Neural network architecture and learning

Toby Dylan Hocking
toby.hocking@nau.edu
toby.hocking@r-project.org

March 22, 2021

Fully connected multi-layer Neural Networks

Computing gradients and learning weights

Convolutional networks and pooling

Supervised learning setup

- Have an input $\mathbf{x} \in \mathbb{R}^d$ – a vector of d real numbers.
- And an output y (real number: regression, integer ID: classification).
- Want to learn a prediction function $f(\mathbf{x}) = y$ that will work on a new input.
- In a neural network with $L-1$ hidden layers the function f is defined using composition of L functions, $f(x) = f^{(L)}[\ldots f^{(1)}[x]] \in \mathbb{R}$.

Each function is matrix multiplication and activation

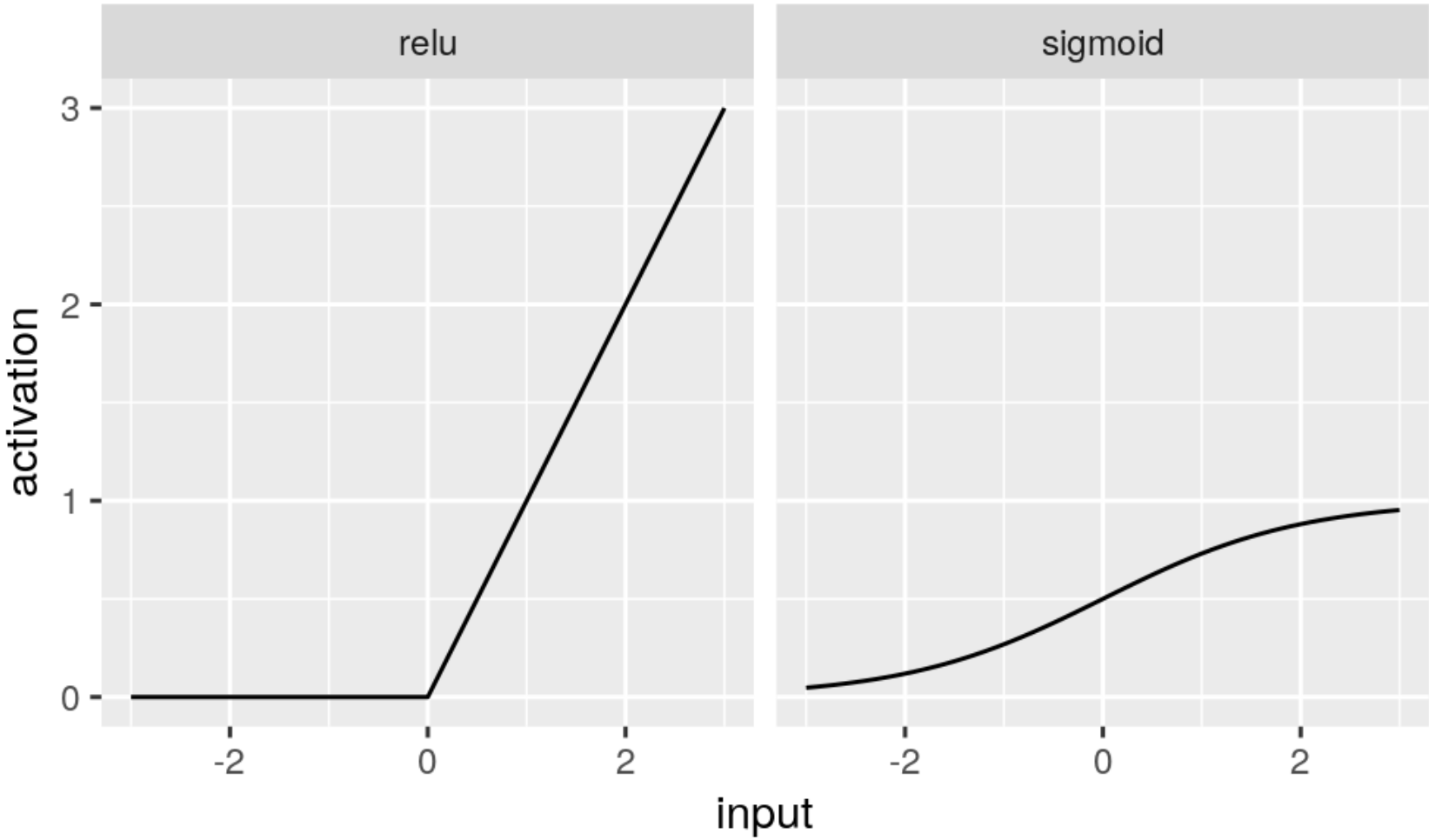
- Prediction function $f(x) = f^{(L)}[\cdots f^{(1)}[x]] \in \mathbb{R}$.
- Each function $l \in \{1, \ldots, L\}$ is a matrix multiplication followed by an activation function: $f^{(l)}[z] = \sigma^{(l)}[W^{(l)}z]$ where $W^{(l)} \in \mathbb{R}^{u^{(l)} \times u^{(l-1)}}$ is a weight matrix to learn, and $z \in \mathbb{R}^{u^{(l-1)}}$ is the input vector to that layer.
- In regression the last activation function must return a real number prediction so it is fixed to the identity: $\sigma^{(L)}[z] = z$.
- The other activation functions must be non-linear, e.g. logistic/sigmoid $\sigma(z) = 1/(1 + \exp(-z))$ or rectified linear units (ReLU)

$$\sigma(z) = \begin{cases} z & \text{if } z > 0, \\ 0 & \text{else.} \end{cases}$$

Non-linear activation functions

$$\sigma(z) = \begin{cases} z & \text{if } z > 0, \\ 0 & \text{else.} \end{cases}$$

$$\sigma(z) = 1/(1 + \exp(-z))$$



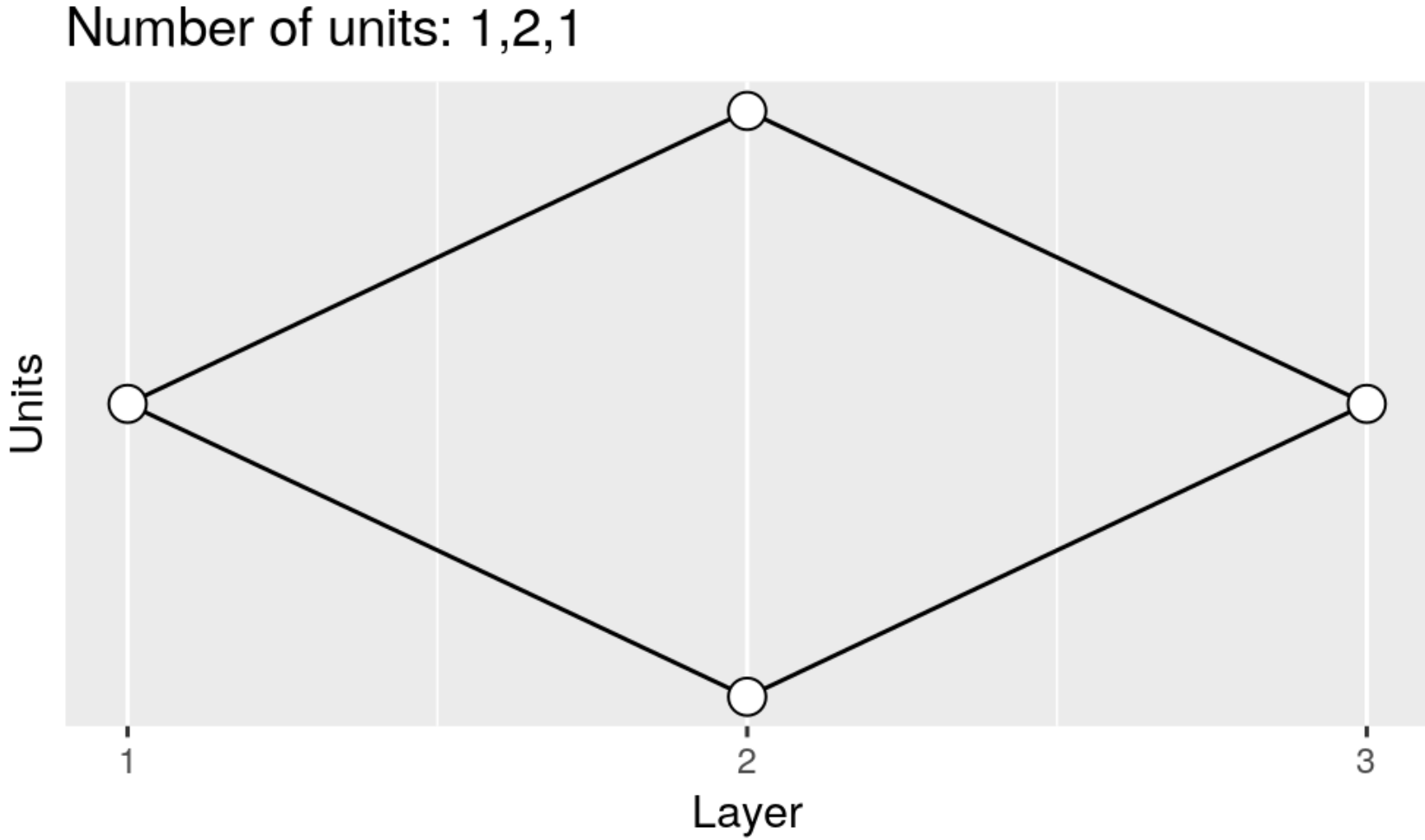
Network size

For binary classification with inputs $x \in \mathbb{R}^d$, the overall neural network architecture is $(u^{(0)} = d, u^{(1)}, \ldots, u^{(L-1)}, u^{(L)} = 1)$, where $u^{(1)}, \ldots, u^{(L-1)} \in \mathbb{Z}_+$ are positive integers (hyper-parameters that control the number of units in each hidden layer, and the size of the parameter matrices $W^{(l)}$).

- First layer size $u^{(0)}$ is fixed to input size.
- Last layer size $u^{(L)}$ is fixed to output size.
- Number of layers and hidden layer sizes $u^{(1)}, \ldots, u^{(L-1)}$ must be chosen (by you).

Network diagrams

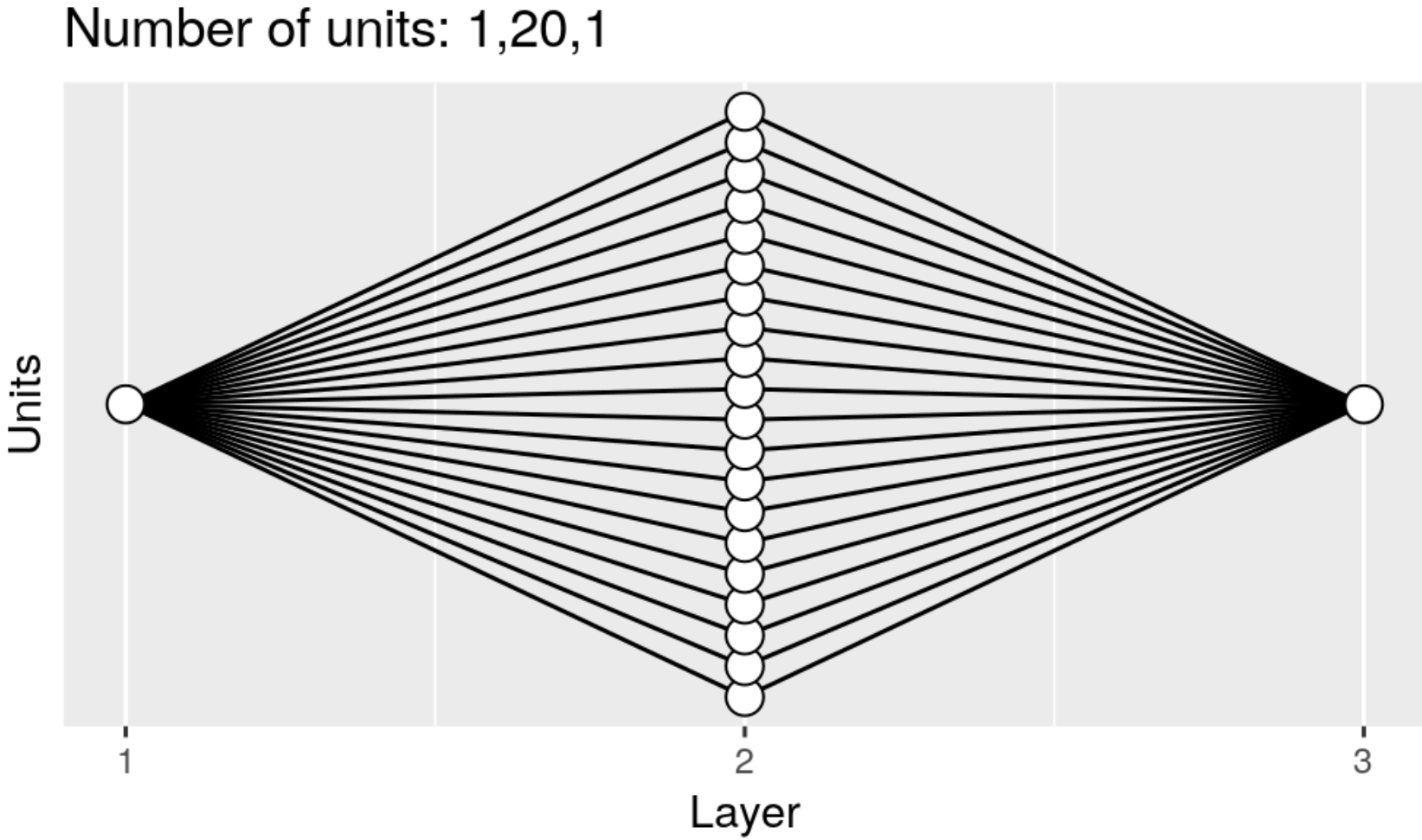
Neural network diagrams show how each hidden unit (node) is computed by applying the weights (edges) to the values of the hidden units at the previous layer.



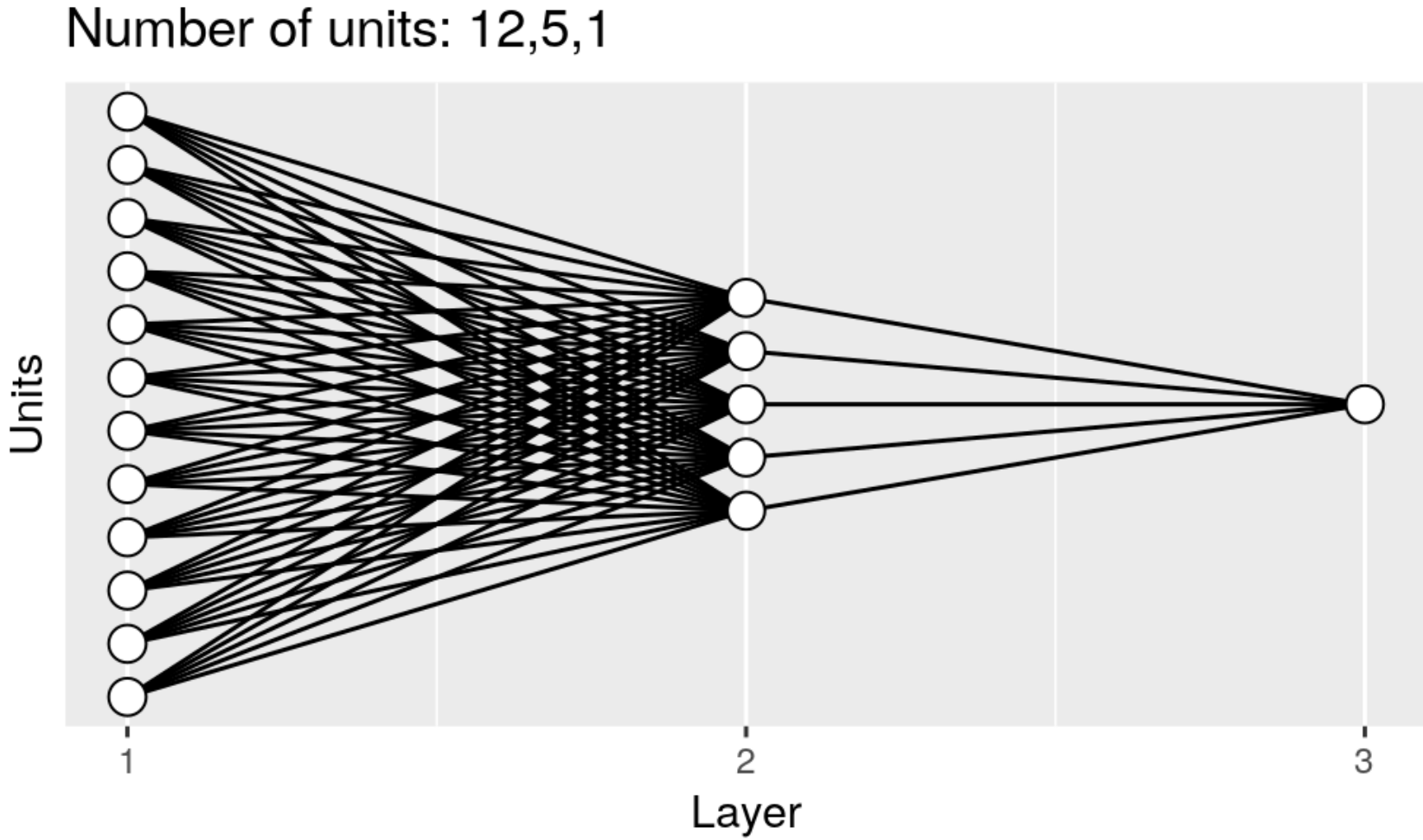
torch code

```
import torch
class Net(torch.nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        n_hidden = 2
        self.act = torch.nn.Sigmoid()
        self.hidden = torch.nn.Linear(1, n_hidden)
        self.out = torch.nn.Linear(n_hidden, 1)
    def forward(self, x):
        x = self.act(self.hidden(x))
        x = self.out(x)
        return x
```

Network diagrams



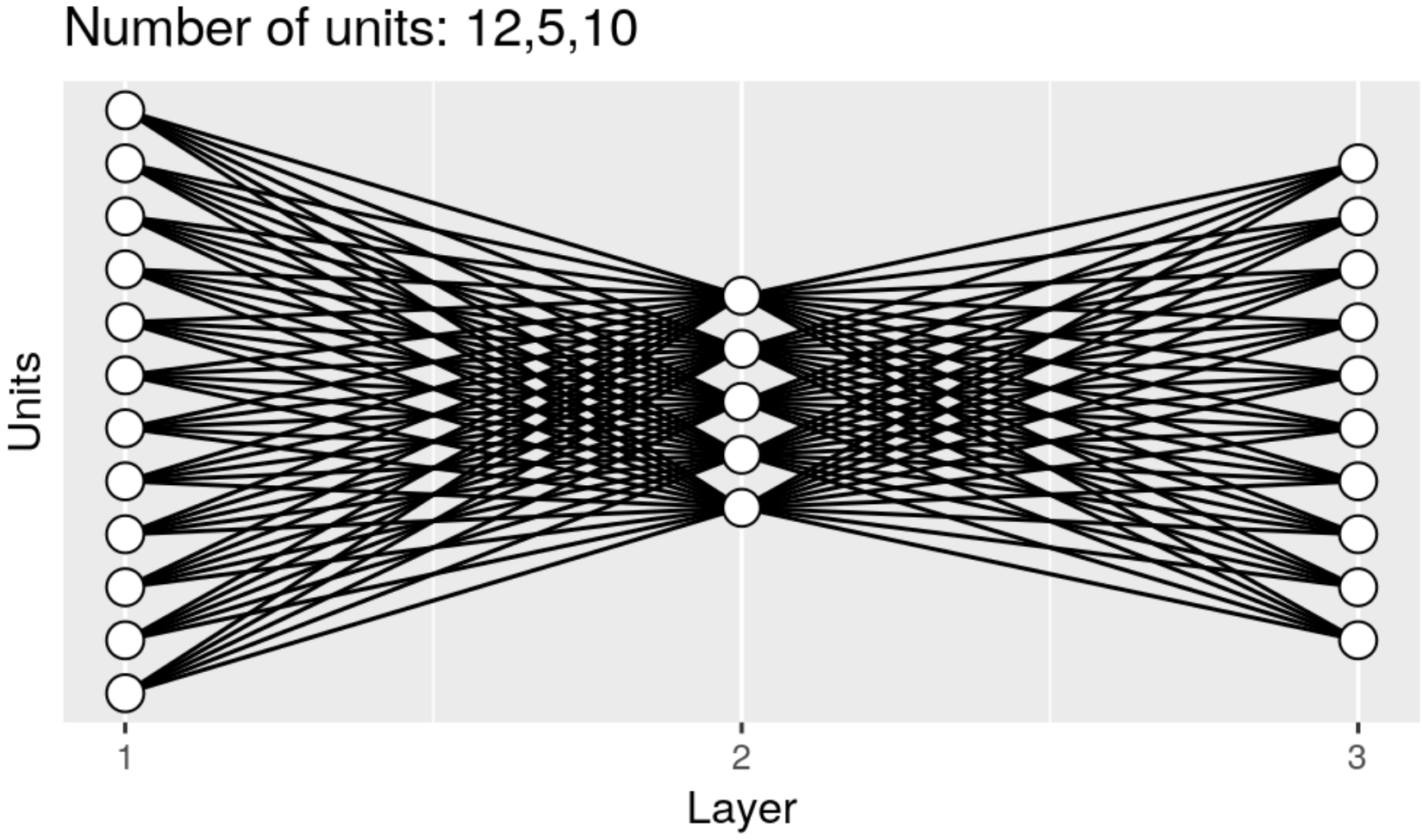
Network diagrams



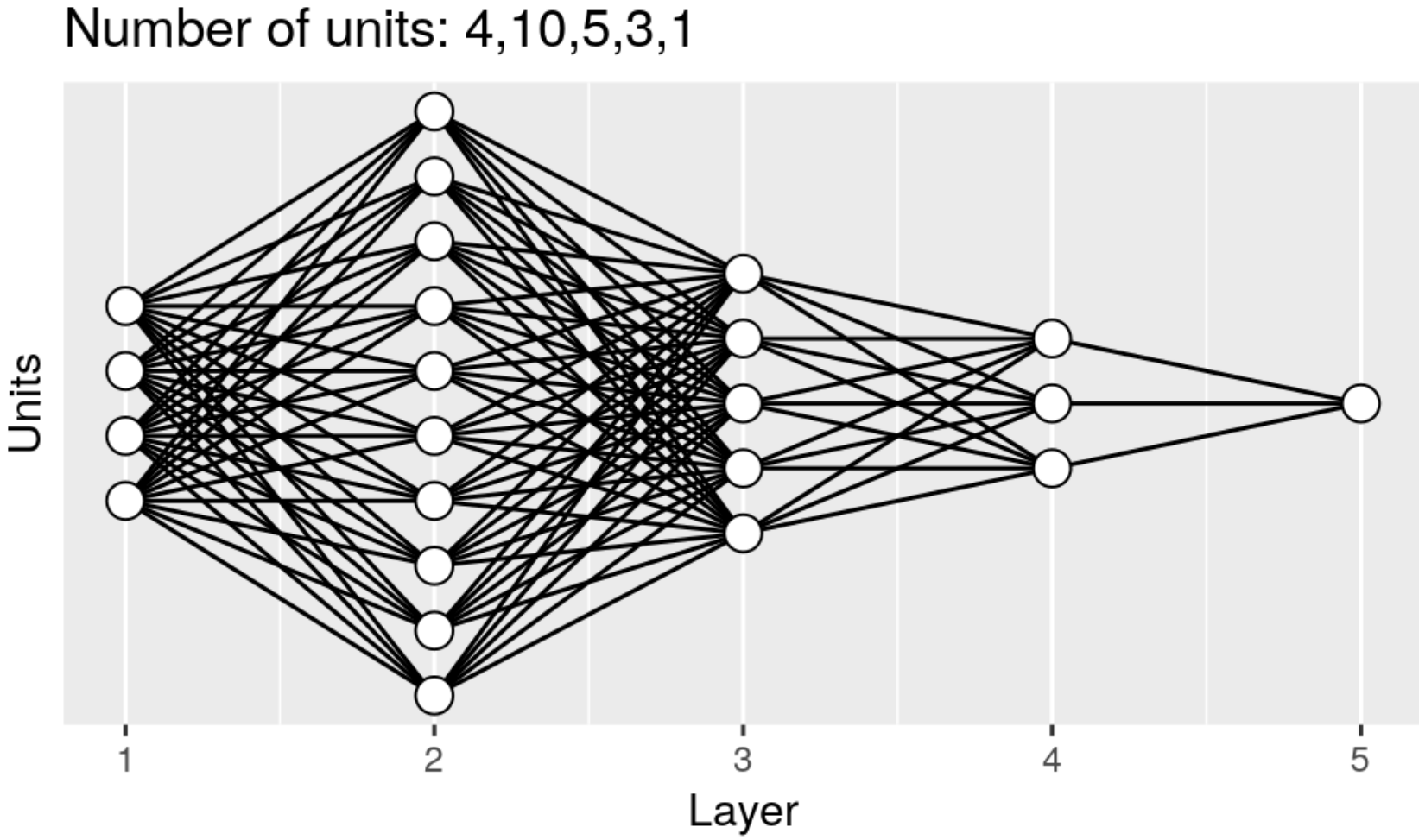
torch code

```
import torch
class Net(torch.nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        n_hidden = 5
        self.act = torch.nn.Sigmoid()
        self.hidden = torch.nn.Linear(12, n_hidden)
        self.out = torch.nn.Linear(n_hidden, 1)
    def forward(self, x):
        x = self.act(self.hidden(x))
        x = self.out(x)
        return x
```

Network diagrams



Network diagrams



Units in each layer

We can write the units at each layer as $h^{(0)}, h^{(1)}, \ldots, h^{(L-1)}, h^{(L)}$ where

- $h^{(0)} = x \in \mathbb{R}^d$ is an input feature vector,
- and $h^{(L)} \in \mathbb{R}$ is the predicted output.

For each layer $l \in \{1, \ldots, L\}$ we have:

$$h^{(l)} = f^{(l)}\left[h^{(l-1)}\right] = \sigma^{(l)}\left[W^{(l)} h^{(l-1)}\right].$$

Total number of parameters to learn is $\sum_{l=1}^{L} u^{(l)} u^{(l-1)}$.
Quiz: how many parameters in a neural network for $d = 10$ inputs/features with one hidden layer with $u = 100$ units? (one output unit, ten output units)

Fully connected multi-layer Neural Networks

Computing gradients and learning weights

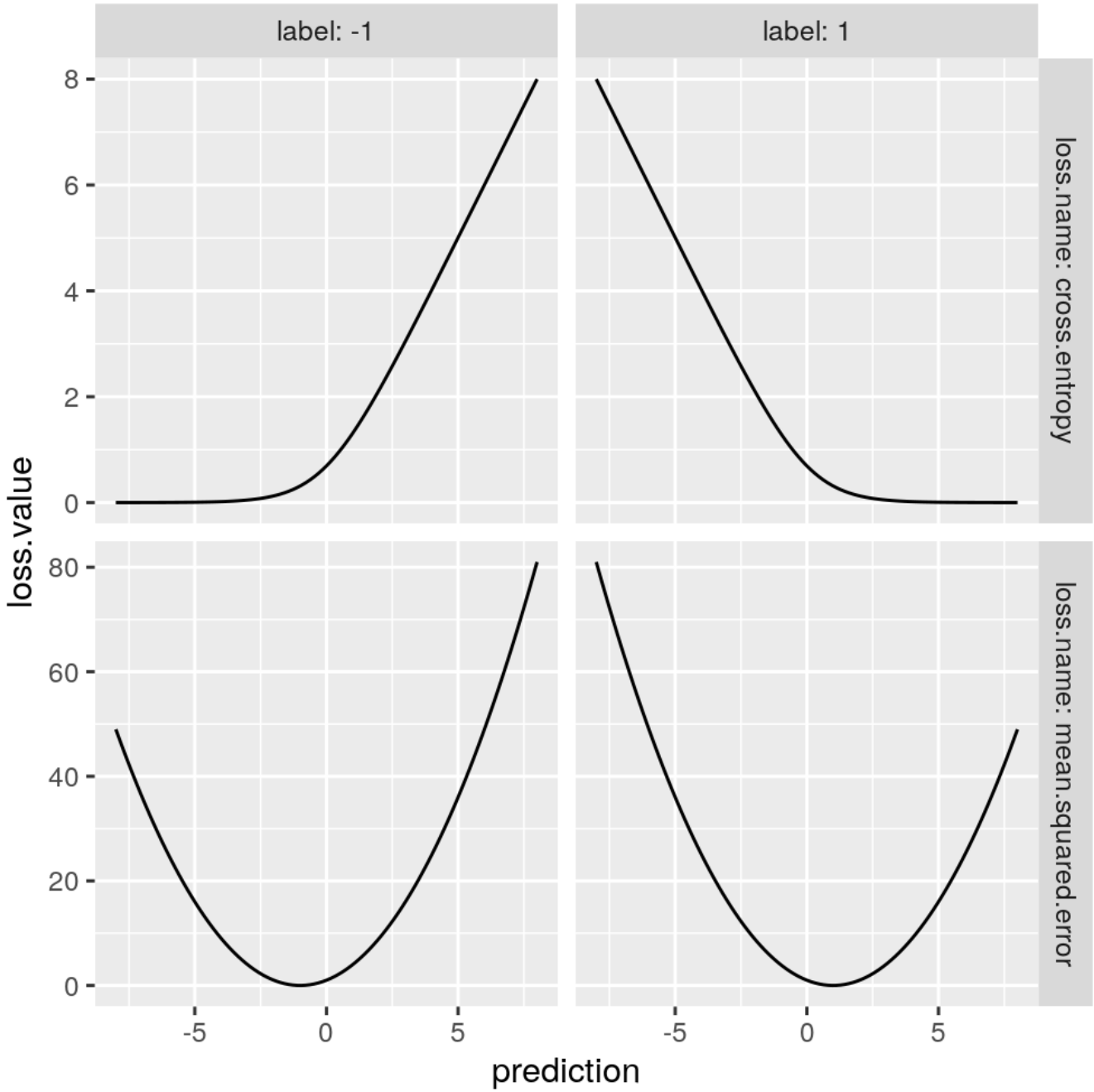
Convolutional networks and pooling

Gradient descent learning

Basic idea of gradient descent learning algorithm is to iteratively update weights $\mathbf{W} = [W^{(1)}, \ldots, W^{(L)}]$ to improve predictions on the subtrain set.

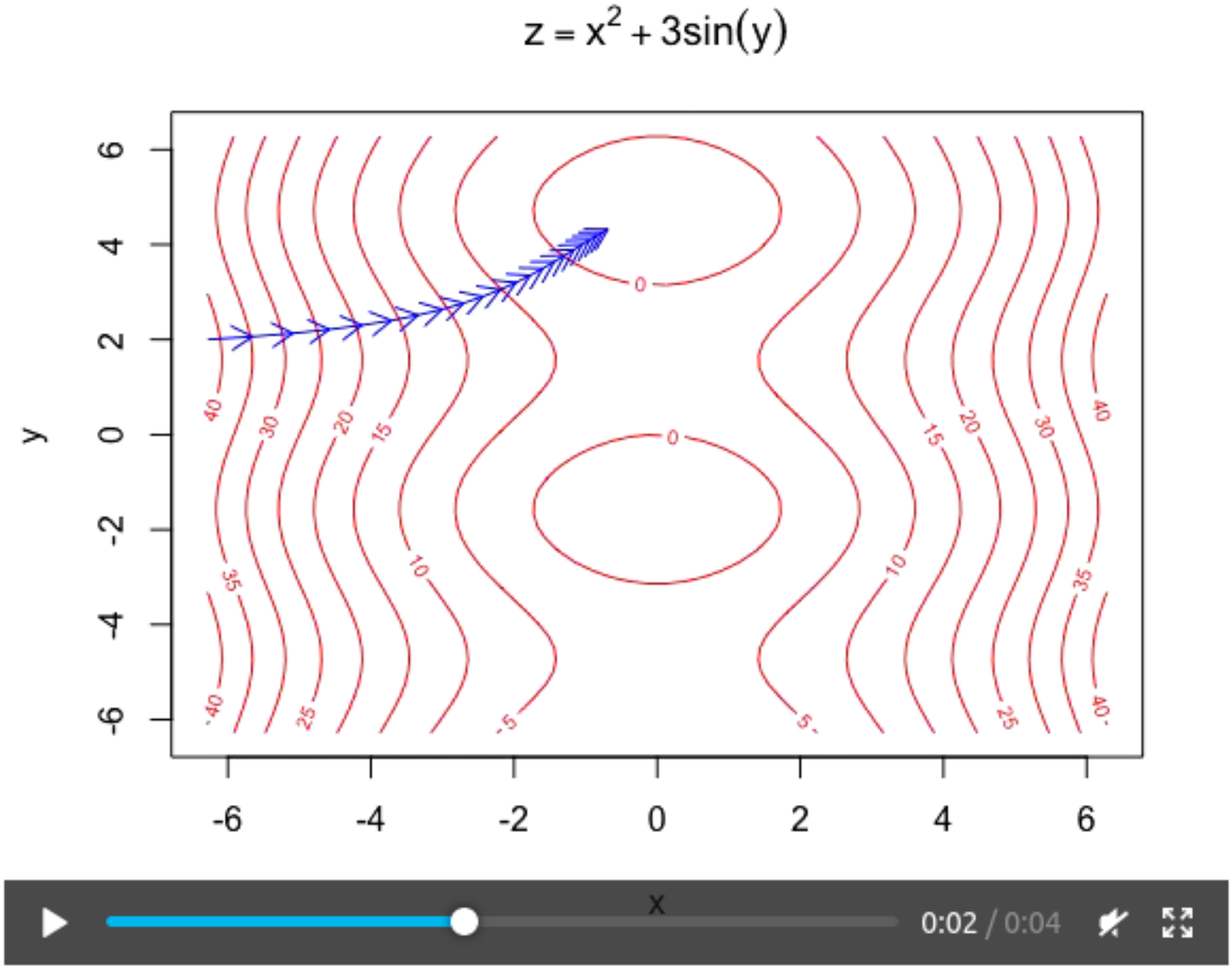
- Need to define a loss function $\mathcal{L}(\mathbf{W})$ which is differentiable, and takes small values for good predictions.
- Typically for regression we use the mean squared error, and for binary classification we use the logistic (cross entropy) loss.
- The gradient $\nabla\mathcal{L}(\mathbf{W})$ is a function which tells us the local direction where the loss is most increasing.

Loss functions



Gradient descent animations

https://yihui.org/animation/example/grad-desc/



Basic full gradient descent algorithm

- Initialize weights $\mathbf{W}_0$ at some random values near zero (more complicated initializations possible).
- Since we want to decrease the loss, we take a step α in the opposite direction of the gradient,

$$\mathbf{W}_t = \mathbf{W}_{t-1} - \alpha \nabla \mathcal{L}(\mathbf{W}_{t-1})$$

- This is the **full** gradient method: batch size $= n =$ subtrain set size, so 1 step per epoch/iteration.

```
optimizer = torch.optim.SGD(net.parameters(), lr=0.03)
optimizer.zero_grad()
predictions = net(subtrain_inputs)
subtrain_loss = criterion(predictions, subtrain_outputs)
subtrain_loss.backward()
optimizer.step()
```

Stochastic gradient descent algorithm

- Initialize weights $\mathbf{W}$ at some random values near zero (more complicated initializations possible).
- for each epoch t from 1 to max epochs:
- for each batch i from 1 to n:
- Let $\mathcal{L}(\mathbf{W}, \mathbf{X}_i, \mathbf{y}_i)$ be the loss with respect to the single observation in batch i.

$$\mathbf{W} \leftarrow \mathbf{W} - \alpha \nabla \mathcal{L}(\mathbf{W}, \mathbf{X}_i, \mathbf{y}_i)$$

- This is the **stochastic** gradient method: batch size $= 1$, so there are n steps per epoch.

```
optimizer = torch.optim.SGD(net.parameters(), lr=0.03)
optimizer.zero_grad()
prediction = net(one_input)
one_loss = criterion(prediction, one_output)
one_loss.backward()
optimizer.step()
```

Batch (stochastic) gradient descent algorithm

- Input: batch size b.
- Initialize weights $\mathbf{W}$ at some random values near zero (more complicated initializations possible).
- for each epoch t from 1 to max epochs:
- for each batch i from 1 to $\lceil n/b \rceil$:
- Let $\mathcal{L}(\mathbf{W}, \mathbf{X}_i, \mathbf{y}_i)$ be the loss with respect to the b observations in batch i.

$$\mathbf{W} \leftarrow \mathbf{W} - \alpha \nabla \mathcal{L}(\mathbf{W}, \mathbf{X}_i, \mathbf{y}_i)$$

- This is the **(mini)batch** stochastic gradient method: batch size $= b$, so there are $\lceil n/b \rceil$ steps per epoch.

```
optimizer = torch.optim.SGD(net.parameters(), lr=0.03)
optimizer.zero_grad()
prediction = net(batch_inputs)
batch_loss = criterion(prediction, batch_outputs)
batch_loss.backward()
optimizer.step()
```

Forward propagation

Forward propagation is the computation of hidden units $h^{(1)}, \ldots, h^{(L)}$ given the inputs x and current parameters $W^{(1)}, \ldots, W^{(L)}$.

```
def forward(self, x):
    x = self.act(self.hidden(x))
    x = self.out(x)
    return x
```

(start from input, apply weights and activation in each layer until predicted output is computed)

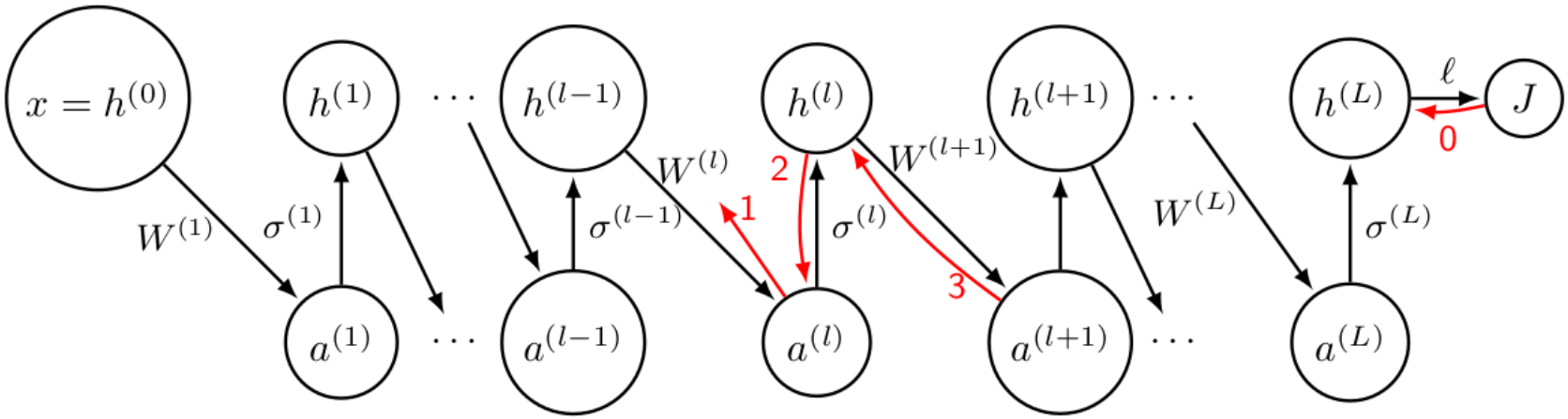
Back propagation

Back propagation is the computation of gradients given current parameters and hidden units.

- Start from loss function, compute gradient, send it to last layer, use chain rule, send gradient to previous layer, finally end up at first layer.
- Result is gradients with respect to all weights in all layers.
- Modern frameworks like torch do this using automatic differentiation based on your definition of the forward method and the loss function.

```
net = Net()
criterion = torch.nn.MSELoss()
optimizer = torch.optim.LBFGS(net.parameters(), lr=0.03)
optimizer.zero_grad()
pred = net(input_X_features)
loss = criterion(pred, output_y_labels)
loss.backward()
```

Computation graph



For each layer $l \in \{1, \ldots, L\}$ we have:

$$\begin{aligned} a^{(l)} &= W^{(l)} h^{(l-1)}, \\ h^{(l)} &= \sigma^{(l)} \left[a^{(l)} \right]. \end{aligned}$$

There are essentially four rules for computing gradients during backpropagation (0-3).

Backprop rules

The rules 0–3 for backprop (from loss backwards):

Rule 0 computes $\nabla_{h^{(L)}} J$, which depends on the choice of the loss function ℓ.

Rule 1 computes $\nabla_{W^{(l)}} J$ using $\nabla_{a^{(l)}} J$, for any $l \in \{1, \ldots, L\}$

$$\nabla_{w_k^{(l)}} J = (\nabla_{a^{(l)}} J) \left(h^{(l-1)}\right)^T \tag{1}$$

Rule 2 computes $\nabla_{a^{(l)}} J$ using $\nabla_{h^{(l)}} J$, for any $l \in \{1, \ldots, L\}$.

$$\nabla_{a^{(l)}} J = (\nabla_{h^{(l)}} J) \odot \left(\nabla_{a^{(l)}} h^{(l)}\right) \tag{2}$$

Rule 3 computes $\nabla_{h^{(l)}} J$ using $\nabla_{a^{(l+1)}} J$, for any $l \in \{1, \ldots, L-1\}$.

$$\nabla_{h^{(l)}} J = (\nabla_{a^{(l+1)}} J) \left(W^{(l+1)}\right)^T \tag{3}$$

Fully connected multi-layer Neural Networks

Computing gradients and learning weights

Convolutional networks and pooling

Convolution is a linear operator for spatial data

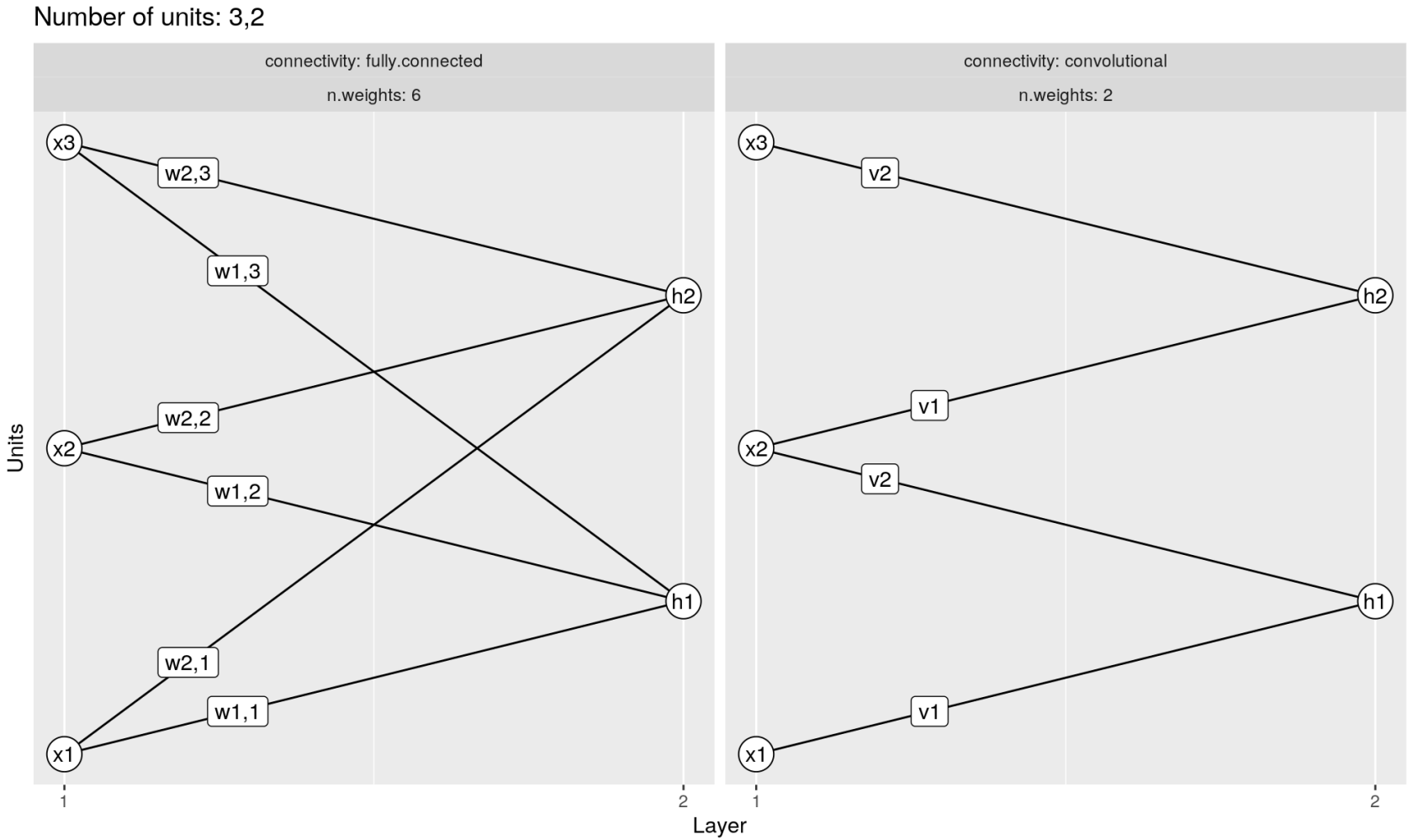
Useful for data which have spatial dimension(s) such as time series (1 dim) or images (2 dim). Simple example with 1 dim:

- $\mathbf{x} = [x_1, \ldots, x_D]$ is an input vector (array of D data).
- $\mathbf{v} = [v_1, \ldots, v_P]$ is a kernel (array of P parameters / weights to learn), $P < D$.
- $\mathbf{h} = [h_1, \ldots, h_U]$ is an output vector of $U = D - P + 1$ hidden units. Convolution (actually cross-correlation) is used to define each hidden unit: $\forall u \in \{1, \ldots, U\}, h_u = \sum_{p=1}^{P} v_p x_{u+p-1}$.
- EX: $D = 3$ inputs, $P = 2$ parameters $\Rightarrow U = 2$ output units:

$$
\begin{aligned}
h_1 &= v_1 x_1 + v_2 x_2 \text{ (convolutional=sparse+shared)} \\
h_2 &= v_1 x_2 + v_2 x_3 \text{ (convolutional=sparse+shared)} \\
h_1 &= w_{1,1} x_1 + w_{1,2} x_2 + w_{1,3} x_3 \text{ (fully connected/dense)} \\
h_2 &= w_{2,1} x_1 + w_{2,2} x_2 + w_{2,3} x_3 \text{ (fully connected/dense)}
\end{aligned}
$$

- Compare with fully connected – convolutional means weights are shared among outputs, and some are zero/sparse.

Difference in connectivity and weight sharing



Matrix interpretation of convolution

$$\begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} \\ w_{2,1} & w_{2,2} & w_{2,3} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \text{ (fully connected)}$$

$$\begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} v_1 & v_2 & 0 \\ 0 & v_1 & v_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \text{ (convolutional)}$$

- Weight sharing: same weights used to compute different output units.
- Sparsity: zeros in weight matrix.

```
torch.nn.Conv1d(
  in_channels=1,
  out_channels=1,
  kernel_size=2)
```

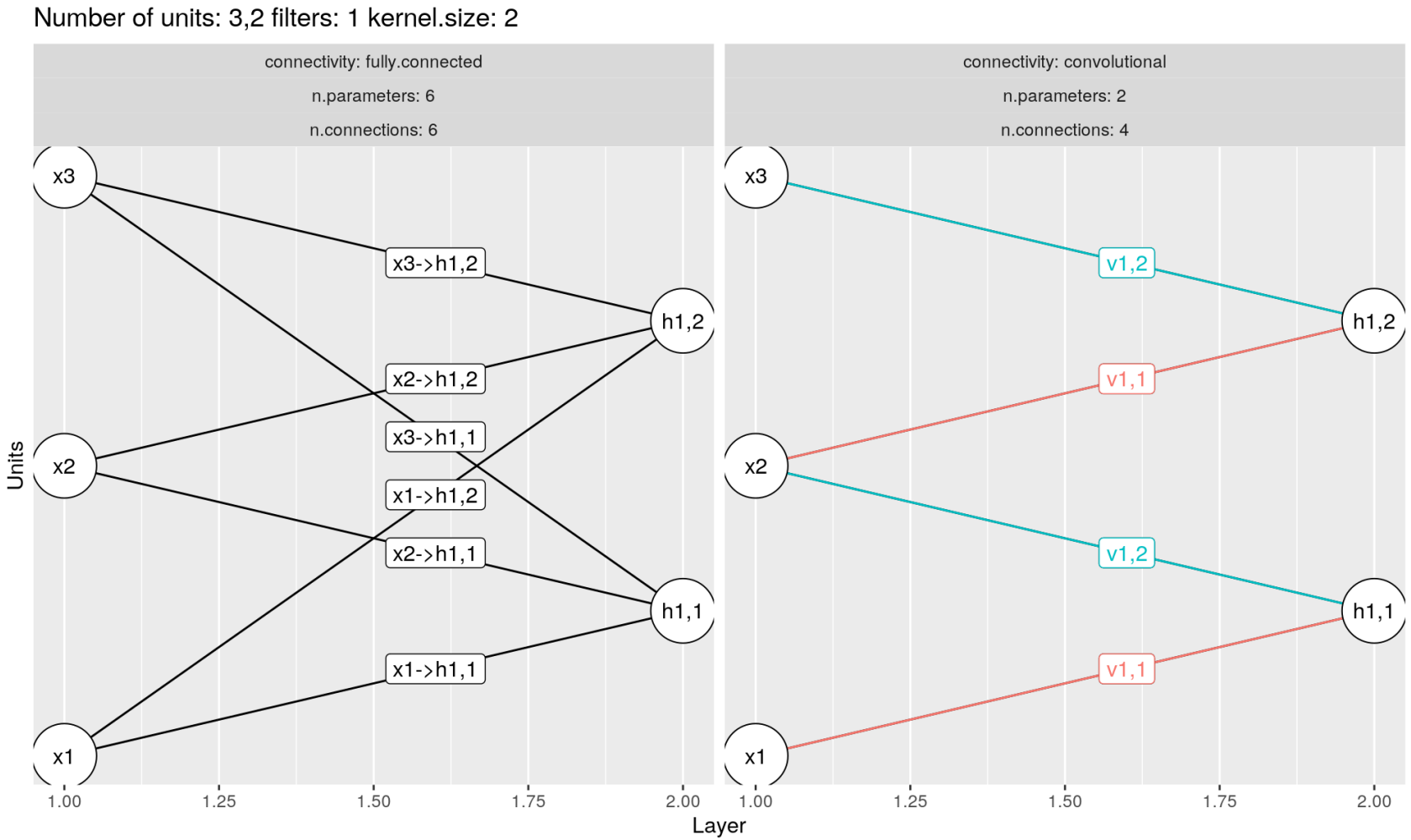
Multiple kernels/filters or sets of weights

- $x = [x_1, \ldots, x_D]$ is an input vector (array of D data).
- $v = \begin{bmatrix} v_{1,1} & \cdots & v_{1,P} \\ \vdots & \ddots & \vdots \\ v_{K,1} & \cdots & v_{K,P} \end{bmatrix}$ is a matrix of K kernels, each row is an array of P parameters / weights to learn, $P < D$.
- $h = \begin{bmatrix} h_{1,1} & \ldots & h_{1,U} \\ \vdots & \ddots & \vdots \\ h_{K,1} & \cdots & h_{K,U} \end{bmatrix}$ is an output matrix of hidden units. Each row is computing by applying a kernel to the input:

$$\forall u \in \{1, \ldots, U\}, \forall k \in \{1, \ldots, K\}, h_{k,u} = \sum_{p=1}^{P} v_{k,p} x_{u+p-1}$$

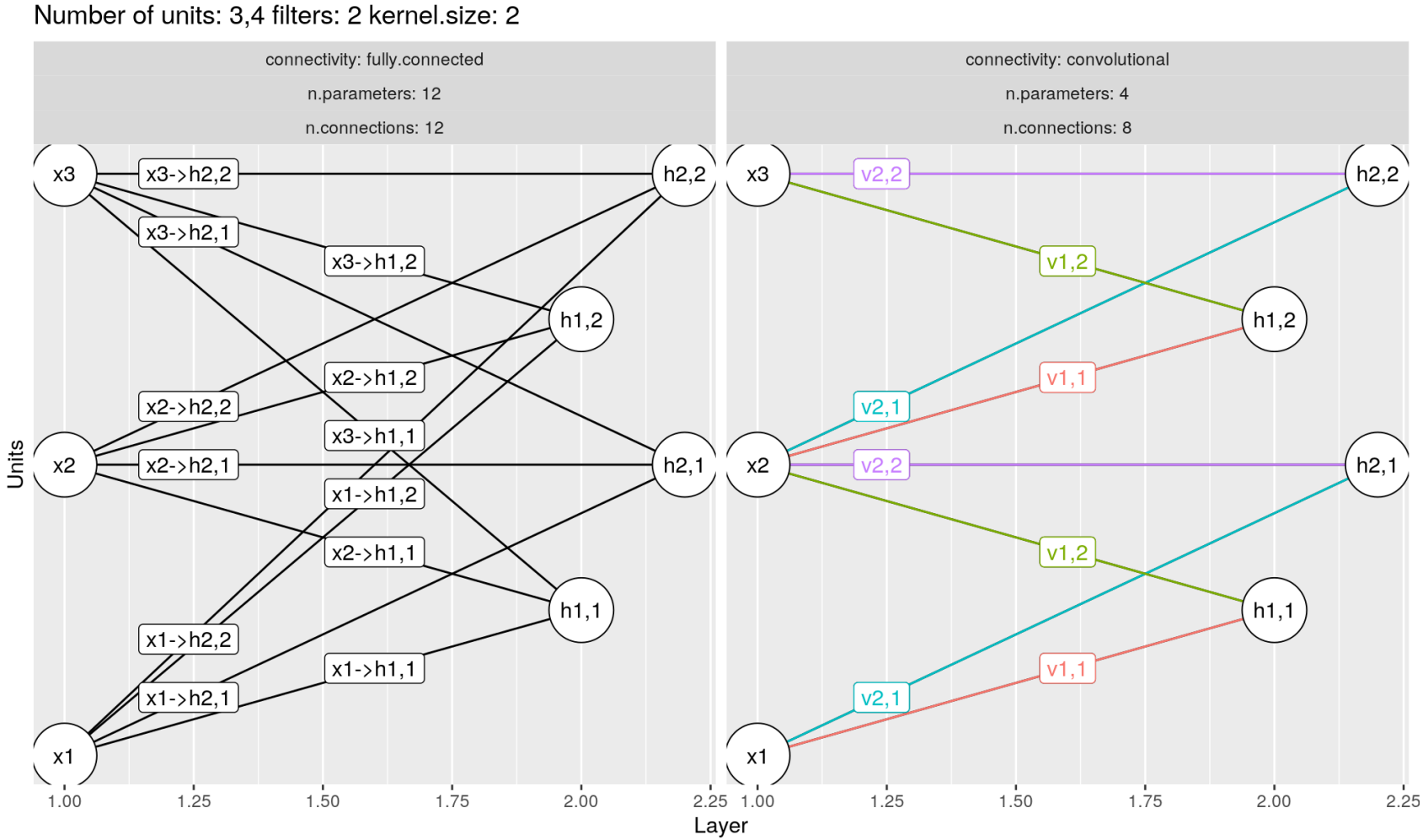
- EX in previous slide: $D = 3$ inputs, $P = 2$ parameters per kernel, $K = 2$ kernels $\Rightarrow U = 2$ output units per kernel, 4 output units total.

Colors for unique weight parameters to learn




```
torch.nn.Conv1d(in_channels=1,
  out_channels=1, kernel_size=2)
```

Colors for unique weight parameters to learn




```
torch.nn.Conv1d(in_channels=1,
  out_channels=2, kernel_size=2)
```

Matrix interpretation of convolution

$$\begin{bmatrix} h_{1,1} \\ h_{1,2} \\ h_{2,1} \\ h_{2,2} \end{bmatrix} = \begin{bmatrix} v_{1,1} & v_{1,2} & 0 \\ 0 & v_{1,1} & v_{1,2} \\ v_{2,1} & v_{2,2} & 0 \\ 0 & v_{2,1} & v_{2,2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \text{ (convolutional)}$$

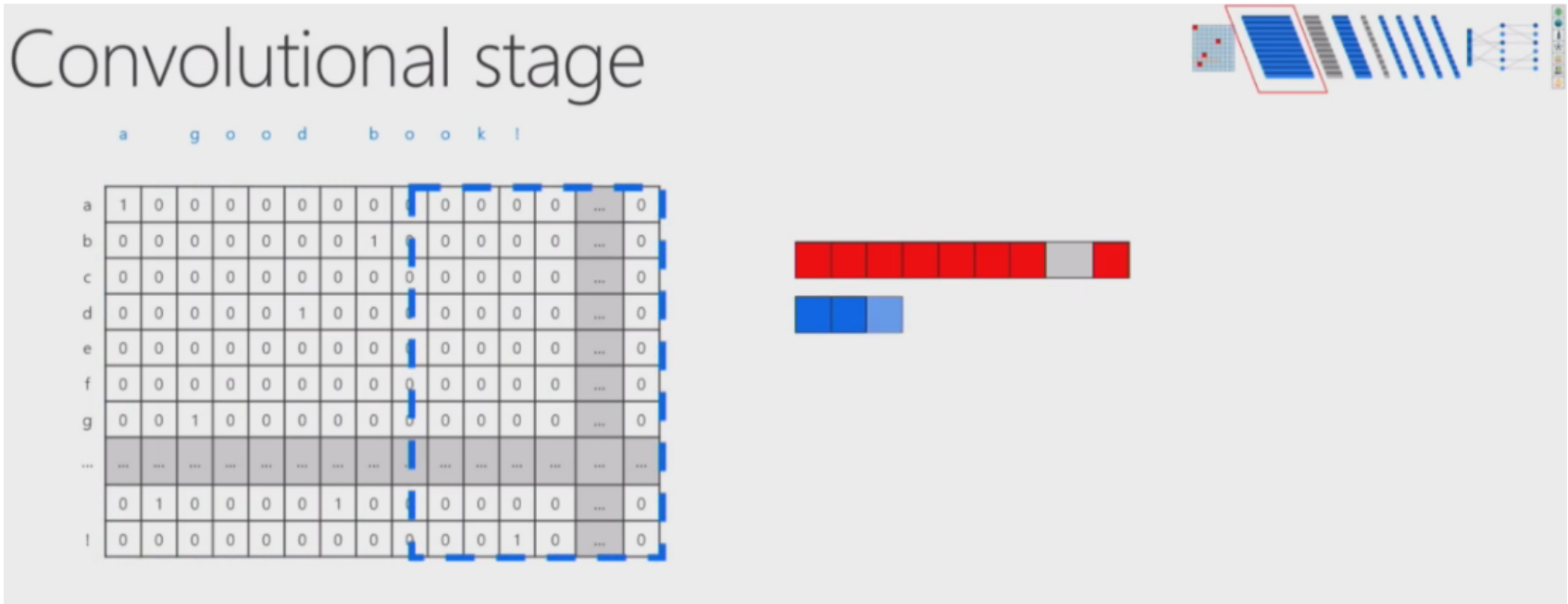
- Weight sharing: same weights used to compute different output units.
- Sparsity: zeros in weight matrix.

```
torch.nn.Conv1d(
  in_channels=1,
  out_channels=2,
  kernel_size=2)
```

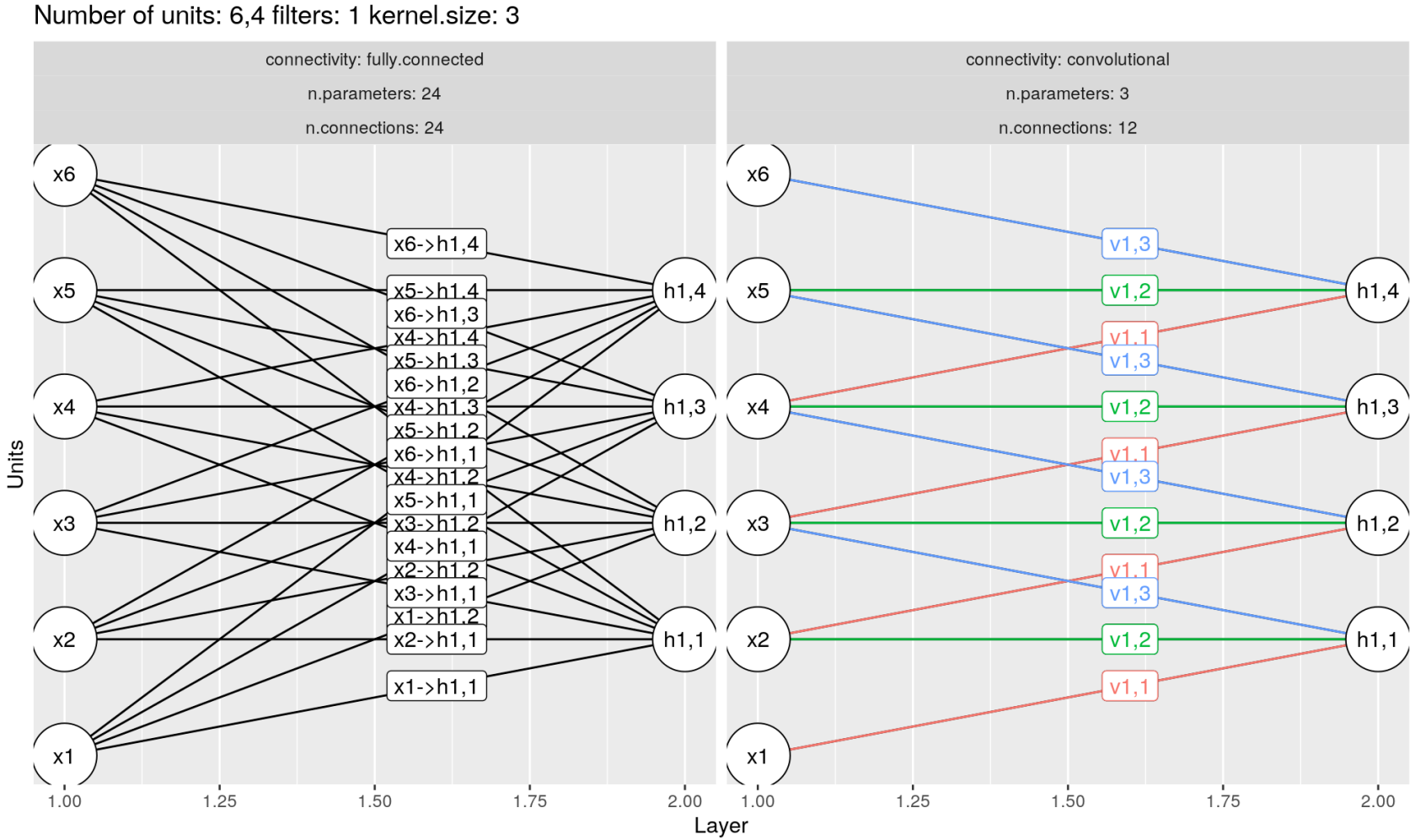
Video about convolution

`https://github.com/tdhock/useR2017-debrief`

Angus Taylor's talk at useR 2017.

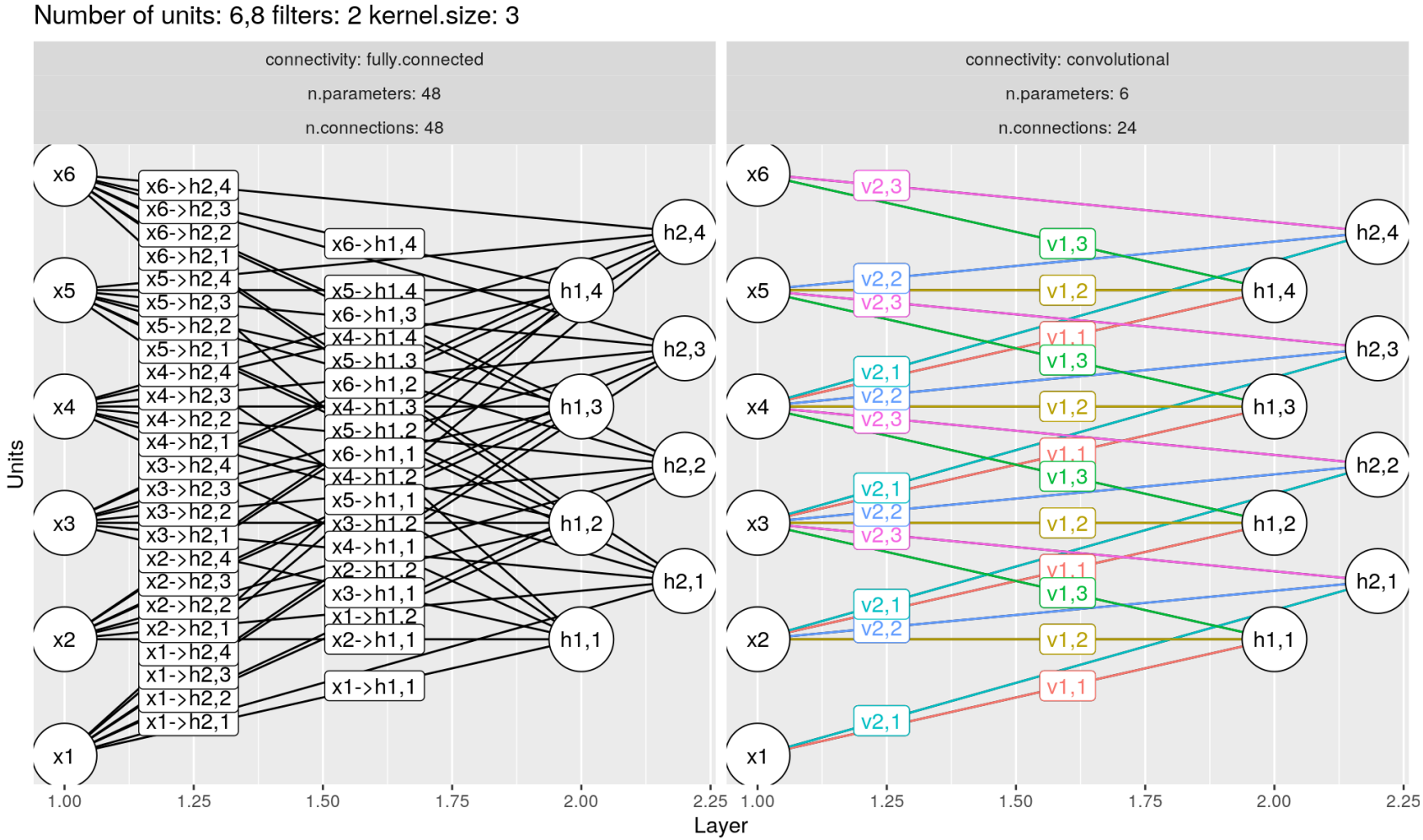


A more complex example (one filter)




```
torch.nn.Conv1d(in_channels=1,
  out_channels=1, kernel_size=3)
```

A more complex example (two filters)




```
torch.nn.Conv1d(in_channels=1,
  out_channels=2, kernel_size=3)
```

Architecture exercises

1D Convolution: if there are $D = 10$ inputs and $U = 5$ outputs,

- how many parameters to learn in a fully connected layer?
- a single kernel in a convolutional layer,
  1. how many parameters are there to learn?
  2. how many connections in the network diagram representation?
  3. how many zeros in the weight matrix representation?

2D Convolution: if you have a 10 x 10 pixel input image, and you apply 5 x 5 kernel,

1. How many parameters are there to learn in each filter?
2. How many parameters total if there are 20 filters?
3. How many output units per filter?
4. How many output units total using 20 filters?

Computation exercises

$$\mathbf{x} = \begin{bmatrix} 0 \\ 3 \\ 10 \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} -1 & 2 & -3 \\ 4 & -5 & 6 \end{bmatrix}$$

$$\mathbf{k} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$$

If $\mathbf{x}$ is an input vector,

1. what is the output vector when using $\mathbf{w}$ as the weight matrix in a fully connected layer?
2. what is the output vector when using $\mathbf{k}$ as the kernel in a 1d convolutional layer?

Pooling

Typical order of application in a layer is

1. Weight matrix multiplication (learned via gradient descent).
2. Activation, nonlinear function (not learned).
3. Pooling, reducing the number of units (not learned).

What is pooling?

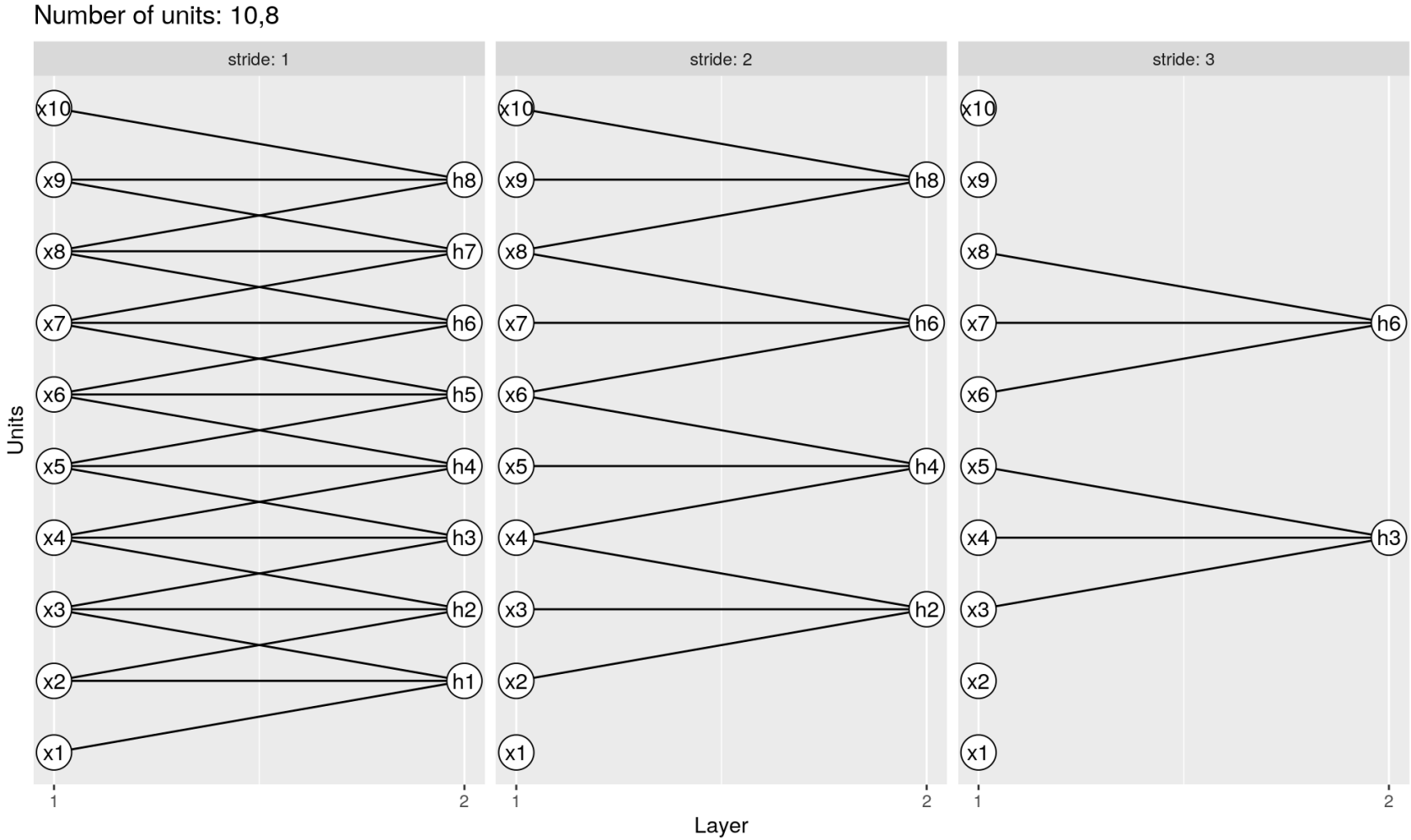
- Main purpose: reducing time/space during learning/prediction.
- Like convolution in that you apply some operation in a window over inputs; each window creates a single output unit.
- In convolution the operation is **multiplication** of inputs in window and corresponding weights, then **addition** to combine results in window.
- In pooling the operation is **mean or max** over all inputs in window.
- Pooling typically used over spatial dimension (independently for each channel/filter).

Stride

- This is the offset between windows on which the convolution/pooling is computed.
- Another technique for reducing number of output units and therefore time/space required during learning/prediction.
- In previous slides we were using stride of 1 (adjacent windows overlap each other and have redundant information).
- Often stride is set to kernel size (no overlapping windows) — this is the default in torch.

```
torch.nn.MaxPool1d(kernel_size=2, stride=2)
```

Stride diagram




```
torch.nn.MaxPool1d(kernel_size=3, stride=1)
torch.nn.MaxPool1d(kernel_size=3, stride=2)
torch.nn.MaxPool1d(kernel_size=3, stride=3)
```

Architecture exercises

1D Convolution: if there are $D = 20$ inputs and you have a kernel of size 5 with stride 5,

1. how many parameters are there to learn?
2. how many output units are there?
3. how many connections in the network diagram representation?
4. how many zeros in the weight matrix representation?

2D Pooling: if you have a 10 x 10 pixel input image, and you apply a 5 x 5 max pooling kernel with stride 5, how many output units are there?

Computation exercises

If $\mathbf{x} = [0, 3, 10, -2, 5, 1]$ is an input vector,
and $\mathbf{k} = [-2, 1]$ is a kernel,

1. what is the output vector when doing mean pooling with a stride of 2?
2. what is the output vector when doing max pooling with a stride of 3?
3. what is the output vector when using $\mathbf{k}$ as the kernel in a 1d convolutional layer with a stride of 2?