Backpropagation: It's Just The Chain Rule

$$f(x_i, W, b) = Wx_i + b$$

$$\ell(x_i, y_i, W, b) = \max\left(0, 1 - y_i(Wx_i + b)\right) \quad \text{(SVM loss)}$$

// dx here = $\frac{d\ell}{dx_i}$

forward():

$$p = Wx_i$$

$$q = p + b$$

$$\ell = \max(0, 1 - y_i q)$$

Compute predictions and loss function

backward():

$$dq = -y_i \text{ if } y_i q < 1 \text{ else } 0$$

$$dp = dq \cdot 1 \quad \left(\frac{\partial \ell}{\partial q} \cdot \frac{dq}{dp}\right)$$

$$dW = dp \cdot x_i$$

$$db = dq \cdot 1$$

Compute gradients of parameters wrt loss function

Want to update W, b to make ℓ smaller.

To do gradient descent, we need:

$$\frac{\partial \ell}{\partial W} = \frac{\partial \ell}{\partial q} \cdot \frac{\partial q}{\partial p} \cdot \frac{\partial p}{\partial W} = \begin{cases} -y_i x_i & \text{if } y_i q < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial \ell}{\partial b} = \frac{\partial \ell}{\partial q} \cdot \frac{\partial q}{\partial b} = \begin{cases} -y_i & \text{if } y_i q < 1 \\ 0 & \text{otherwise} \end{cases}$$

Assume $-y_i x_i < 0$ - how do we update our parameters W, b?

step size/learning rate

$$W \leftarrow W - \alpha \frac{\partial \ell}{\partial W}$$

$$= W - \alpha(-y_i x_i)$$

$$= W + \alpha y_i x_i$$

$$b \leftarrow b - \alpha \frac{\partial \ell}{\partial b}$$

$$= b + \alpha y_i$$

y_i \ x_i	+	−
+	+	−
−	−	+