Loss and Optimization

Loss Functions

- How “bad” is our model?

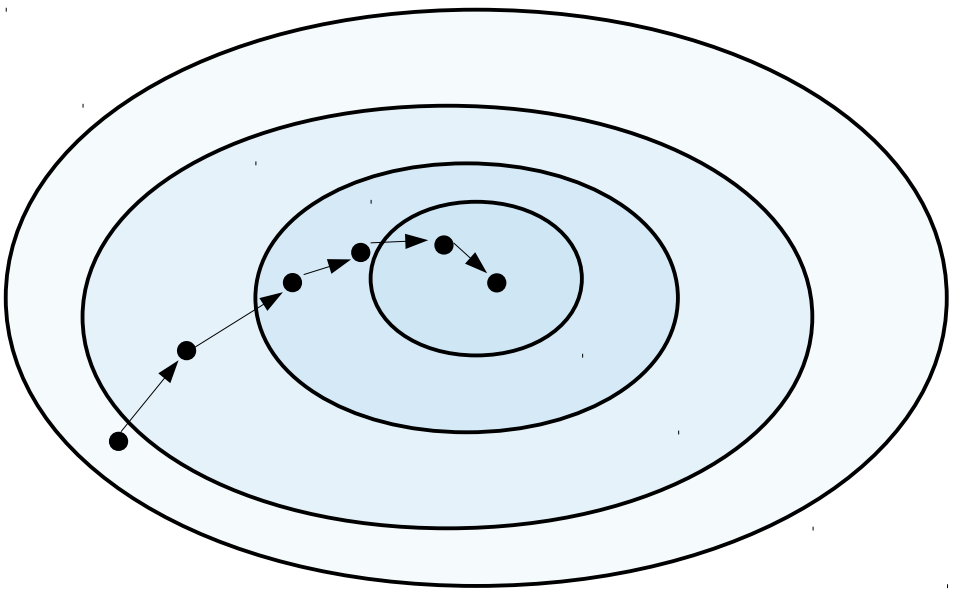
$$L(\theta) = \sum_i \ell(\theta; x_i, y_i)$$

Example: least-squares regression $L(\theta) = \sum_i \left(y_i - f(\theta; x_i) \right)^2$

Classification: Negative log-likelihood $L(\theta) = \sum_i -\log\left(p(y_i | \theta, x_i) \right)$

Optimization – Gradient Descent

How do we find $\operatorname{argmin}_\theta L(\theta)$



Initialize random θ

N iterations:

Compute $\nabla_\theta L(\theta)$

$\theta \leftarrow \theta - \epsilon \nabla_\theta L(\theta)$

Gradients

- How do we compute $\nabla_\theta L(\theta)$?

$$\nabla_\theta L(\theta) = \nabla_\theta \left(\sum_i \ell(\theta \mid x_i, y_i) \right) = \sum_i \nabla_\theta \ell(\theta \mid x_i, y_i)$$

Squared error:

$$\ell(\theta \mid x_i, y_i) = \left((\mathrm{W}\, \mathrm{x}_i + b) - y_i \right)^2$$

$$\nabla_{\mathrm{W}} \left((\mathrm{W}\, \mathrm{x}_i + b) - y_i \right)^2 = 2 \left((\mathrm{W}\, \mathrm{x}_i + b) - y_i \right) \mathrm{x}_i$$

$$\nabla_{\mathrm{b}} \left((\mathrm{W}\, \mathrm{x}_i + b) - y_i \right)^2 = 2 \left((\mathrm{W}\, \mathrm{x}_i + b) - y_i \right)$$

Negative log-likelihood:

$$\ell(\theta \mid x_i, y_i) = -\log \left(\mathrm{softmax} \left(\mathrm{W}\, \mathrm{x}_i + \mathrm{b} \right) \right)_{y_i}$$

$$\nabla_{\mathrm{W}_j} \left(-\log \left(\mathrm{softmax} \left(\mathrm{W}\, \mathrm{x}_i + \mathrm{b} \right) \right)_{y_i} \right) = \left(\mathrm{softmax} \left(\mathrm{W}\, \mathrm{x}_i + \mathrm{b} \right)_j - [y_i = j] \right) \mathrm{x}_i$$

$$\nabla_{\mathrm{b}_j} \left(-\log \left(\mathrm{softmax} \left(\mathrm{W}\, \mathrm{x}_i + \mathrm{b} \right) \right)_{y_i} \right) = \mathrm{softmax} \left(\mathrm{W}\, \mathrm{x}_i + \mathrm{b} \right)_j - [y_i = j]$$

Gradients – Computation Graphs

$$\left(y-\left(W\,x+b\right)\right)^2$$

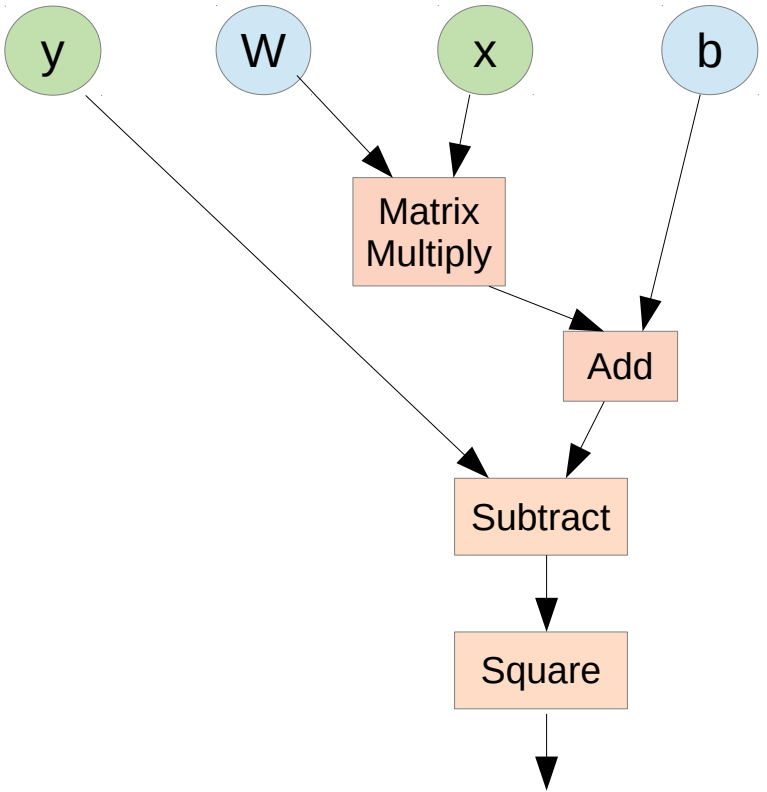
sq(sub(y, add(matmul(W, x), b)))

$\frac{\partial}{\partial W}$ sq(sub(y, add(matmul(W, x), b)))

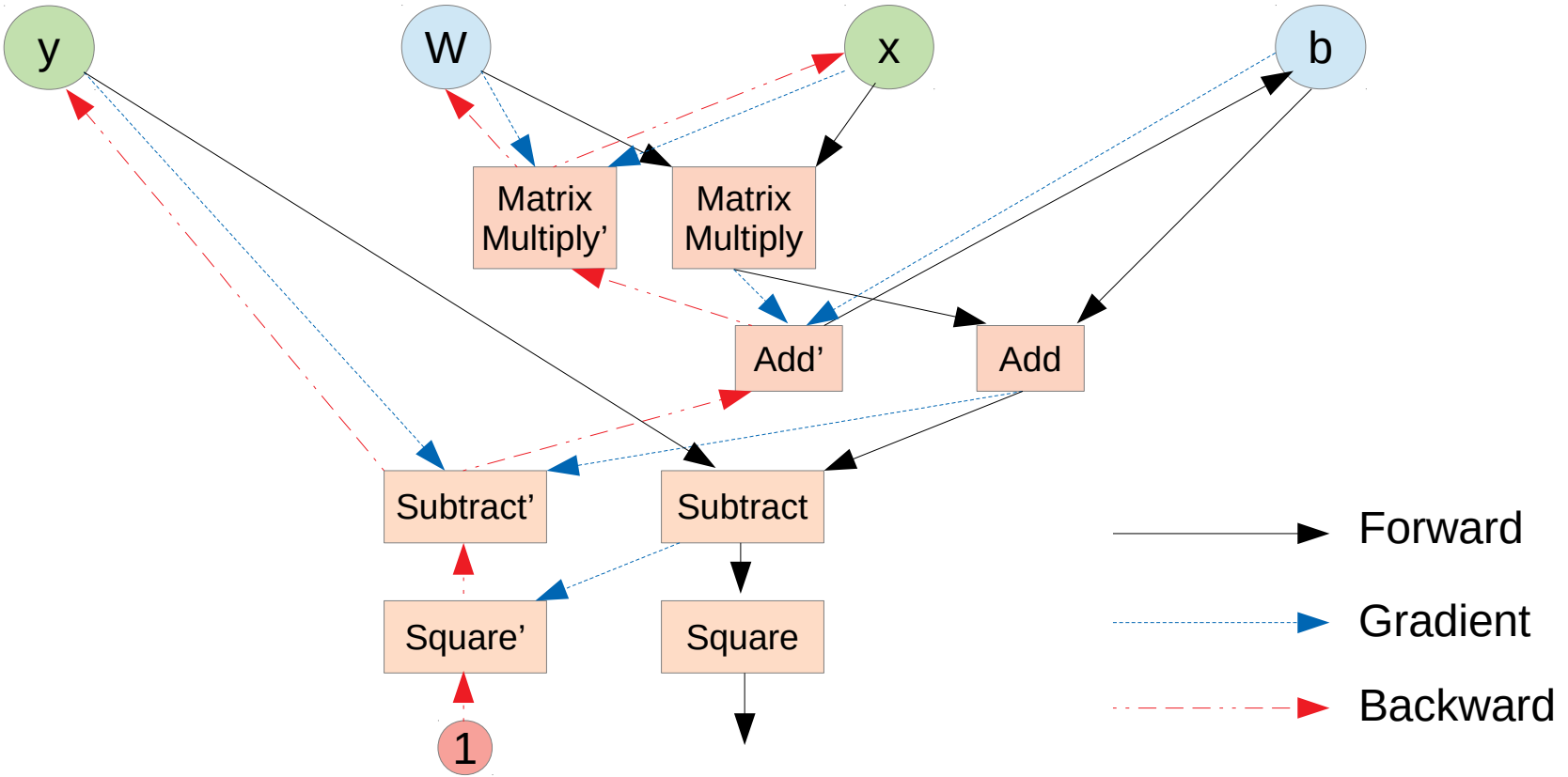
= sq'(sub(y, add(matmul(W, x), b))) *
$\frac{\partial}{\partial W}$ sub(y, add(matmul(W, x), b))

=...

= sq'(sub(y, add(matmul(W, x), b))) *
sub'(y, add(matmul(W, x), b)) *
add'(matmul(W, x), b) *
matmul'(W, x)

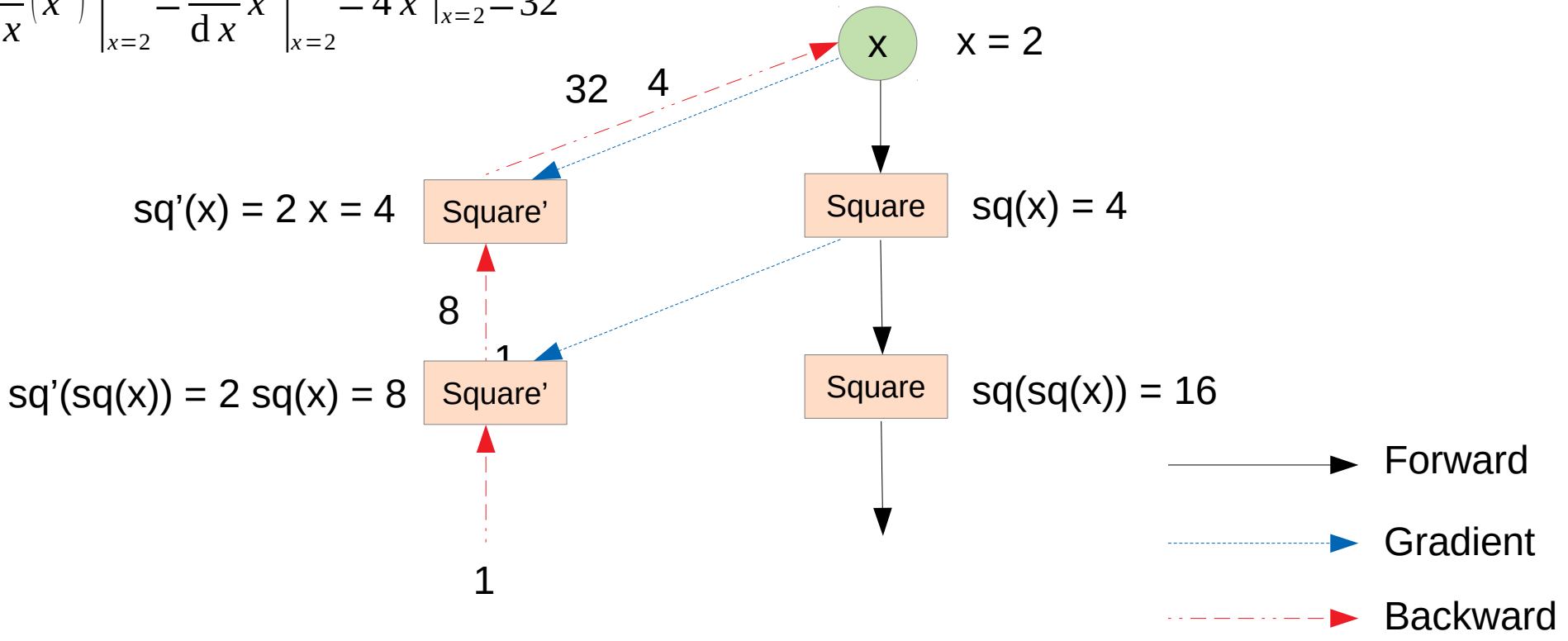


Gradients – Computation Graphs



Gradients – Computation Graphs

$$\left.\frac{\mathrm{d}}{\mathrm{d}\,x}\left(x^2\right)^2\right|_{x=2} = \left.\frac{\mathrm{d}}{\mathrm{d}\,x}x^4\right|_{x=2} = \left.4\,x^3\right|_{x=2} = 32$$



Learning Rate

$$\theta \leftarrow \theta - \epsilon \nabla_\theta L(\theta)$$