Worked Example: Backpropagation on XOR

Learning goals

By the end of this handout, you should be able to:

1. identify inputs, weights, biases, pre-activations, activations, predictions, and targets in a small neural network;
2. trace the chain rule from the output back to the hidden layer; and
3. update parameters using gradient descent.

This example performs only two stochastic-gradient-descent updates. It is a trace of the algorithm, not evidence that the network has already solved XOR.

1. XOR and the network

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

We use a `2-2-1` network: two inputs, two hidden units, and one output unit.

```
                hidden layer                          output

x1 --w11--> z1 <--w12-- x2 --> sigmoid --> h1 --v1--+
                        + b1                        +--> zo --sigmoid--> yhat
x1 --w21--> z2 <--w22-- x2 --> sigmoid --> h2 --v2--+
                        + b2                        bo
```

The first index identifies the hidden unit; the second identifies the input. Here z is a **pre-activation** (a weighted sum plus bias). The output z_o is also called the output **logit**: it plays the same linear-score role as $w^\mathsf{T}x + b$ in logistic regression. Hidden-layer values z_1, z_2 are pre-activations inside the network; they are not predictions.

$$\begin{aligned} z_1 &= w_{11}x_1 + w_{12}x_2 + b_1, \quad & h_1 &= \text{sigmoid}(z_1), \\ z_2 &= w_{21}x_1 + w_{22}x_2 + b_2, \quad & h_2 &= \text{sigmoid}(z_2), \\ z_o &= v_1h_1 + v_2h_2 + b_o, & \hat{y} &= \text{sigmoid}(z_o). \end{aligned}$$

$$\text{sigmoid}(t) = \frac{1}{1+e^{-t}}, \qquad \frac{d}{dt}\,\text{sigmoid}(t) = \text{sigmoid}(t)(1-\text{sigmoid}(t)), \qquad E = \frac{1}{2}(y-\hat{y})^2.$$

Here y is the actual target value and $\hat{y}$ is the predicted value. The learning rate is $\eta = 0.5$.

2. Initial parameters

Parameter	Initial value
w_{11}, w_{12}	(0.15, 0.20)
w_{21}, w_{22}	(-0.10, 0.05)
b_1, b_2	(0.10, -0.05)
v_1, v_2	(0.30, -0.25)
b_o	(0.05)

3. Backpropagation equations

First compute the output error signal:

$$\delta_o = \frac{\partial E}{\partial z_o} = (\hat{y} - y)\hat{y}(1 - \hat{y}).$$

Then compute the output-layer gradients:

$$\frac{\partial E}{\partial v_j} = \delta_o h_j, \qquad \frac{\partial E}{\partial b_o} = \delta_o.$$

Send the error signal to hidden unit (j):

$$\delta_j = \frac{\partial E}{\partial z_j} = \delta_o v_j h_j (1 - h_j).$$

Finally:

$$\frac{\partial E}{\partial w_{ji}} = \delta_j x_i, \qquad \frac{\partial E}{\partial b_j} = \delta_j.$$

Every parameter is updated using:

$$\theta_{new} = \theta_{old} - \eta \frac{\partial E}{\partial \theta}.$$

4. Update round 1: sample x_1, x_2, y=(0,1,1)

Forward pass

$$
\begin{aligned}
z_1 &= 0.15(0) + 0.20(1) + 0.10 = 0.30, \\
h_1 &= \text{sigmoid}(0.30) = 0.574443, \\
z_2 &= -0.10(0) + 0.05(1) - 0.05 = 0.00, \\
h_2 &= \text{sigmoid}(0.00) = 0.500000, \\
z_o &= 0.30(0.574443) - 0.25(0.500000) + 0.05 = 0.097333, \\
\hat{y} &= \text{sigmoid}(0.097333) = 0.524314.
\end{aligned}
$$

Round 1 forward trace

Step	Equation	Values used	Calculation	Result	Meaning
1	$z_1 = w_{11}x_1 + w_{12}x_2 + b_1$	$w_{11} = 0.15, w_{12} = 0.20, b_1 = 0.10, x_1 = 0, x_2 = 1$	$0.15(0) + 0.20(1) + 0.10$	0.300000	Linear score entering hidden unit 1
2	$h_1 = \text{sigmoid}(z_1)$	$z_1 = 0.300000$	sigmoid(0.300000)	0.574443	Output of hidden unit 1
3	$z_2 = w_{21}x_1 + w_{22}x_2 + b_2$	$w_{21} = -0.10, w_{22} = 0.05, b_2 = -0.05, x_1 = 0, x_2 = 1$	$-0.10(0) + 0.05(1) - 0.05$	0.000000	Linear score entering hidden unit 2
4	$h_2 = \text{sigmoid}(z_2)$	$z_2 = 0.000000$	sigmoid(0.000000)	0.500000	Output of hidden unit 2
5	$z_o = v_1h_1 + v_2h_2 + b_o$	$v_1 = 0.30, v_2 = -0.25, b_o = 0.05, h_1 = 0.574443, h_2 = 0.500000$	$0.30(0.574443) - 0.25(0.500000) + 0.05$	0.097333	Output linear score/logit
6	$\hat{y} = \text{sigmoid}(z_o)$	$z_o = 0.097333$	sigmoid(0.097333)	0.524314	Predicted y; compare with actual $y = 1$

The loss is:

$$
E = \frac{1}{2}(1 - 0.524314)^2 = 0.113139.
$$

Backward pass

$$
\delta_o = (0.524314 - 1)(0.524314)(1 - 0.524314) = -0.118640.
$$

$$
\begin{aligned}
\frac{\partial E}{\partial v_1} &= -0.068152, \quad \frac{\partial E}{\partial v_2} = -0.059320, \quad \frac{\partial E}{\partial b_o} = -0.118640, \\
\delta_1 &= -0.008701, \qquad \delta_2 = 0.007415.
\end{aligned}
$$

Round 1 backward trace

Step	Equation	Values used	Calculation	Result	Meaning
1	$\delta_o = (\hat{y} - y)\hat{y}(1 - \hat{y})$	$\hat{y} = 0.524314, y = 1$	$(0.524314-1)(0.524314)(1-0.524314)$	-0.118640	Loss signal at output pre-activation
2	$\partial E/\partial v_1 = \delta_o h_1$	$\delta_o = -0.118640, h_1 = 0.574443$	$(-0.118640)(0.574443)$	-0.068152	How v_1 affects the loss
3	$\partial E/\partial v_2 = \delta_o h_2$	$\delta_o = -0.118640, h_2 = 0.500000$	$(-0.118640)(0.500000)$	-0.059320	How v_2 affects the loss
4	$\partial E/\partial b_o = \delta_o$	$\delta_o = -0.118640$	-0.118640	-0.118640	How b_o affects the loss
5	$\delta_1 = \delta_o v_1 h_1(1 - h_1)$	$\delta_o = -0.118640, v_1 = 0.30, h_1 = 0.574443$	$(-0.118640)(0.30)$ $(0.574443)(0.425557)$ $(-0.118640)(-0.25)$	-0.008701	Loss signal at hidden unit 1
6	$\delta_2 = \delta_o v_2 h_2(1 - h_2)$	$\delta_o = -0.118640, v_2 = -0.25, h_2 = 0.500000$	$(0.500000)(0.500000)$	0.007415	Loss signal at hidden unit 2
7	$\partial E/\partial w_{12} = \delta_1 x_2$	$\delta_1 = -0.008701, x_2 = 1$	$(-0.008701)(1)$	-0.008701	Weight from x_2 to hidden unit 1
8	$\partial E/\partial w_{22} = \delta_2 x_2$	$\delta_2 = 0.007415, x_2 = 1$	$(0.007415)(1)$	0.007415	Weight from x_2 to hidden unit 2
9	$\partial E/\partial b_1 = \delta_1,\ \partial E/\partial b_2 = \delta_2$	$\delta_1 = -0.008701, \delta_2 = 0.007415$	Copy the two hidden deltas	-0.008701 0.007415	Hidden-bias gradients

Because $x_1 = 0$, both $\partial E/\partial w_{11}$ and $\partial E/\partial w_{21}$ are zero in this update. For the remaining input-to-hidden parameters:

$$\frac{\partial E}{\partial w_{12}} = -0.008701, \quad \frac{\partial E}{\partial w_{22}} = 0.007415, \quad \frac{\partial E}{\partial b_1} = -0.008701, \quad \frac{\partial E}{\partial b_2} = 0.007415.$$

Parameter update

Parameter	Old	Gradient	New
w_{11}	0.150000	0.000000	0.150000
w_{12}	0.200000	-0.008701	0.204350
b_1	0.100000	-0.008701	0.104350
w_{21}	-0.100000	0.000000	-0.100000
w_{22}	0.050000	0.007415	0.046292
b_2	-0.050000	0.007415	-0.053708
v_1	0.300000	-0.068152	0.334076
v_2	-0.250000	-0.059320	-0.220340
b_o	0.050000	-0.118640	0.109320

5. Update round 2: sample x_1, x_2, y=(1,1,0)

Use the new parameters from round 1.

Forward pass

$$
\begin{aligned}
z_1 &= 0.15(1) + 0.204350(1) + 0.104350 = 0.458701,\\
h_1 &= \text{sigmoid}(0.458701) = 0.612706,\\
z_2 &= -0.10(1) + 0.046292(1) - 0.053708 = -0.107415,\\
h_2 &= \text{sigmoid}(-0.107415) = 0.473172,\\
z_o &= 0.334076(0.612706) - 0.220340(0.473172) + 0.109320 = 0.209752,\\
\hat{y} &= \text{sigmoid}(0.209752) = 0.552247.
\end{aligned}
$$

Round 2 forward trace

Step	Equation	Values used	Calculation	Result	Meaning
1	$z_1 = w_{11}x_1 + w_{12}x_2 + b_1$	$w_{11} = 0.15, w_{12} = 0.204350, b_1 = 0.104350, x_1 = 1, x_2 = 1$	$0.15(1) + 0.204350(1) + 0.104350$	0.458701	Linear score entering hidden unit 1
2	$h_1 = \text{sigmoid}(z_1)$	$z_1 = 0.458701$	sigmoid(0.458701)	0.612706	Output of hidden unit 1
3	$z_2 = w_{21}x_1 + w_{22}x_2 + b_2$	$w_{21} = -0.10, w_{22} = 0.046292, b_2 = -0.053708, x_1 = 1, x_2 = 1$	$-0.10(1) + 0.046292(1) - 0.053708$	−0.107415	Linear score entering hidden unit 2
4	$h_2 = \text{sigmoid}(z_2)$	$z_2 = -0.107415$	sigmoid(−0.107415)	0.473172	Output of hidden unit 2
5	$z_o = v_1h_1 + v_2h_2 + b_o$	$v_1 = 0.334076, v_2 = -0.220340, b_o = 0.109320, h_1 = 0.612706, h_2 = 0.473172$	$0.334076(0.612706) - 0.220340(0.473172) + 0.109320$	0.209752	Output linear score/logit
6	$\hat{y} = \text{sigmoid}(z_o)$	$z_o = 0.209752$	sigmoid(0.209752)	0.552247	Predicted y; compare with actual $y = 0$

Since $y = 0$, the prediction is too high:

$$E = \frac{1}{2}(0 - 0.552247)^2 = 0.152488.$$

Backward pass and update

$$
\begin{aligned}
\delta_o &= (0.552247)(0.552247)(1 - 0.552247) = 0.136554,\\
\delta_1 &= (0.136554)(0.334076)(0.612706)(1 - 0.612706) = 0.010825,\\
\delta_2 &= (0.136554)(-0.220340)(0.473172)(1 - 0.473172) = -0.007500.
\end{aligned}
$$

Round 2 backward trace

Step	Equation	Values used	Calculation	Result	Meaning
1	$\delta_o = (\hat{y} - y)\hat{y}(1 - \hat{y})$	$\hat{y} = 0.552247, y = 0$	$(0.552247-0)(0.552247)(1-0.552247)$	0.136554	Positive signal: prediction is above actual target
2	$\delta_1 = \delta_o v_1 h_1(1 - h_1)$	$\delta_o = 0.136554, v_1 = 0.334076, h_1 = 0.612706$	$(0.136554)(0.334076)$ $(0.612706)(0.387294)$ $(0.136554)(-0.220340)$	0.010825	Loss signal at hidden unit 1
3	$\delta_2 = \delta_o v_2 h_2(1 - h_2)$	$\delta_o = 0.136554, v_2 = -0.220340, h_2 = 0.473172$	$(0.473172)(0.526828)$ $0.136554(0.612706)$	-0.007500 0.083668	Loss signal at hidden unit 2
4	$\partial E/\partial v_1 = \delta_o h_1,\ \partial E/\partial v_2 = \delta_o h_2$	$\delta_o = 0.136554, h_1 = 0.612706, h_2 = 0.473172$	$0.136554(0.473172)$	0.064614	Gradients for output weights
5	$\partial E/\partial b_o = \delta_o$	$\delta_o = 0.136554$	Copy δ_o	0.136554 0.010825	Gradient for output bias
6	$\partial E/\partial w_{ji} = \delta_j x_i$	$x_1 = x_2 = 1, \delta_1 = 0.010825, \delta_2 = -0.007500$	$\partial E/\partial w_{1i} = 0.010825,\ \partial E/\partial w_{2i} = -0.007500$	-0.007500	Each input-to-hidden weight receives its hidden delta

The final values after the second update are:

Parameter	After round 1	Gradient in round 2	After round 2
w_{11}	0.150000	0.010825	0.144587
w_{12}	0.204350	0.010825	0.198938
b_1	0.104350	0.010825	0.098938
w_{21}	-0.100000	-0.007500	-0.096250
w_{22}	0.046292	-0.007500	0.050043
b_2	-0.053708	-0.007500	-0.049957
v_1	0.334076	0.083668	0.292242
v_2	-0.220340	0.064614	-0.252647
b_o	0.109320	0.136554	0.041043

6. Takeaway

```
forward:  x → z → activation → prediction → loss
backward: loss → output delta → hidden delta → gradients
update:   parameter ← parameter − learning rate × gradient
repeat:   process more samples and monitor the training evidence
```

Backpropagation uses the chain rule to determine how the loss changes when each parameter changes. Gradient descent then moves each parameter in the direction that locally reduces the loss.

Check your understanding

1. Why were the gradients of (w_{11}) and (w_{21}) zero in round 1?
2. Why was b_o increased in round 1 but decreased in round 2?
3. Why can two updates not demonstrate that the XOR problem has converged?