EECS 442 Discussion 5

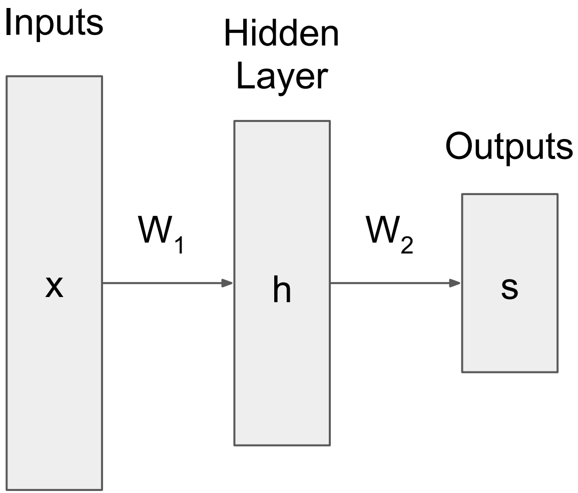
Backpropagation

Discussion Agenda

- Neural Network
- Layers
  - Fully Connected
  - ReLU
- Computational Graphs
  - Forward and Backward Passes
  - Chain Rule
- Backpropagation
- Optimizer
  - SGD with Momentum

Neural Networks

- Models that can learn varying features of data by approximating *almost any* nonlinear function

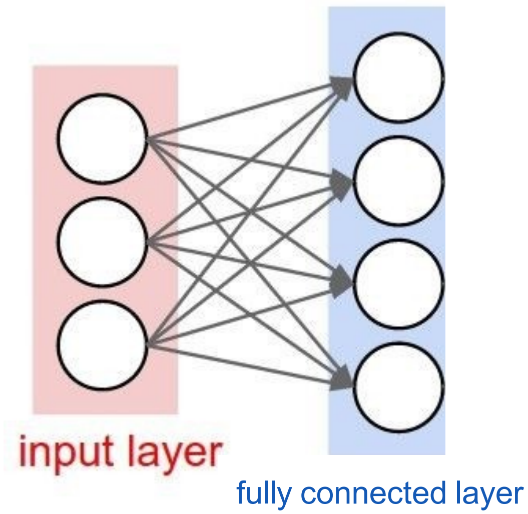


$$f(x) = W_2\, h(W_1 x + b_1) + b_2$$

Fully Connected Layer

For regular neural networks, the most common layer type is the fully-connected layer in which neurons between two adjacent layers are fully pairwise connected, but neurons within a single layer share no connections.

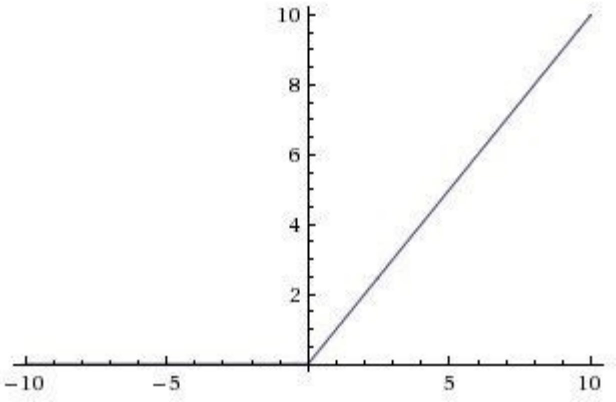
The weight dimension is (3, 4) in the right example.



CS231n Convolutional Neural Networks for Visual Recognition

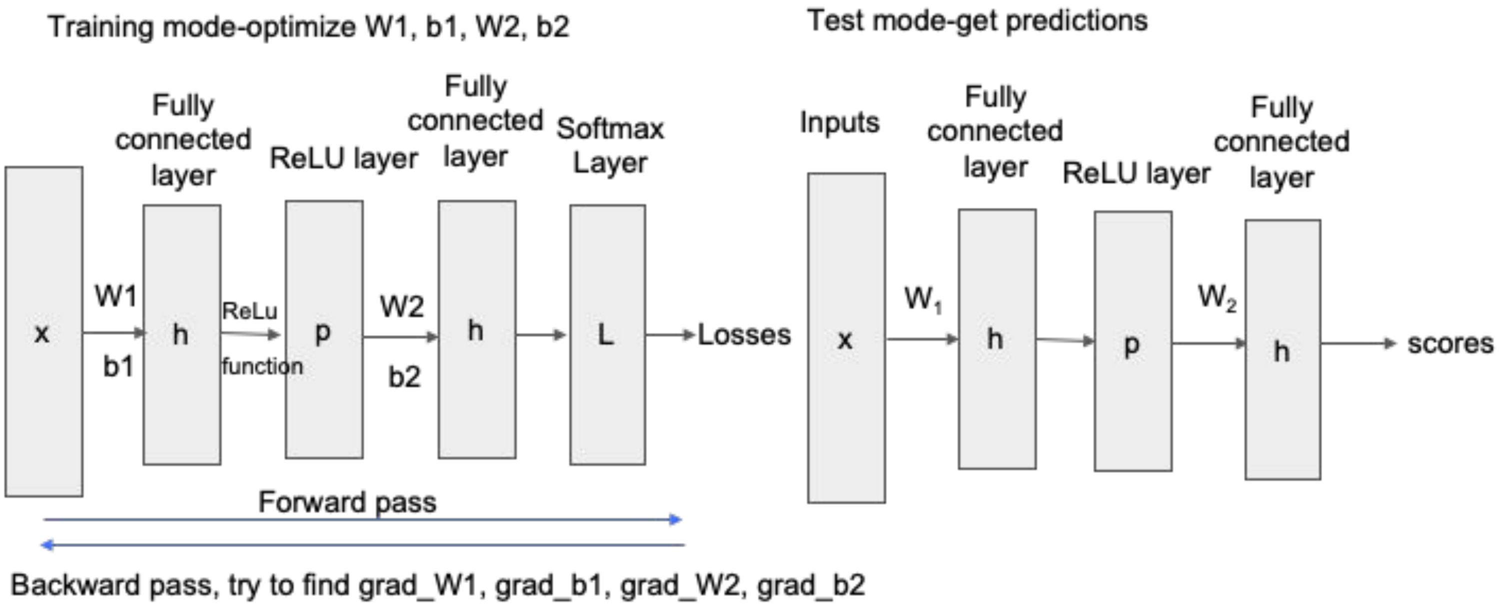
Rectified Linear Unit (ReLU)

- Activation function: introduces non-linearity
- Thresholded at zero
- $f(x) = \max(0, x)$
- Accelerates the convergence of Stochastic Gradient Descent (SGD)
- Simple to implement and fast to compute



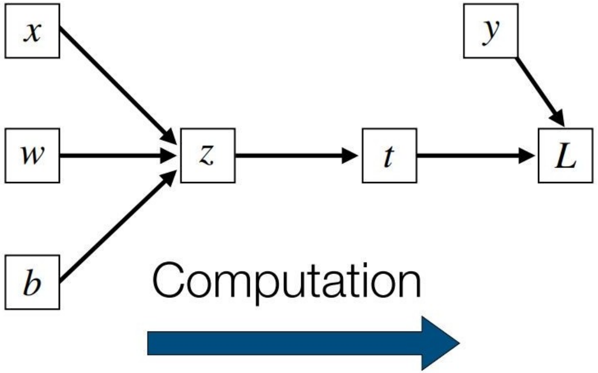
CS231n Convolutional Neural Networks for Visual Recognition

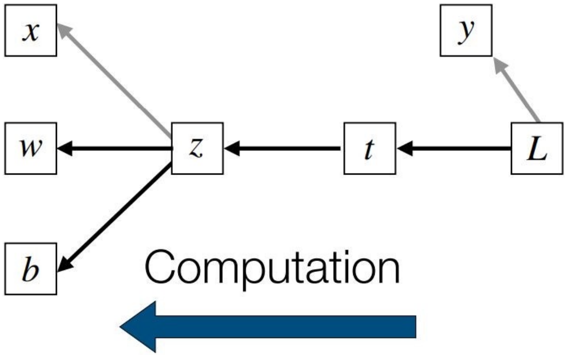
Network Structure for PS4



Computational Graphs

- Computing gradients is infeasible for complex models
  - Need to analytically derive all gradients
- Instead: modularize computation!





Forward and Backward Passes

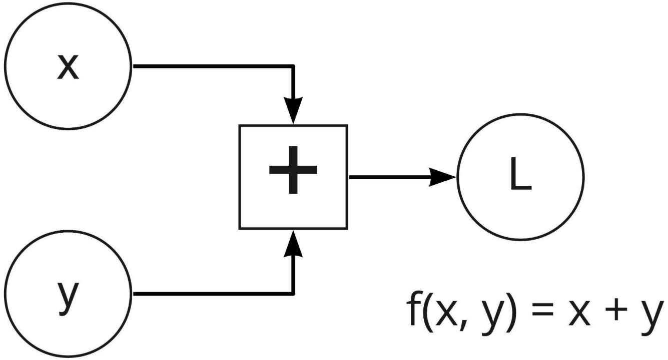
$$f(x, y) = x + y$$

1. **Forward pass:** Compute outputs

$$L = x + y$$

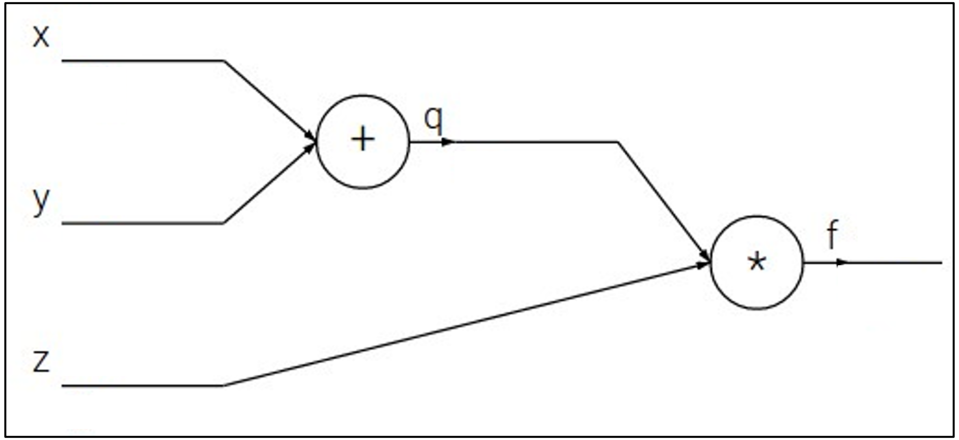
2. **Backward pass:** Compute derivatives

$$\frac{\partial f}{\partial y} = 1 \quad \frac{\partial f}{\partial x} = 1$$



Backpropagation: Simple Example

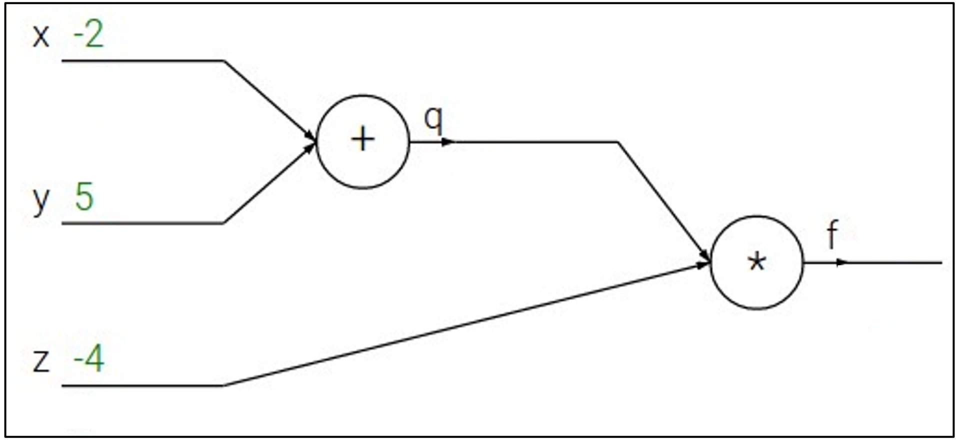
$$f(x, y, z) = (x + y) \cdot z$$



Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

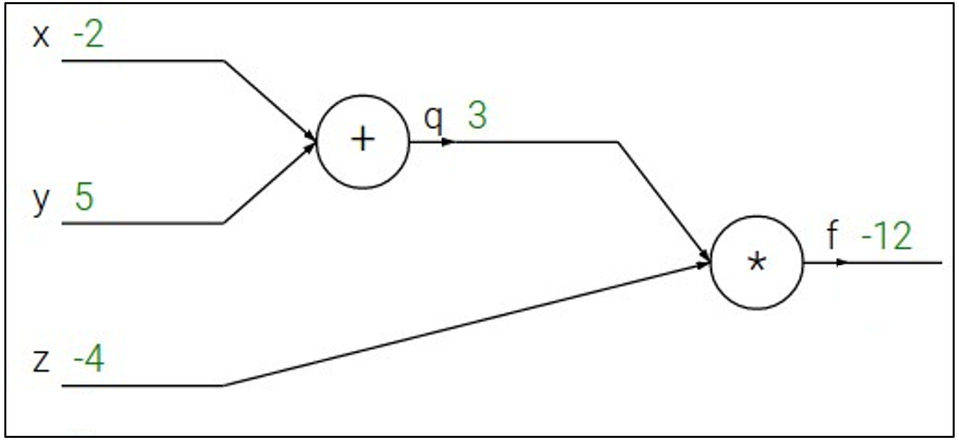
e.g. x = -2, y = 5, z = -4



Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. x = -2, y = 5, z = -4



1. Forward pass: Compute outputs

$$q = x + y \qquad f = q \cdot z$$

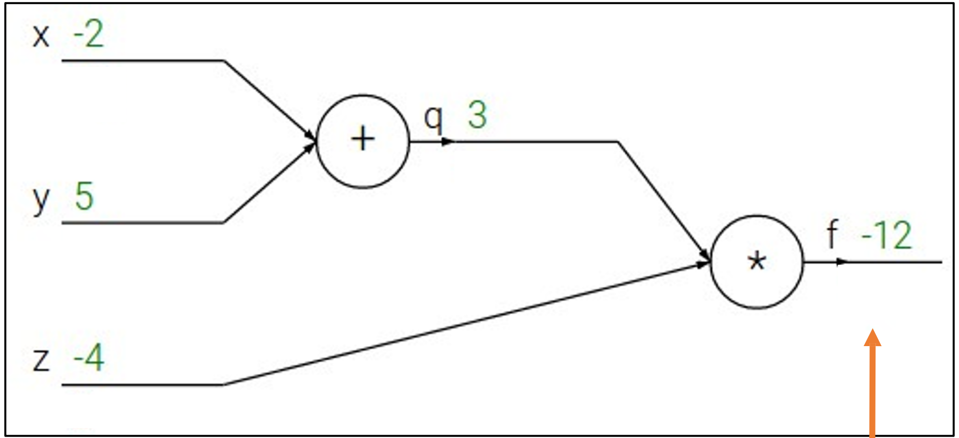
2. Backward pass: Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. x = -2, y = 5, z = -4



$$\boxed{\frac{\partial f}{\partial f}}$$

1. Forward pass: Compute outputs

$$q = x + y \qquad f = q \cdot z$$

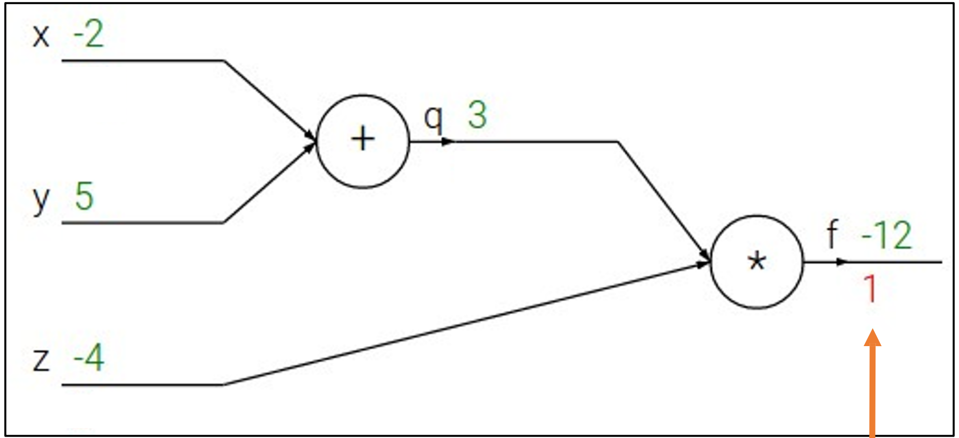
2. Backward pass: Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. x = -2, y = 5, z = -4



1. Forward pass: Compute outputs

$$q = x + y \qquad f = q \cdot z$$

$$\boxed{\frac{\partial f}{\partial f}}$$

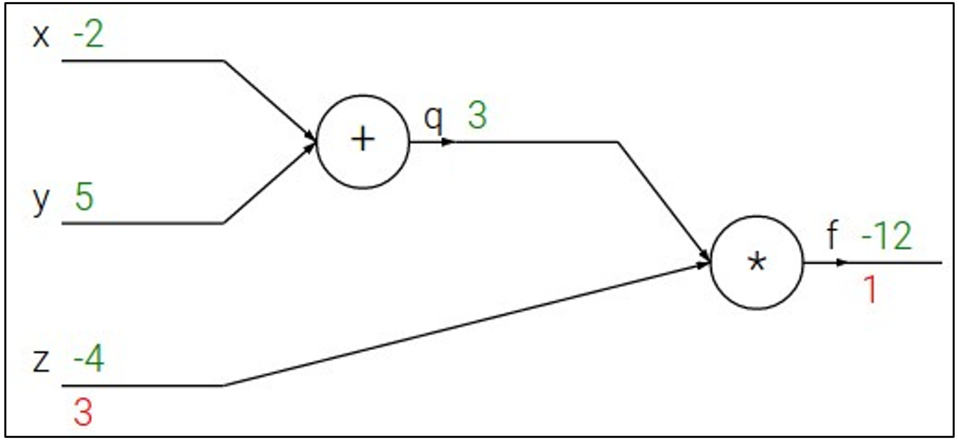
2. Backward pass: Compute derivatives

Want: $\frac{\partial f}{}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. x = -2, y = 5, z = -4



1. Forward pass: Compute outputs

$$q = x + y \qquad f = q \cdot z$$

2. Backward pass: Compute derivatives

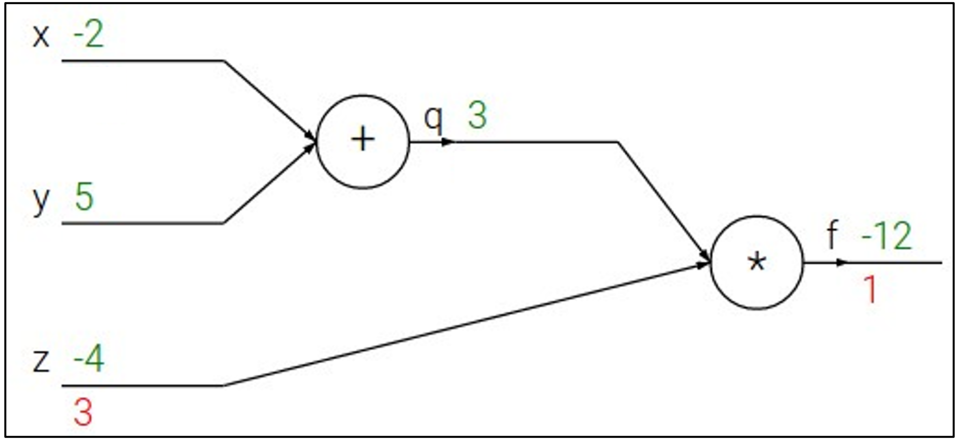
Want: $\frac{\partial f}{}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

$$\boxed{\frac{\partial f}{\partial z}}$$

Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. x = -2, y = 5, z = -4



1. Forward pass: Compute outputs

$$q = x + y \qquad \boxed{f = q \cdot z}$$

2. Backward pass: Compute derivatives

Want: $\frac{\partial f}{\quad}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

$$\boxed{\frac{\partial f}{\partial z} = q}$$

Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

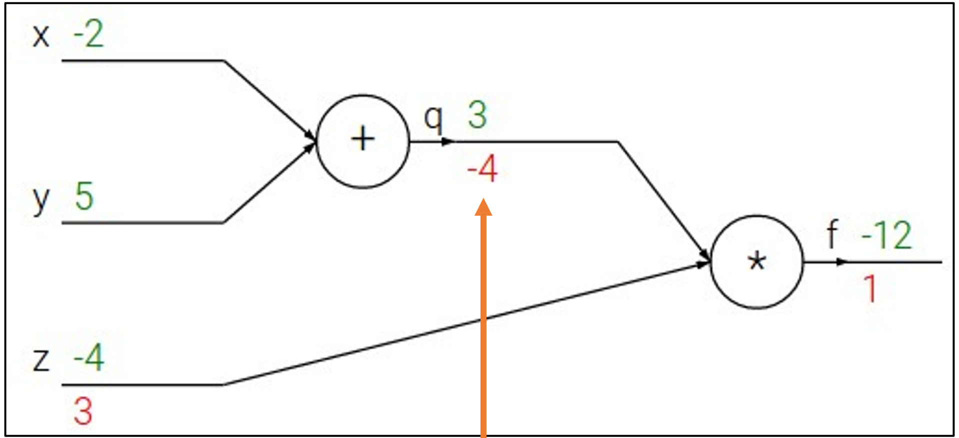
e.g. x = -2, y = 5, z = -4

1. Forward pass: Compute outputs

$$q = x + y \qquad f = q \cdot z$$

2. Backward pass: Compute derivatives

Want: $\frac{\partial f}{\quad}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

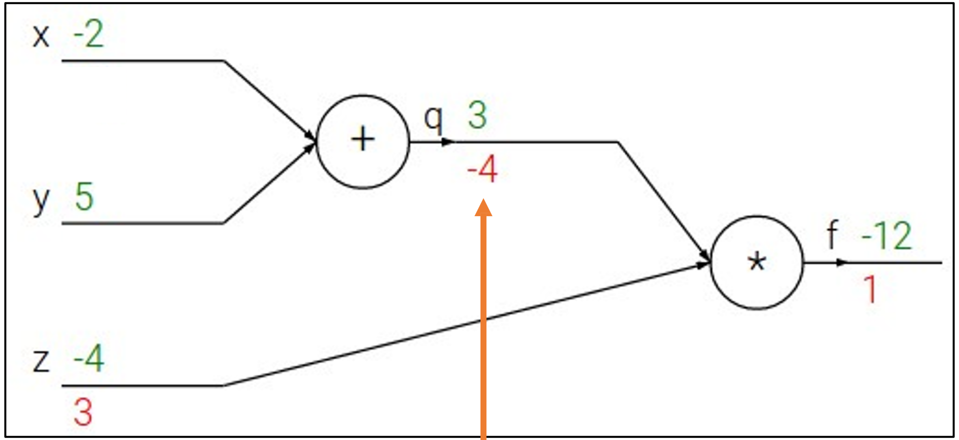


$$\frac{\partial f}{\partial q}$$

Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. x = -2, y = 5, z = -4



1. Forward pass: Compute outputs

$$q = x + y \qquad \boxed{f = q \cdot z}$$

$$\boxed{\frac{\partial f}{\partial q} = z}$$

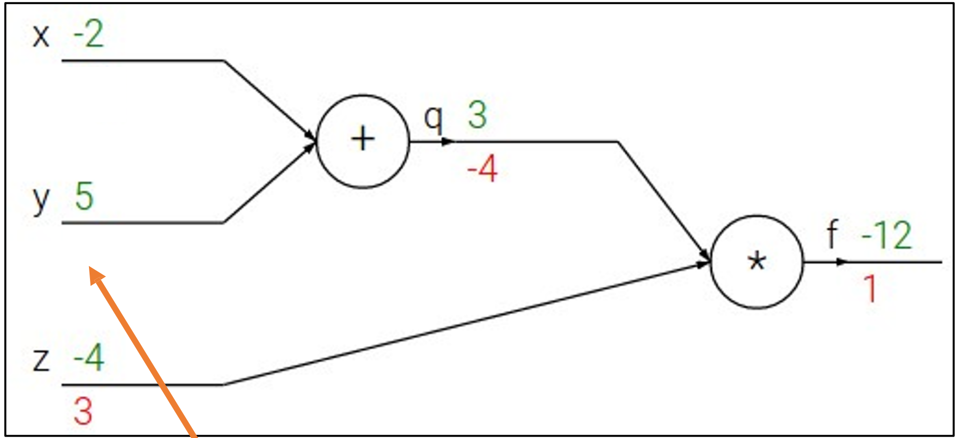
2. Backward pass: Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. x = -2, y = 5, z = -4



1. Forward pass: Compute outputs

$$q = x + y \qquad f = q \cdot z$$

2. Backward pass: Compute derivatives

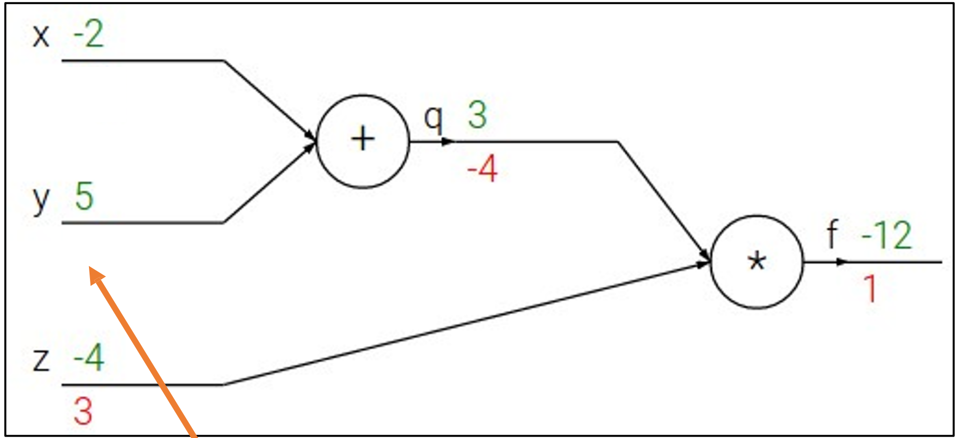
Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

$$\boxed{\frac{\partial f}{\partial y}}$$

Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. x = -2, y = 5, z = -4



1. Forward pass: Compute outputs

$$q = x + y \qquad f = q \cdot z$$

2. Backward pass: Compute derivatives

Want: $\dfrac{\partial f}{\partial x}, \dfrac{\partial f}{\partial y}, \dfrac{\partial f}{\partial z}$

Chain Rule

$$\boxed{\frac{\partial f}{\partial y} = \frac{\partial q}{\partial y}\frac{\partial f}{\partial q}}$$

Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

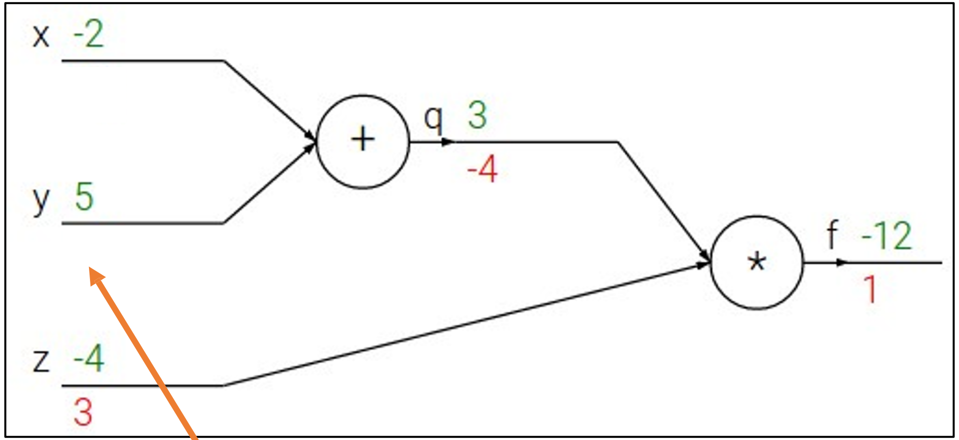
e.g. x = -2, y = 5, z = -4

1. Forward pass: Compute outputs

$$\boxed{q = x + y} \qquad f = q \cdot z$$

2. Backward pass: Compute derivatives

Want: $\dfrac{\partial f}{\partial x}, \dfrac{\partial f}{\partial y}, \dfrac{\partial f}{\partial z}$



Chain Rule

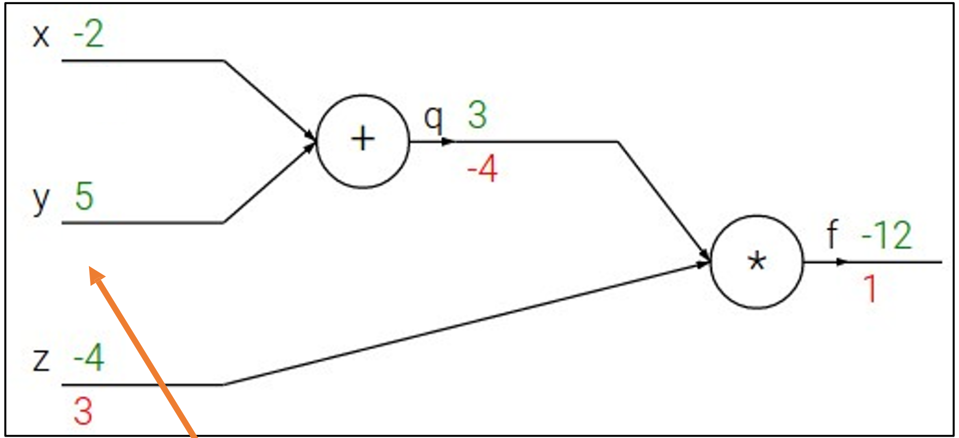
$$\boxed{\frac{\partial f}{\partial y} = \frac{\partial q}{\partial y}\frac{\partial f}{\partial q}}$$

Downstream Gradient **Local Gradient** **Upstream Gradient**

Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. x = -2, y = 5, z = -4



1. Forward pass: Compute outputs

$$\boxed{q = x + y} \qquad f = q \cdot z$$

2. Backward pass: Compute derivatives

Want: $\dfrac{\partial f}{}, \dfrac{\partial f}{\partial y}, \dfrac{\partial f}{\partial z}$

Chain Rule

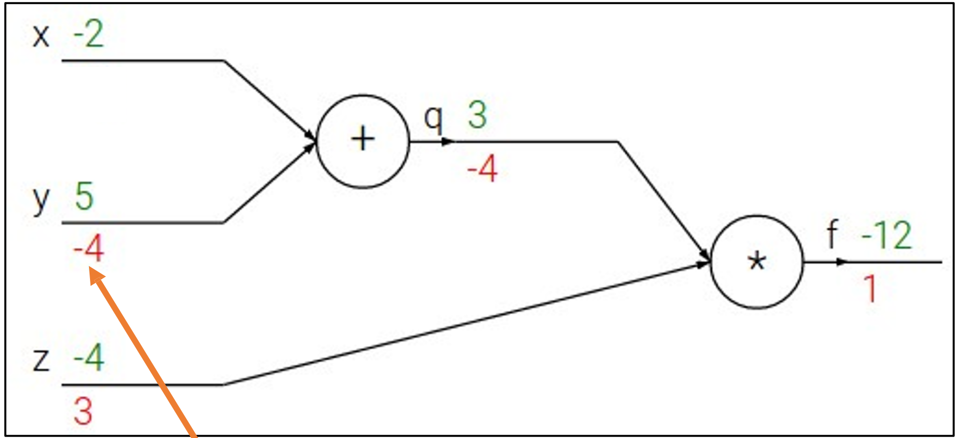
$$\boxed{\frac{\partial f}{\partial y} = \frac{\partial q}{\partial y}\frac{\partial f}{\partial q}} \qquad \boxed{\frac{\partial q}{\partial y} = 1}$$

Downstream **Local Gradient** **Upstream**

Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. x = -2, y = 5, z = -4



1. Forward pass: Compute outputs

$$\boxed{q = x + y} \qquad f = q \cdot z$$

2. Backward pass: Compute derivatives

Want: $\dfrac{\partial f}{\partial x}, \dfrac{\partial f}{\partial y}, \dfrac{\partial f}{\partial z}$

Chain Rule

$$\boxed{\frac{\partial f}{\partial y} = \frac{\partial q}{\partial y}\frac{\partial f}{\partial q}} \qquad \boxed{\frac{\partial q}{\partial y} = 1}$$

Downstream Gradient **Local Gradient** **Upstream Gradient**

Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

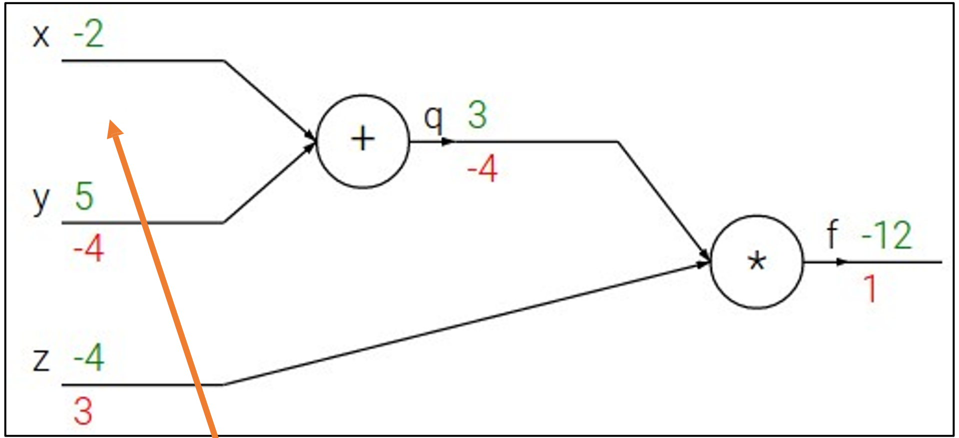
e.g. x = -2, y = 5, z = -4

1. Forward pass: Compute outputs

$$\boxed{q = x + y} \qquad f = q \cdot z$$

2. Backward pass: Compute derivatives

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



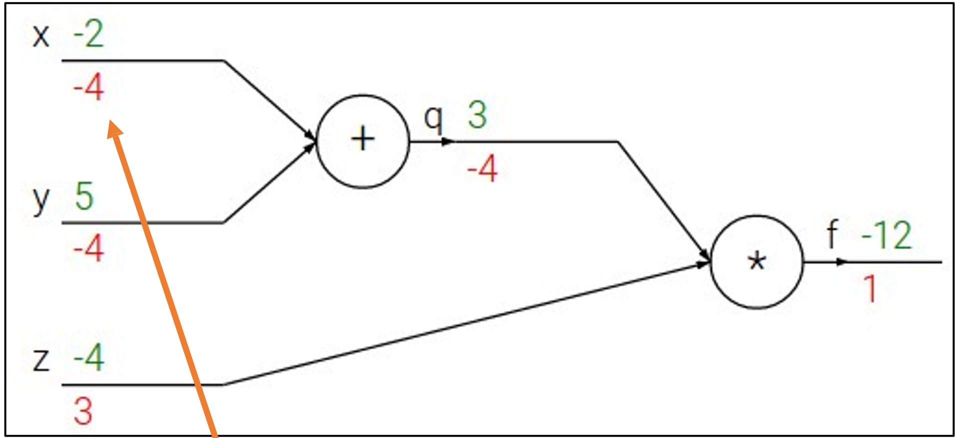
Chain Rule

$$\boxed{\underbrace{\frac{\partial f}{\partial x}}_{\text{Downstream Gradient}} = \underbrace{\frac{\partial q}{\partial x}}_{\text{Local Gradient}} \underbrace{\frac{\partial f}{\partial q}}_{\text{Upstream Gradient}}} \qquad \boxed{\frac{\partial q}{\partial x} = 1}$$

Backpropagation: Simple Example

$$f(x, y, z) = (x + y) \cdot z$$

e.g. x = -2, y = 5, z = -4



1. Forward pass: Compute outputs

$$\boxed{q = x + y} \qquad f = q \cdot z$$

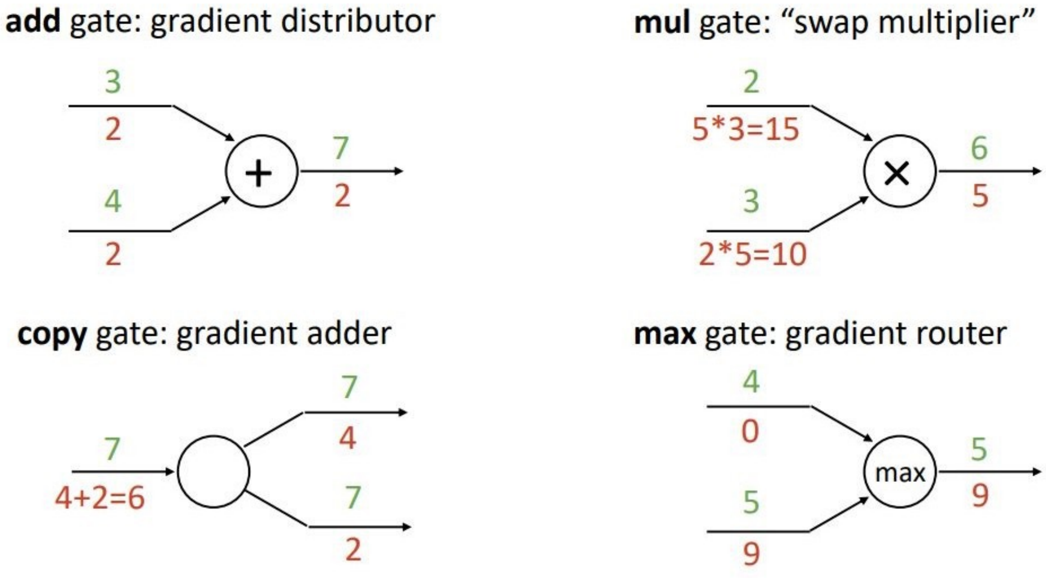
2. Backward pass: Compute derivatives

Want: $\dfrac{\partial f}{\partial x}, \dfrac{\partial f}{\partial y}, \dfrac{\partial f}{\partial z}$

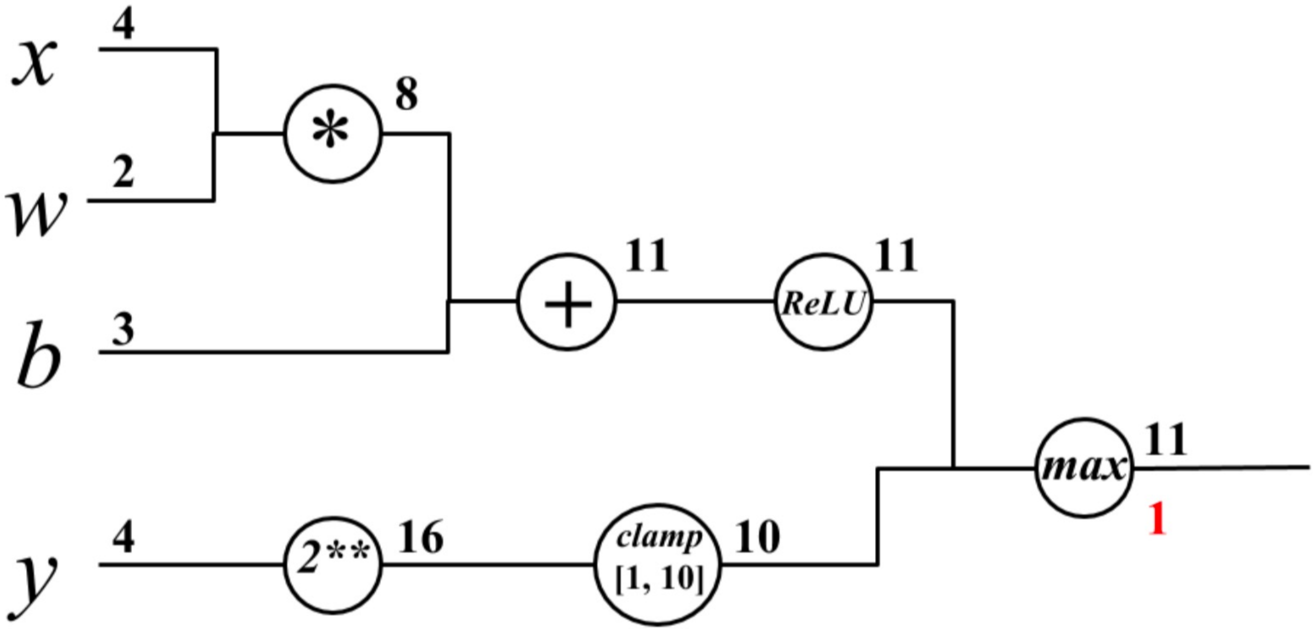
Chain Rule

$$\boxed{\underbrace{\frac{\partial f}{\partial x}}_{\text{Downstream Gradient}} = \underbrace{\frac{\partial q}{\partial x}}_{\text{Local Gradient}} \underbrace{\frac{\partial f}{\partial q}}_{\text{Upstream Gradient}}} \qquad \boxed{\frac{\partial q}{\partial x} = 1}$$

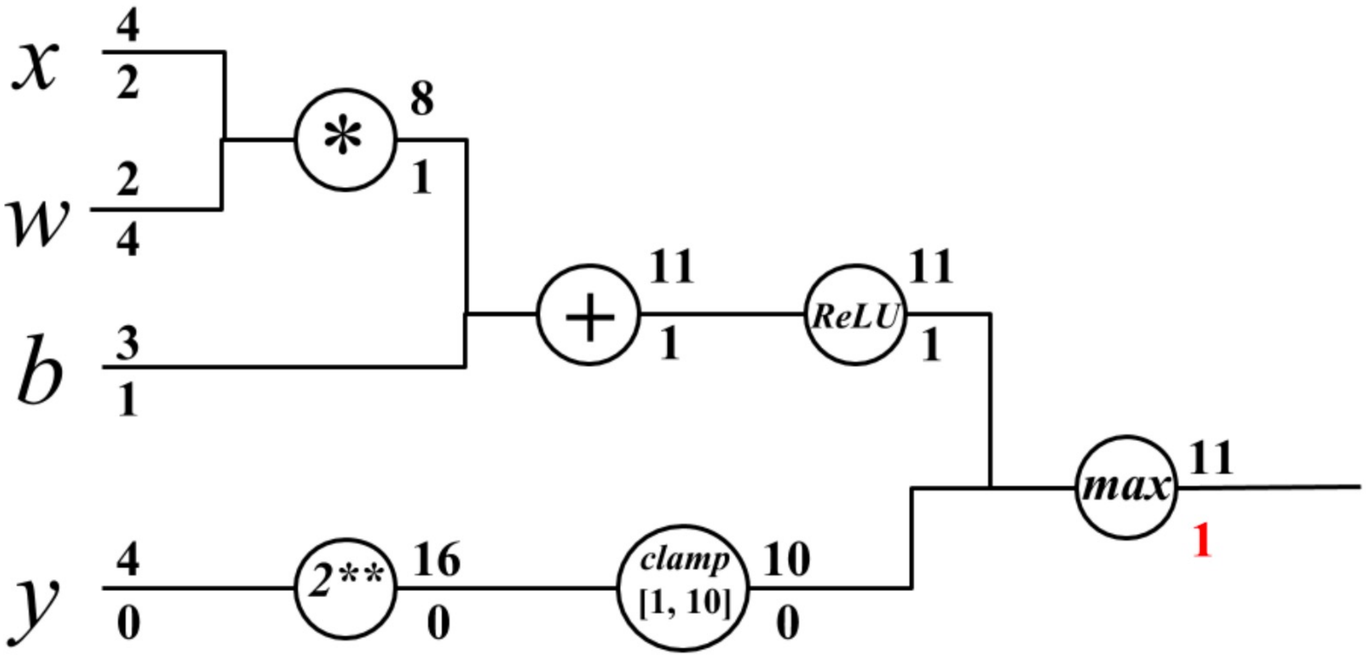
Backpropagation of some common operations



Practice 1

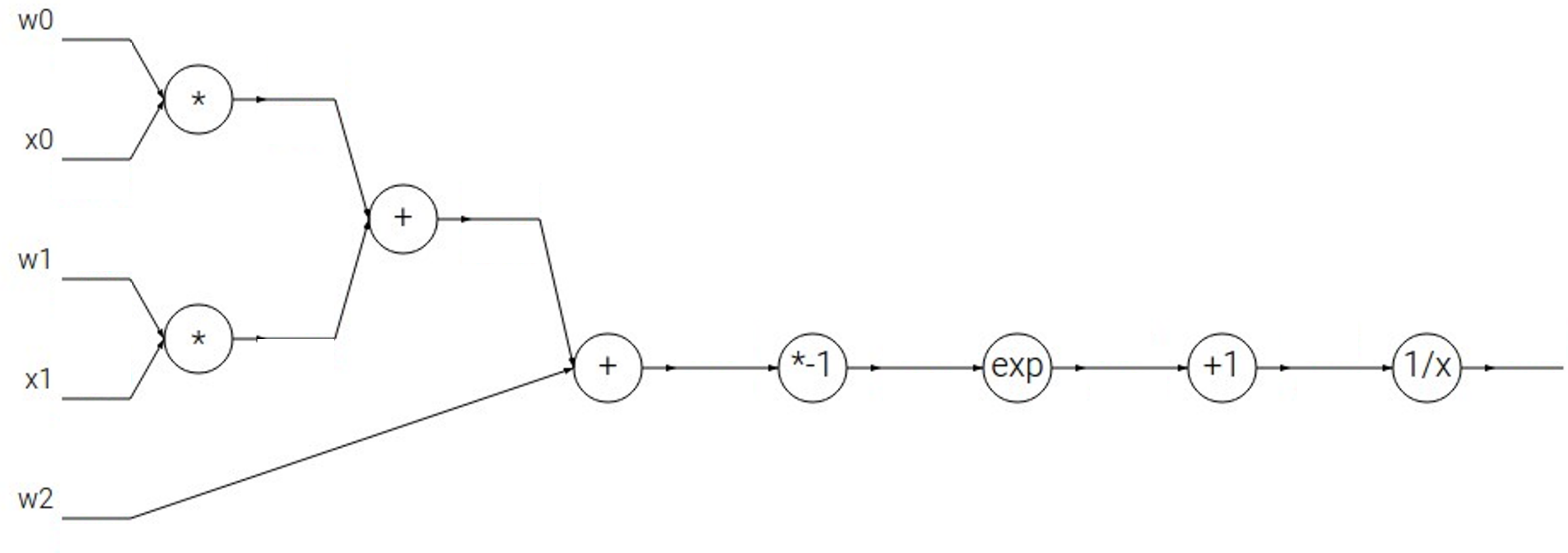


Practice 1



Another Example

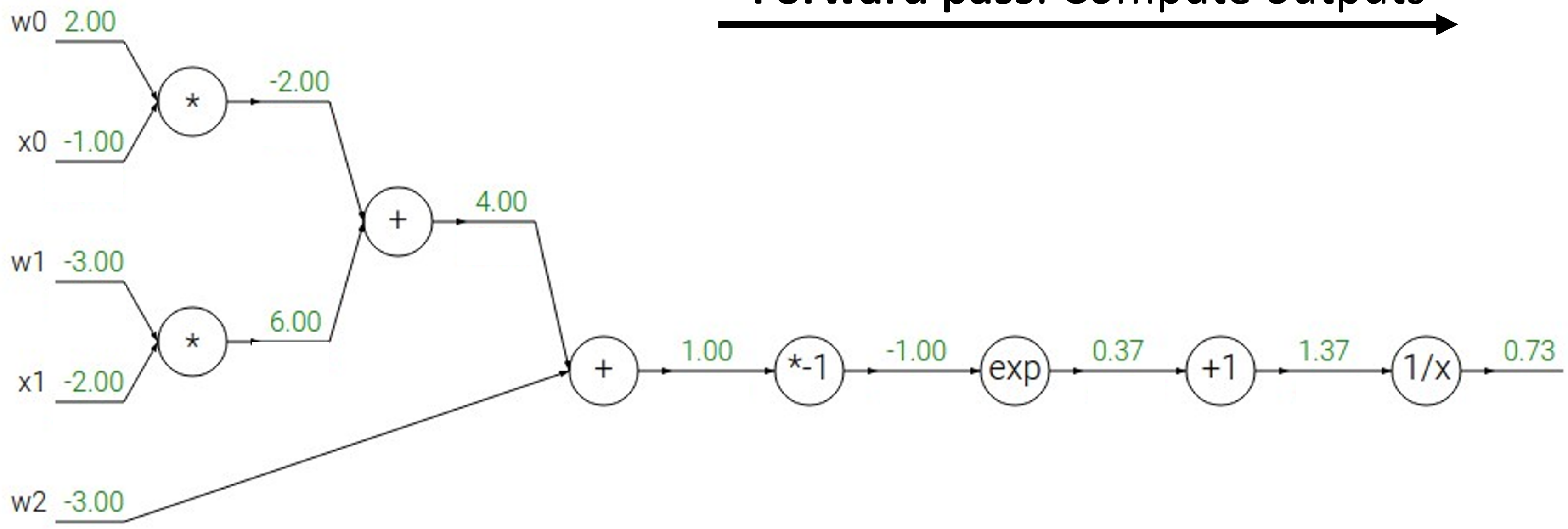
$$f(x,w) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}}$$



Another Example

$$f(x,w) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

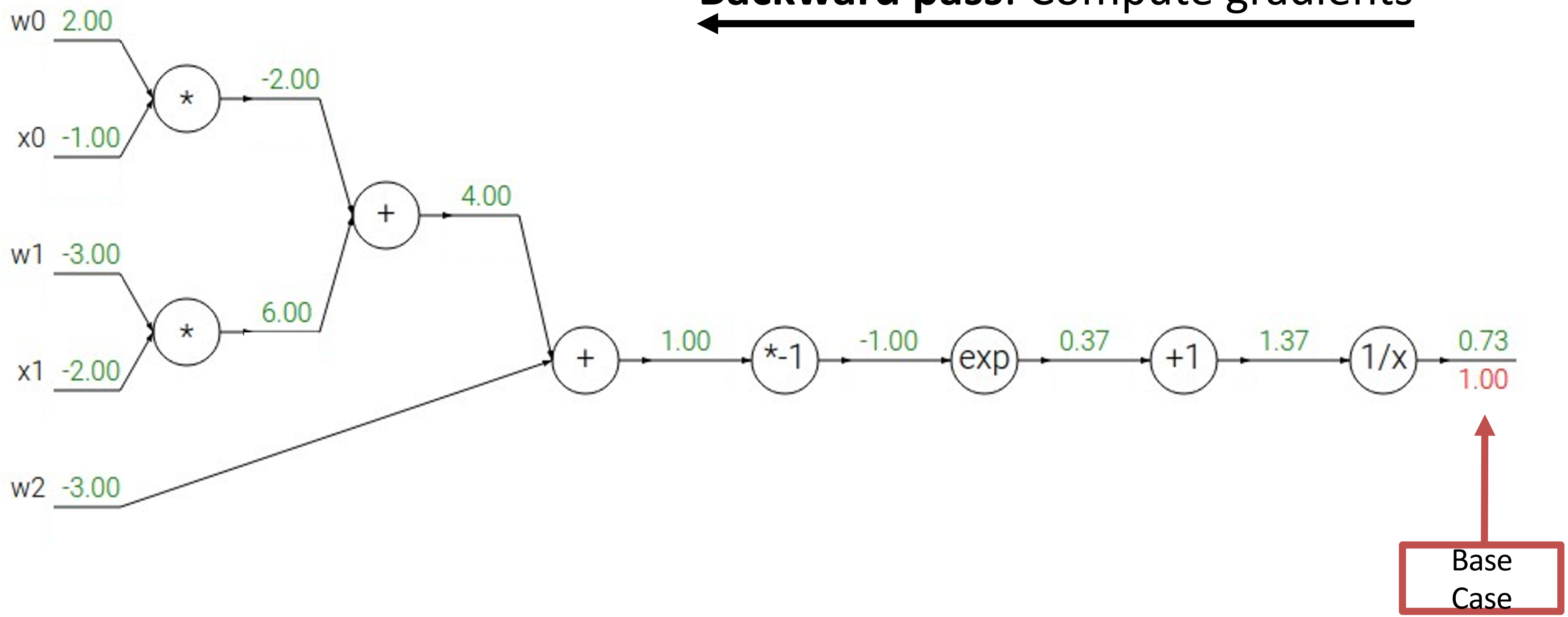
Forward pass: Compute outputs



Another Example

$$f(x,w) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

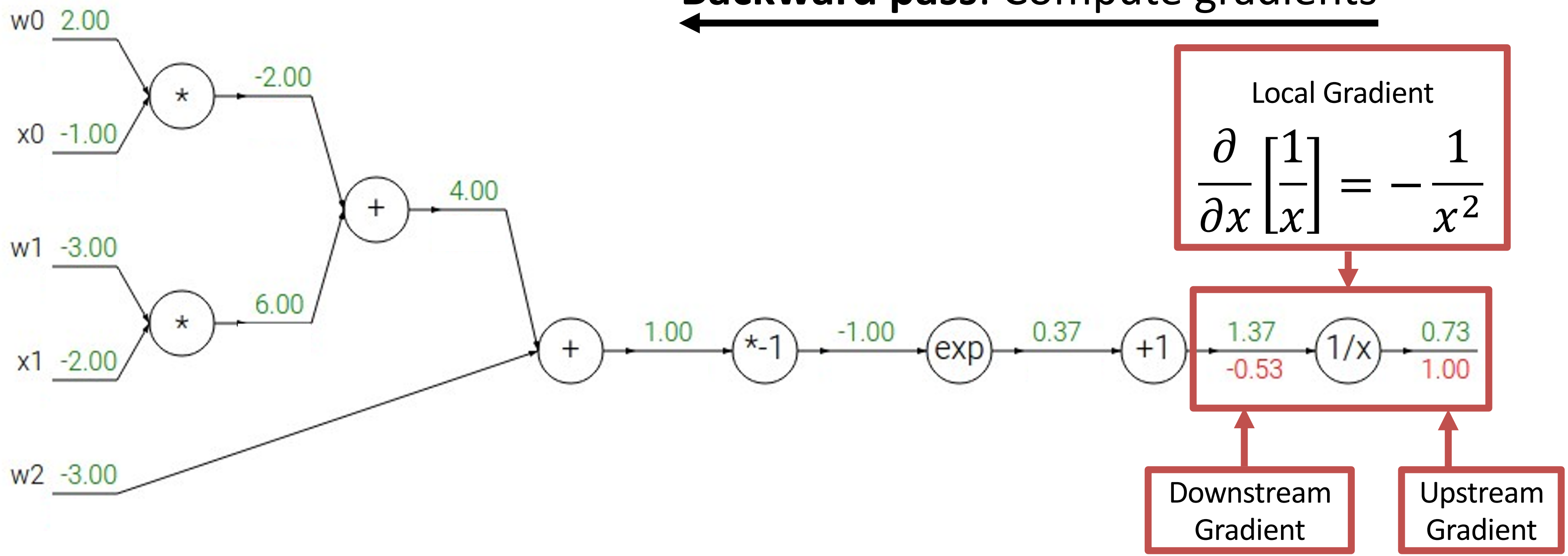
Backward pass: Compute gradients



Another Example

$$f(x,w) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

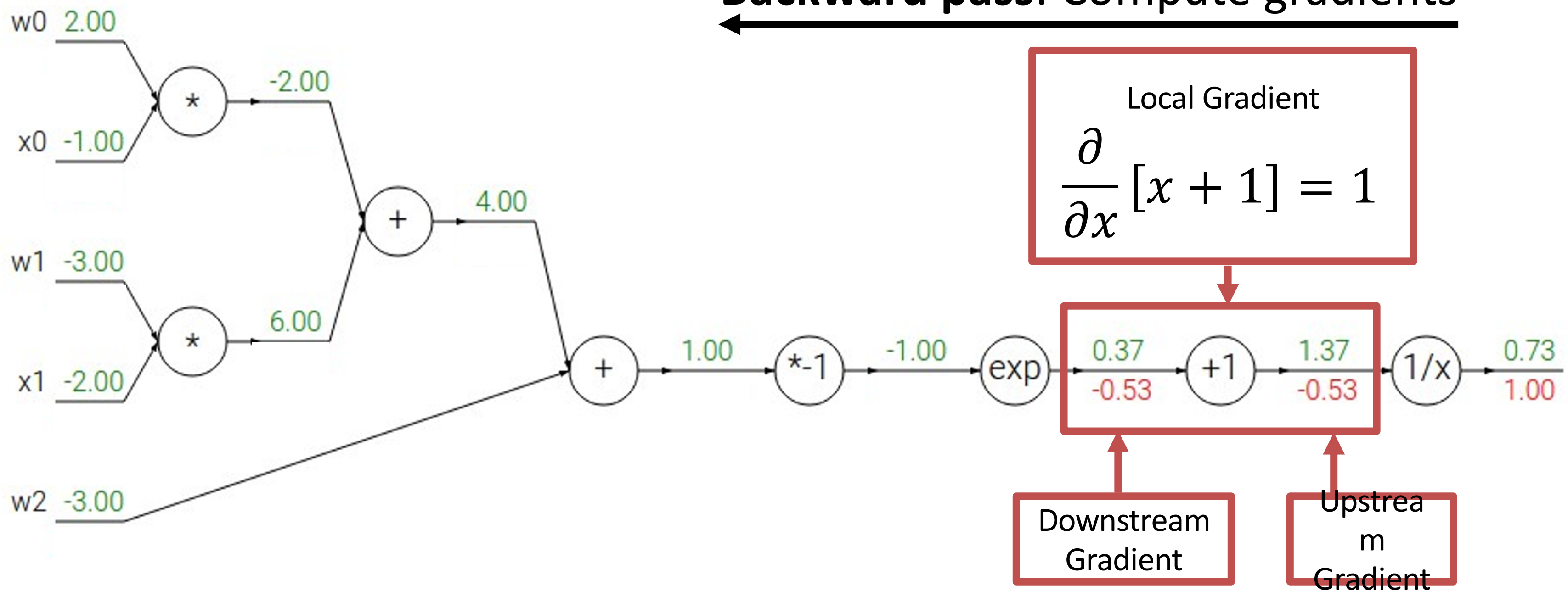
Backward pass: Compute gradients



Another Example

$$f(x,w) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

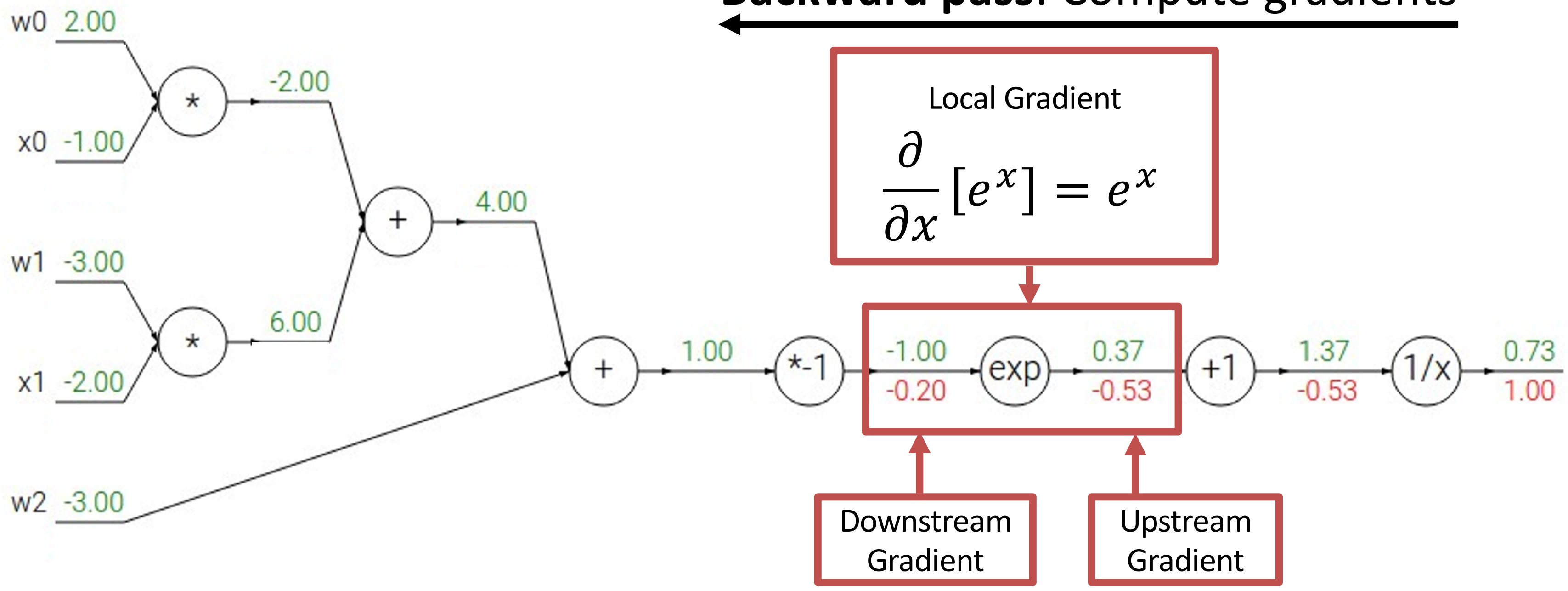
Backward pass: Compute gradients



Another Example

$$f(x,w) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

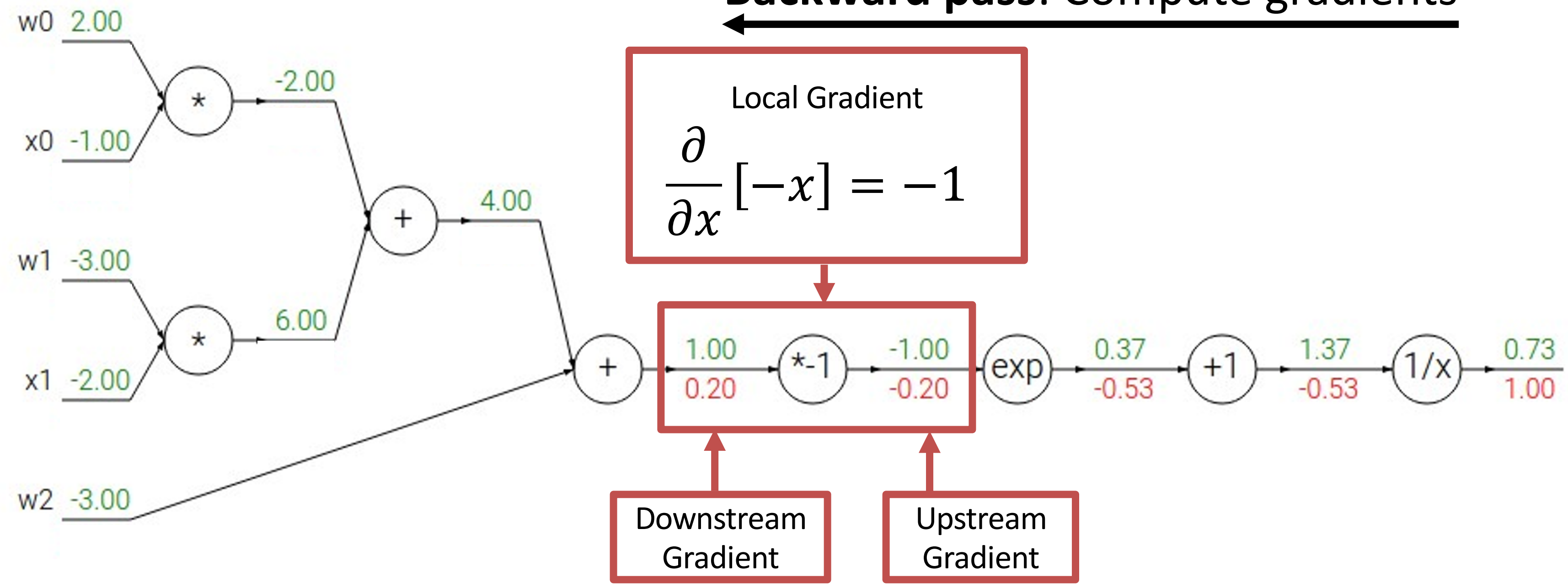
Backward pass: Compute gradients



Another Example

$$f(x,w) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

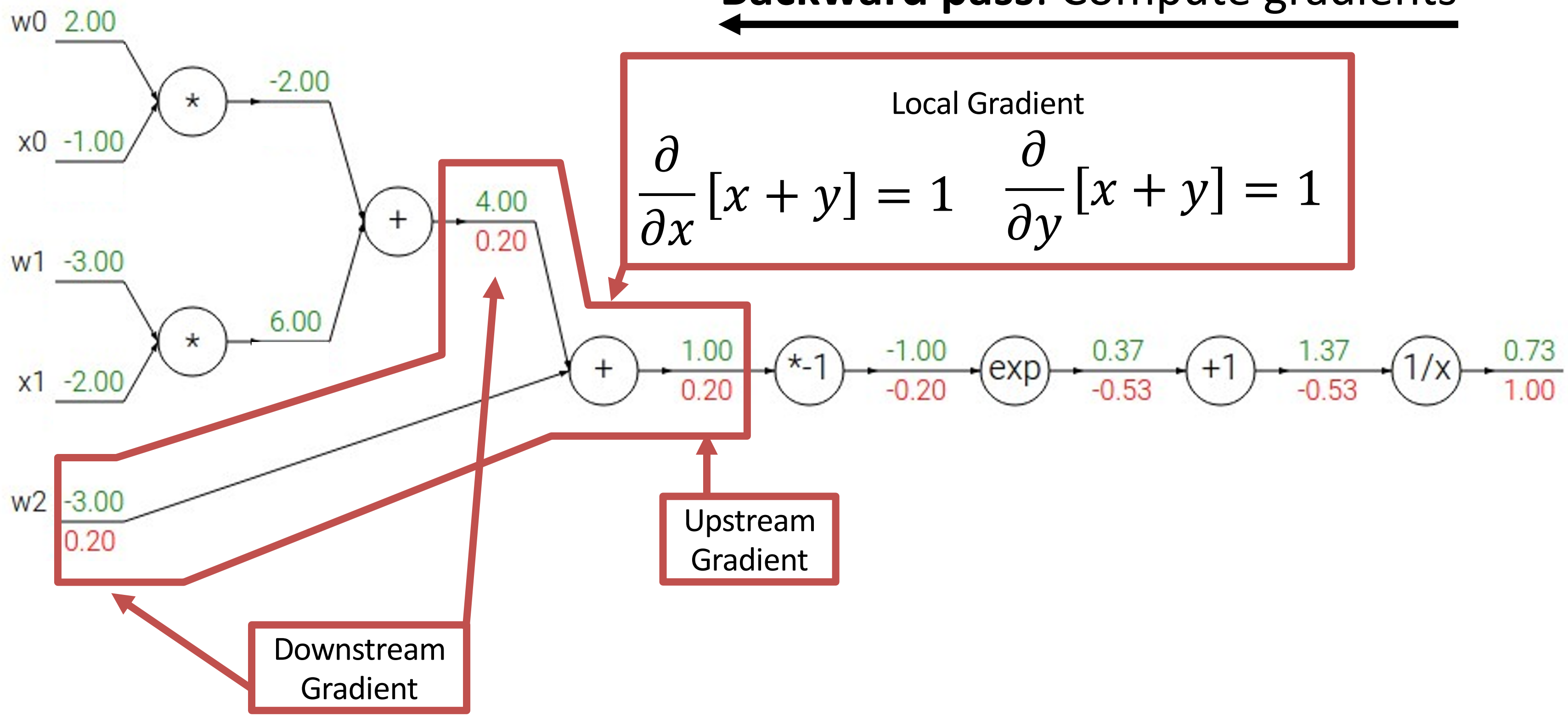
Backward pass: Compute gradients



Another Example

$$f(x,w) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

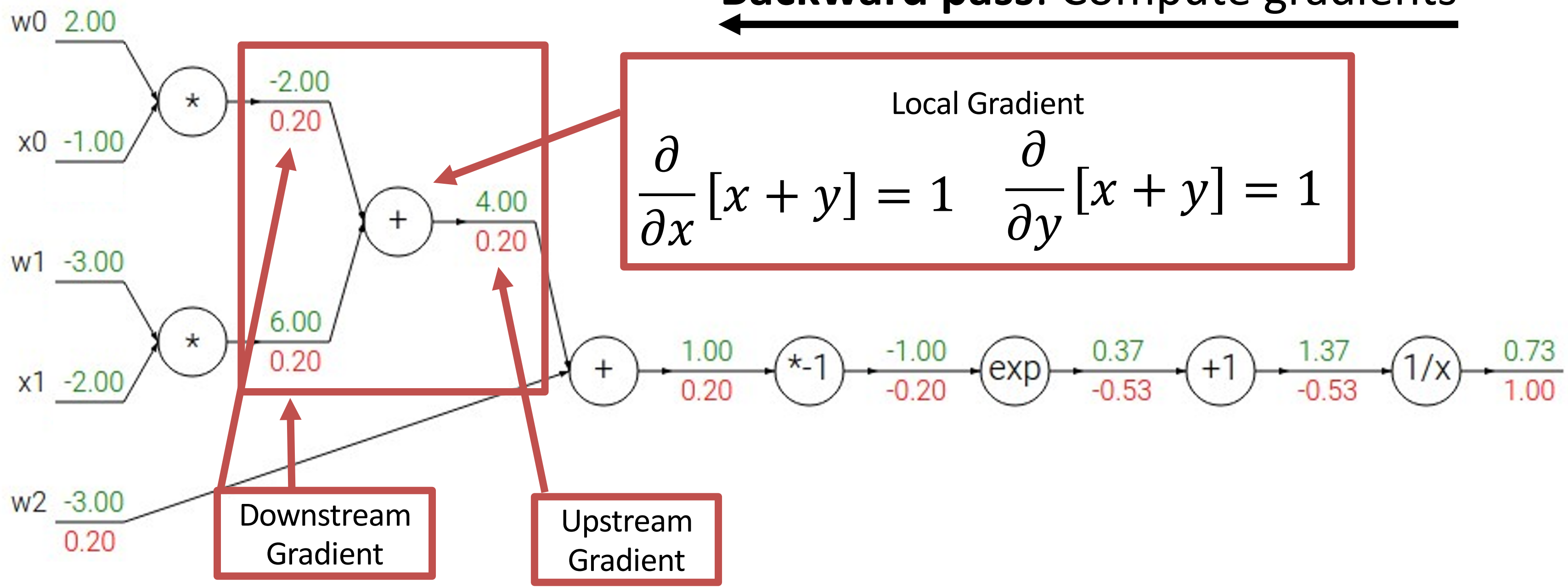
Backward pass: Compute gradients



Another Example

$$f(x,w) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

Backward pass: Compute gradients



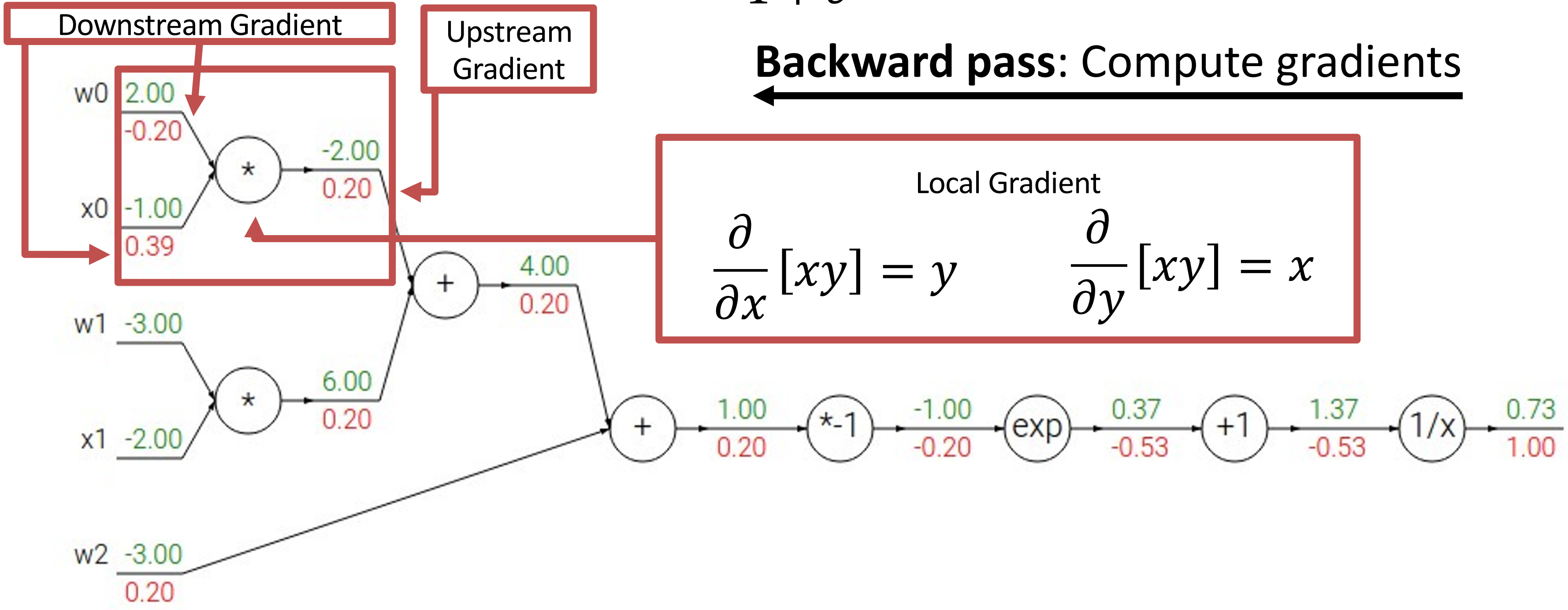
Local Gradient

$$\frac{\partial}{\partial x}[x+y] = 1 \qquad \frac{\partial}{\partial y}[x+y] = 1$$

Another Example

$$f(x,w) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

Backward pass: Compute gradients

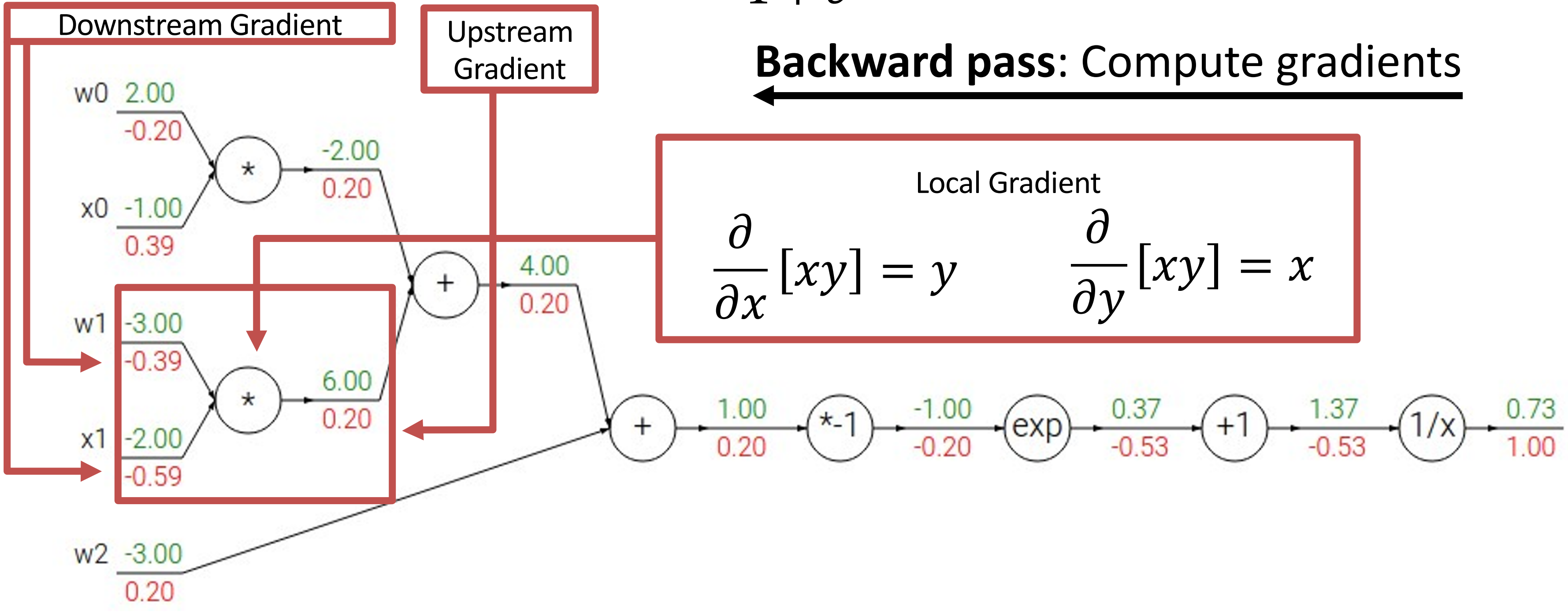


$$\frac{\partial}{\partial x}[xy] = y \qquad \frac{\partial}{\partial y}[xy] = x$$

Another Example

$$f(x,w) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$

Backward pass: Compute gradients



Local Gradient

$$\frac{\partial}{\partial x}[xy] = y \qquad \frac{\partial}{\partial y}[xy] = x$$

SGD + Momentum

Stochastic Gradient Descent

Select training samples instead of looping over all training examples

$$\nabla J(\theta) \approx \frac{1}{|B|} \sum_{i \in B} \nabla L(x_i, y_i, \theta)$$

Update rule

One iteration of gradient descent:

$$\theta^{t+1} = \theta^t - \underbrace{\eta_t}_{\text{learning rate}} \nabla_\theta J(\theta)\Big|_{\theta=\theta^t}$$

where B is a minibatch: a random subset of examples

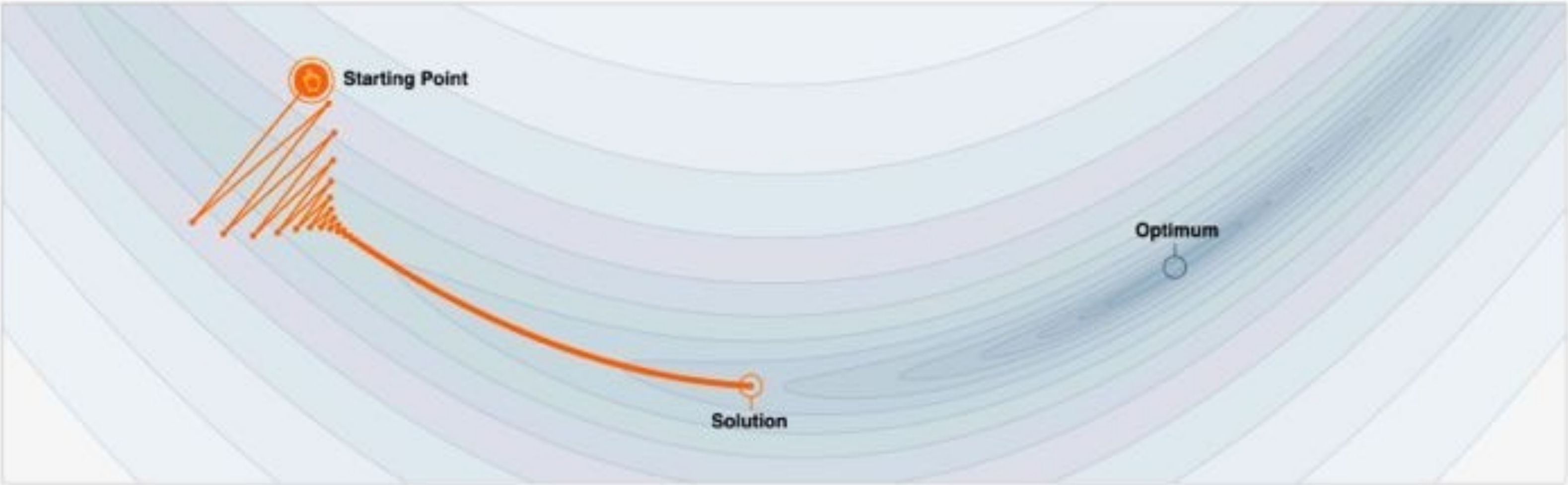
Momentum

SGD with Momentum has the following update rule

$$v \leftarrow dW + \beta * v$$
$$W \leftarrow W - \text{learning\_rate} * v$$

where beta is a scalar in range [0,1], dW is the gradient of the network parameter W, v is velocity initialized as all zeros.

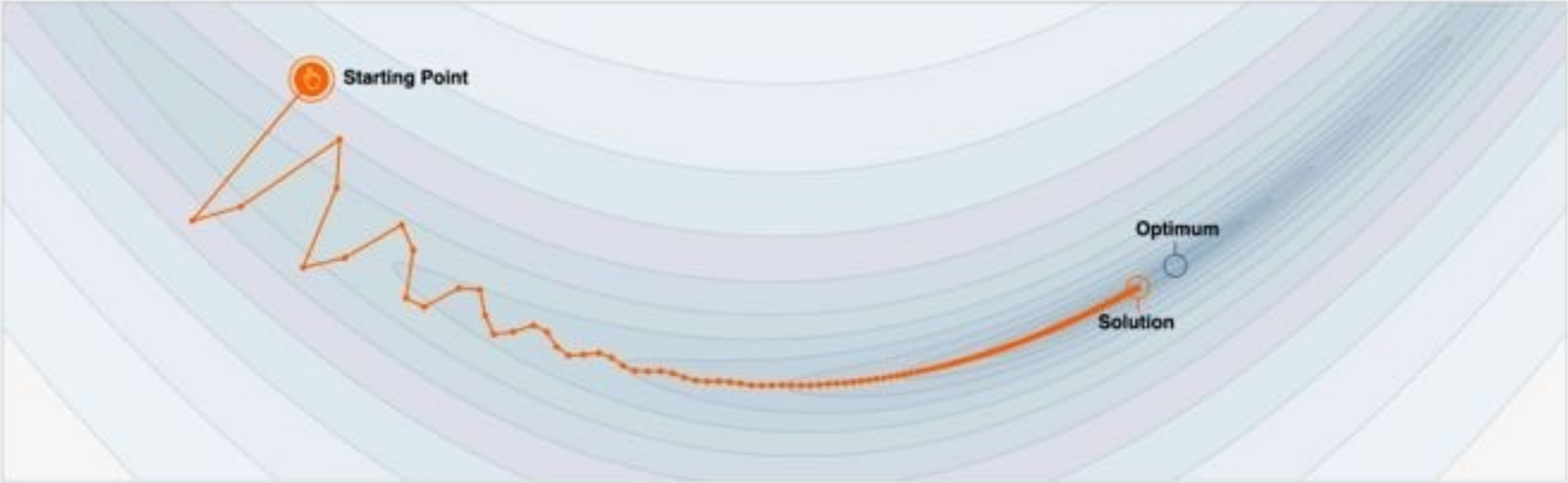
No Momentum



$\beta = 0$ i.e. disable momentum

Source: https://distill.pub/2017/momentum

With Momentum



$\beta = 0.99$

Source: https://distill.pub/2017/momentum