Universität Konstanz

03 | Shallow Neural Nets

Max Pellert (https://mpellert.at)

Deep Learning for the Social Sciences

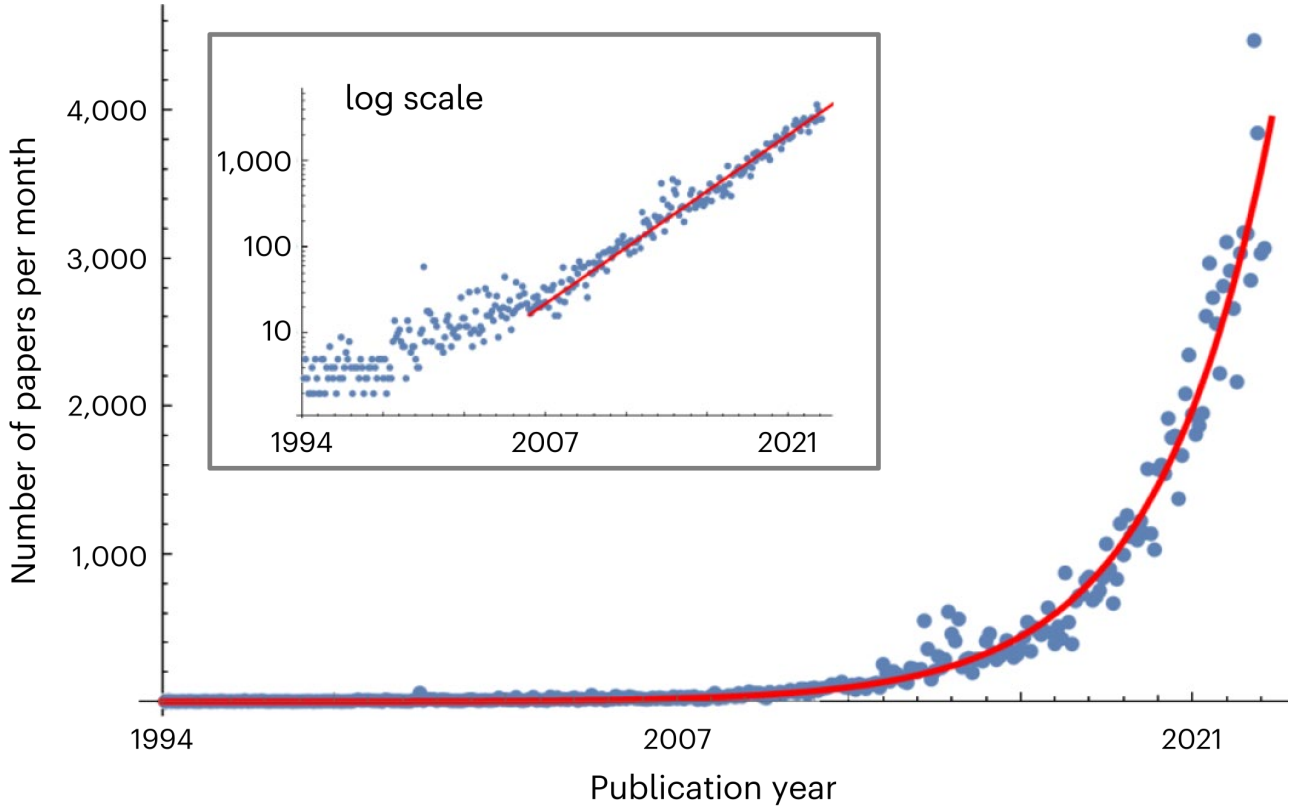


Fig. 1 | The number of papers published per month in the arXiv categories of AI and ML are growing exponentially. The doubling rate of papers per month is roughly 23 months, which might lead to problems for publishing in these fields, at some point. The categories are cs.AI, cs.LG, cs.NE and stat.ML.

Krenn, M., Buffoni, L., Coutinho, B., Eppel, S., Foster, J. G., Gritsevskiy, A., Lee, H., Lu, Y., Moutinho, J. P., Sanjabi, N., Sonthalia, R., Tran, N. M., Valente, F., Xie, Y., Yu, R., & Kopp, M. (2023). Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. Nature Machine Intelligence, 5(11), 1326–1335. https://doi.org/10.1038/s42256-023-00735-0

Shallow neural networks

1D regression model is obviously limited

We want to be able to describe input/output relationships that are not lines

We want multiple inputs

We want multiple outputs

Shallow neural networks

Flexible enough to describe arbitrarily complex input/output mappings

Can have as many inputs as we want

Can have as many outputs as we want

1D Linear Regression

$$
\begin{aligned}
y &= \mathrm{f}[x, \boldsymbol{\phi}] \\
&= \phi_0 + \phi_1 x
\end{aligned}
$$

Example shallow network

$$
\begin{aligned}
y &= \mathrm{f}[x, \boldsymbol{\phi}] \\
&= \phi_0 + \phi_1 \mathrm{a}[\theta_{10} + \theta_{11} x] + \phi_2 \mathrm{a}[\theta_{20} + \theta_{21} x] + \phi_3 \mathrm{a}[\theta_{30} + \theta_{31} x]
\end{aligned}
$$

Example shallow network

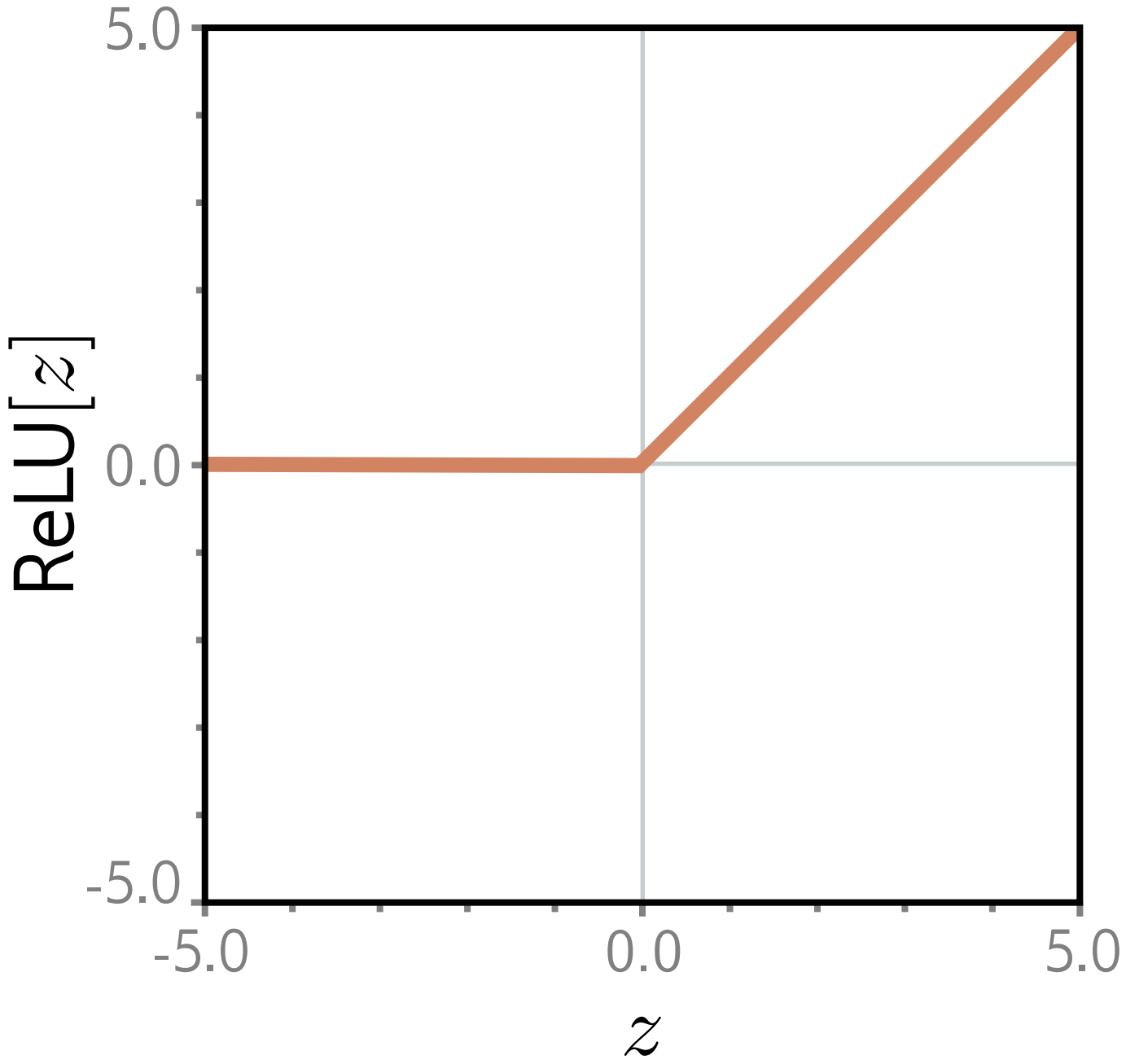
Activation function

$$y = \mathrm{f}[x, \boldsymbol{\phi}]$$
$$= \phi_0 + \phi_1 \mathrm{a}[\theta_{10} + \theta_{11} x] + \phi_2 \mathrm{a}[\theta_{20} + \theta_{21} x] + \phi_3 \mathrm{a}[\theta_{30} + \theta_{31} x]$$

$$\mathrm{a}[z] = \mathrm{ReLU}[z] = \begin{cases} 0 & z < 0 \\ z & z \geq 0 \end{cases}.$$

Rectified Linear Unit

(particular kind of activation function)


5.0
0.0
-5.0
ReLU[z]
-5.0
0.0
5.0
z

Example shallow network

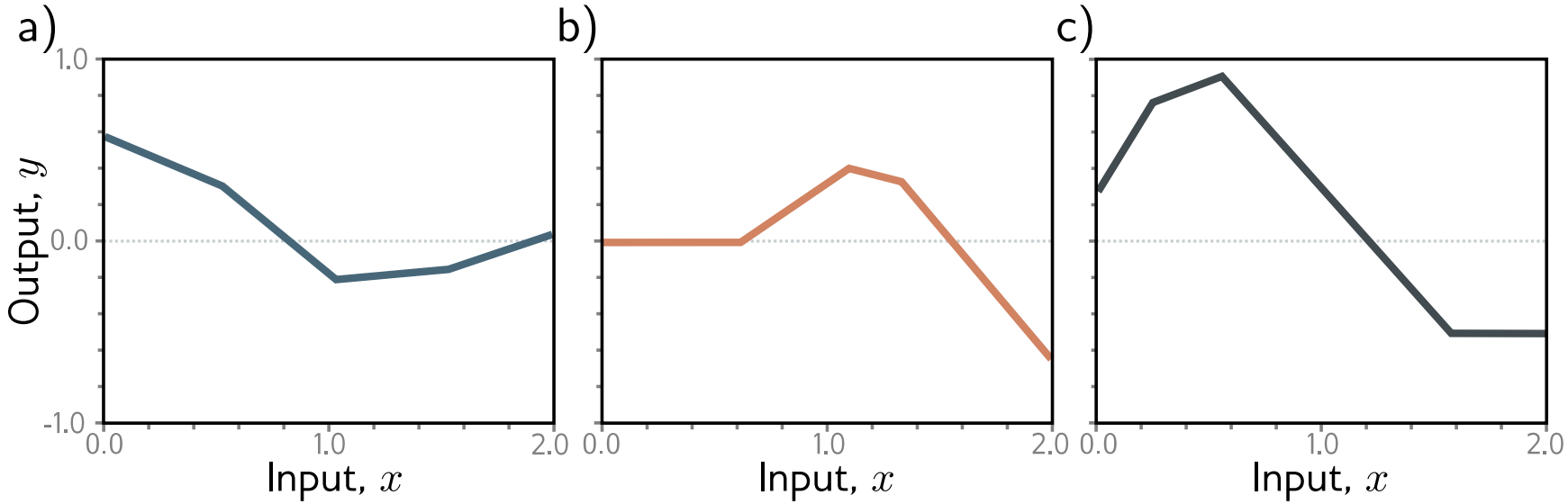
This model has 10 parameters:

$$\boldsymbol{\phi} = \{\phi_0, \phi_1, \phi_2, \phi_3, \theta_{10}, \theta_{11}, \theta_{20}, \theta_{21}, \theta_{30}, \theta_{31}\}$$

- Represents a family of functions
- Parameters determine particular function
- Given parameters can perform **inference** (run equation)
- Given training dataset: $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{I}$
- Define loss function $L[\boldsymbol{\phi}]$ (least squares)
- Change parameters to minimize loss function

Three example shallow networks

$$y = \mathrm{f}[x, \boldsymbol{\phi}]$$
$$= \phi_0 + \phi_1 \mathrm{a}[\theta_{10} + \theta_{11} x] + \phi_2 \mathrm{a}[\theta_{20} + \theta_{21} x] + \phi_3 \mathrm{a}[\theta_{30} + \theta_{31} x]$$



Piecewise linear functions with three joints

$$y = \mathrm{f}[x, \boldsymbol{\phi}]$$
$$= \phi_0 + \phi_1 \mathrm{a}[\theta_{10} + \theta_{11}x] + \phi_2 \mathrm{a}[\theta_{20} + \theta_{21}x] + \phi_3 \mathrm{a}[\theta_{30} + \theta_{31}x]$$

Break down into two parts:

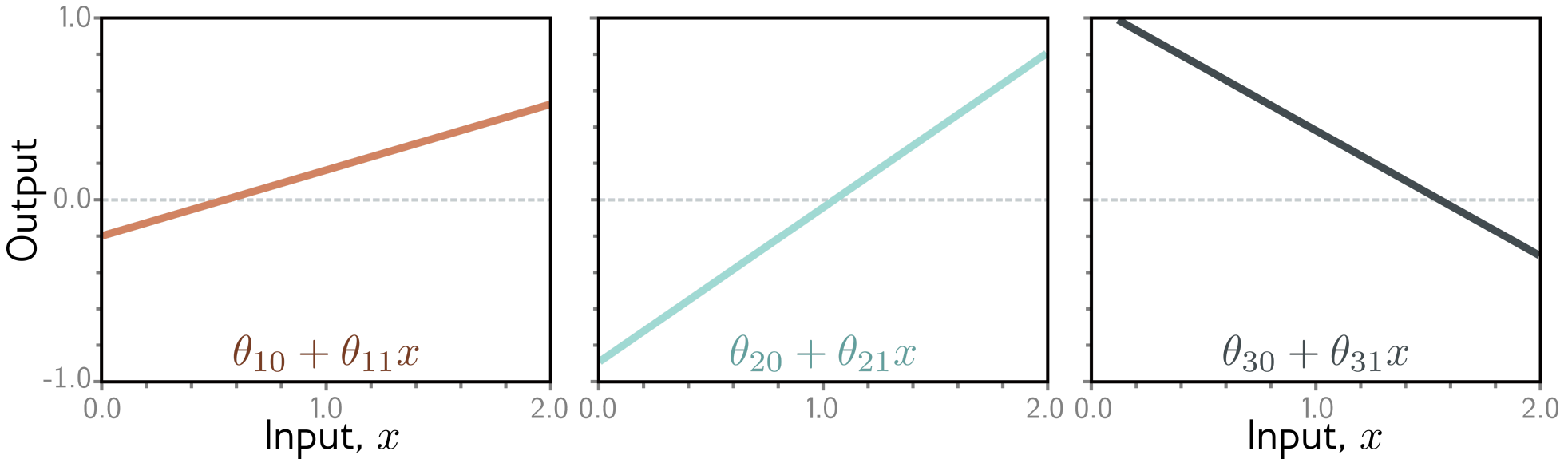
$$y = \phi_0 + \phi_1 h_1 + \phi_2 h_2 + \phi_3 h_3$$

where:

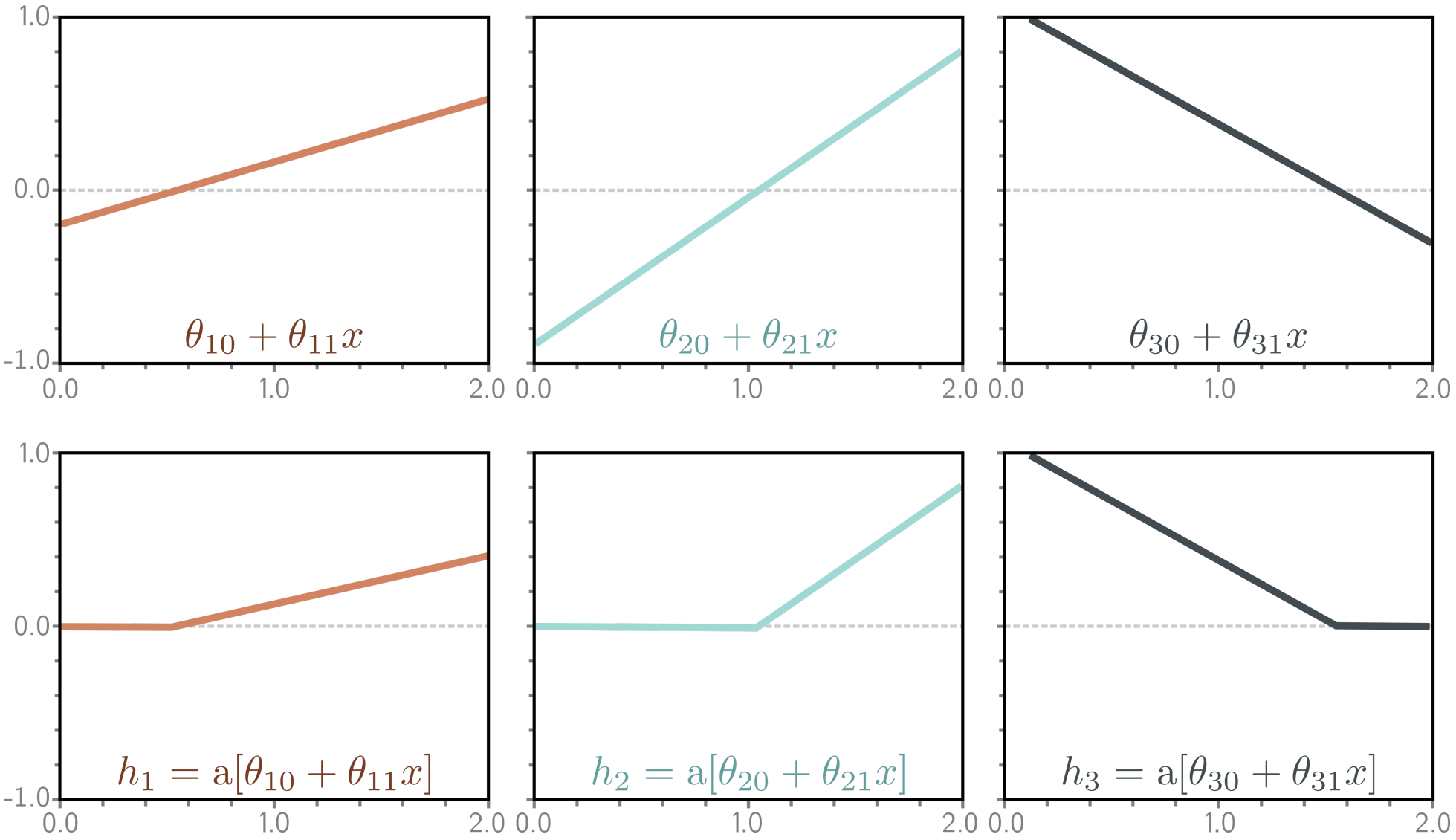
Hidden units

$$\begin{aligned} h_1 &= \mathrm{a}[\theta_{10} + \theta_{11}x] \\ h_2 &= \mathrm{a}[\theta_{20} + \theta_{21}x] \\ h_3 &= \mathrm{a}[\theta_{30} + \theta_{31}x] \end{aligned}$$

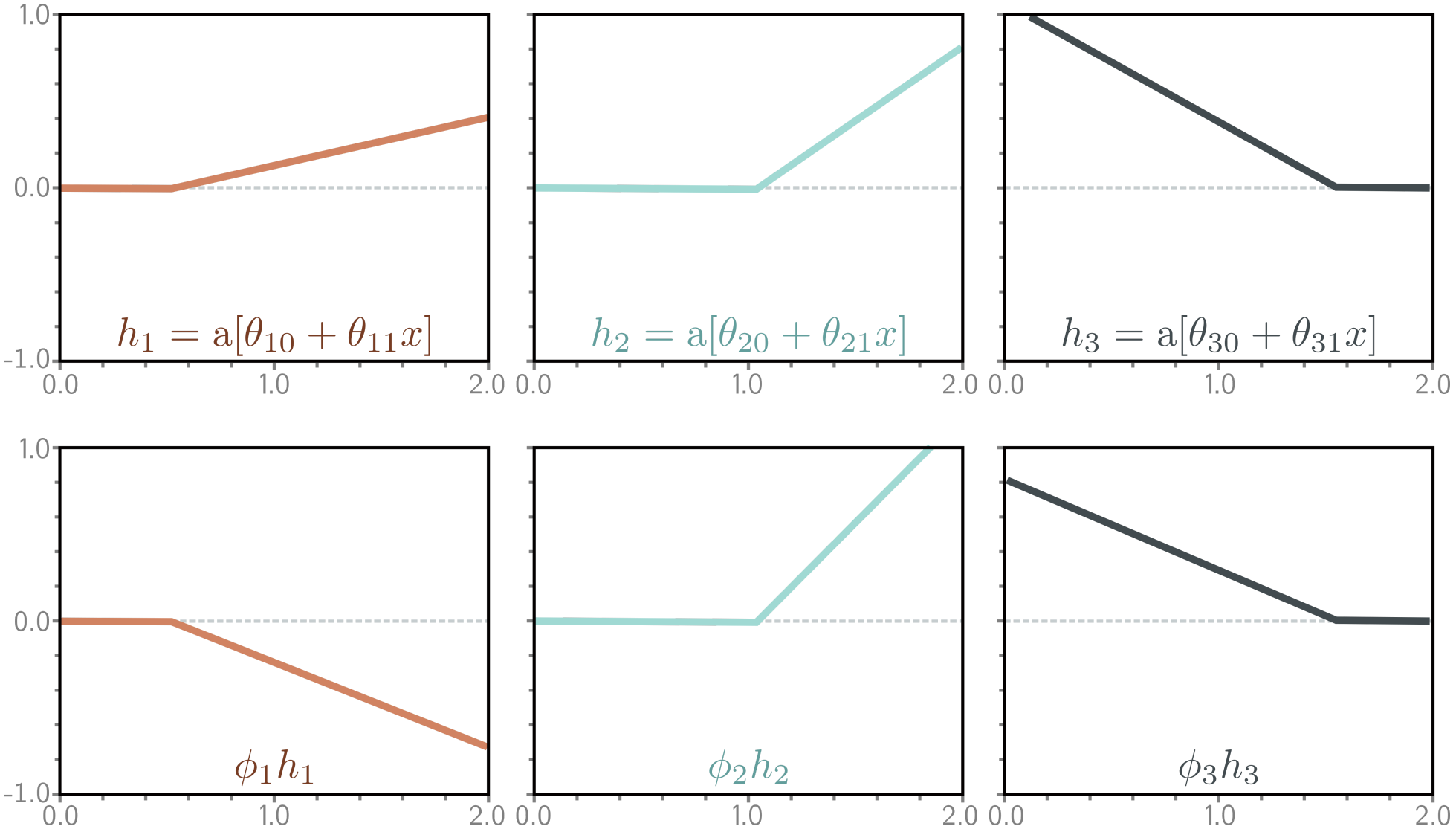
Building up a shallow network



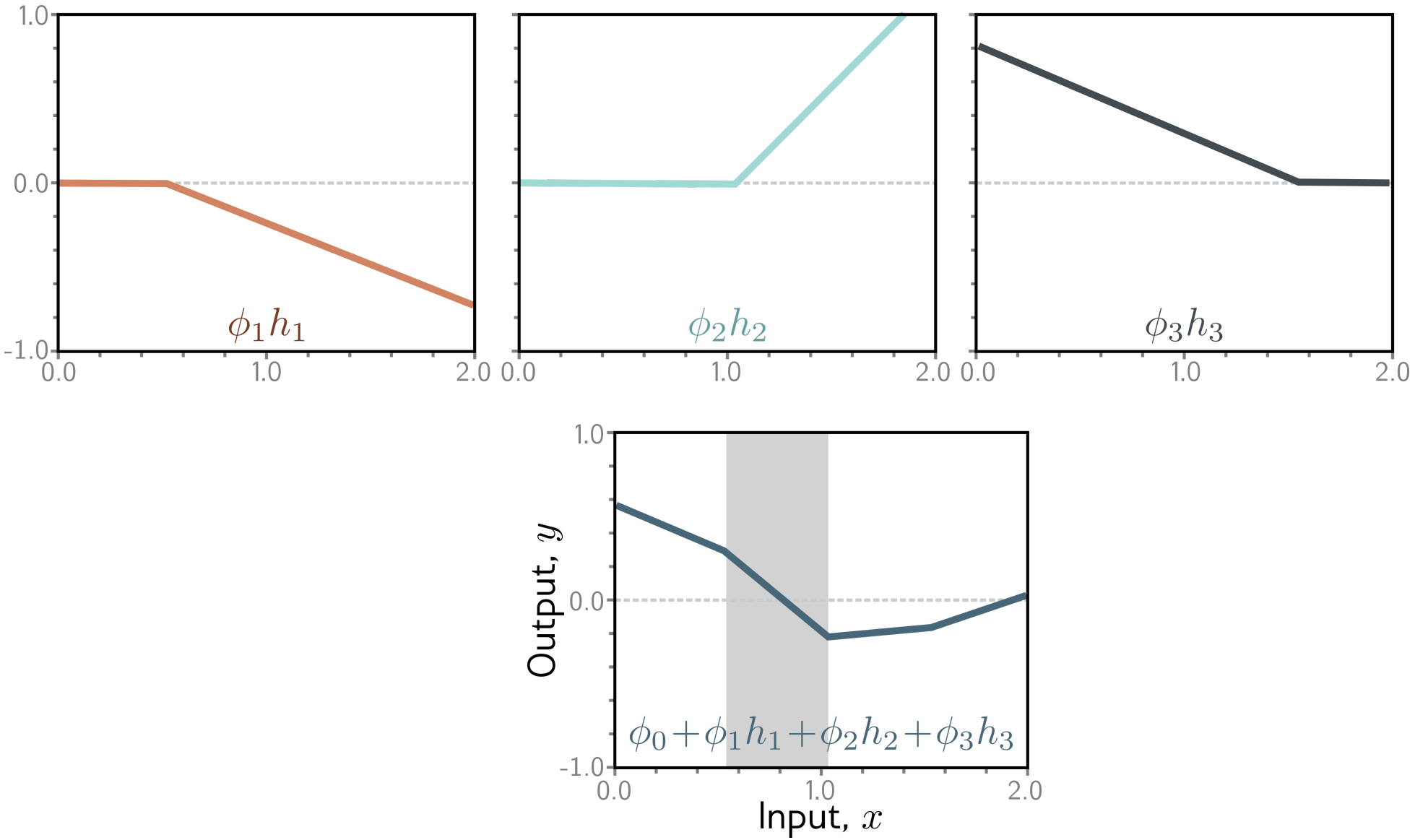
1. Compute the three linear functions



2. Pass through ReLU functions (creates hidden units)

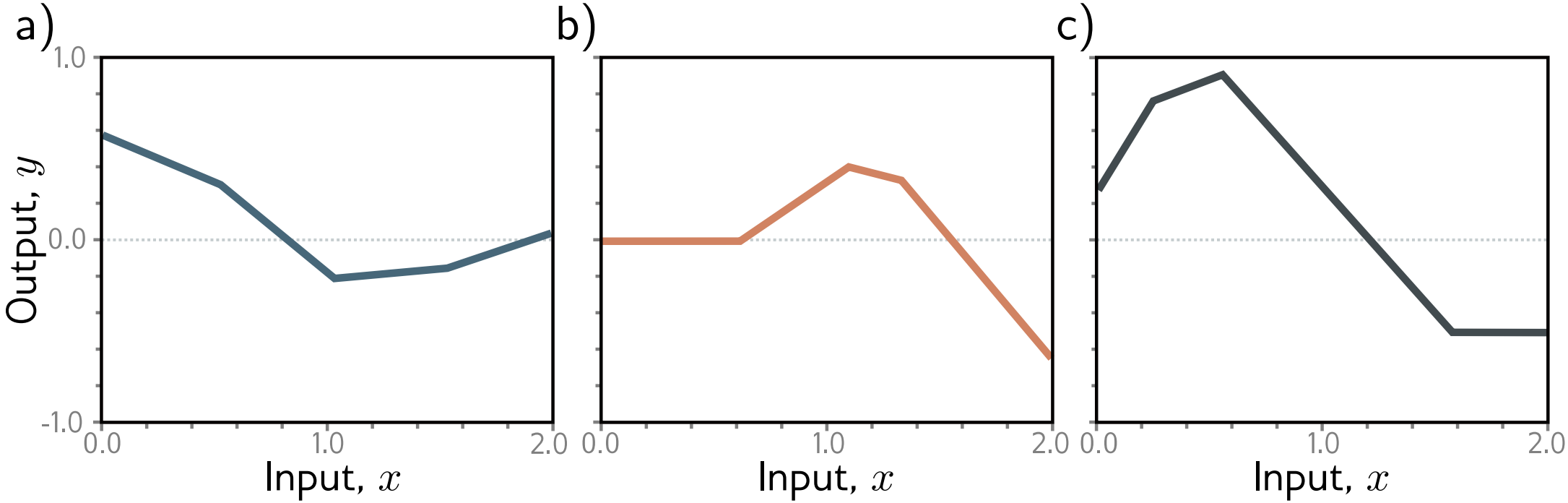


3. Weight the hidden units



4. Sum the weighted hidden units to create output

Other examples of shallow networks

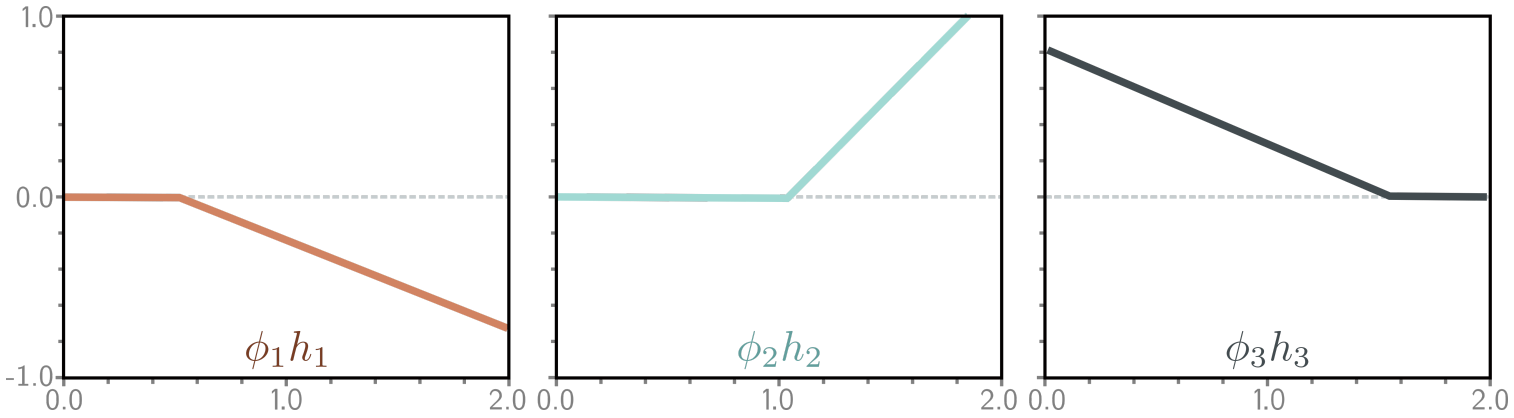


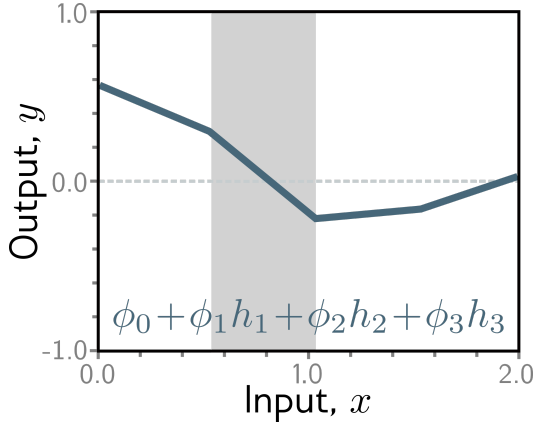
Example shallow network = piecewise linear functions

1 “joint” per ReLU function

Activation patterns

Which hidden units are activated in the shaded region?





Shaded region:

Unit 1 active

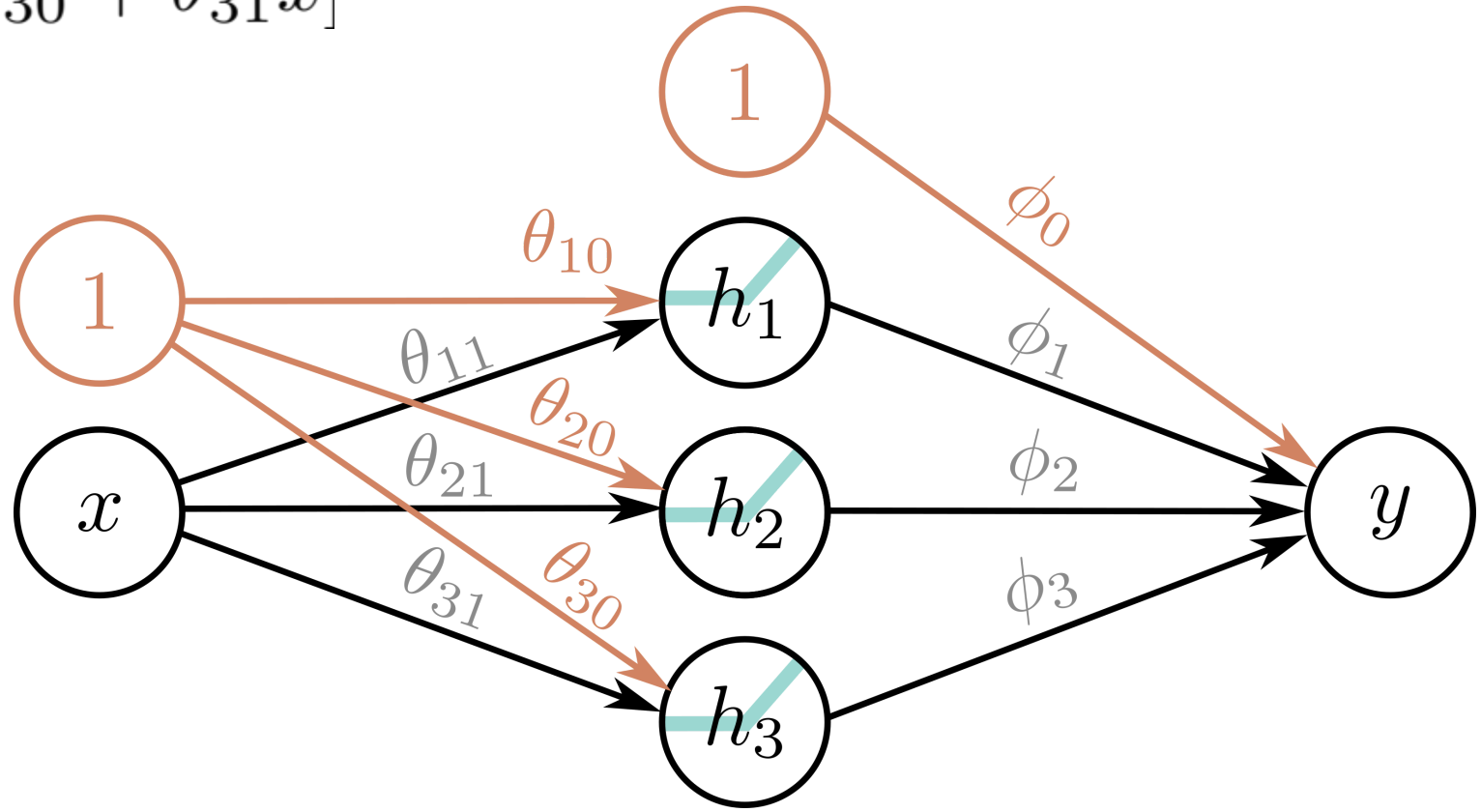
Unit 2 inactive

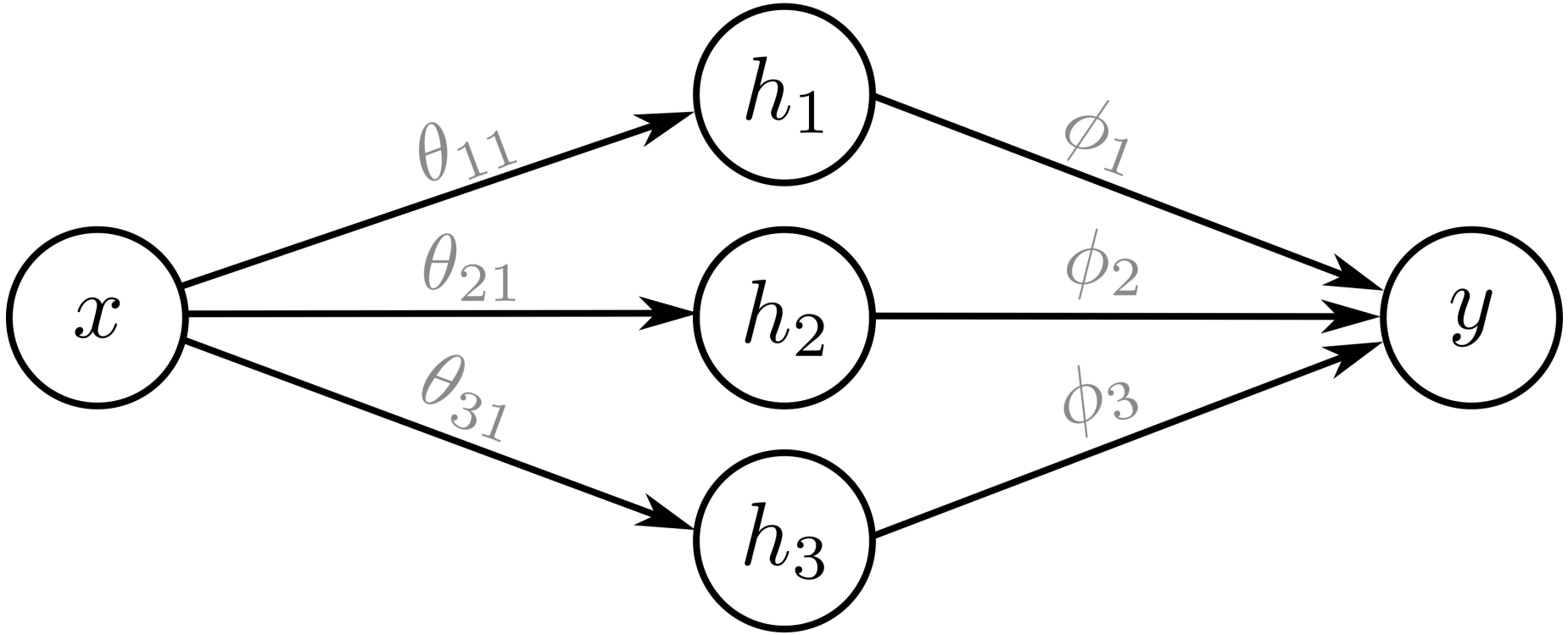
Unit 3 active

Depicting Neural Nets

$$h_1 = \mathrm{a}[\theta_{10} + \theta_{11}x]$$
$$h_2 = \mathrm{a}[\theta_{20} + \theta_{21}x]$$
$$h_3 = \mathrm{a}[\theta_{30} + \theta_{31}x]$$

$$y = \phi_0 + \phi_1 h_1 + \phi_2 h_2 + \phi_3 h_3$$





Each parameter multiplies its source and adds to its target

Universal Approximation Theoreom

With three hidden units, like before:

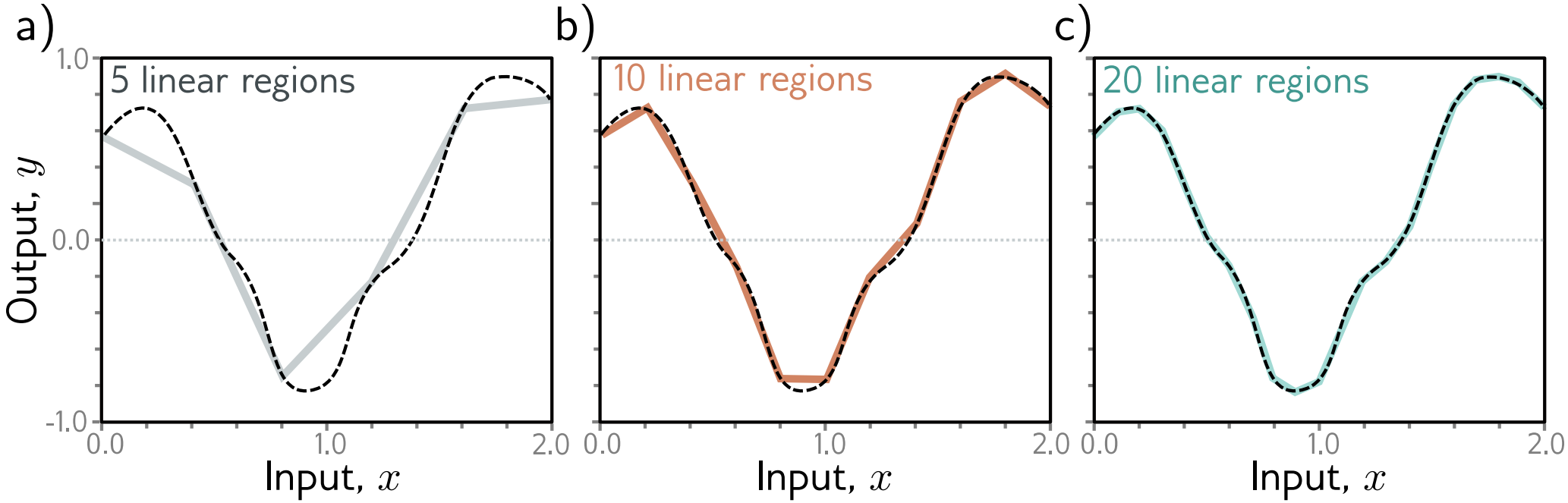
$$h_1 = \mathrm{a}[\theta_{10} + \theta_{11}x]$$
$$h_2 = \mathrm{a}[\theta_{20} + \theta_{21}x] \qquad y = \phi_0 + \phi_1 h_1 + \phi_2 h_2 + \phi_3 h_3$$
$$h_3 = \mathrm{a}[\theta_{30} + \theta_{31}x]$$

With D hidden units:

$$h_d = \mathrm{a}[\theta_{d0} + \theta_{d1}x] \qquad y = \phi_0 + \sum_{d=1}^{D} \phi_d h_d$$

With enough hidden units...

...we can describe any 1D function to arbitrary accuracy



Universal approximation theorem

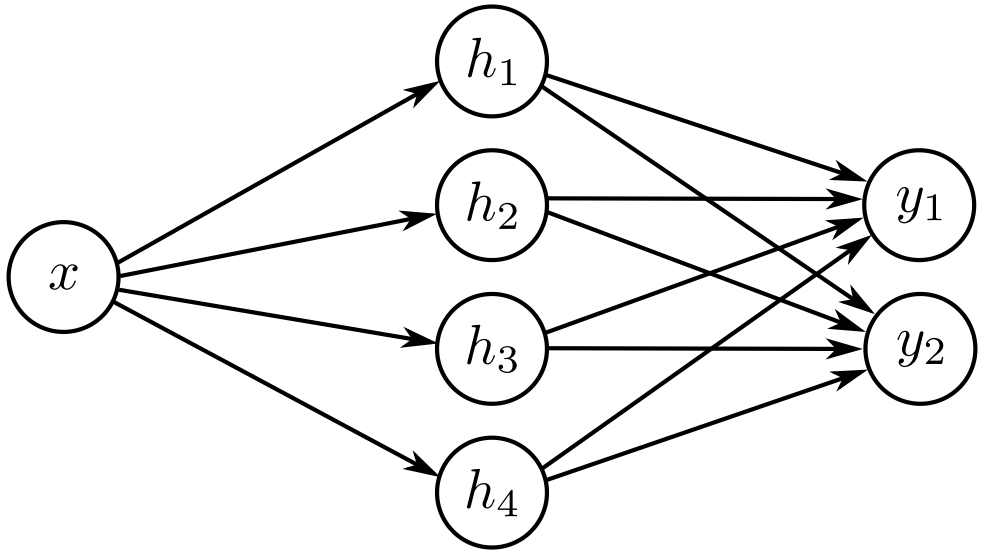
"A formal proof that, with enough hidden units, a shallow neural network can describe any continuous function on a compact subset of $\mathbb{R}^D$ to arbitrary precision"

More than 1 output

1 input, 4 hidden units, 2 outputs

$$h_1 = \mathrm{a}[\theta_{10} + \theta_{11}x]$$
$$h_2 = \mathrm{a}[\theta_{20} + \theta_{21}x]$$
$$h_3 = \mathrm{a}[\theta_{30} + \theta_{31}x]$$
$$h_4 = \mathrm{a}[\theta_{40} + \theta_{41}x]$$

$$y_1 = \phi_{10} + \phi_{11}h_1 + \phi_{12}h_2 + \phi_{13}h_3 + \phi_{14}h_4$$
$$y_2 = \phi_{20} + \phi_{21}h_1 + \phi_{22}h_2 + \phi_{23}h_3 + \phi_{24}h_4$$

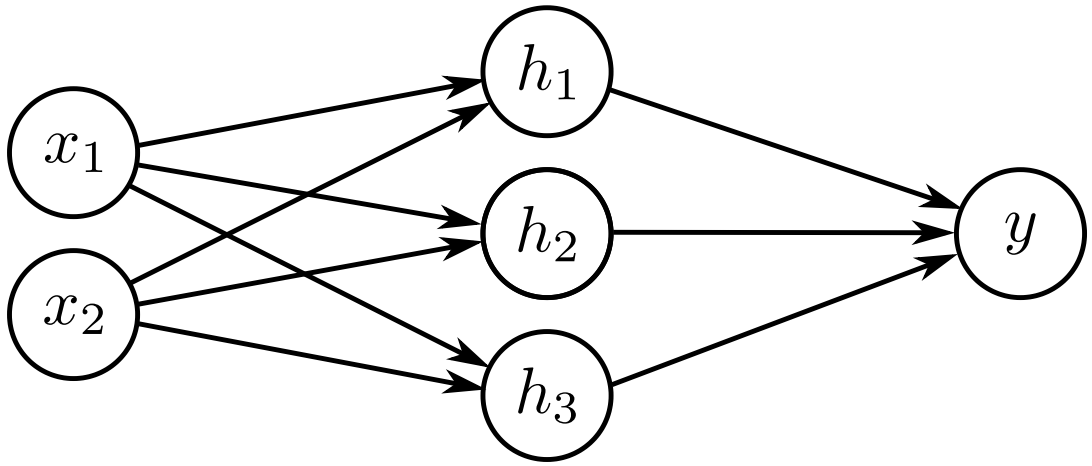


More than 1 input

2 inputs, 3 hidden units, 1 output

$$h_1 = \mathrm{a}[\theta_{10} + \theta_{11}x_1 + \theta_{12}x_2]$$
$$h_2 = \mathrm{a}[\theta_{20} + \theta_{21}x_1 + \theta_{22}x_2]$$
$$h_3 = \mathrm{a}[\theta_{30} + \theta_{31}x_1 + \theta_{32}x_2]$$

$$y = \phi_0 + \phi_1 h_1 + \phi_2 h_2 + \phi_3 h_3$$

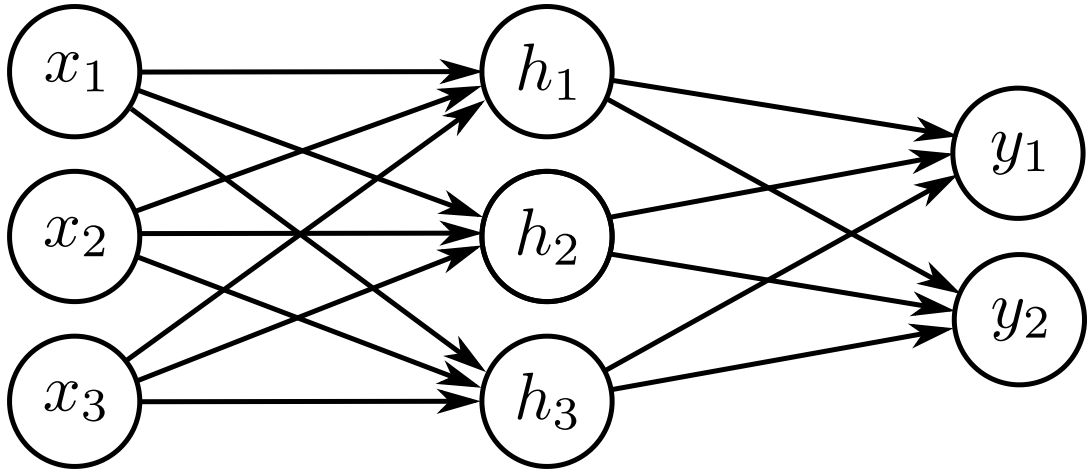


Arbitrary inputs, hidden units, outputs

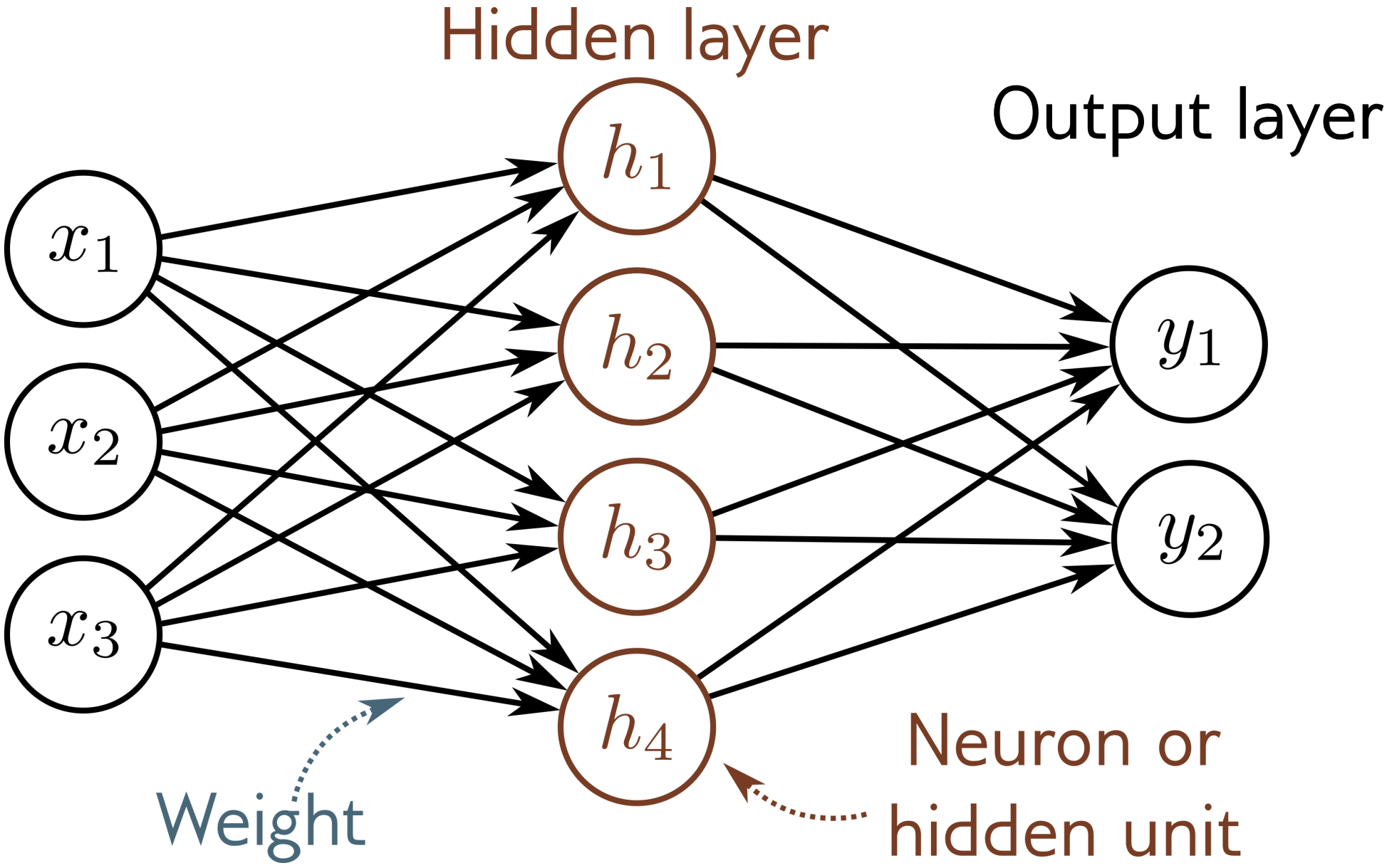
Outputs, D hidden units, and inputs

$$h_d = \mathrm{a}\left[\theta_{d0} + \sum_{i=1}^{D_i} \theta_{di} x_i\right] \qquad y_j = \phi_{j0} + \sum_{d=1}^{D} \phi_{jd} h_d$$

e.g. three inputs, three hidden units, two outputs

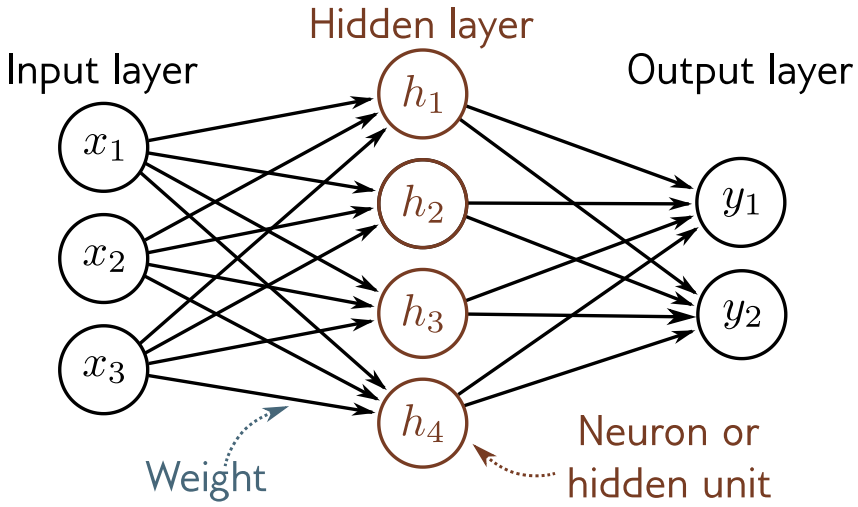


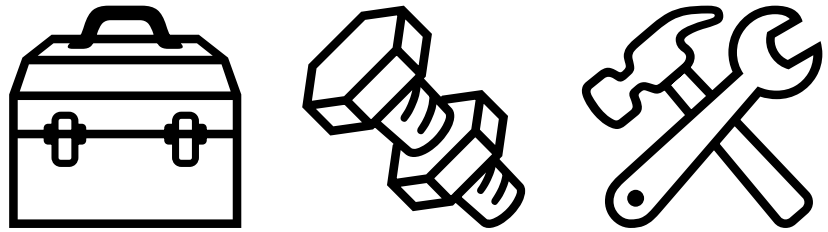
Some nomenclature...


Hidden layer
Output layer
x_1
x_2
x_3
h_1
h_2
h_3
h_4
y_1
y_2
Weight
Neuron or
hidden unit

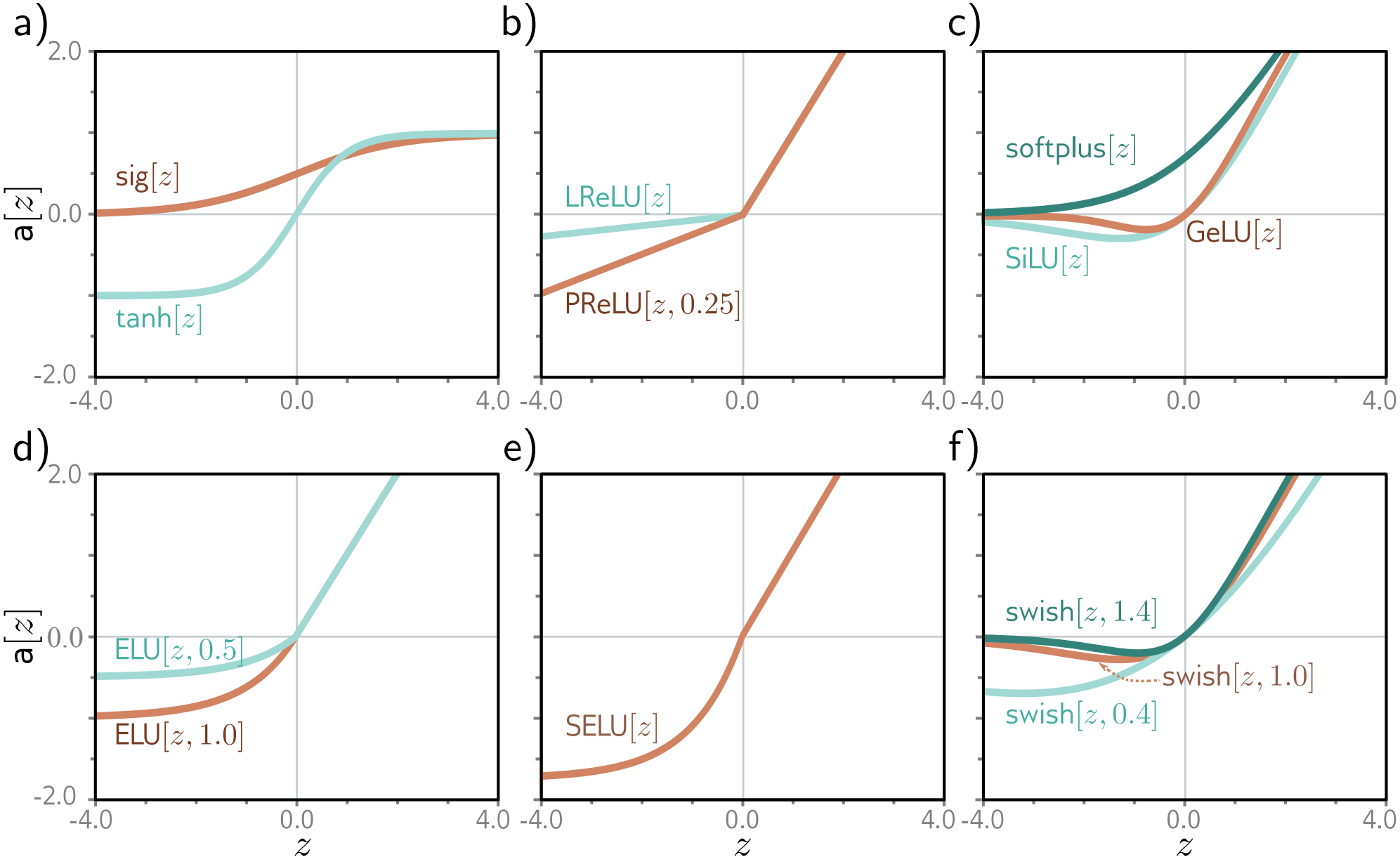
More nomenclature

- Y-offsets = biases
- Slopes = weights
- Everything in one layer connected to everything in the next = fully connected network
- No loops = feedforward network
- Values after ReLU (activation functions) = activations
- Values before ReLU = pre-activations
- One hidden layer = shallow neural network
- More than one hidden layer = deep neural network
- Number of hidden units = network capacity





Other activation functions



Question for next time: What happens if we feed one neural network into another neural network?