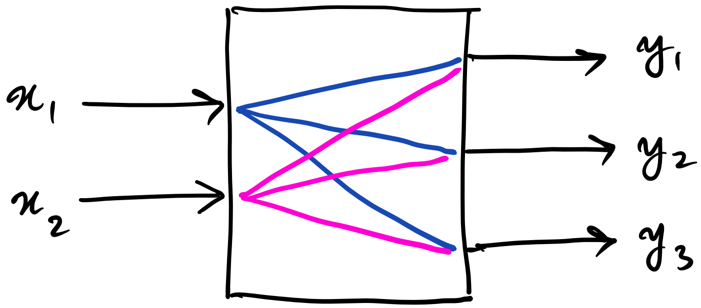
Linear layers



$$y_1 = w_{11} x_1 + w_{21} x_2$$

Re-write this as matrix-vector product

$$\begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}$$

Let's re-write this linear layer to handle mini-batches

$$\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \rightarrow \boxed{} \rightarrow \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \\ y_{31} & y_{23} \end{bmatrix}$$

$\longrightarrow$

Batches

And in the matrix-vector form

Batches

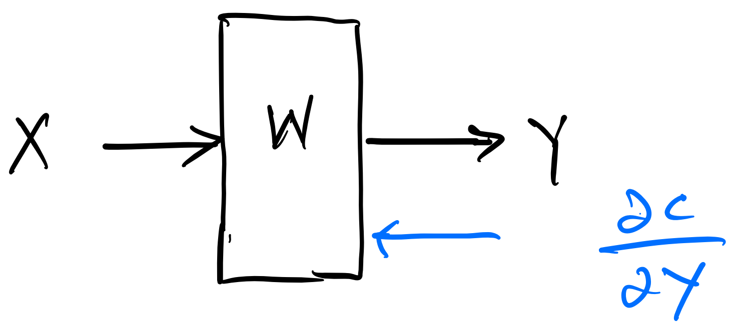
$$\underbrace{\begin{bmatrix} y_{11} & y_{21} & y_{31} \\ y_{12} & y_{22} & y_{32} \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} x_{11} & x_{21} \\ x_{21} & x_{22} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}}_{W}$$

For a linear layer: $Y = XW$ (Forward fn.)

In order to be able to backprop, we need to compute

$$\frac{\partial C}{\partial X} \quad \text{and} \quad \frac{\partial C}{\partial W}$$

Recall that we already have $\frac{\partial C}{\partial Y}$ (this was backpropagated)



In our example, $\frac{\partial c}{\partial Y}$ is the gradient whose size is the size of Y. c is a scalar.

Therefore, $\frac{\partial c}{\partial Y} \in \mathbb{R}^{2 \times 3}$

$$\frac{\partial c}{\partial Y} = \begin{bmatrix} \frac{\partial c}{\partial y_{11}} & \frac{\partial c}{\partial y_{21}} & \frac{\partial c}{\partial y_{31}} \\ \frac{\partial c}{\partial y_{12}} & \frac{\partial c}{\partial y_{22}} & \frac{\partial c}{\partial y_{32}} \end{bmatrix}$$

By applying chain-rule, we get

$$\frac{\partial c}{\partial X} = \frac{\partial c}{\partial Y} \cdot \frac{\partial Y}{\partial X}$$

Following from above

$$\frac{\partial c}{\partial X} = \begin{bmatrix} \frac{\partial c}{\partial x_{11}} & \frac{\partial c}{\partial x_{21}} \\ \frac{\partial c}{\partial x_{12}} & \frac{\partial c}{\partial x_{22}} \end{bmatrix}$$

Let's try to unpack $\frac{\partial Y}{\partial X}$. We begin by considering the one batch first.

$$\begin{bmatrix} y_{11} & y_{21} & y_{31} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} \end{bmatrix} \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}$$

This suggests that $\frac{\partial Y}{\partial X}$ is actually a Jacobian.

$$\frac{\partial Y}{\partial X} = \begin{bmatrix} \frac{\partial y_{11}}{\partial x_{11}} & \frac{\partial y_{11}}{\partial x_{21}} \\ \frac{\partial y_{21}}{\partial x_{11}} & \frac{\partial y_{21}}{\partial x_{21}} \\ \frac{\partial y_{31}}{\partial x_{11}} & \frac{\partial y_{31}}{\partial x_{21}} \end{bmatrix} = \begin{bmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{bmatrix} = W^T$$

Now let's try to dig a little deeper. to see what is $\frac{\partial C}{\partial x_{11}}$?

$$\frac{\partial c}{\partial x_{11}} = \frac{\partial c}{\partial y_{11}}\frac{\partial y_{11}}{\partial x_{11}} + \frac{\partial c}{\partial y_{21}}\frac{\partial y_{21}}{\partial x_{11}} + \frac{\partial c}{\partial y_{31}}\frac{\partial y_{31}}{\partial x_{11}}$$

$$= \begin{bmatrix} \frac{\partial c}{\partial y_{11}} & \frac{\partial c}{\partial y_{21}} & \frac{\partial c}{\partial y_{31}} \end{bmatrix} \begin{bmatrix} \frac{\partial y_{11}}{\partial x_{11}} \\ \frac{\partial y_{21}}{\partial x_{11}} \\ \frac{\partial y_{31}}{\partial x_{11}} \end{bmatrix}$$

We can re-write it for every dimension of a single sample.

$$\begin{bmatrix} \frac{\partial c}{\partial x_{11}} & \frac{\partial c}{\partial x_{21}} \end{bmatrix} = \begin{bmatrix} \frac{\partial c}{\partial y_{11}} & \frac{\partial c}{\partial y_{21}} & \frac{\partial c}{\partial y_{31}} \end{bmatrix} \begin{bmatrix} \frac{\partial y_{11}}{\partial x_{11}} & \frac{\partial y_{11}}{\partial x_{21}} \\ \frac{\partial y_{21}}{\partial x_{11}} & \frac{\partial y_{21}}{\partial x_{21}} \\ \frac{\partial y_{31}}{\partial x_{11}} & \frac{\partial y_{31}}{\partial x_{21}} \end{bmatrix}$$

$$\frac{\partial c}{\partial X} = \frac{\partial c}{\partial Y} W^T$$

This can be easily extended to the minibatch.

$$\begin{bmatrix} \frac{\partial c}{\partial x_{11}} & \frac{\partial c}{\partial x_{21}} \\ \frac{\partial c}{\partial x_{12}} & \frac{\partial c}{\partial x_{22}} \end{bmatrix} = \begin{bmatrix} \frac{\partial c}{\partial y_{11}} & \frac{\partial c}{\partial y_{21}} & \frac{\partial c}{\partial y_{31}} \\ \frac{\partial c}{\partial y_{12}} & \frac{\partial c}{\partial y_{22}} & \frac{\partial c}{\partial y_{32}} \end{bmatrix} W^T$$

We can similarly find $\frac{\partial c}{\partial W}$.

Useful properties of linear layers.

- Global context. Every output depends upon every input.
- Information mixin
- Frequently used as the last layer for many commonly used networks.

- Note that the above description ignores the activation functions.