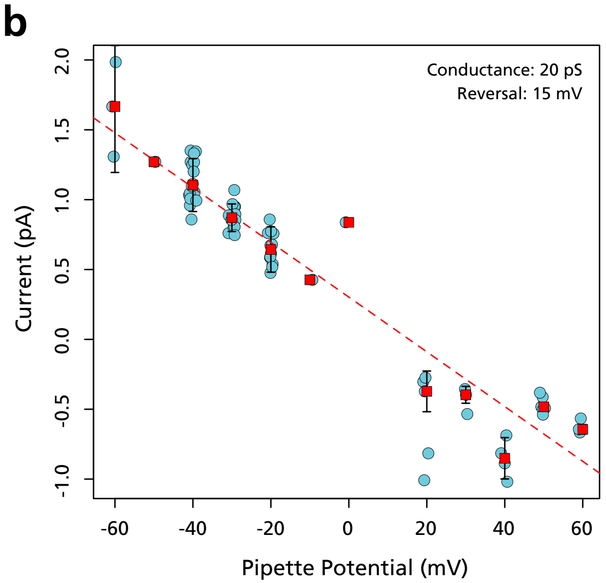
- No labs, lectures or office hour next week

Week 5 Recap

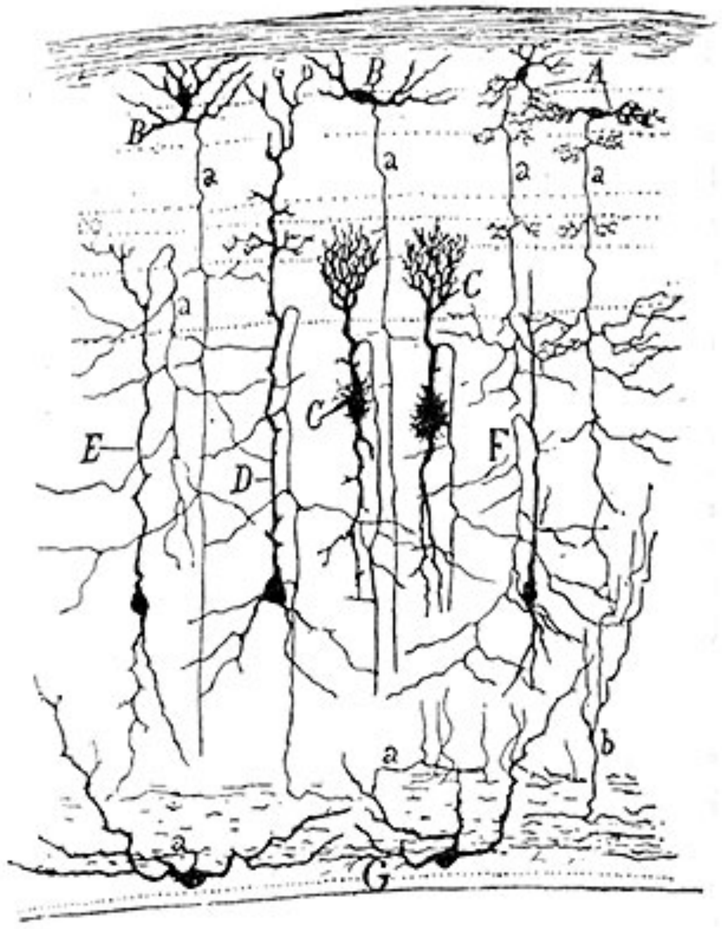
Safety in Numbers

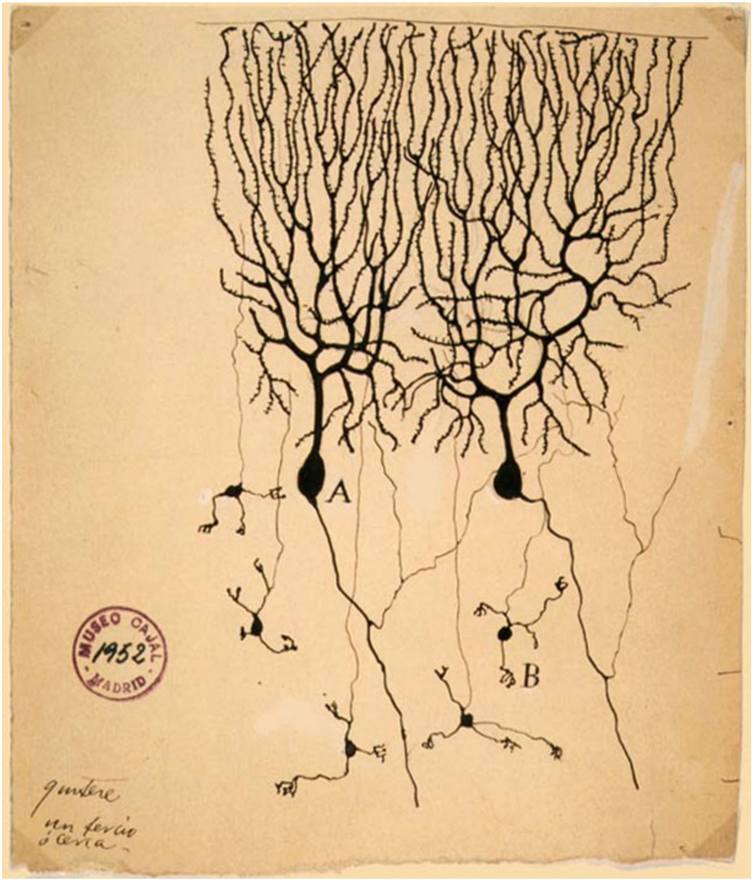

a
b
c
d


a
40
30
-20
-30
-40
-50
-60
5 pA
100 ms

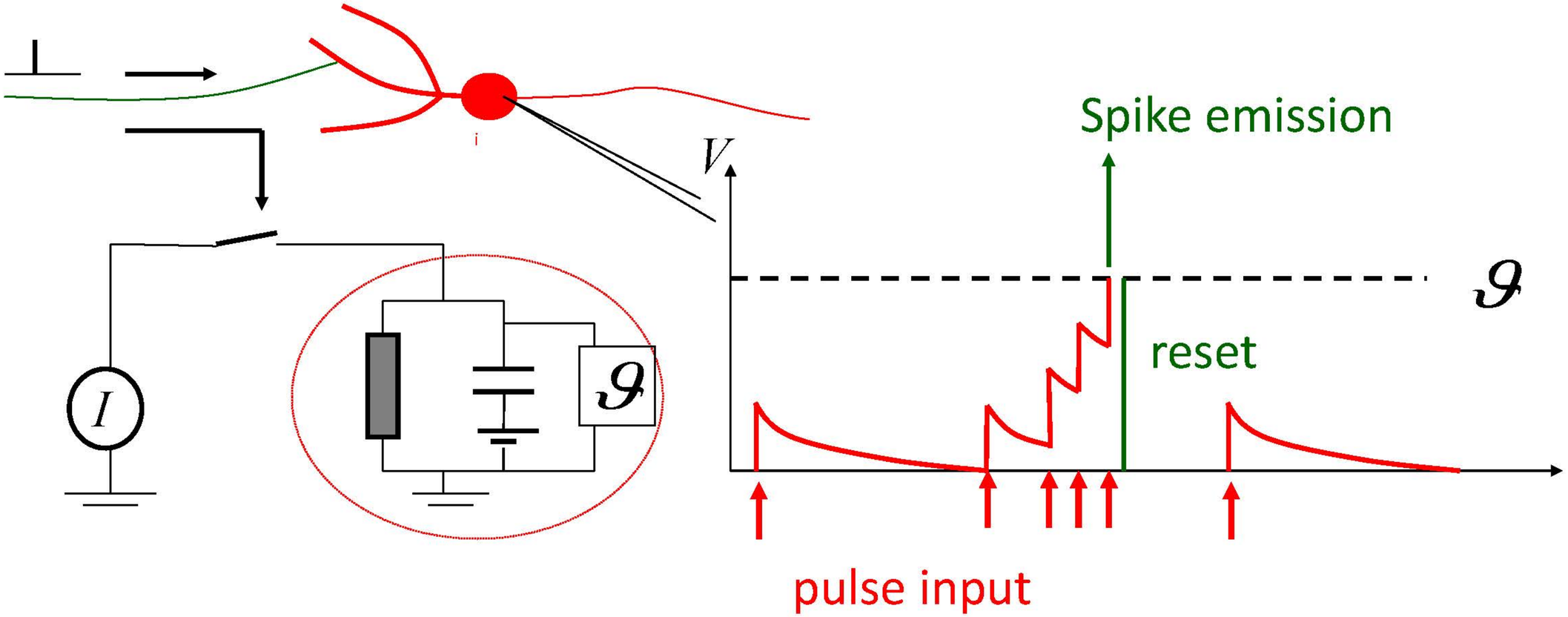

b
Conductance: 20 pS
Reversal: 15 mV
Current (pA)
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-60
-40
-20
0
20
40
60
Pipette Potential (mV)


c
d
a
b


B
B
A
a
a
a
a
a
C
C
E
D
F
a
b
a
G


A
B
MUSEO NAVAL
1952
MADRID

Leaky Integrate-and-Fire Model



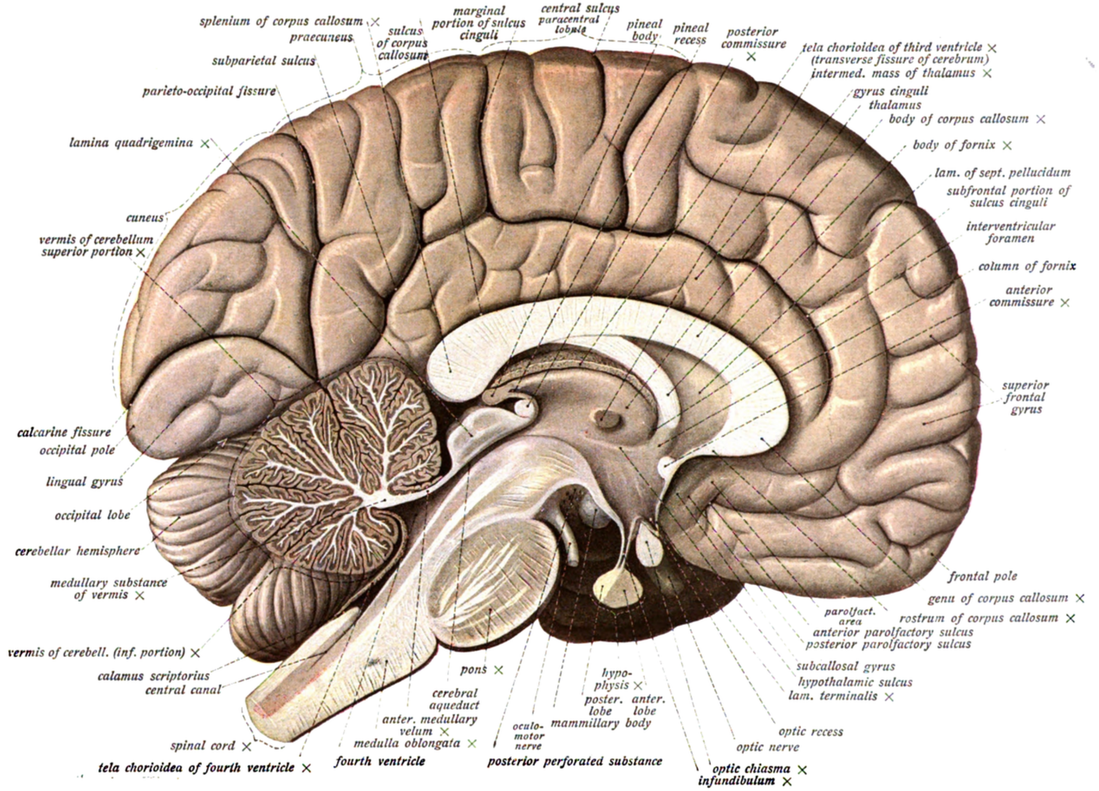
$$\text{fire} = \begin{cases} 1 & \text{if } \mathbf{x} \cdot \mathbf{w} \geq \text{threshold} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{fire} = \begin{cases} 1 & \text{if } \mathbf{x} \cdot \mathbf{w} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

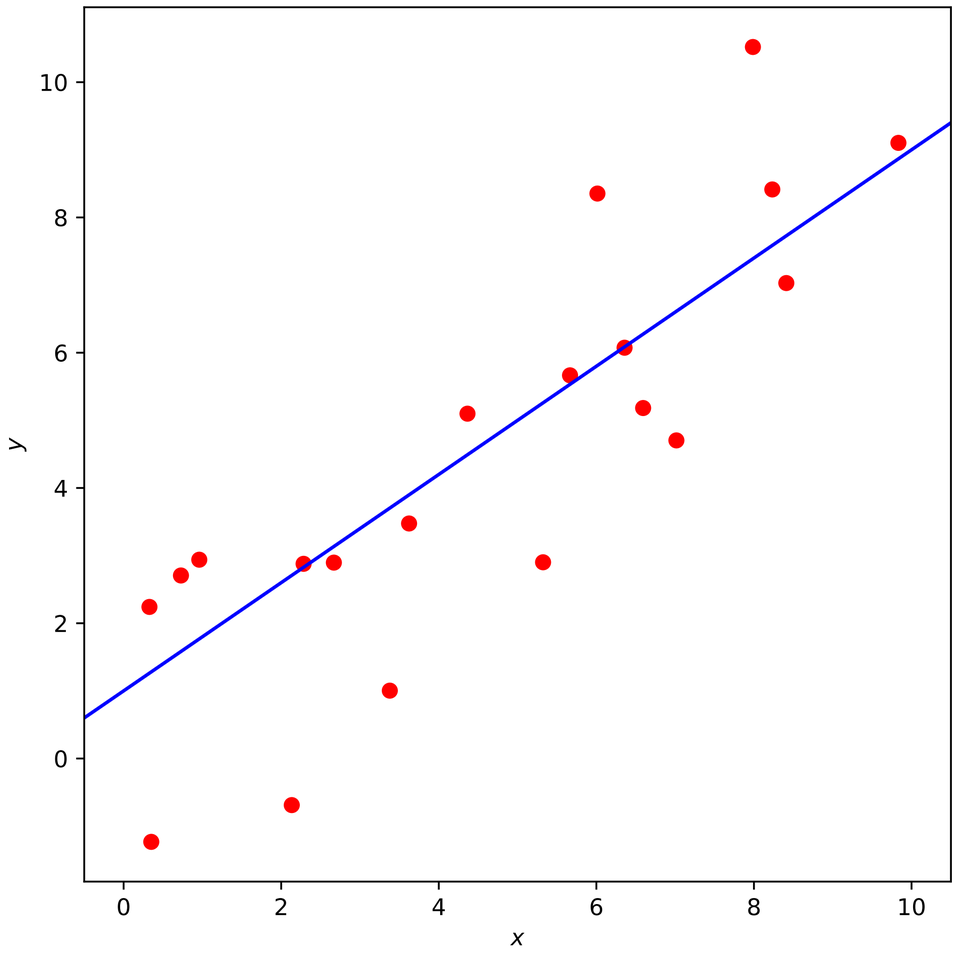
$$\text{fire} = \begin{cases} 1 & \text{if } \mathbf{x} \cdot \mathbf{w} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

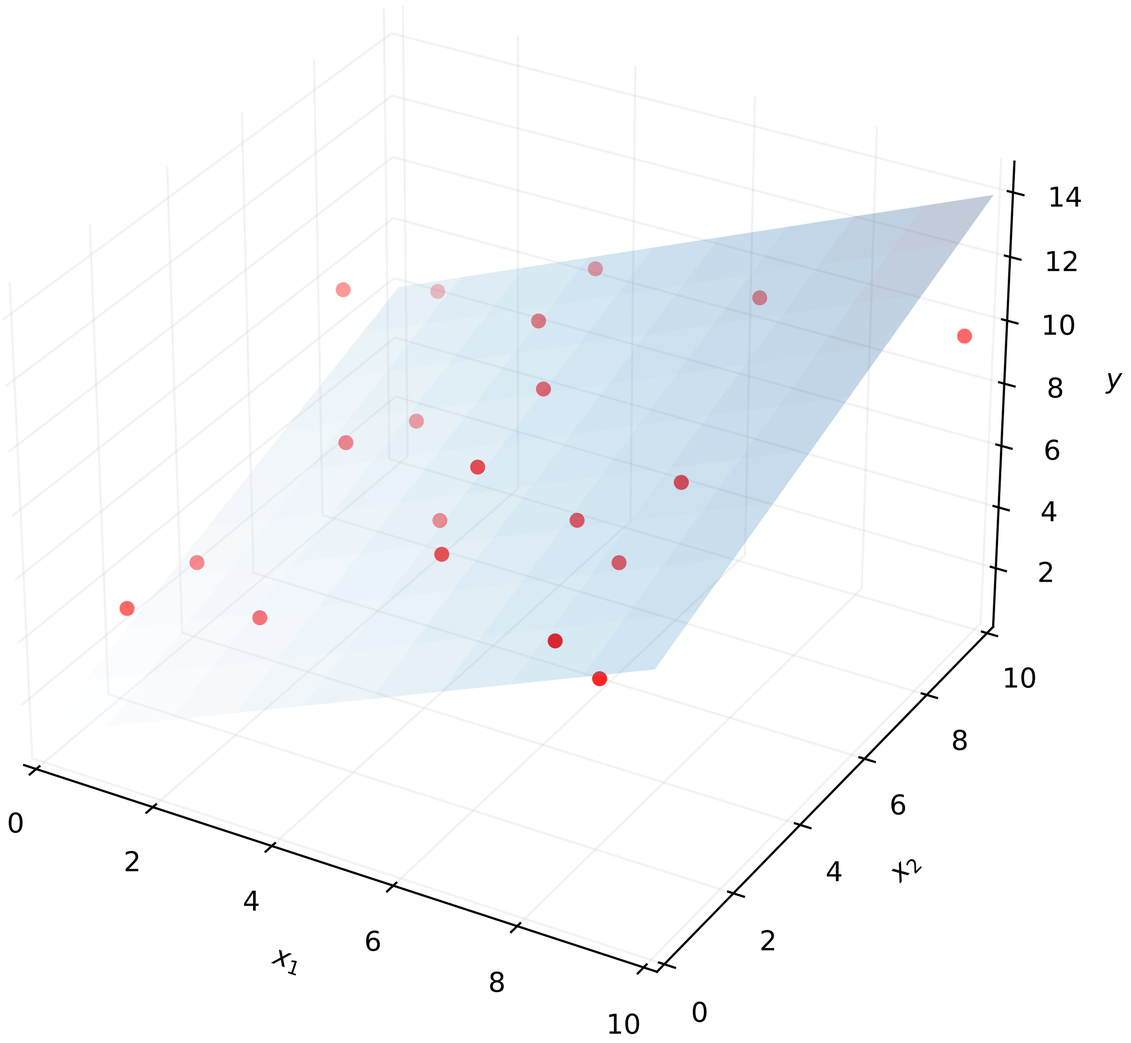
$\neq$

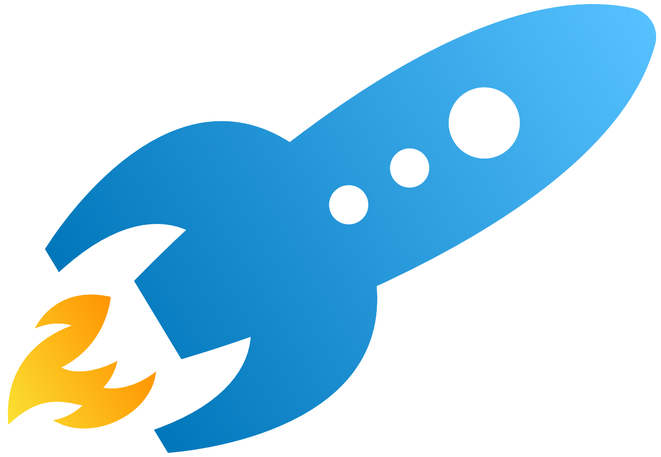
≠


splenium of corpus callosum
praecuneus
subparietal sulcus
parieto-occipital fissure
lamina quadrigemina
cuneus
vermis of cerebellum superior portion
calcarine fissure
occipital pole
lingual gyrus
occipital lobe
cerebellar hemisphere
medullary substance of vermis
vermis of cerebell. (inf. portion)
calamus scriptorius
central canal
spinal cord
tela chorioidea of fourth ventricle
sulcus of corpus callosum
marginal portion of sulcus cinguli
central sulcus
paracentral lobule
pineal body
pineal recess
posterior commissure
tela chorioidea of third ventricle
(transverse fissure of cerebrum)
intermed. mass of thalamus
gyrus cinguli
thalamus
body of corpus callosum
body of fornix
lam. of sept. pellucidum
subfrontal portion of sulcus cinguli
interventricular foramen
column of fornix
anterior commissure
superior frontal gyrus
frontal pole
genu of corpus callosum
rostrum of corpus callosum
parolfact. area
anterior parolfactory sulcus
posterior parolfactory sulcus
subcallosal gyrus
hypothalamic sulcus
lam. terminalis
optic recess
optic nerve
optic chiasma
infundibulum
hypo-physis
poster. lobe
anter. lobe
mammillary body
oculo-motor nerve
posterior perforated substance
pons
cerebral aqueduct
anter. medullary velum
medulla oblongata
fourth ventricle

$$\text{fire} = \begin{cases} 1 & \text{if } \mathbf{x} \cdot \mathbf{w} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

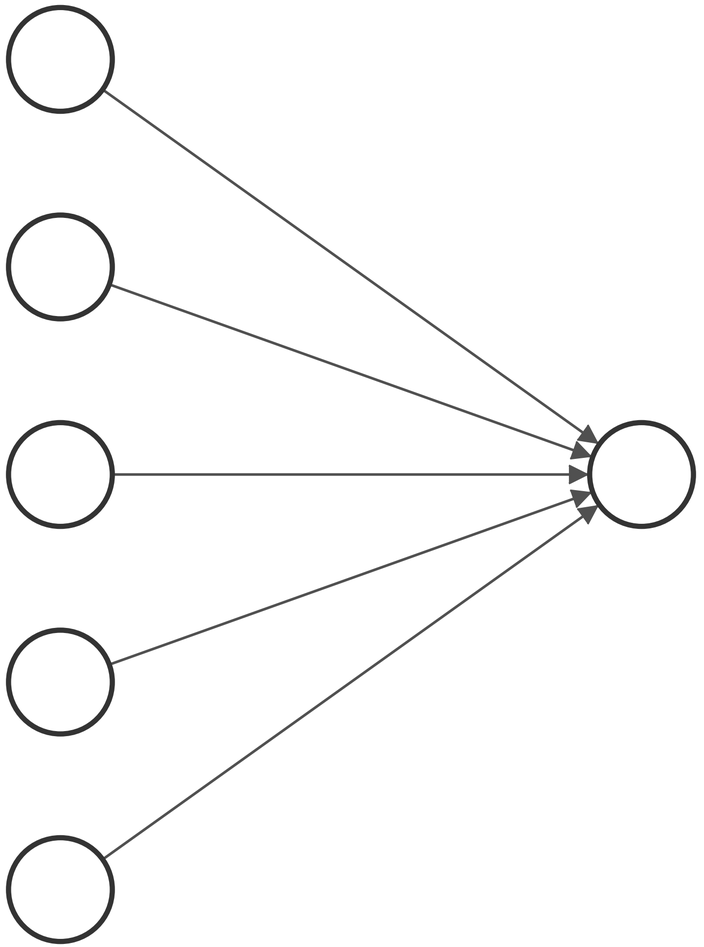

y
x
0
2
4
6
8
10

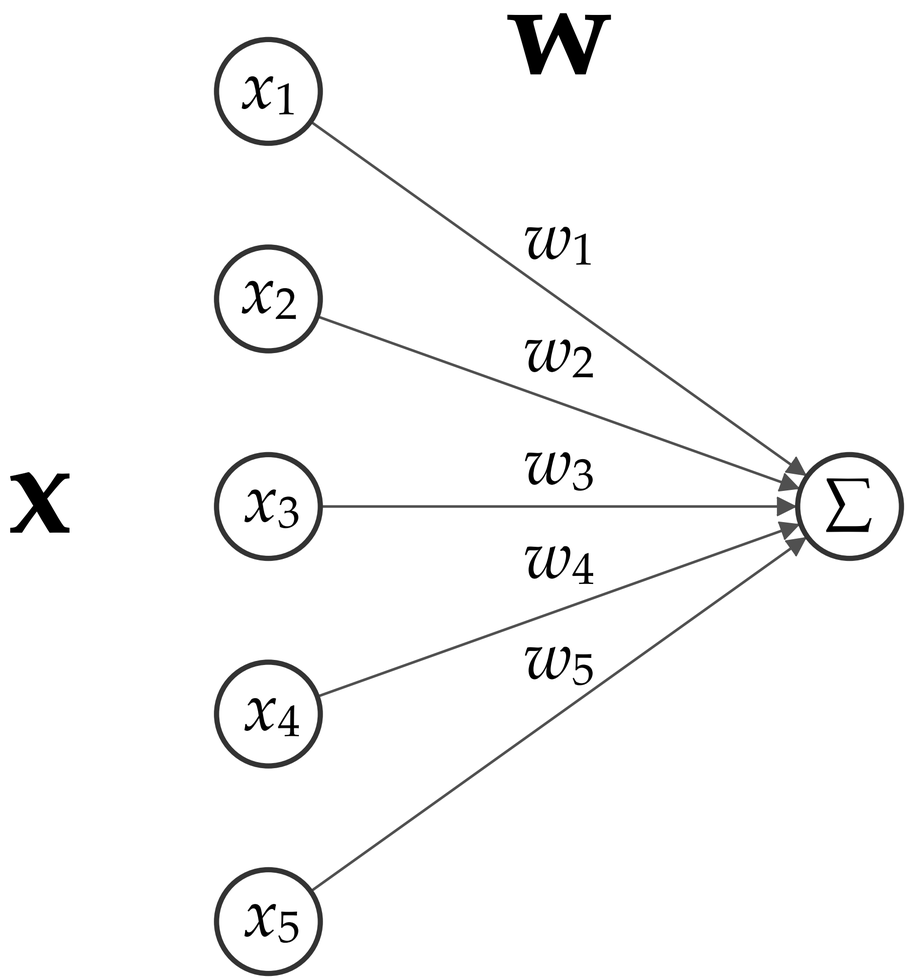

y
x_1
x_2
0
2
4
6
8
10
12
14

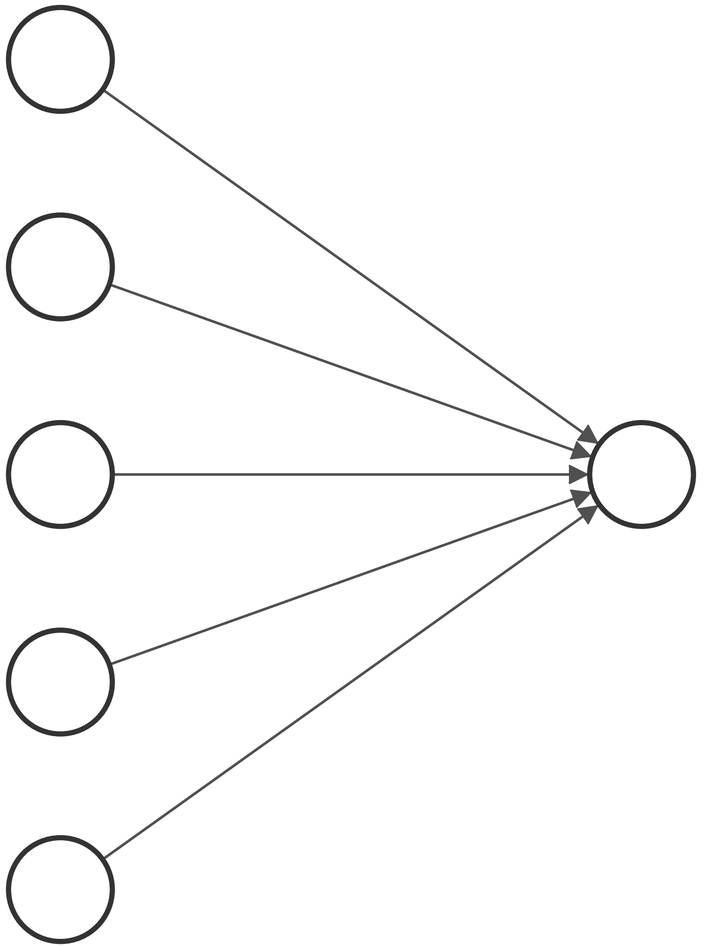


$$\mathbf{x} \cdot \mathbf{w} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

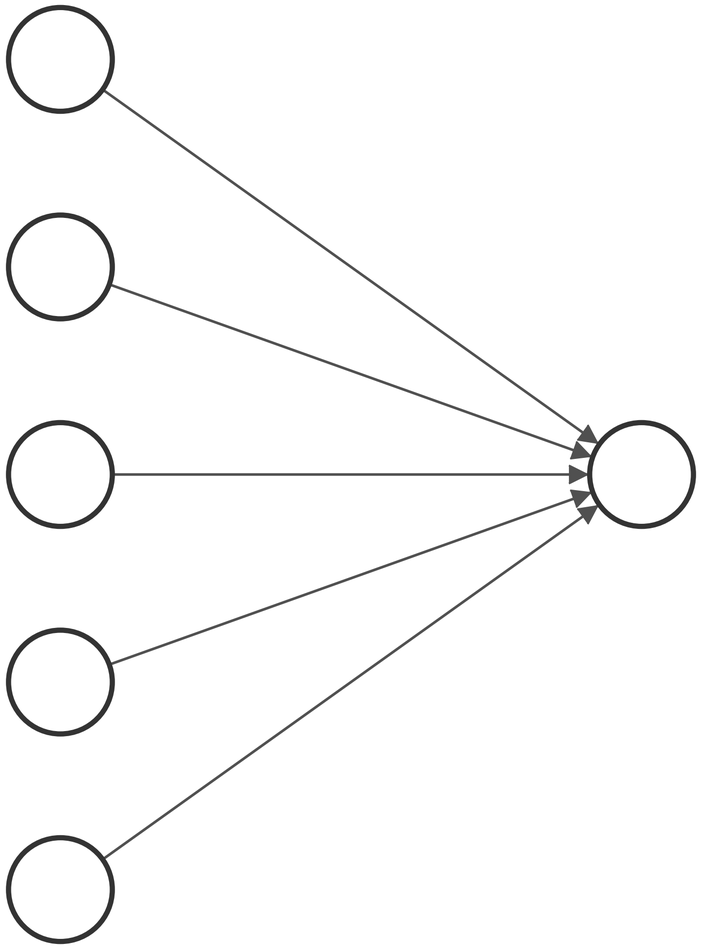
$$\mathbf{x}^\top \mathbf{w} = \begin{bmatrix} x_1 & x_2 & \dots & x_d \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

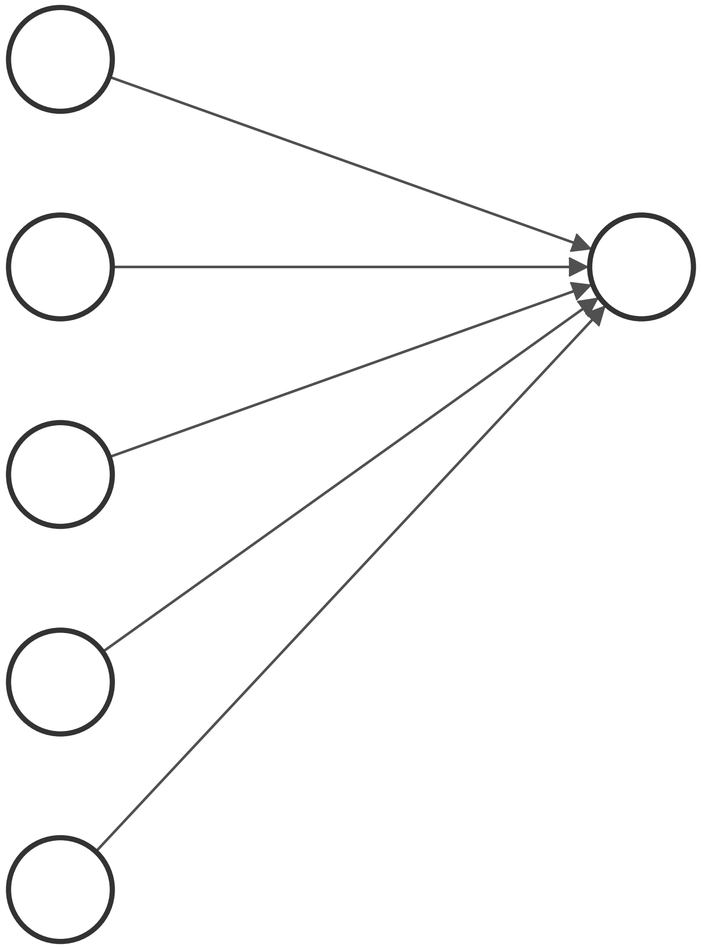


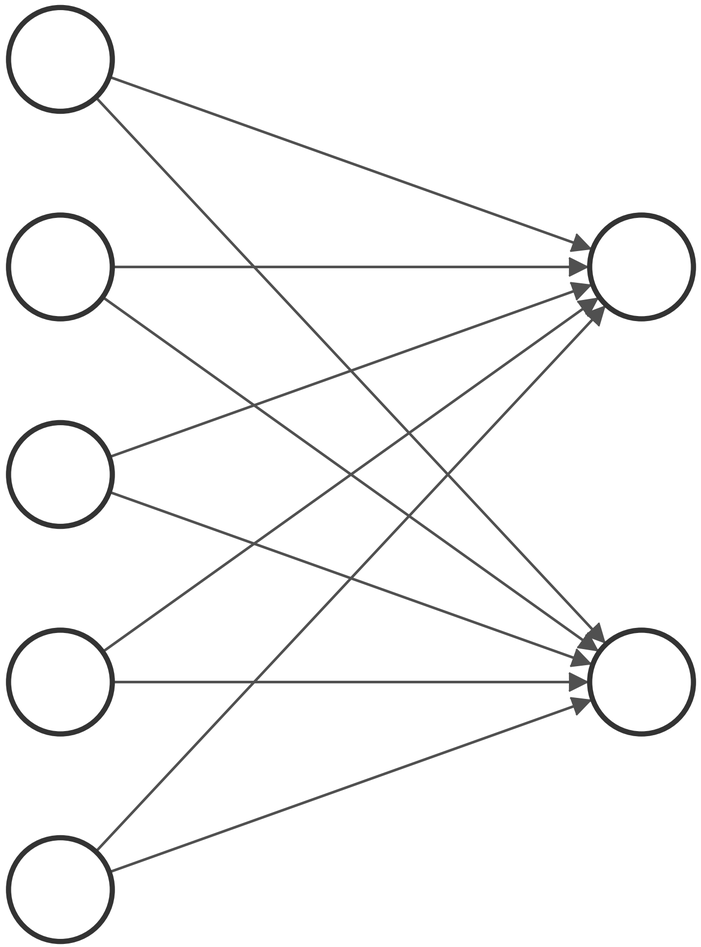

W
x
x_1
x_2
x_3
x_4
x_5
w_1
w_2
w_3
w_4
w_5
Σ



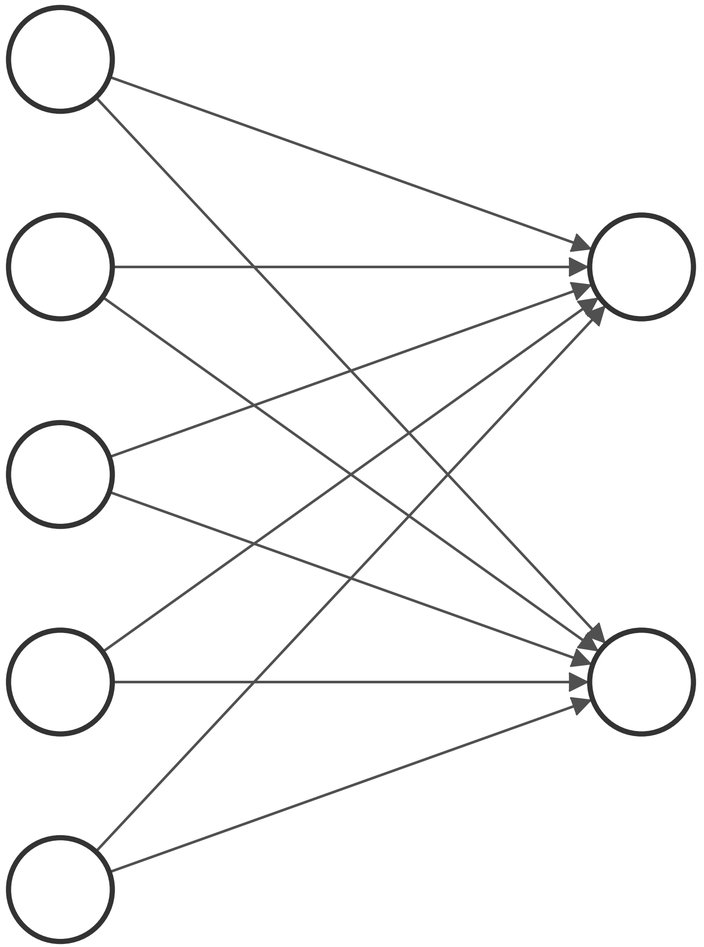
$$\mathbf{x}^\top \mathbf{w} = \begin{bmatrix} x_1 & x_2 & \dots & x_d \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

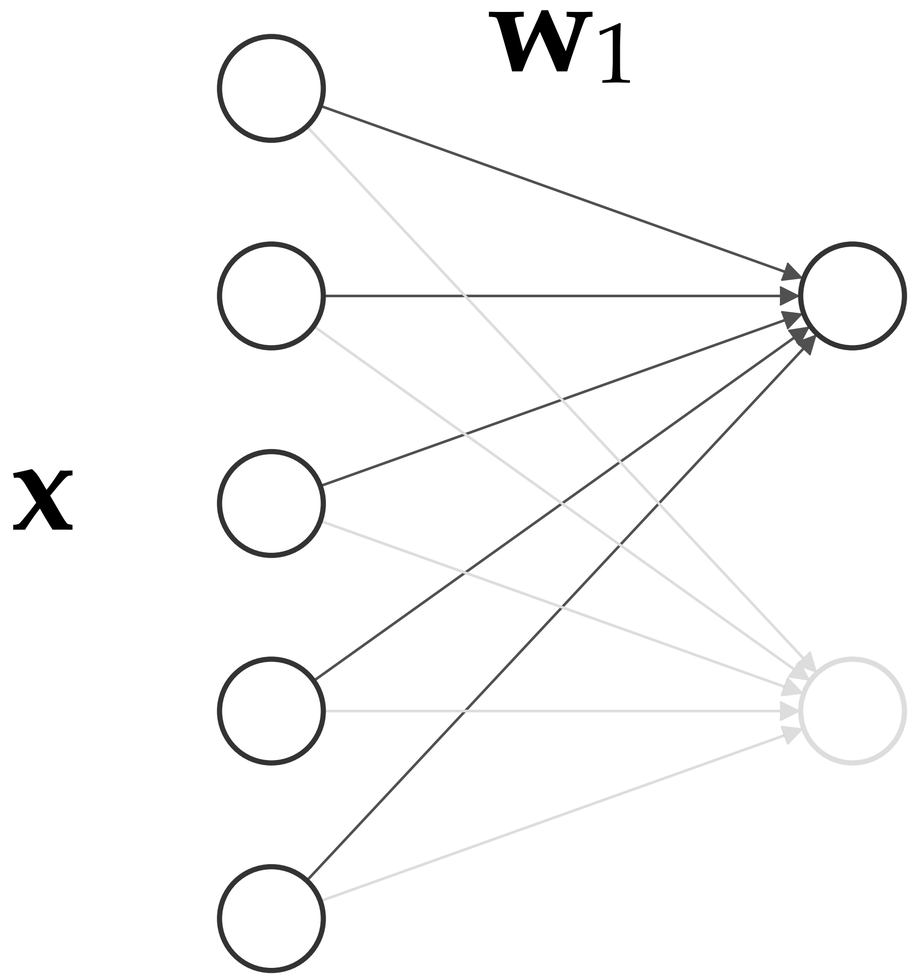


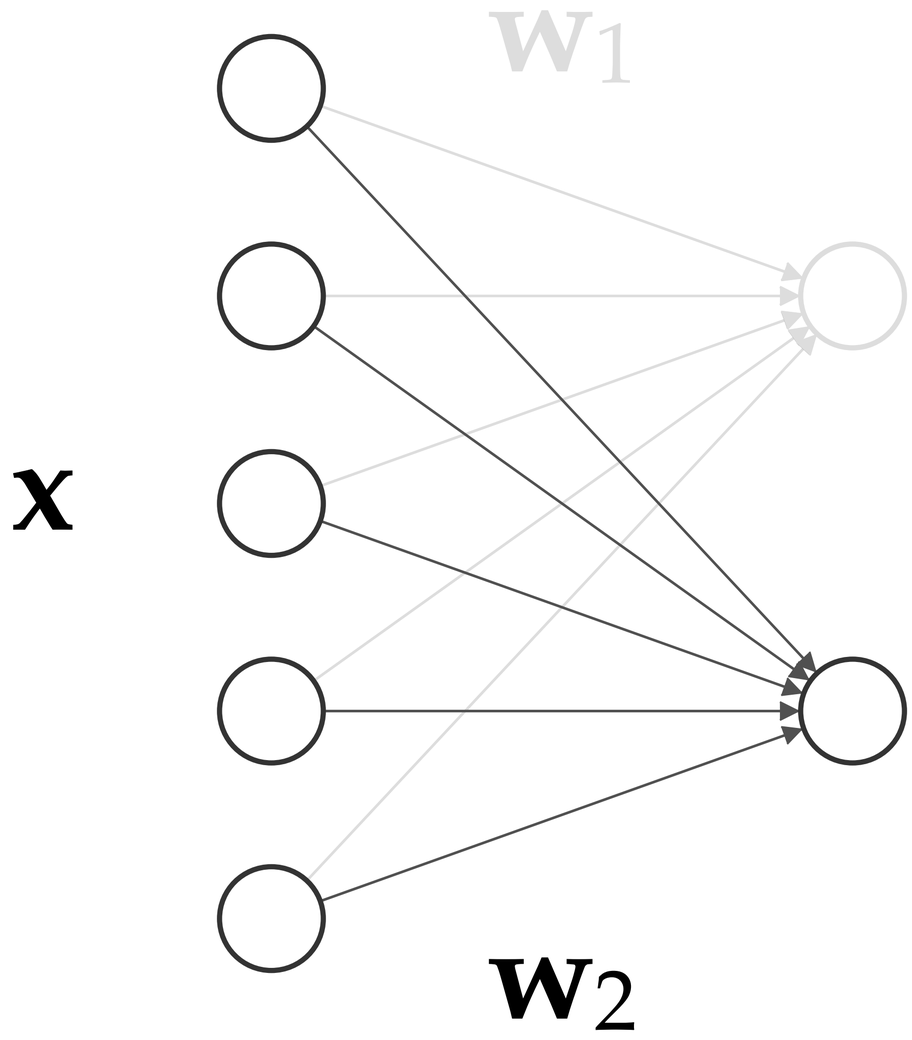


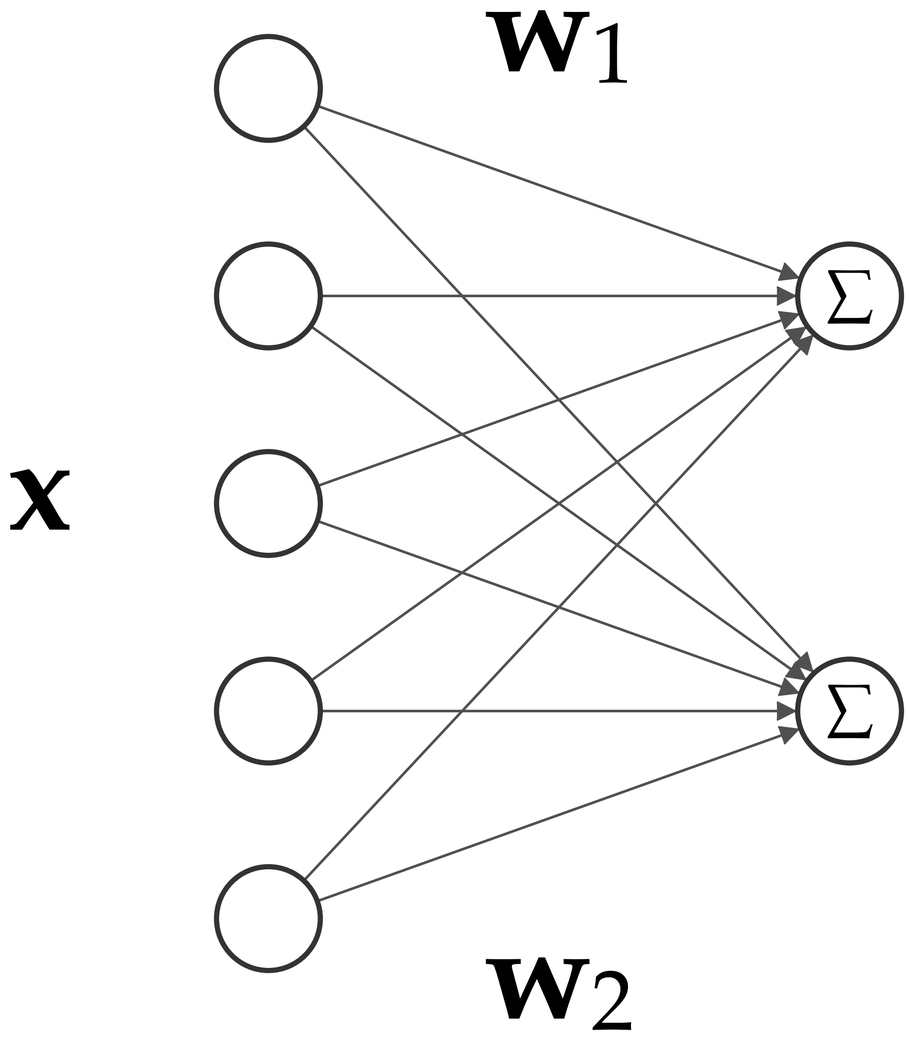


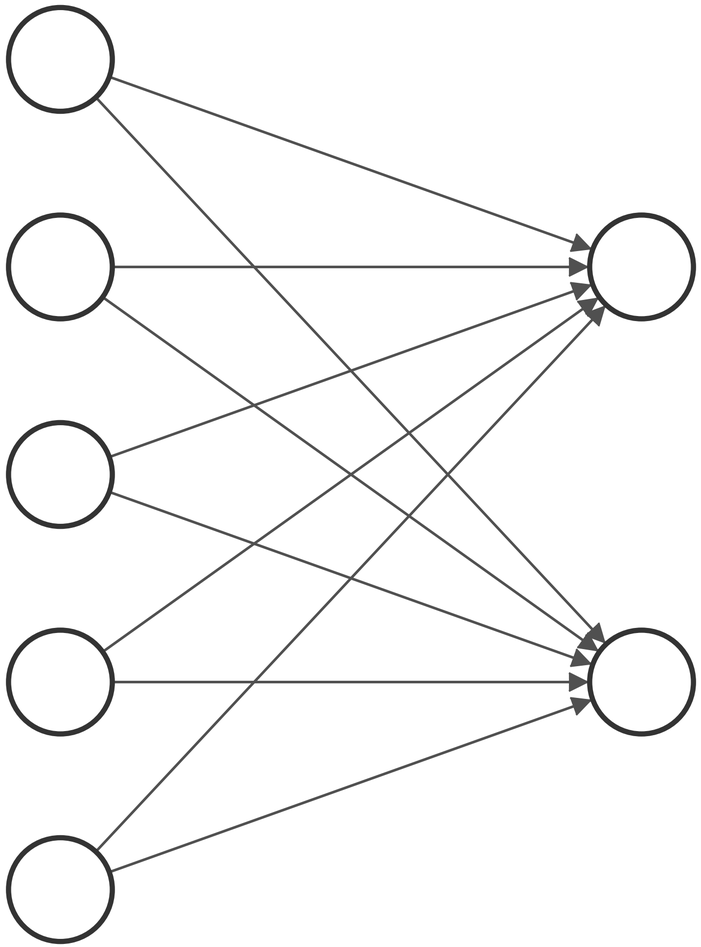
$\mathbf{x}$




$\mathbf{W}_1$
$\mathbf{x}$


$\mathbf{W}_1$
$\mathbf{x}$
$\mathbf{W}_2$


$\mathbf{W}_1$
$\mathbf{x}$
Σ
Σ
$\mathbf{W}_2$



$$\mathbf{x}^\top \mathbf{w} = \begin{bmatrix} x_1 & x_2 & \dots & x_d \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

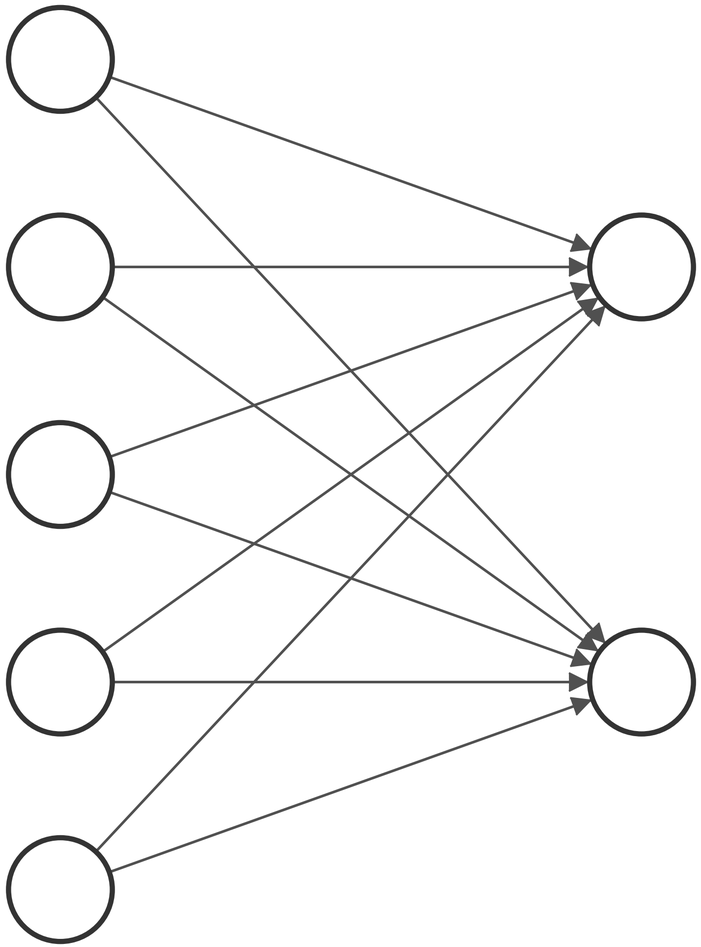
$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 \end{bmatrix}$$

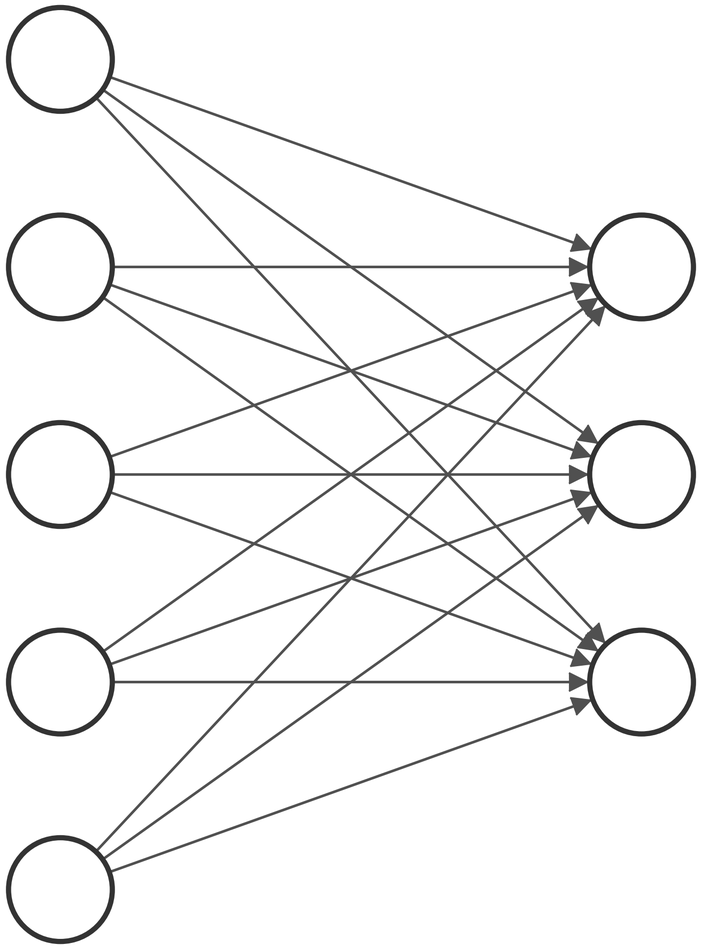
$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 \end{bmatrix} = \begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \\ \vdots & \vdots \\ w_{d,1} & w_{d,2} \end{bmatrix}$$

$$\mathbf{x}^\top \mathbf{W} = \begin{bmatrix} x_1 & x_2 & \dots & x_d \end{bmatrix} \begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \\ \vdots & \vdots \\ w_{d,1} & w_{d,2} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{x} \cdot \mathbf{w}_1 & \mathbf{x} \cdot \mathbf{w}_2 \end{bmatrix}$$



d



k

$$\mathbf{x}^\top \mathbf{W} = \begin{bmatrix} x_1 & x_2 & \dots & x_d \end{bmatrix} \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,k} \\ w_{2,1} & w_{2,2} & \dots & w_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ w_{d,1} & w_{d,2} & \dots & w_{d,k} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{x} \cdot \mathbf{w}_1 & \mathbf{x} \cdot \mathbf{w}_2 & \dots & \mathbf{x} \cdot \mathbf{w}_k \end{bmatrix}$$

↓ Feature dimensions ↓

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{bmatrix}$$

↑ Samples ↑

$$\mathbf{XW} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,d} \\ x_{2,1} & x_{2,2} & \dots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,d} \end{bmatrix} \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,k} \\ w_{2,1} & w_{2,2} & \dots & w_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ w_{d,1} & w_{d,2} & \dots & w_{d,k} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{x}_1 \cdot \mathbf{w}_1 & \mathbf{x}_1 \cdot \mathbf{w}_2 & \dots & \mathbf{x}_1 \cdot \mathbf{w}_k \\ \mathbf{x}_2 \cdot \mathbf{w}_1 & \mathbf{x}_2 \cdot \mathbf{w}_2 & \dots & \mathbf{x}_2 \cdot \mathbf{w}_k \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n \cdot \mathbf{w}_1 & \mathbf{x}_n \cdot \mathbf{w}_2 & \dots & \mathbf{x}_n \cdot \mathbf{w}_k \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{y}}_1^\top \\ \hat{\mathbf{y}}_2^\top \\ \vdots \\ \hat{\mathbf{y}}_n^\top \end{bmatrix} = \hat{\mathbf{Y}}$$

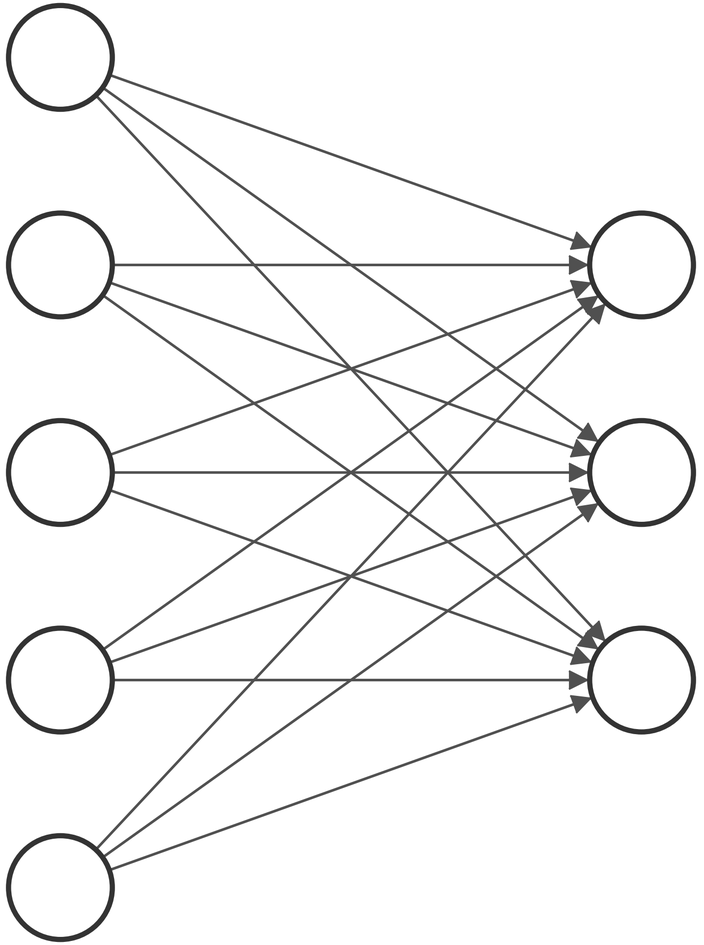
$$\mathbf{X} \in \mathbb{R}^{n \times d}$$

$$\mathbf{W} \in \mathbb{R}^{d \times k}$$

$$\hat{\mathbf{Y}} \in \mathbb{R}^{n \times k}$$

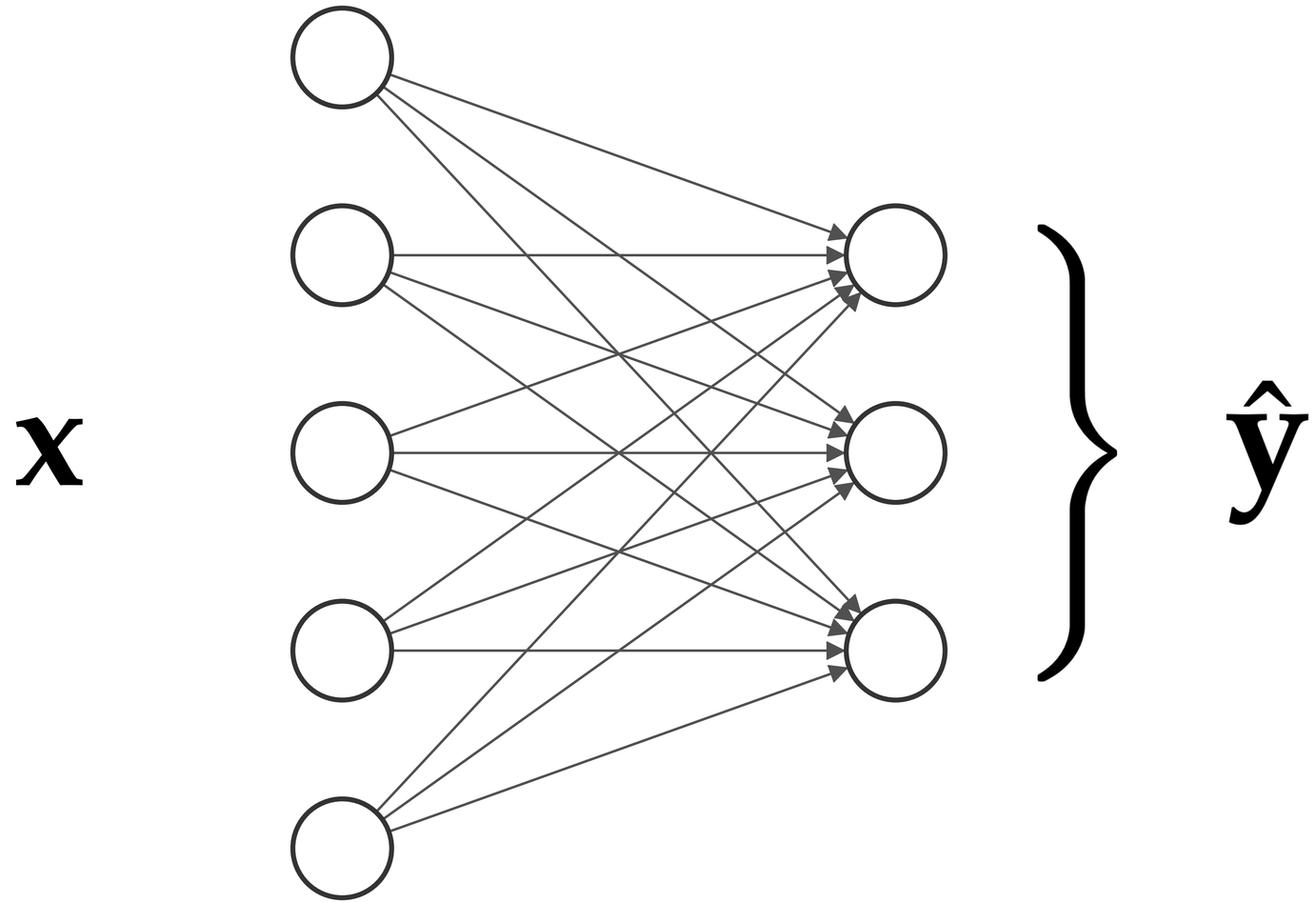
d

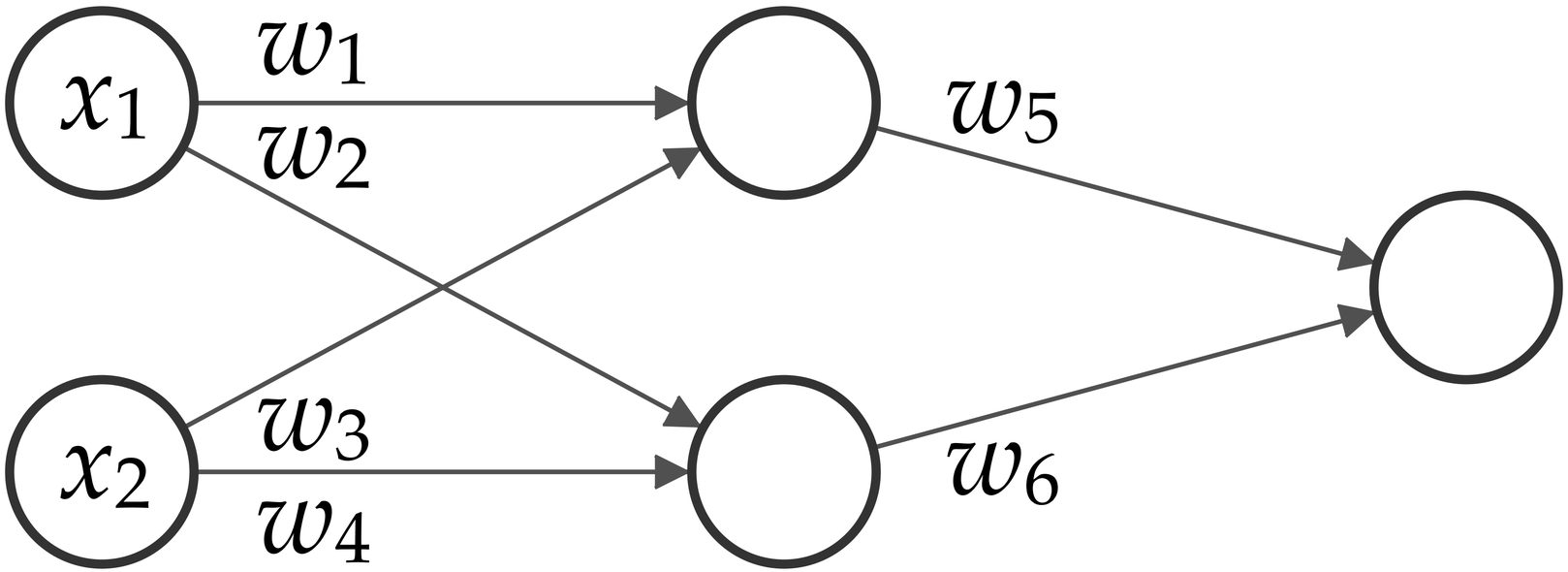
fan_in



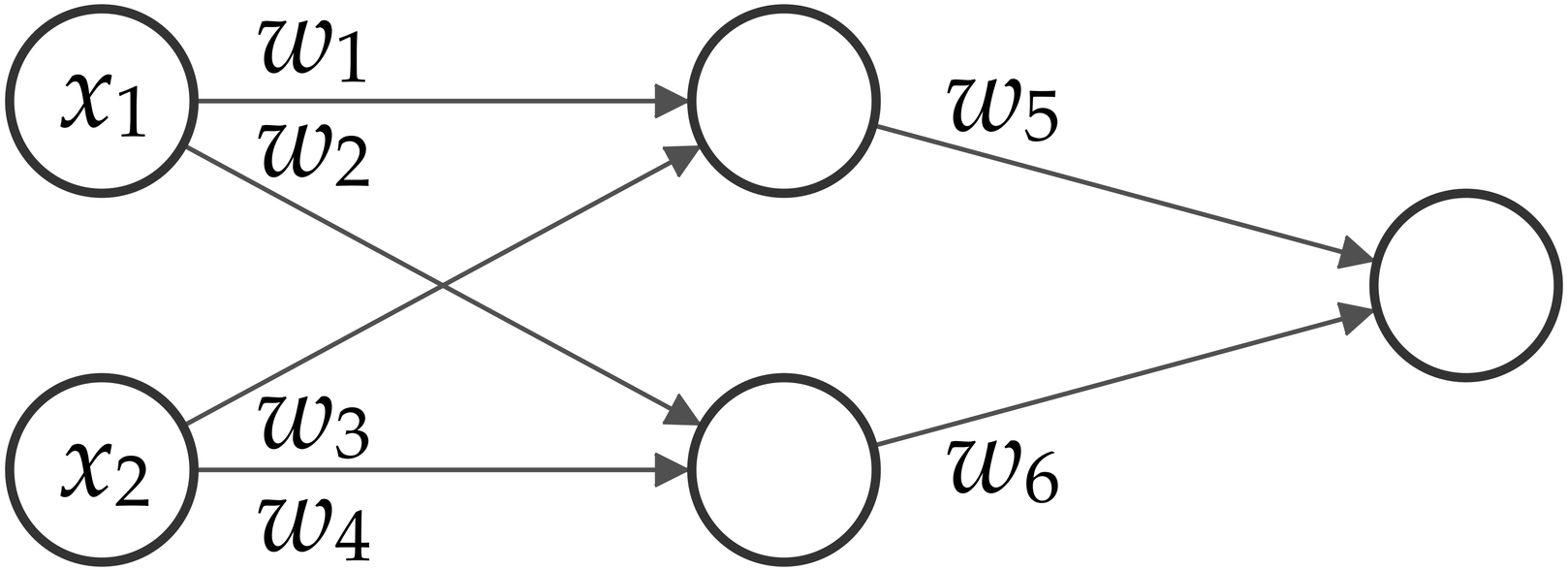
k

fan_out


x
ŷ


x_1
x_2
w_1
w_2
w_3
w_4
w_5
w_6

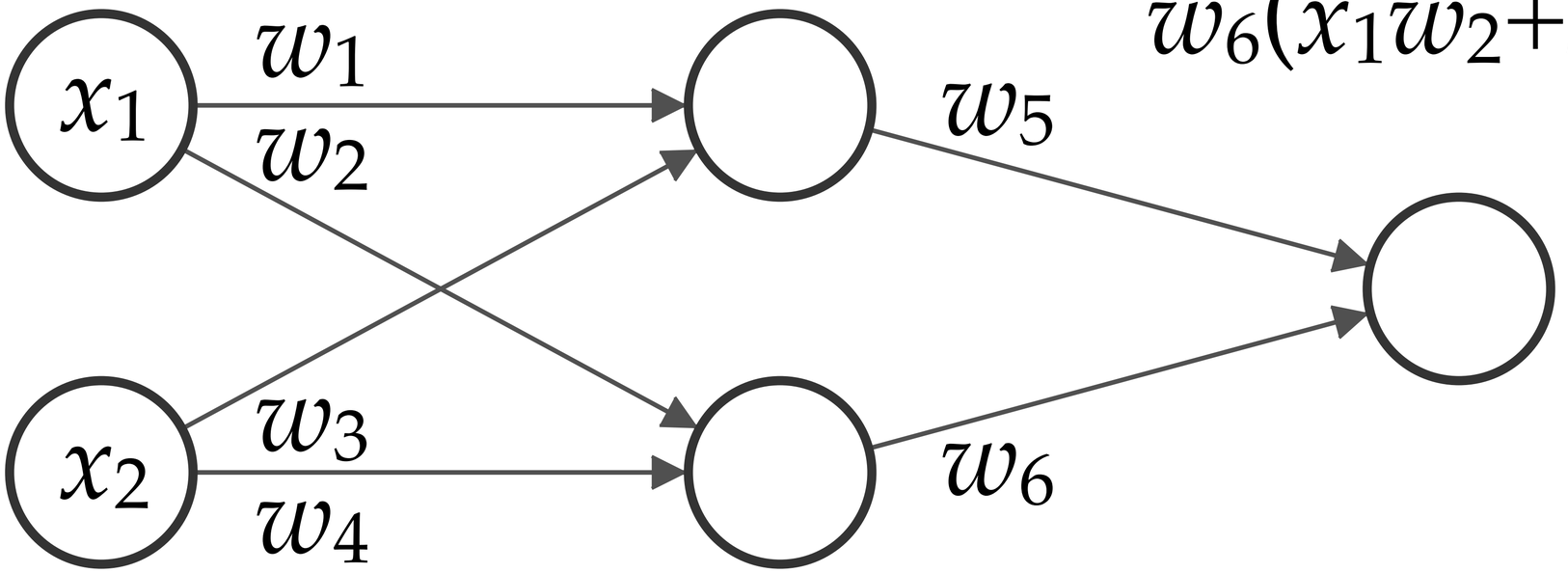
$x_1 w_1 + x_2 w_3$



$x_1 w_2 + x_2 w_4$

$$w_5(x_1w_1+x_2w_3)$$
$$+$$
$$w_6(x_1w_2+x_2w_4)$$

$$x_1w_1+x_2w_3$$



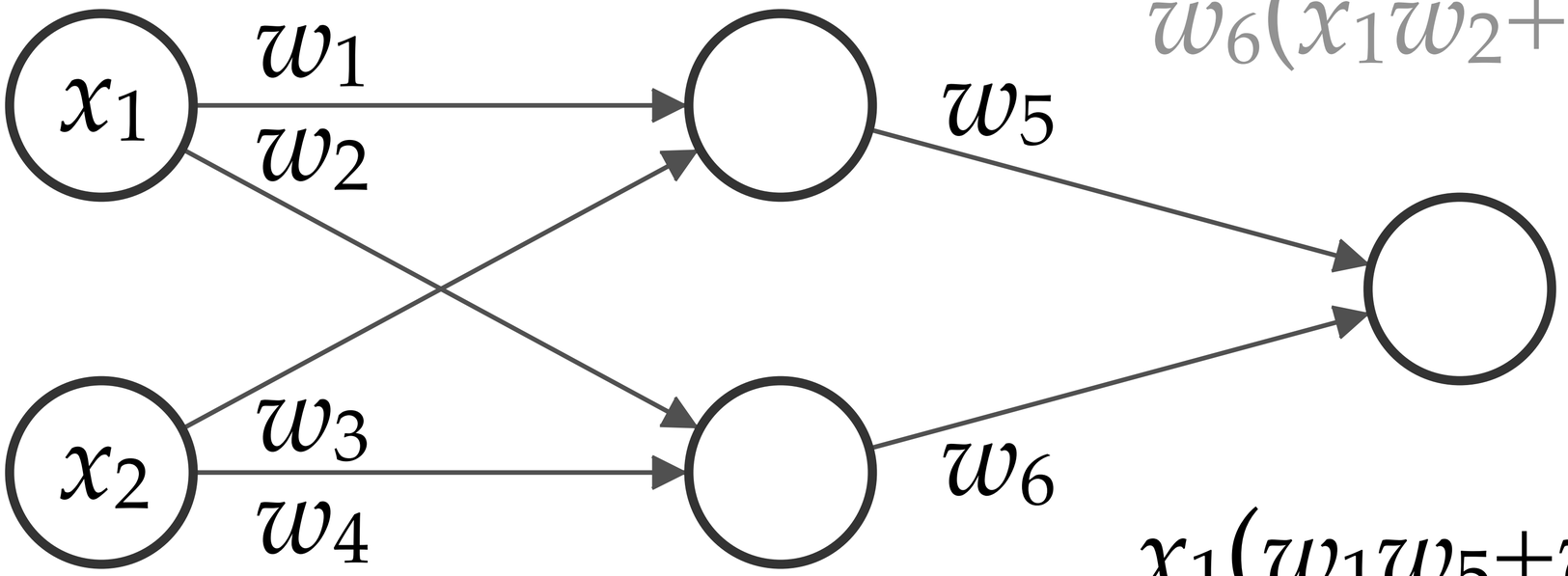
$$x_1w_2+x_2w_4$$

$w_5(x_1w_1+x_2w_3)$

$+$

$w_6(x_1w_2+x_2w_4)$

$x_1w_1+x_2w_3$

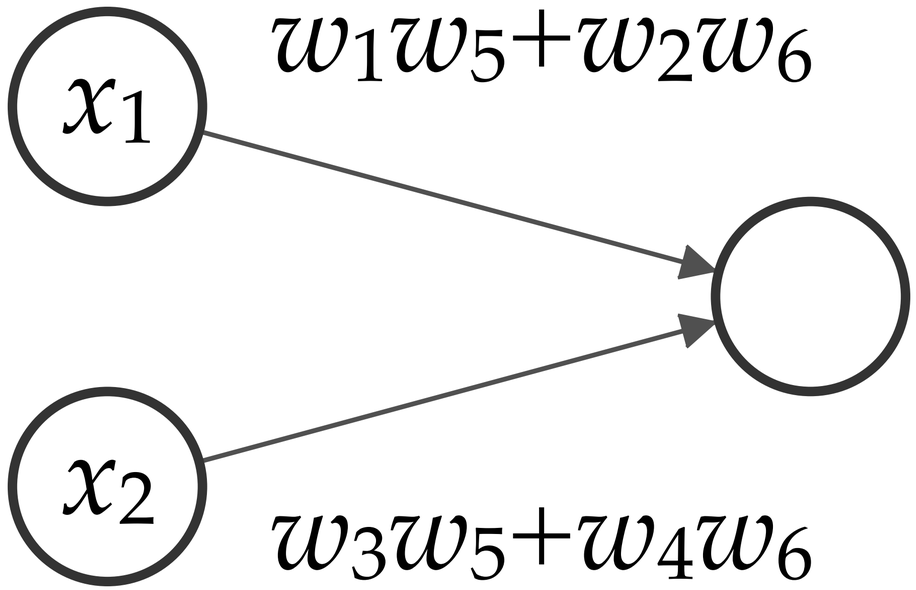


$x_1w_2+x_2w_4$

$x_1(w_1w_5+w_2w_6)$

$+$

$x_2(w_3w_5+w_4w_6)$


$w_1w_5+w_2w_6$
x_1
x_2
$w_3w_5+w_4w_6$

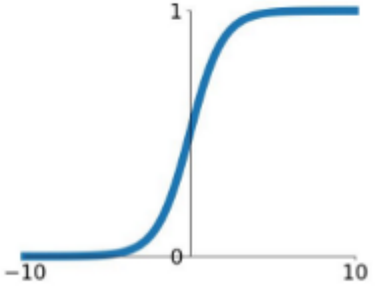
$$\begin{aligned}\hat{\mathbf{y}} &= (((((\mathbf{x}\mathbf{W}_1)\mathbf{W}_2)\ldots)\mathbf{W}_k) \\ &= \mathbf{x}(\mathbf{W}_1\mathbf{W}_2\ldots\mathbf{W}_k) \\ &\equiv \mathbf{x}\mathbf{W}\end{aligned}$$

$$\text{fire} = \begin{cases} 1 & \text{if } \mathbf{x} \cdot \mathbf{w} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{\mathbf{y}} = a_k(...a_2(a_1(\mathbf{x}\mathbf{W}_1)\mathbf{W}_2)...\mathbf{W}_k)$$

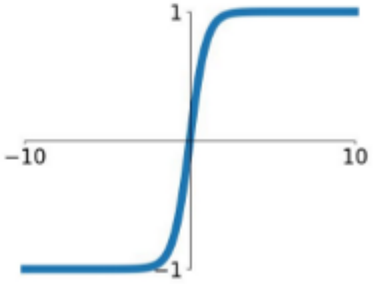
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



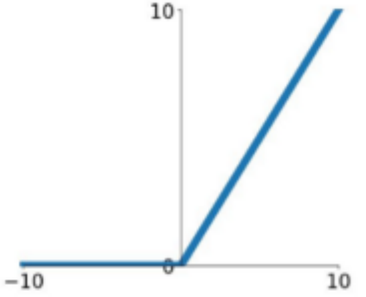
tanh

$$\tanh(x)$$



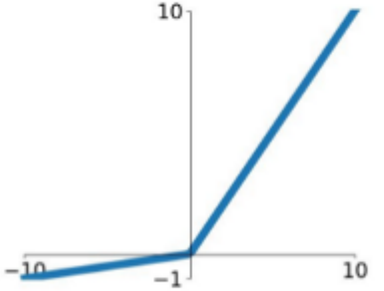
ReLU

$$\max(0, x)$$



Leaky ReLU

$$\max(0.1x, x)$$

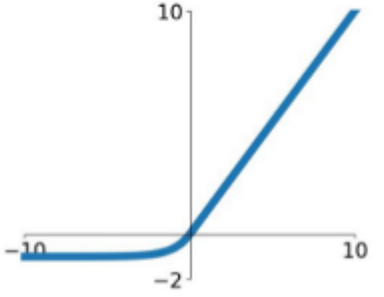


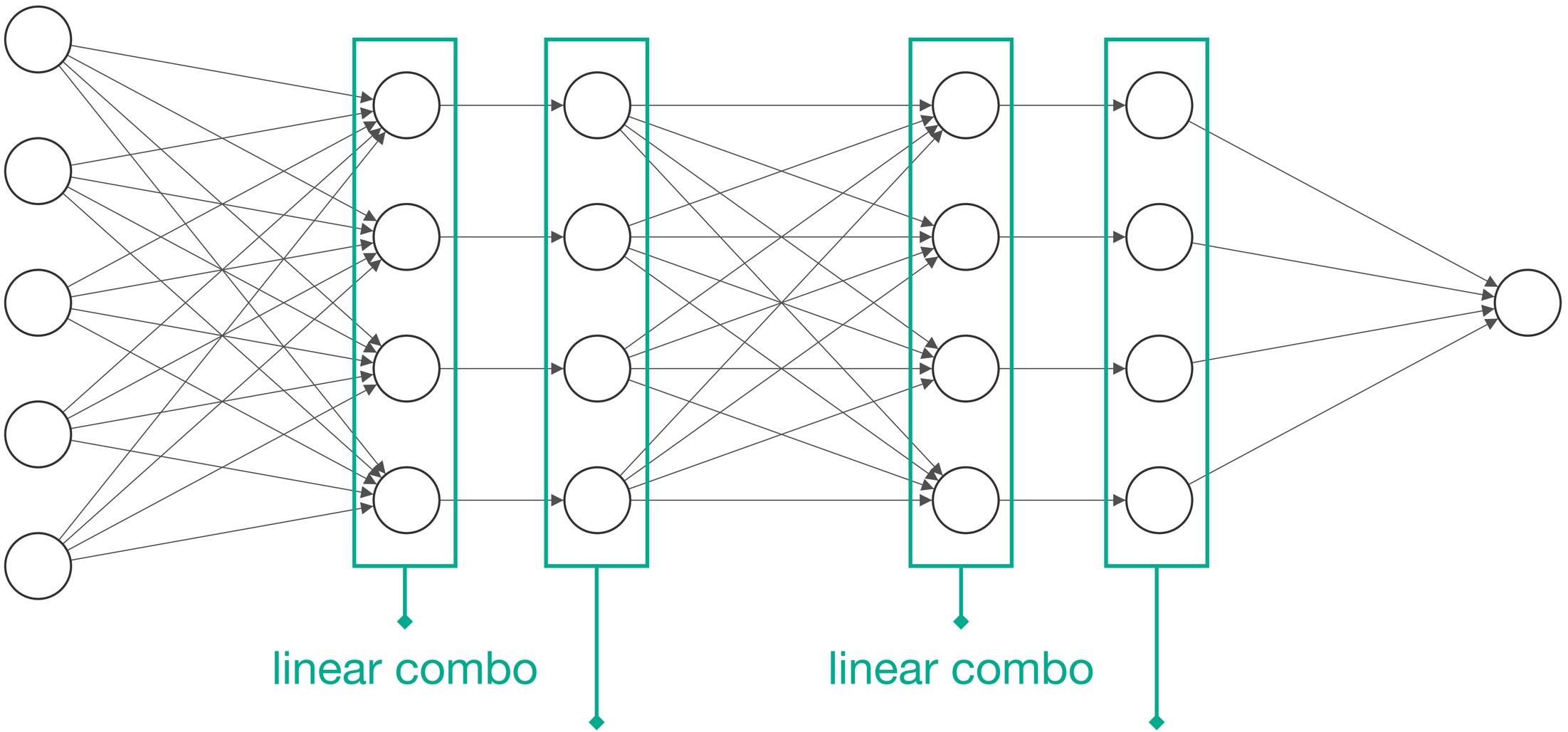
Maxout

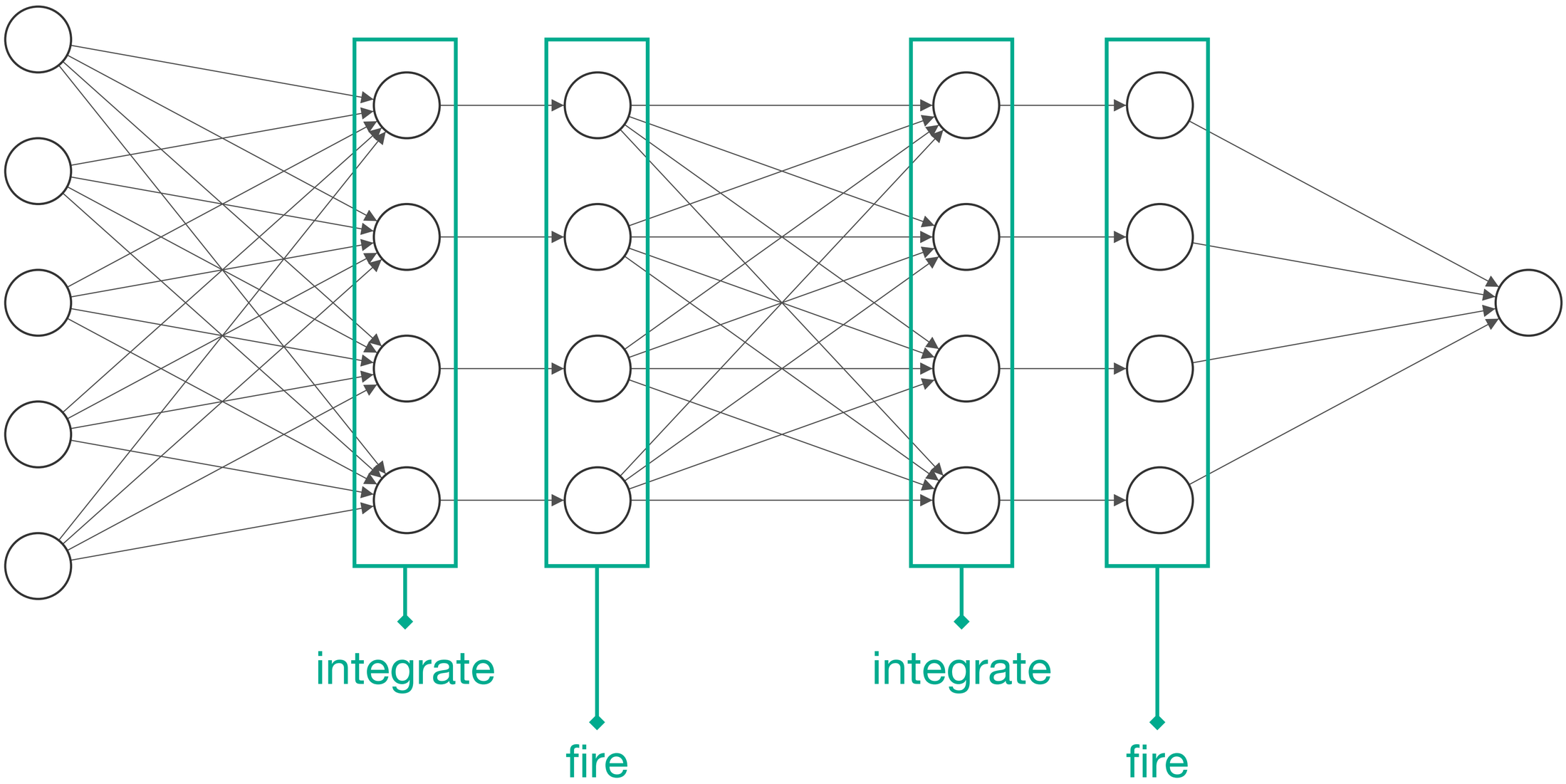
$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

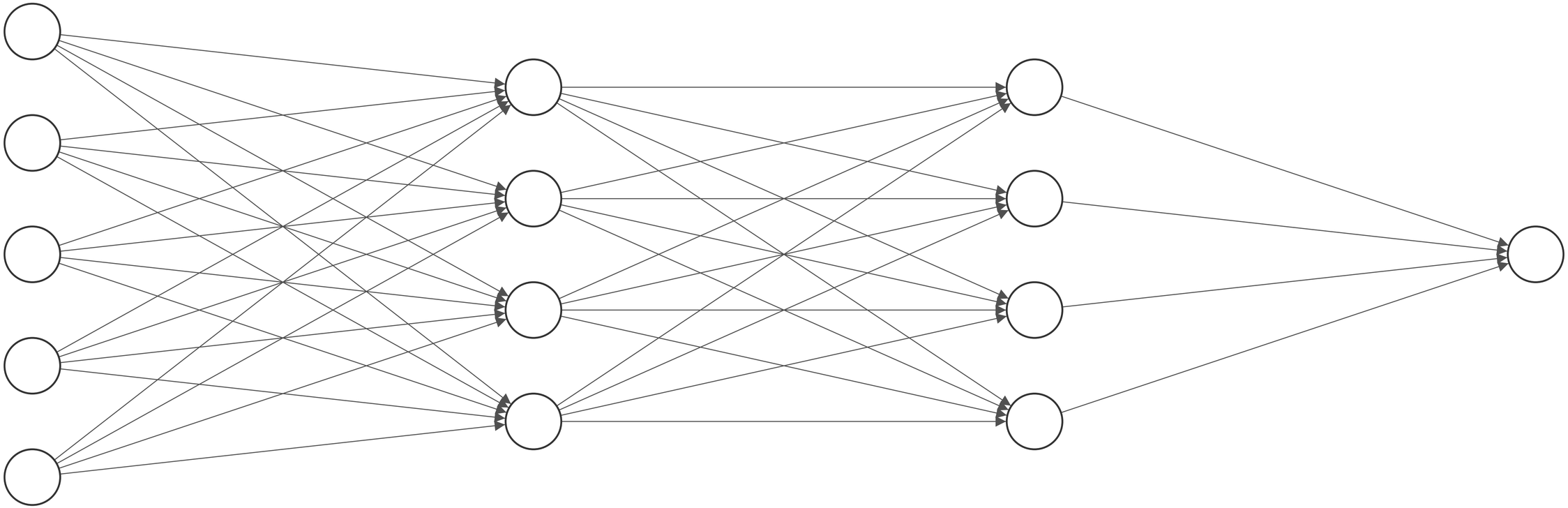
ELU

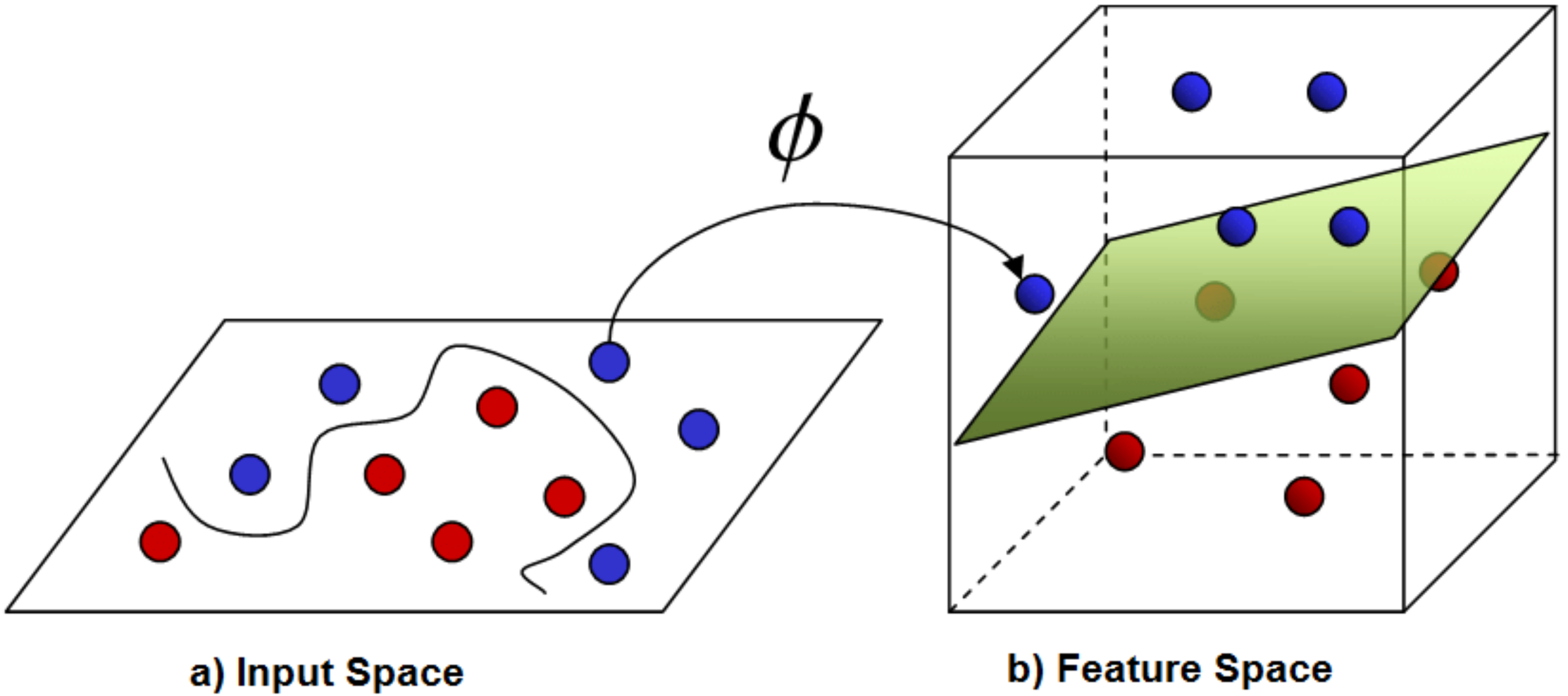
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$

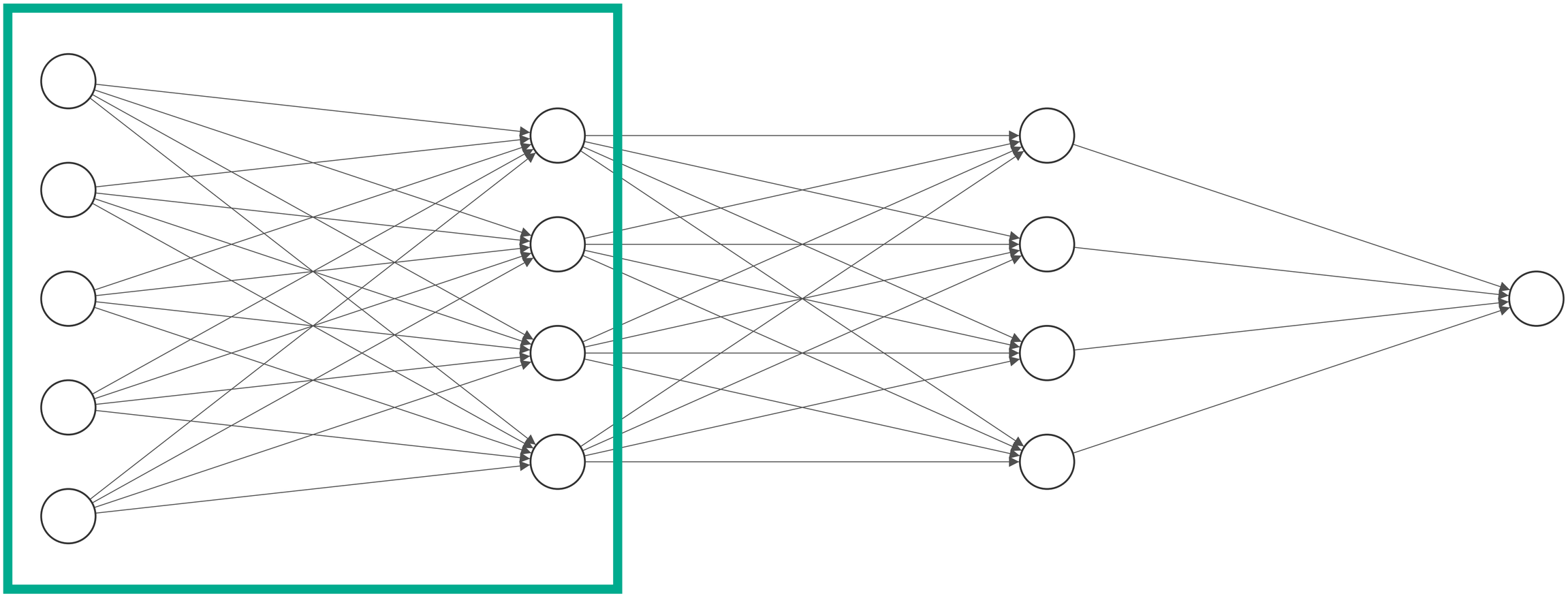


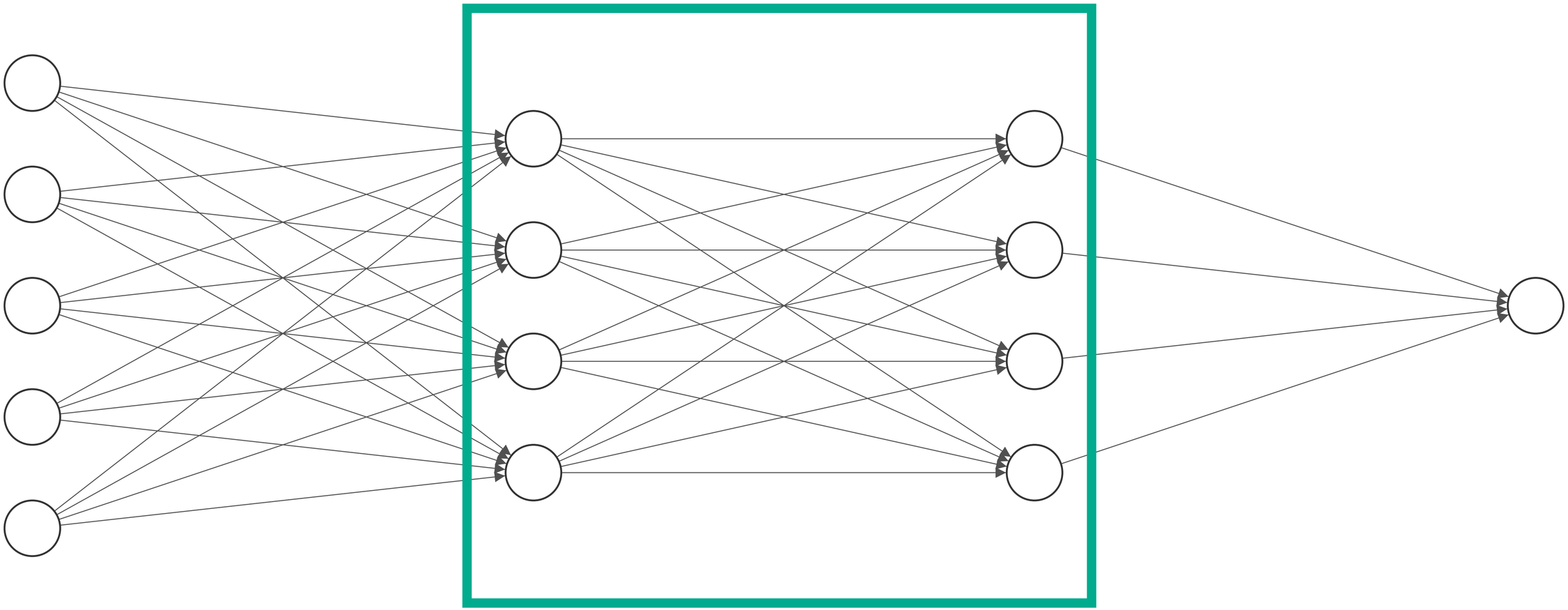

linear combo
non-linear activation
linear combo
non-linear activation

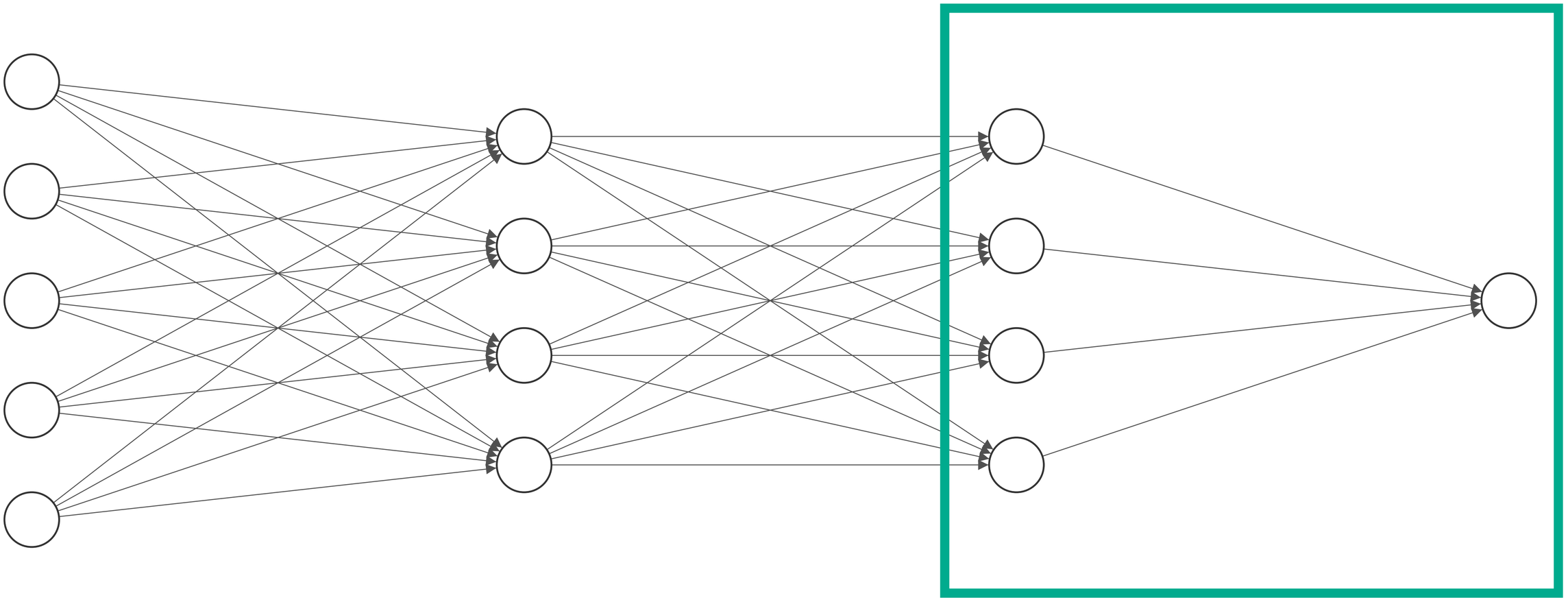

integrate
fire
integrate
fire

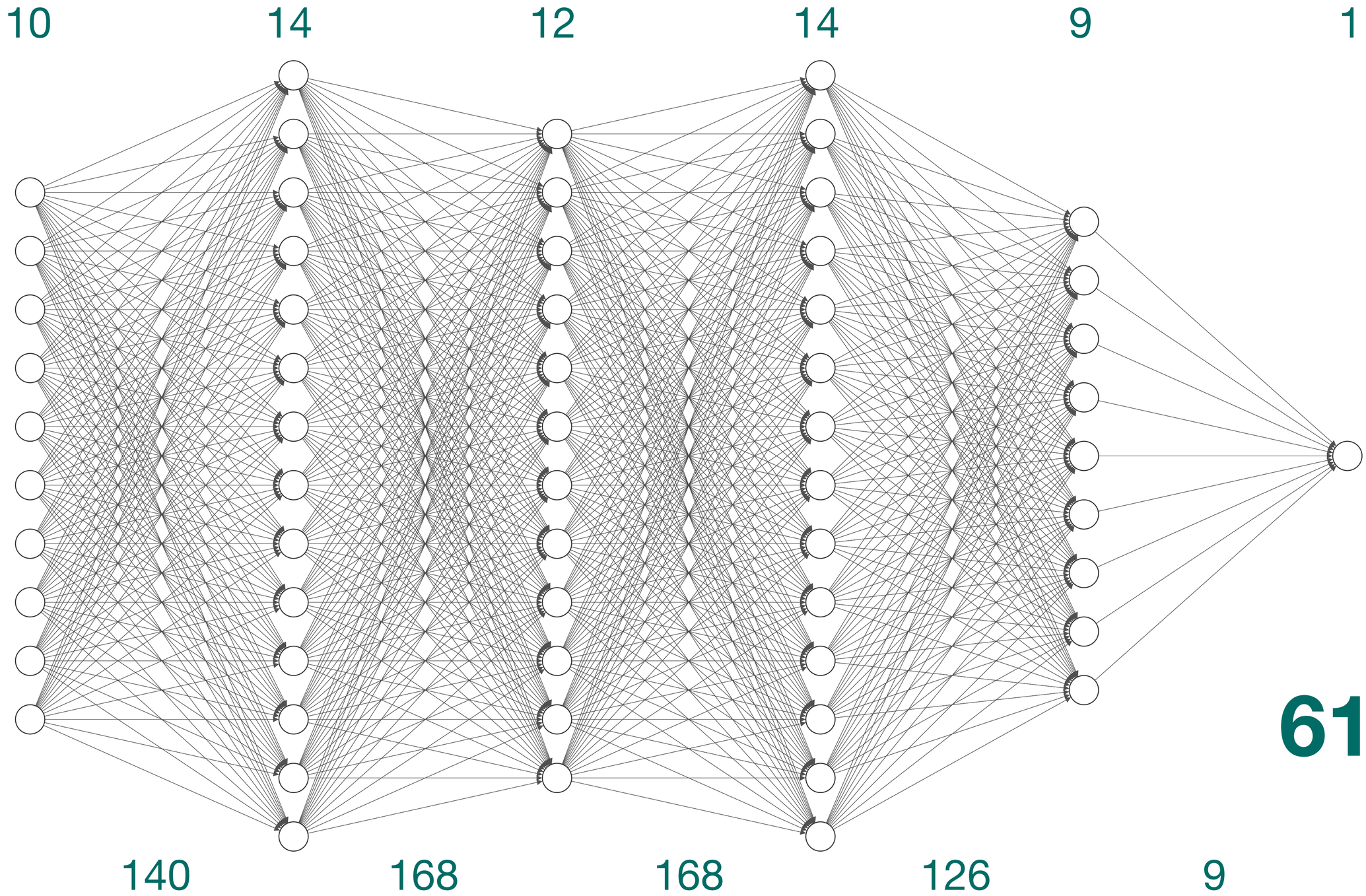


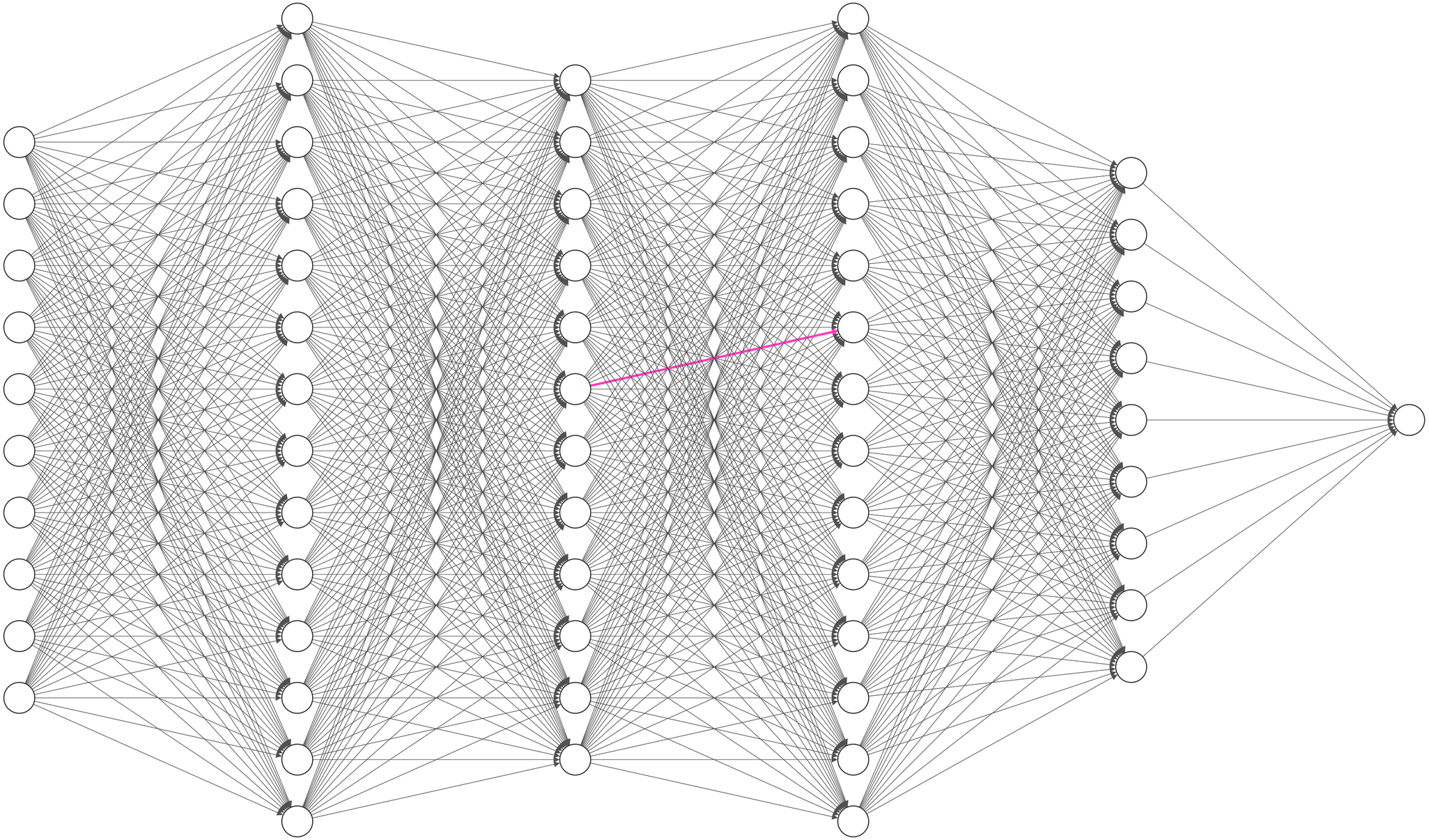

ϕ
a) Input Space
b) Feature Space








10
14
12
14
9
1
140
168
168
126
9
611



analytic vs numerical

analytic optimisation

$$\mathbf{w}^\star = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

numerical optimisation

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \nabla_{\mathbf{w}} L(\mathbf{f}, \mathbf{X}, \mathbf{y}, \mathbf{w})$$

$$\nabla_{\mathbf{w}} L = \begin{bmatrix} \frac{\partial L}{\partial w_1} \\ \frac{\partial L}{\partial w_2} \\ \vdots \\ \frac{\partial L}{\partial w_m} \end{bmatrix}$$

$$\frac{\partial L}{\partial w_i} = \lim_{\delta \to 0} \frac{L(..., w_i + \delta, ...) - L(..., w_i, ...)}{\delta}$$

numerical gradient

$$\frac{\partial L}{\partial w_i} \approx \frac{L(..., w_i + \delta, ...) - L(..., w_i, ...)}{\delta}$$

analytic gradient

$$
\begin{aligned}
\frac{\partial}{\partial w_j} l(\mathbf{w}) &= \frac{\partial}{\partial w_j}\left[y \log(\sigma(\mathbf{x}\cdot\mathbf{w})) + (1-y)\log(1-\sigma(\mathbf{x}\cdot\mathbf{w}))\right] \\
&= \left[y \frac{1}{\sigma(\mathbf{x}\cdot\mathbf{w})} \frac{\partial}{\partial w_j}\sigma(\mathbf{x}\cdot\mathbf{w})\right] + \left[(1-y)\frac{1}{1-\sigma(\mathbf{x}\cdot\mathbf{w})}\left(-\frac{\partial}{\partial w_j}\sigma(\mathbf{x}\cdot\mathbf{w})\right)\right] \\
&= \left[y \frac{1}{\sigma(\mathbf{x}\cdot\mathbf{w})} \frac{\partial}{\partial w_j}\sigma(\mathbf{x}\cdot\mathbf{w})\right] - \left[(1-y)\frac{1}{1-\sigma(\mathbf{x}\cdot\mathbf{w})} \frac{\partial}{\partial w_j}\sigma(\mathbf{x}\cdot\mathbf{w})\right] \\
&= \left[y \frac{1}{\sigma(\mathbf{x}\cdot\mathbf{w})} - (1-y)\frac{1}{1-\sigma(\mathbf{x}\cdot\mathbf{w})}\right] \frac{\partial}{\partial w_j}\sigma(\mathbf{x}\cdot\mathbf{w}) \\
&= \left[y \frac{1}{\sigma(\mathbf{x}\cdot\mathbf{w})} - (1-y)\frac{1}{1-\sigma(\mathbf{x}\cdot\mathbf{w})}\right] \sigma(\mathbf{x}\cdot\mathbf{w})(1-\sigma(\mathbf{x}\cdot\mathbf{w}))\frac{\partial}{\partial w_j}(\mathbf{x}\cdot\mathbf{w}) \\
&= (y(1-\sigma(\mathbf{x}\cdot\mathbf{w})) - (1-y)\sigma(\mathbf{x}\cdot\mathbf{w}))\cdot x_j \\
&= (y - y\sigma(\mathbf{x}\cdot\mathbf{w}) - \sigma(\mathbf{x}\cdot\mathbf{w}) + y\sigma(\mathbf{x}\cdot\mathbf{w}))\cdot x_j \\
&= (y - \sigma(\mathbf{x}\cdot\mathbf{w}))\cdot x_j \\
&= (y - \hat{y})x_j
\end{aligned}
$$

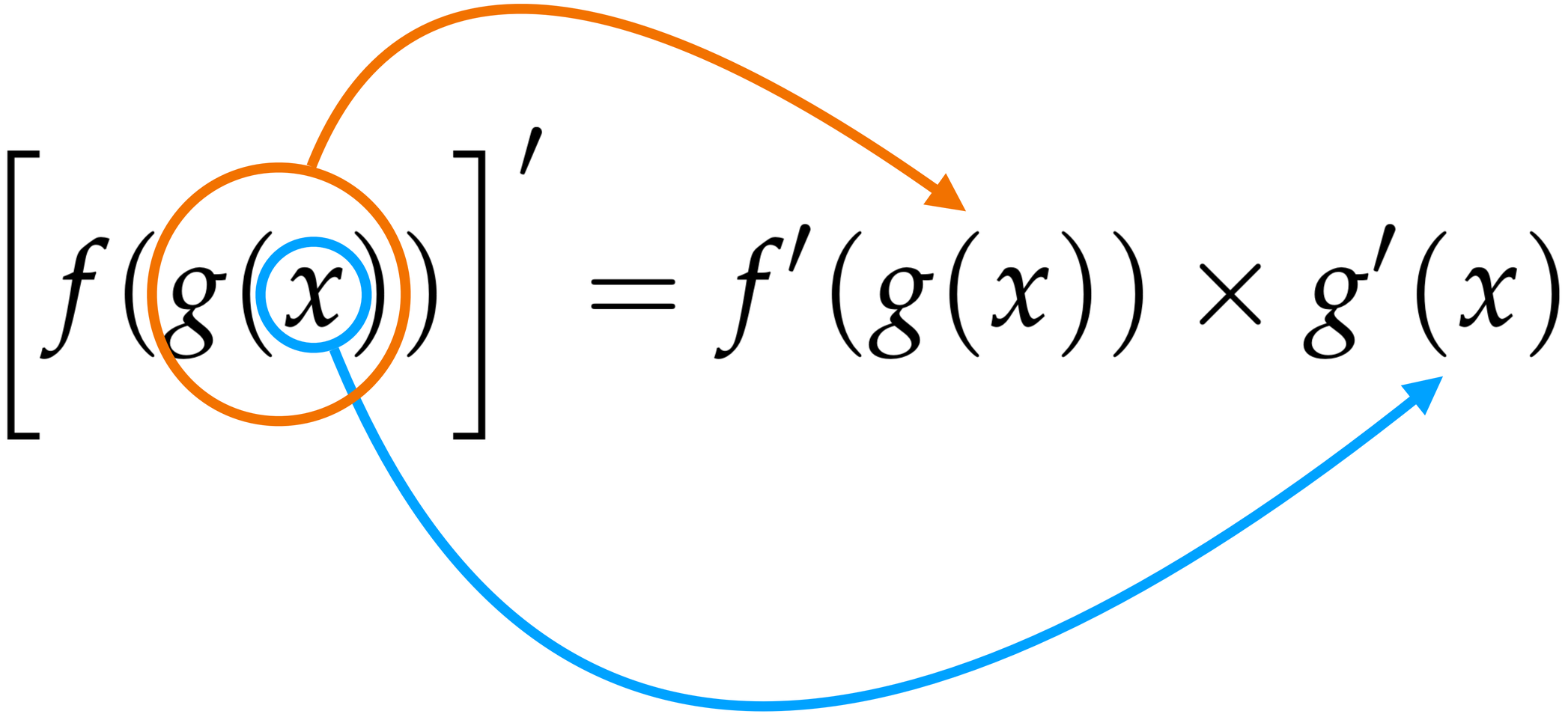
the chain rule

$$\frac{dy}{dx} = \frac{dy}{du} \times \frac{du}{dx}$$

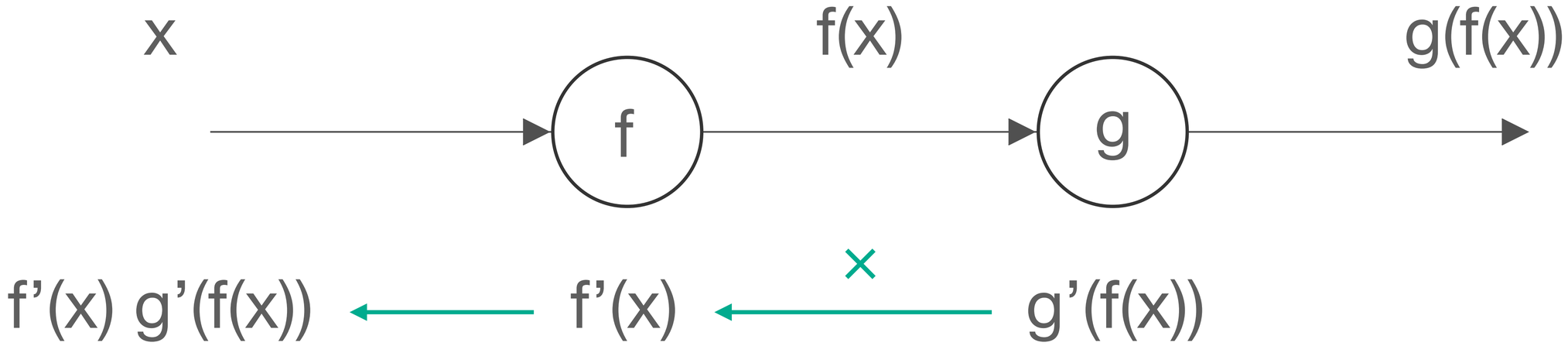
$$(f \circ g)' = (f' \circ g) \times g'$$

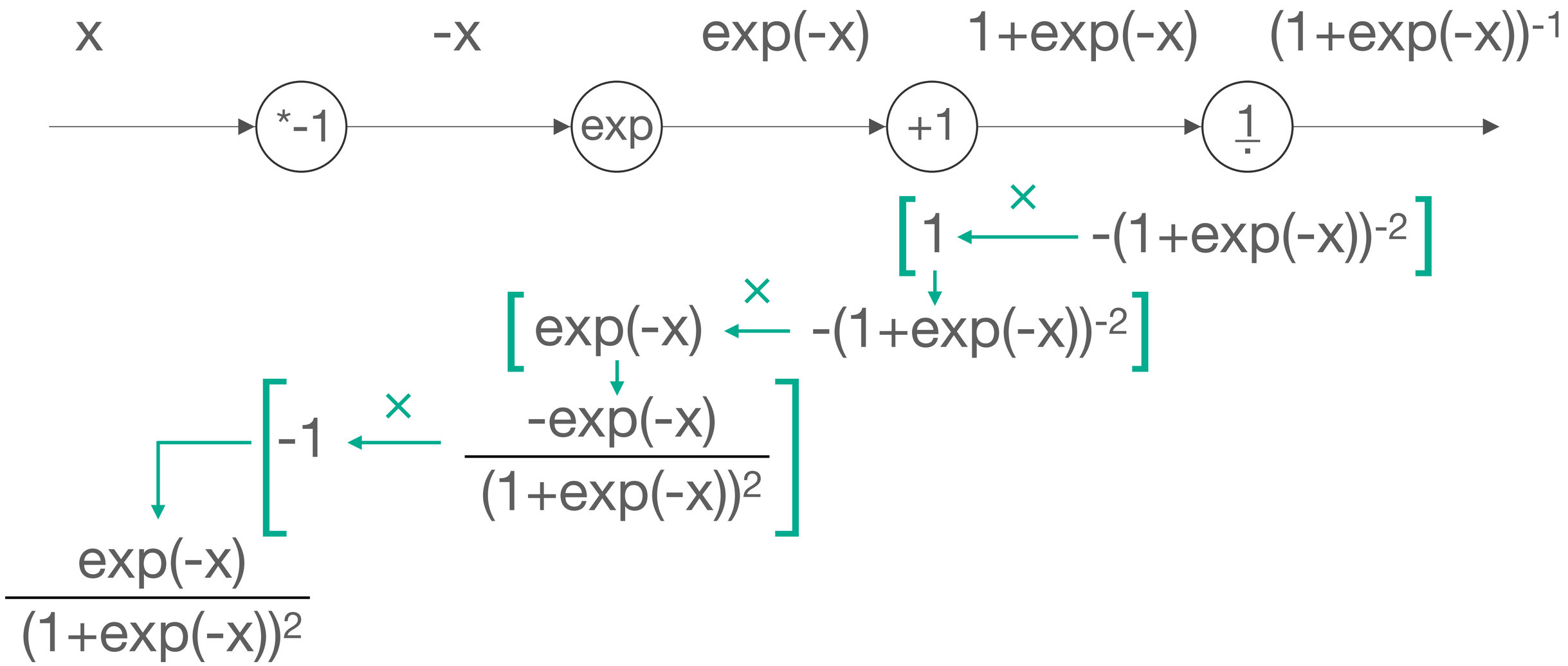
the chain rule

gradients multiply

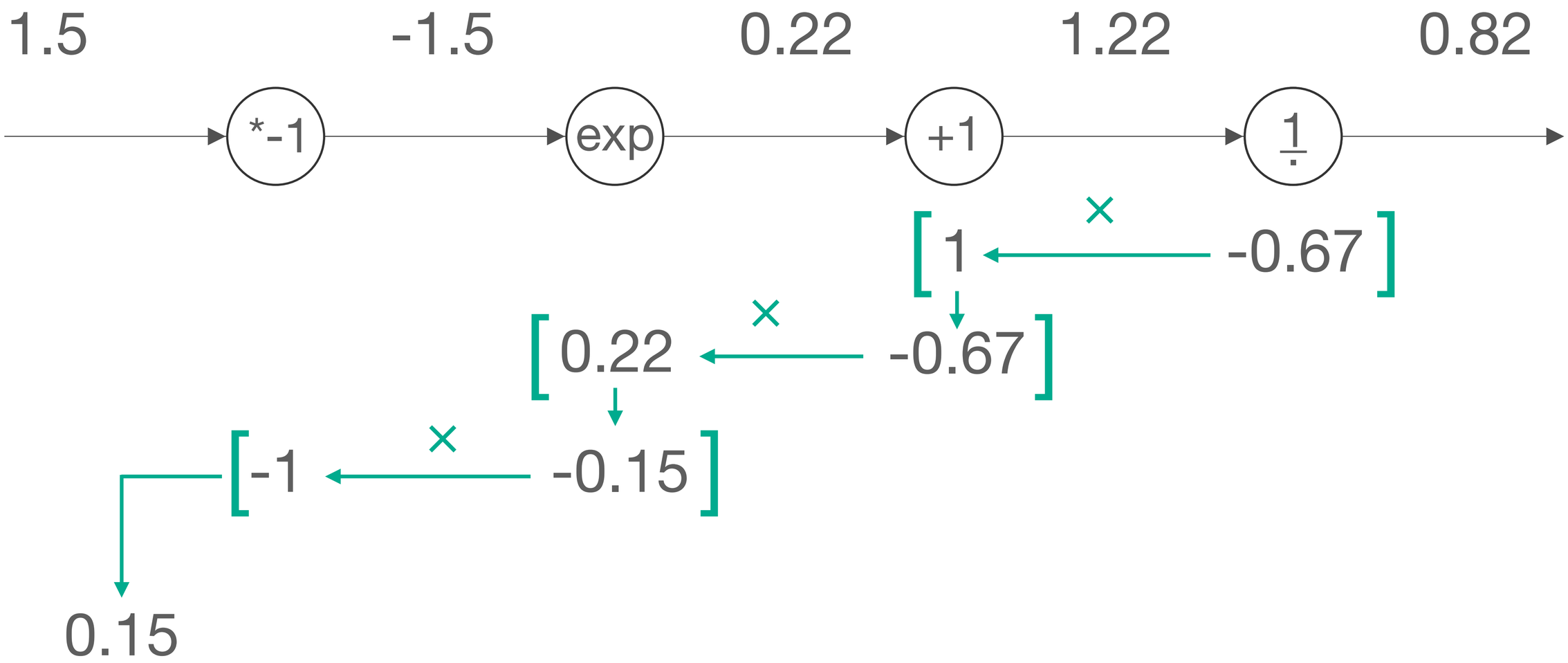

$\left[f(g(x)) \right]' = f'(g(x)) \times g'(x)$

$$\Big[f(g(h(x)))\Big]' = f'(g(h(x))) \times g'(h(x)) \times h'(x)$$


x
f(x)
g(f(x))
f
g
×
f’(x) g’(f(x))
f’(x)
g’(f(x))


x
-x
exp(-x)
1+exp(-x)
(1+exp(-x))^-1
*-1
exp
+1
1/·
1
×
-(1+exp(-x))^-2
exp(-x)
-1
-exp(-x)
(1+exp(-x))^2
exp(-x)
(1+exp(-x))^2

gradients are just numbers!


1.5
-1.5
0.22
1.22
0.82
*-1
exp
+1
1
1
-0.67
0.22
-0.67
-1
-0.15
0.15

```
out1 = f1(x)
out2 = f2(out1)
out3 = f3(out2)
out4 = f4(out3)
loss = f5(out4)

dloss_out4 = df5(out4)
dloss_out3 = df4(out3) * dloss_out4
dloss_out2 = df3(out2) * dloss_out3
dloss_out1 = df2(out1) * dloss_out2
dloss_x = df1(x) * dloss_out1
```

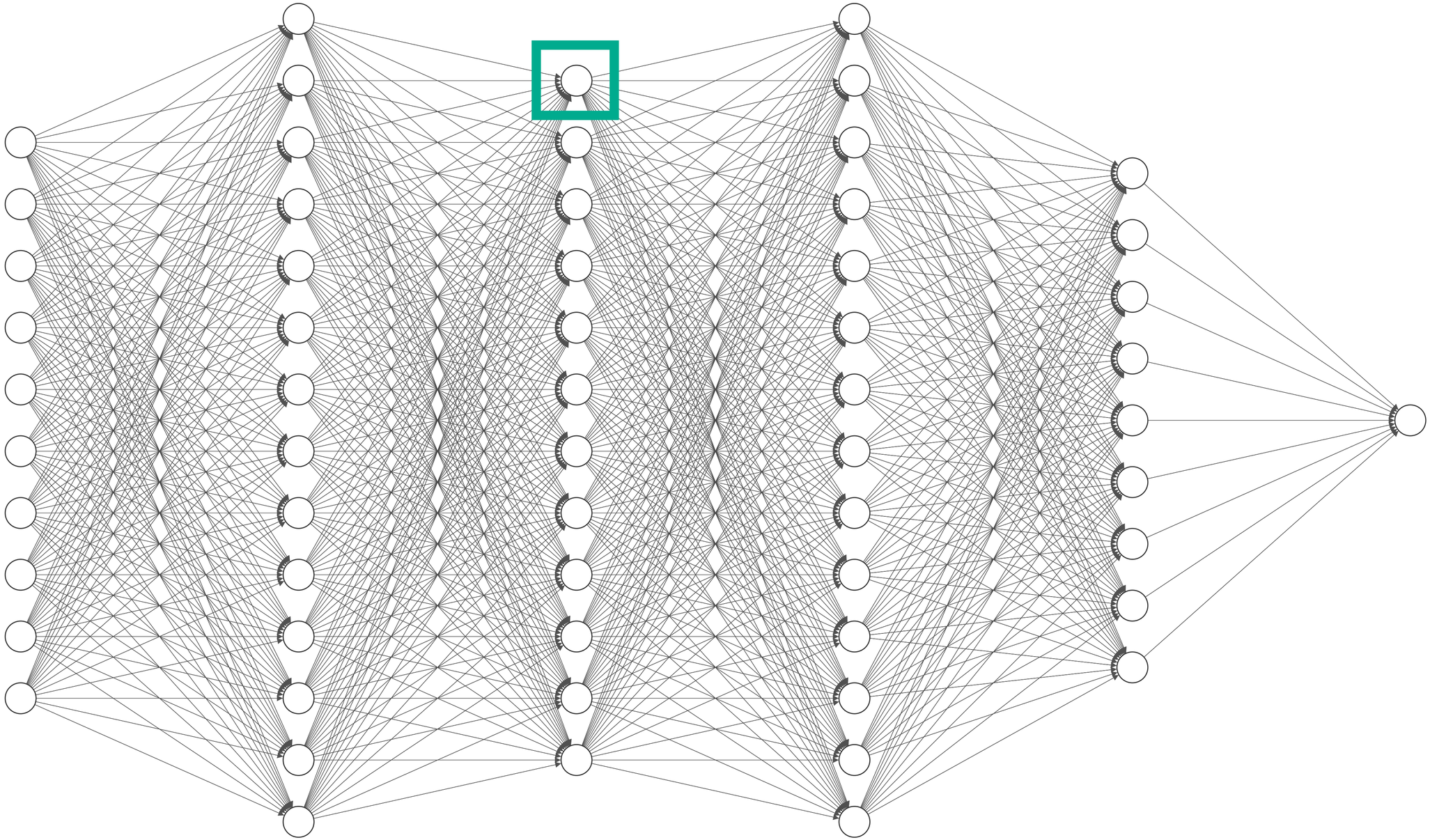
```
out1 = f1(x)
out2 = f2(out1)
out3 = f3(out2)
out4 = f4(out3)
loss = f5(out4)


dloss_out4 = df5(out4)
dloss_out3 = df4(out3) * dloss_out4
dloss_out2 = df3(out2) * dloss_out3
dloss_out1 = df2(out1) * dloss_out2
dloss_x = df1(x) * dloss_out1
```

local gradient

downstream gradient

analytic gradients, numerically

backpropagation



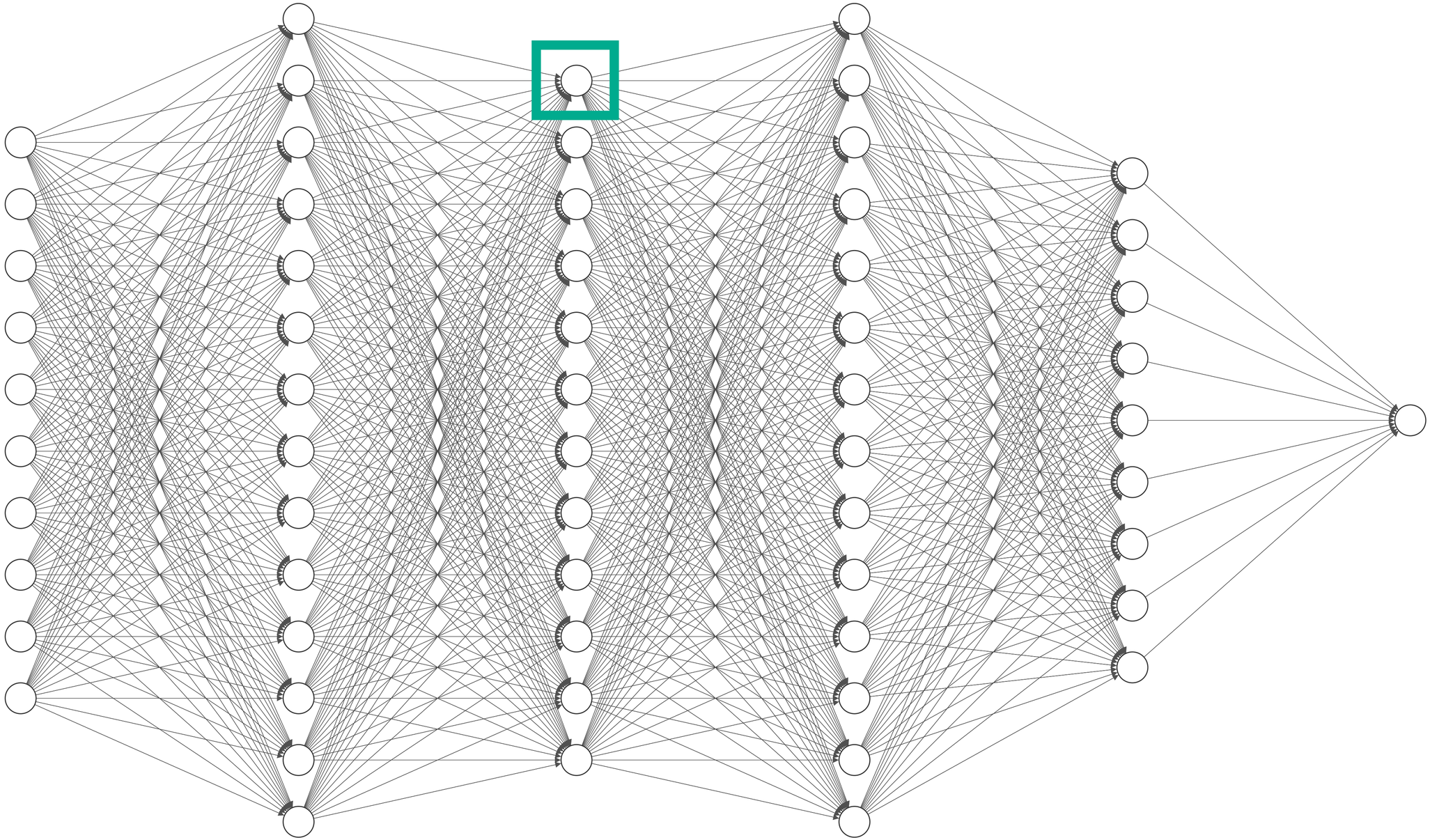
$$z = \mathbf{x} \cdot \mathbf{w}$$

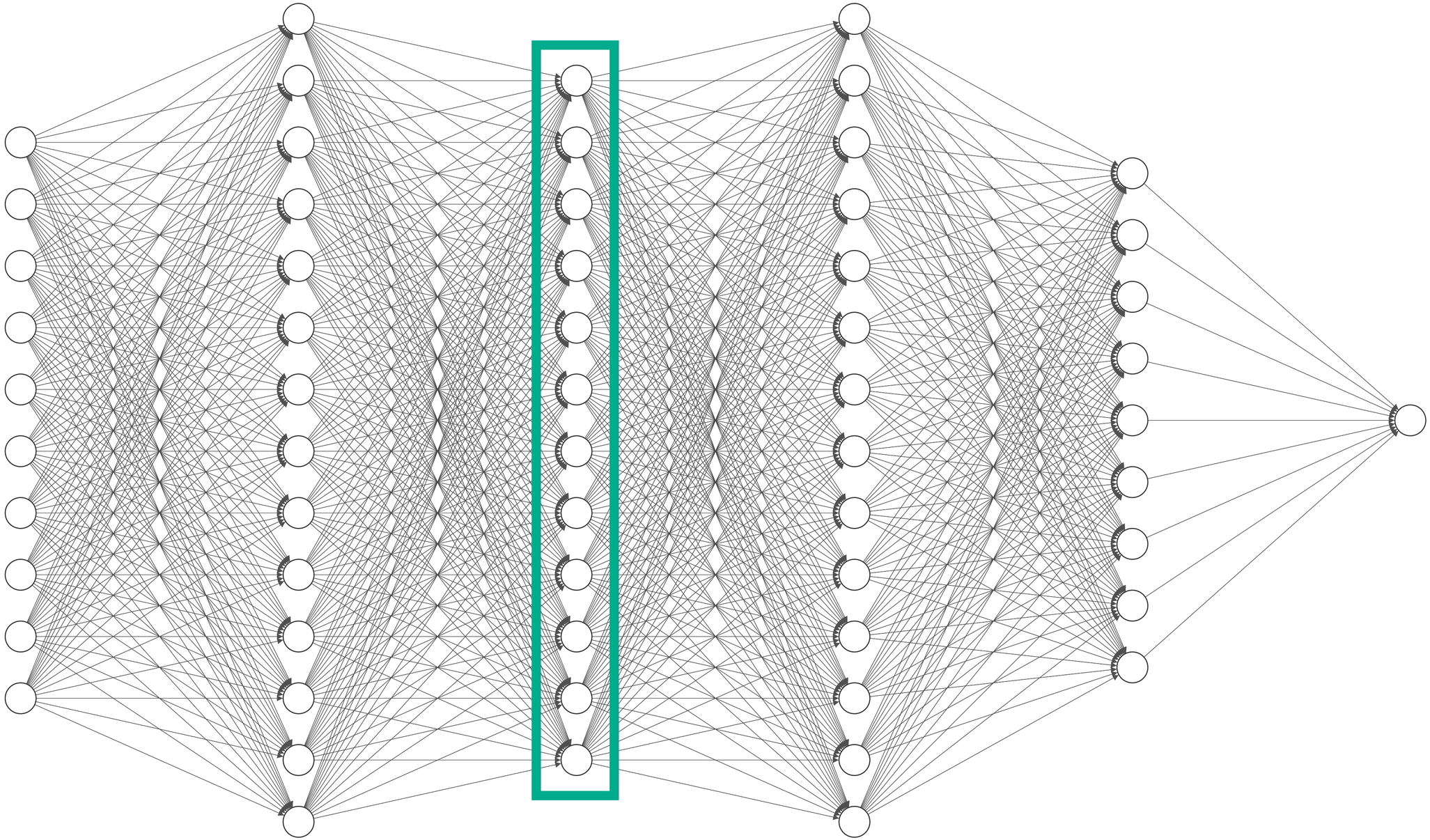
$$\nabla_{\mathbf{x}} z = \mathbf{w}$$

$$\nabla_{\mathbf{w}} z = \mathbf{x}$$

$$a(\mathbf{x}) = [f(x_1), f(x_2), \ldots, f(x_d)]$$

$$\nabla_{\mathbf{x}} a = [f'(x_1), f'(x_2), \ldots, f'(x_d)]$$





forward pass

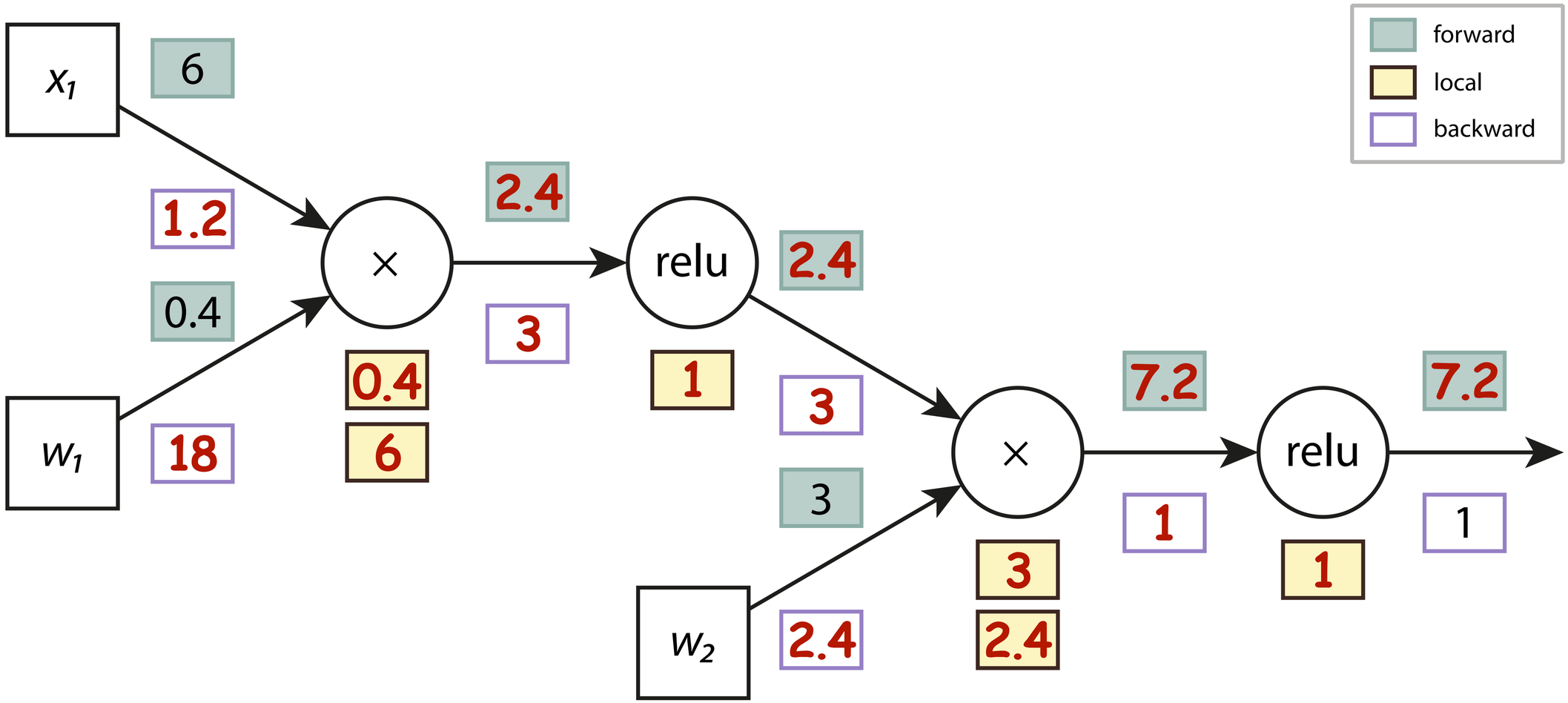
evaluate model on data
calculate local gradients
save ins and outs
calculate final loss

backward pass

calculate final loss gradient
propagate gradients backward
multiply downstream by local

optimise

update weights with gradients


forward
local
backward
x_1
6
1.2
×
2.4
relu
2.4
0.4
3
0.4
1
3
w_1
18
6
×
7.2
relu
7.2
3
1
1
w_2
2.4
3
2.4

Questions?

Next: More Neural Networks