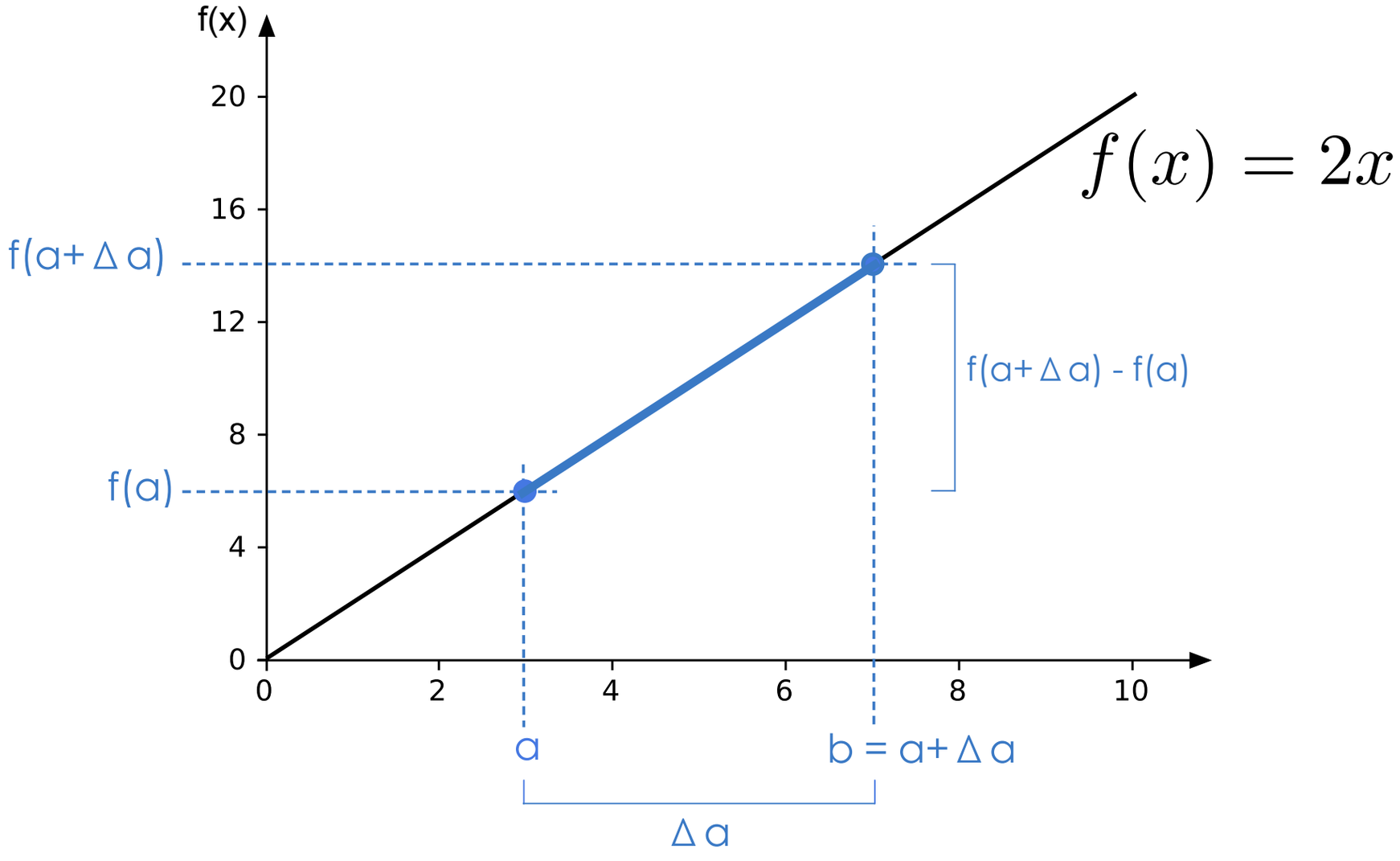
# Recapping Derivative Rules

1. Online, batch, and minibatch mode
2. Relation between perceptron and linear regression
3. An iterative training algorithm for linear regression
4. **(Optional) Calculus refresher I: Derivatives**
5. (Optional) Calculus refresher II: Gradients
6. Understanding gradient descent
7. Training an adaptive linear neuron (Adaline)

# Differential Calculus Refresher

Derivative of a function = "rate of change" = "slope"




$$\text{Slope} = \frac{f(a+\Delta a) - f(a)}{a + \Delta a - a} = \frac{f(a+\Delta a) - f(a)}{\Delta a}$$

# Function Derivative

$$f'(x) = \frac{df}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

Example 1: $f(x) = 2x$

$$\begin{aligned}
\frac{df}{dx} &= \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{2x + 2\Delta x - 2x}{\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{2\Delta x}{\Delta x} \\
&= \lim_{\Delta x \to 0} 2.
\end{aligned}$$

# Numerical vs Analytical/Symbolical Derivatives

$$f'(x) = \frac{df}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

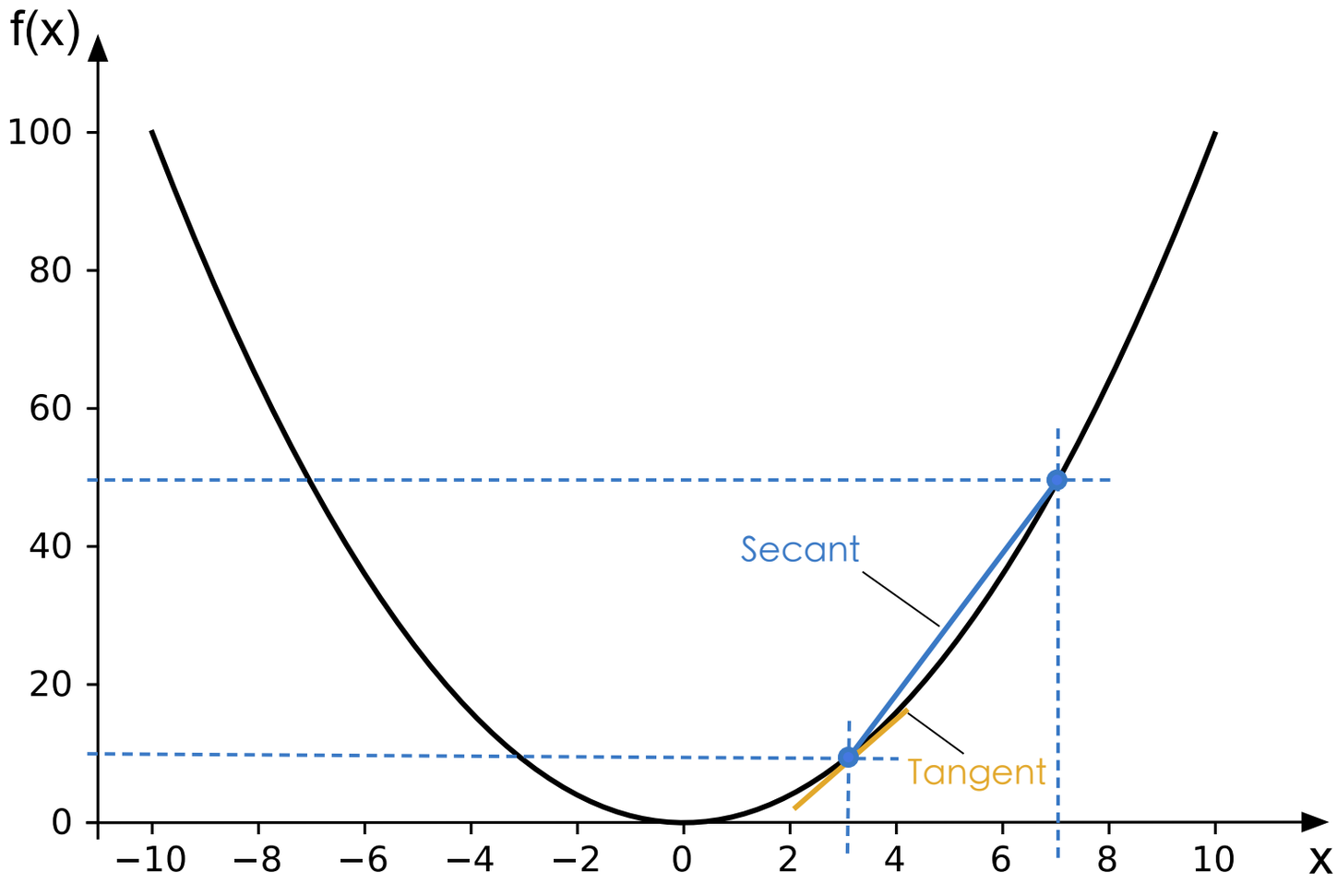
Example 2: $f(x) = x^2$

$$\begin{aligned}
\frac{df}{dx} &= \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{x^2 + 2x\Delta x + (\Delta x)^2 - x^2}{\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{2x\Delta x + (\Delta x)^2}{\Delta x} \\
&= \lim_{\Delta x \to 0} 2x + \Delta x.
\end{aligned}$$

# Numerical vs Analytical/Symbolical Derivatives

Conceptually, we obtained the derivative $\frac{d}{dx}x^2 = 2x$

By approximating the slope (tangent) by a secant between two points (as before)



# A Cheatsheet for You (1)

| | Function $f(x)$ | Derivative with respect to $x$ |
|---|---|---|
| 1 | $a$ | $0$ |
| 2 | $x$ | $1$ |
| 3 | $ax$ | $a$ |
| 4 | $x^2$ | $2x$ |
| 5 | $x^a$ | $ax^{a-1}$ |
| 6 | $a^x$ | $\log(a)a^x$ |
| 7 | $\log(x)$ | $1/x$ |
| 8 | $\log_a(x)$ | $1/(x\log(a))$ |
| 9 | $\sin(x)$ | $\cos(x)$ |
| 10 | $\cos(x)$ | $-\sin(x)$ |
| 11 | $\tan(x)$ | $\sec^2(x)$ |

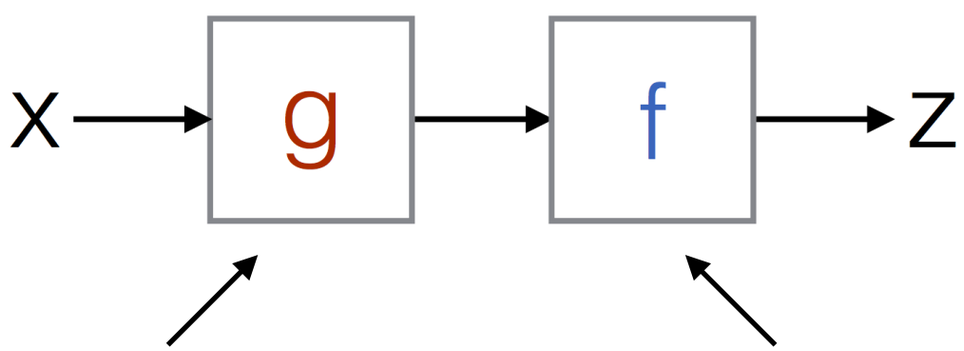
# A Cheatsheet for You (2)

| | Function | Derivative |
|---|---|---|
| Sum Rule | $f(x) + g(x)$ | $f'(x) + g'(x)$ |
| Difference Rule | $f(x) - g(x)$ | $f'(x) - g'(x)$ |
| Product Rule | $f(x)g(x)$ | $f'(x)g(x) + f(x)g'(x)$ |
| Quotient Rule | $f(x)/g(x)$ | $[g(x)f'(x) - f(x)g'(x)]/[g(x)]^2$ |
| Reciprocal Rule | $1/f(x)$ | $-[f'(x)]/[f(x)]^2$ |
| Chain Rule | $f(g(x))$ | $f'(g(x))g'(x)$ |

# Chain Rule & "Computation Graph" Intuition

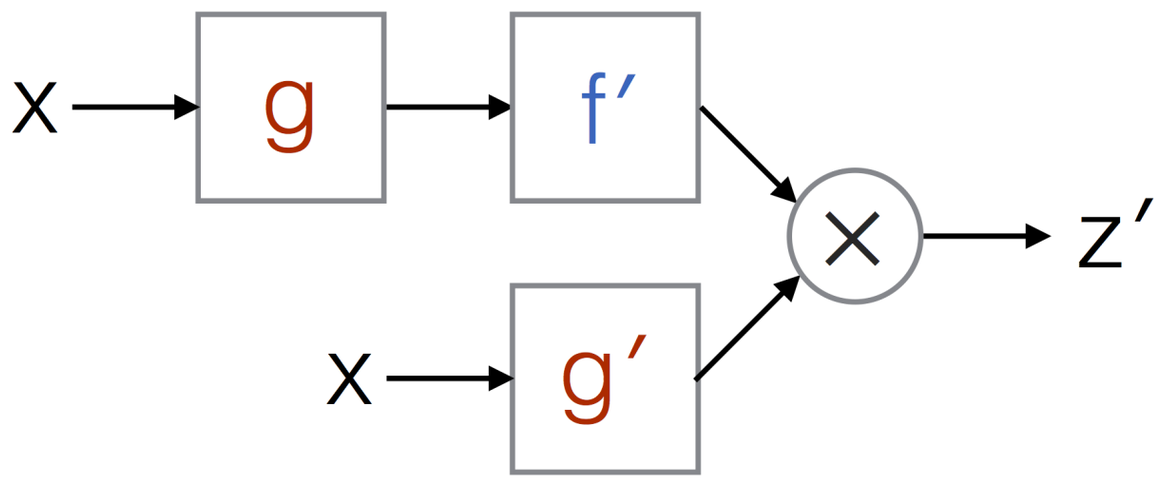
$$F(x) = f(\,g(x)\,) = z$$

Decomposition of some (nested) function:



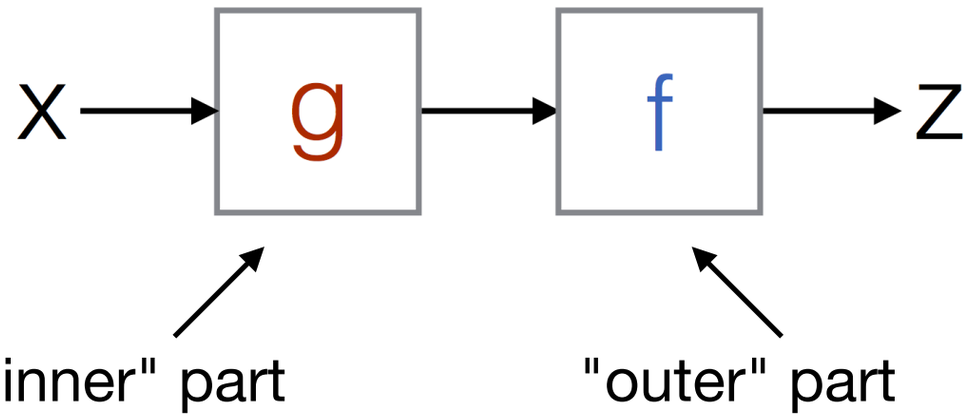

$$F'(x) = f'(\,g(x)\,)\,g'(x) = z'$$

Derivative of that nested function:



# Chain Rule & "Computation Graph" Intuition

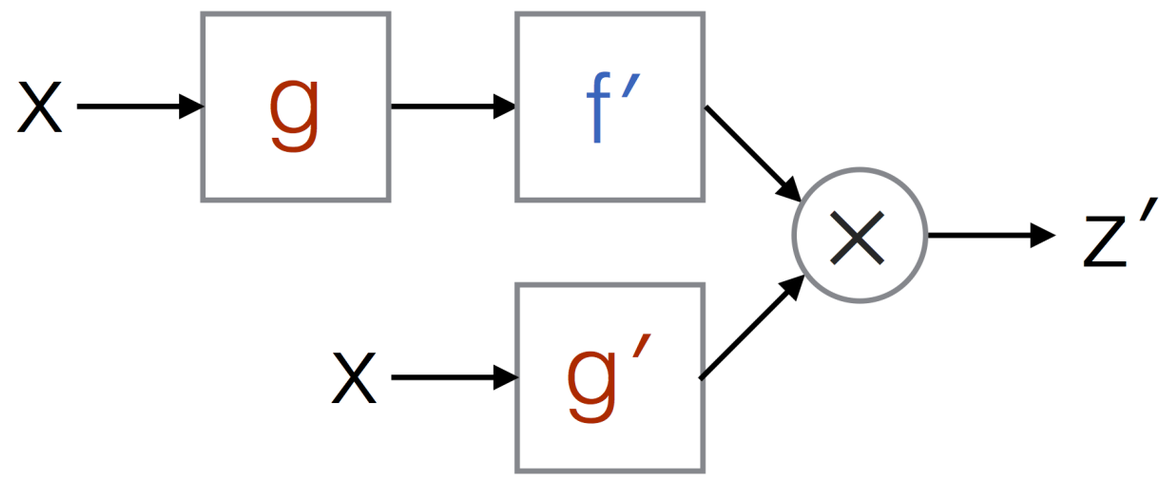
$$F(x) = f(\,g(x)\,) = z$$




Later, we will see that PyTorch can do that automatically for us :) (PyTorch literally keeps a computation graph in the background)

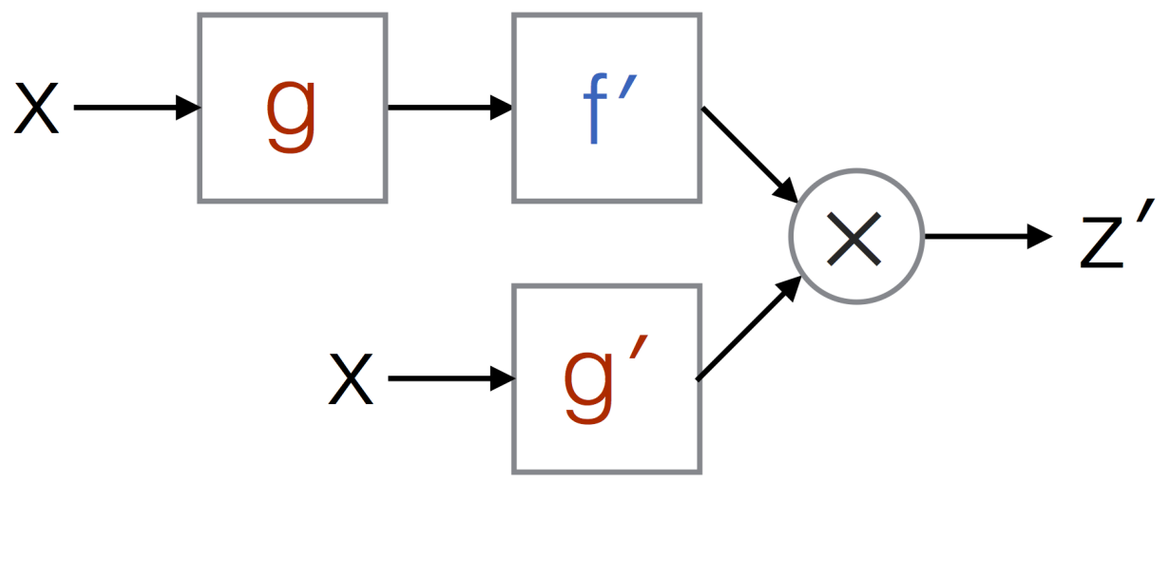
Also, PyTorch can compute the derivatives of most (differentiable) functions automatically

$$F'(x) = f'(\,g(x)\,)\,g'(x) = z'$$



# Chain Rule & "Computation Graph" Intuition

$$F'(x) = f'(\ g(x)\ )\ g'(x) = z'$$




In text, for efficiency, we will mostly use the Leibniz notation:

$$\frac{d}{dx}\left[f(g(x))\right] = \frac{df}{dg} \cdot \frac{dg}{dx}$$

# Chain Rule Example

$$\frac{d}{dx}\left[f(g(x))\right] = \frac{df}{dg} \cdot \frac{dg}{dx}$$

Example: $f(x) = \log(\sqrt{x})$

substituting $\frac{df}{dx} = \frac{d}{dg}\log(g) \cdot \frac{d}{dx}\sqrt{x}$

with $\frac{d}{dg}\log(g) = \frac{1}{g} = \frac{1}{\sqrt{x}}$ and $\frac{d}{dx}x^{1/2} = \frac{1}{2}x^{-1/2} = \frac{1}{2\sqrt{x}}$

leads us to the solution $\frac{df}{dx} = \frac{1}{\sqrt{x}} \cdot \frac{1}{2\sqrt{x}} = \frac{1}{2x}$

# Chain Rule for Arbitrarily Long Function Compositions

$$F(x) = f(g(h(u(v(x)))))$$

$$\frac{dF}{dx} = \frac{d}{dx}F(x) = \frac{d}{dx}f(g(h(u(v(x)))))$$

$$= \frac{df}{dg} \cdot \frac{dg}{dh} \cdot \frac{dh}{du} \cdot \frac{du}{dv} \cdot \frac{dv}{dx}$$

# Chain Rule for Arbitrarily Long Function Compositions

$$\frac{dF}{dx} = \frac{d}{dx}F(x) = \frac{d}{dx}f(g(h(u(v(x)))))$$
$$= \frac{df}{dg} \cdot \frac{dg}{dh} \cdot \frac{dh}{du} \cdot \frac{du}{dv} \cdot \frac{dv}{dx}$$

Also called "reverse mode" as we start with the outer function. In neural nets, this will be from right to left.

We could also start from the inner parts ("forward mode")

$$\frac{dv}{dx} \cdot \frac{du}{dv} \cdot \frac{dh}{du} \cdot \frac{dg}{dh} \cdot \frac{df}{dg}$$

- Backpropagation (covered later) is basically "reverse" mode auto-differentiation
- It is cheaper than forward mode if we work with gradients, since then we have matrix-"vector" multiplications instead of matrix multiplications

# Gradients: Derivatives of Multivariable Functions

1. Online, batch, and minibatch mode
2. Relation between perceptron and linear regression
3. An iterative training algorithm for linear regression
4. (Optional) Calculus refresher I: Derivatives
5. **(Optional) Calculus refresher II: Gradients**
6. Understanding gradient descent
7. Training an adaptive linear neuron (Adaline)

# Gradients: Derivatives of Multivariable* Functions

***note that in some fields, the terms "multivariable" and "multivariate" are used interchangeably,**
**but here, we really mean "multivariable" because "multivariate" means "multiple outputs", which is**
**not the case here -- similarly, in most DL applications output one prediction value, or one prediction value per training example**

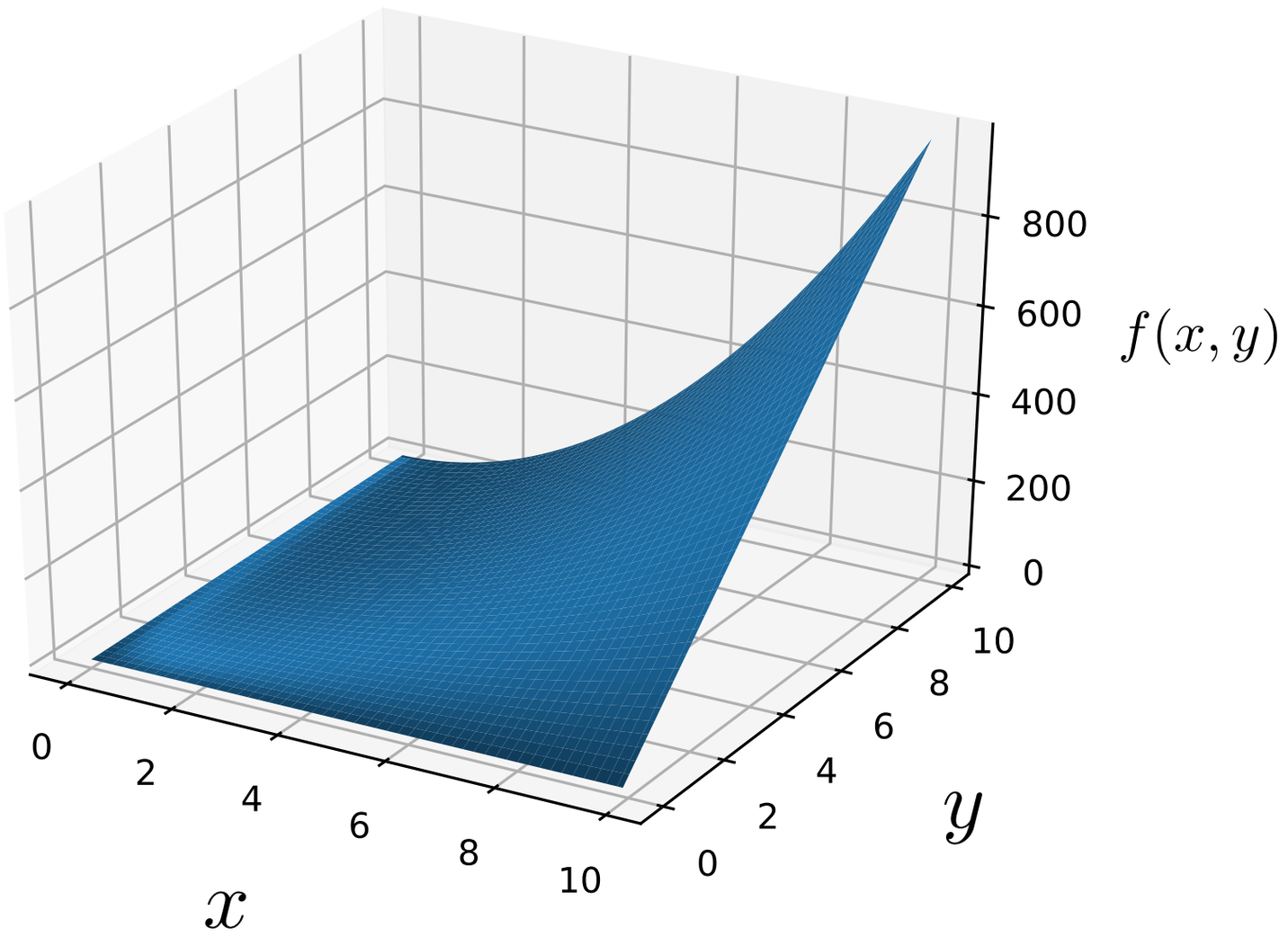
$$f(x, y, z, ...)$$

$$\nabla f = \begin{bmatrix} \partial f / \partial x \\ \partial f / \partial y \\ \partial f / \partial z \\ \vdots \end{bmatrix}$$

For gradients, we use the "partial" symbol to denote partial derivatives; more of a notational convention and the concept is the same as before when we were computing ordinary derivatives (denoted them as "d")

# Gradients: Derivatives of Multivariable Functions

Example: $f(x, y) = x^2 y + y$



# Gradients: Derivatives of Multivariable Functions

Example: $f(x, y) = x^2 y + y$

$$\nabla f(x, y) = \begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \end{bmatrix},$$

where

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x} x^2 y + y = 2xy$$

(via the power rule and constant rule), and

$$\frac{\partial f}{\partial y} = \frac{\partial}{\partial y} x^2 y + y = x^2 + 1.$$

So, the gradient of the function $f$ is defined as

$$\nabla f(x, y) = \begin{bmatrix} 2xy \\ x^2 + 1 \end{bmatrix}.$$

# Gradients & the Multivariable Chain Rule

Suppose we have a composite function like this:

$$f\big(g(x), h(x)\big)$$

Remember the regular chain rule for a single input:

$$\frac{d}{dx}\big[f(g(x))\big] = \frac{df}{dg} \cdot \frac{dg}{dx}$$

For two inputs, we now have

$$\frac{d}{dx}\big[f(g(x), h(x))\big] = \frac{\partial f}{\partial g} \cdot \frac{dg}{dx} + \frac{\partial f}{\partial h} \cdot \frac{dh}{dx}$$

# Gradients & the Multivariable Chain Rule

$$f\big(g(x),h(x)\big)$$

$$\frac{d}{dx}\big[f(g(x),h(x))\big] = \frac{\partial f}{\partial g}\cdot\frac{dg}{dx} + \frac{\partial f}{\partial h}\cdot\frac{dh}{dx}$$

Example:

$$f\big(g,h\big) = g^2h + h$$

where $g(x) = 3x,$ and $h(x) = x^2$

$$\frac{\partial f}{\partial g} = 2gh \qquad \frac{\partial f}{\partial h} = g^2 + 1$$

$$\frac{dg}{dx} = \frac{d}{dx}3x = 3 \qquad \frac{dh}{dx} = \frac{d}{dx}x^2 = 2x$$

$$\begin{aligned}\frac{d}{dx}\big[f(g(x))\big] &= [2gh\cdot 3] + [(g^2+1)\cdot 2x] \\ &= 2xg^2 + 6gh + 2x\end{aligned}$$

# Gradients & the Multivariable Chain Rule

# in Vector Form

$$f\big(g(x), h(x)\big)$$

$$\begin{aligned}\frac{d}{dx}\left[f(g(x), h(x))\right] &= \frac{\partial f}{\partial g}\cdot\frac{dg}{dx} + \frac{\partial f}{\partial h}\cdot\frac{dh}{dx} \\ &= \nabla f \cdot \mathbf{v}'(x).\end{aligned}$$

Where

$$\mathbf{v}(x) = \begin{bmatrix} g(x) \\ h(x) \end{bmatrix} \qquad \mathbf{v}'(x) = \frac{d}{dx}\begin{bmatrix} g(x) \\ h(x) \end{bmatrix} = \begin{bmatrix} dg/dx \\ dh/dx \end{bmatrix}$$

Putting it together:

$$\nabla f \cdot \mathbf{v}'(x) = \begin{bmatrix} \partial f/\partial g \\ \partial f/\partial h \end{bmatrix} \cdot \begin{bmatrix} dg/dx \\ dh/dx \end{bmatrix} = \frac{\partial f}{\partial g}\cdot\frac{dg}{dx} + \frac{\partial f}{\partial h}\cdot\frac{dh}{dx}$$

# The Jacobian (Matrix)

$$\mathbf{f}(x_1, x_2, ..., x_m) = \begin{bmatrix} f_1\left(x_1, x_2, x_3, \cdots x_m\right) \\ f_2\left(x_1, x_2, x_3, \cdots x_m\right) \\ f_3\left(x_1, x_2, x_3, \cdots x_m\right) \\ \vdots \\ f_m\left(x_1, x_2, x_3, \cdots x_m\right) \end{bmatrix}$$

$$J\left(x_1, x_2, x_3, \cdots x_m\right) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_2}{\partial x_3} & \cdots & \frac{\partial f_2}{\partial x_m} \\ \frac{\partial f_3}{\partial x_1} & \frac{\partial f_3}{\partial x_2} & \frac{\partial f_3}{\partial x_3} & \cdots & \frac{\partial f_3}{\partial x_m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \frac{\partial f_m}{\partial x_3} & \cdots & \frac{\partial f_m}{\partial x_m} \end{bmatrix}$$

# The Jacobian (Matrix)

$$\mathbf{f}(x_1, x_2, ..., x_m) = \begin{bmatrix} f_1\left(x_1, x_2, x_3, \cdots x_m\right) \\ f_2\left(x_1, x_2, x_3, \cdots x_m\right) \\ f_3\left(x_1, x_2, x_3, \cdots x_m\right) \\ \vdots \\ f_m\left(x_1, x_2, x_3, \cdots x_m\right) \end{bmatrix}$$

$$J\left(x_1, x_2, x_3, \cdots x_m\right) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \frac{\partial f_1}{\partial x_3} & \cdots & \frac{\partial f_1}{\partial x_m} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_2}{\partial x_3} & \cdots & \frac{\partial f_2}{\partial x_m} \\ \frac{\partial f_3}{\partial x_1} & \frac{\partial f_3}{\partial x_2} & \frac{\partial f_3}{\partial x_3} & \cdots & \frac{\partial f_3}{\partial x_m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \frac{\partial f_m}{\partial x_3} & \cdots & \frac{\partial f_m}{\partial x_m} \end{bmatrix} \left(\nabla f_1\right)^\top$$