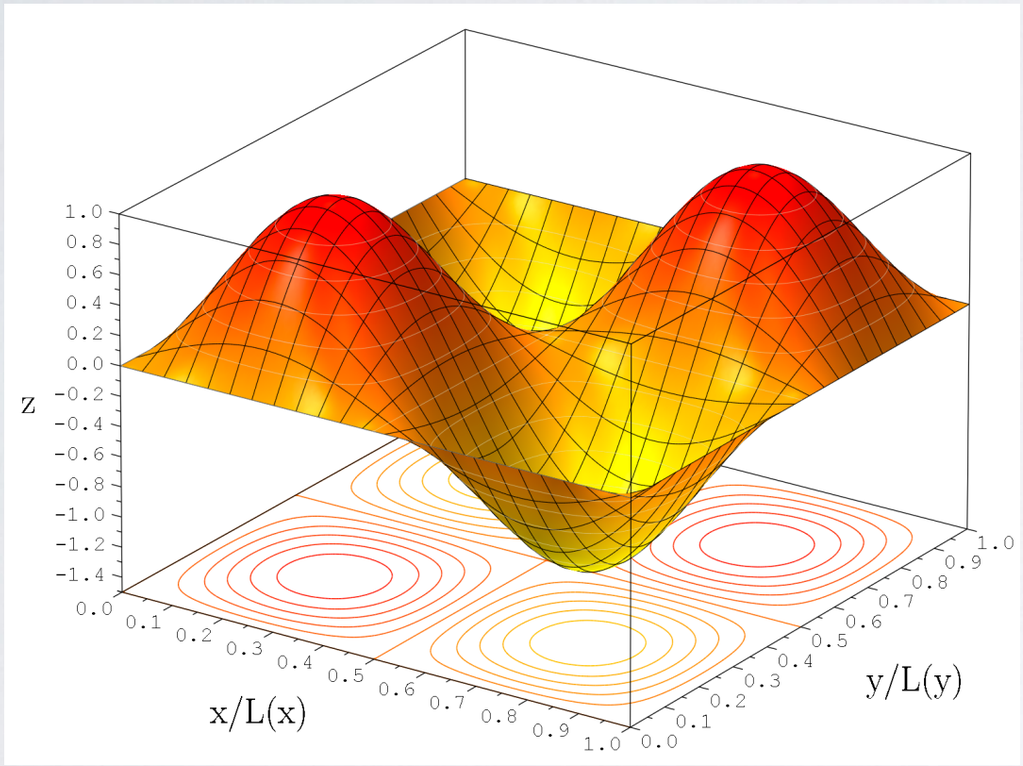
MANY-VARIABLE CALCULUS

CONTINUOUS GRADIENTS

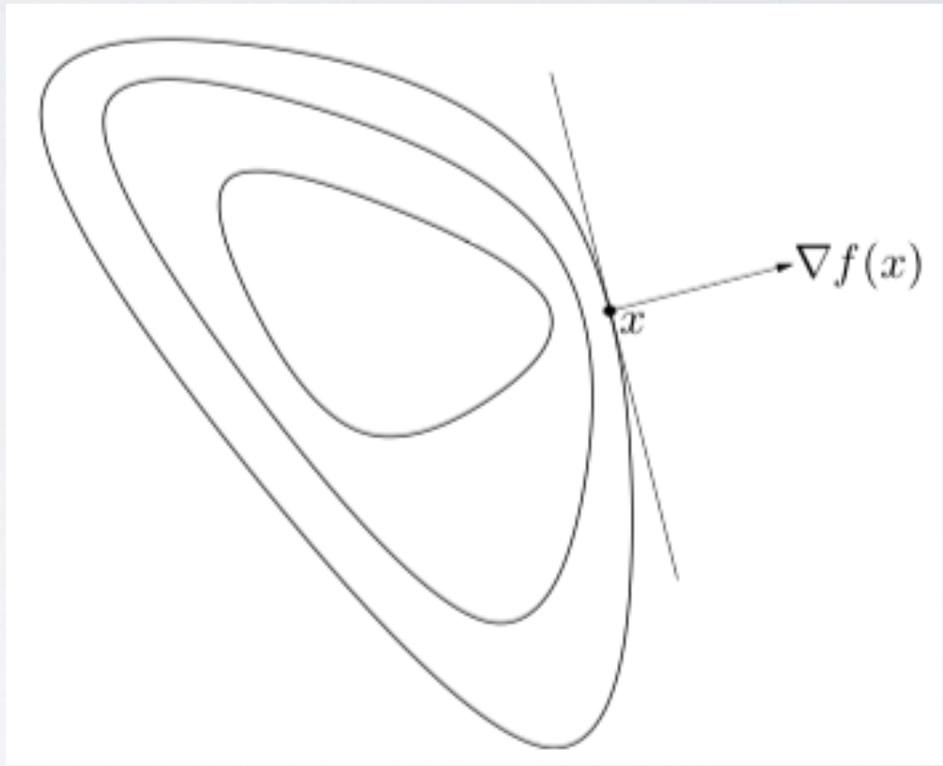
WHY WE NEED GRADIENT

First-order conditions



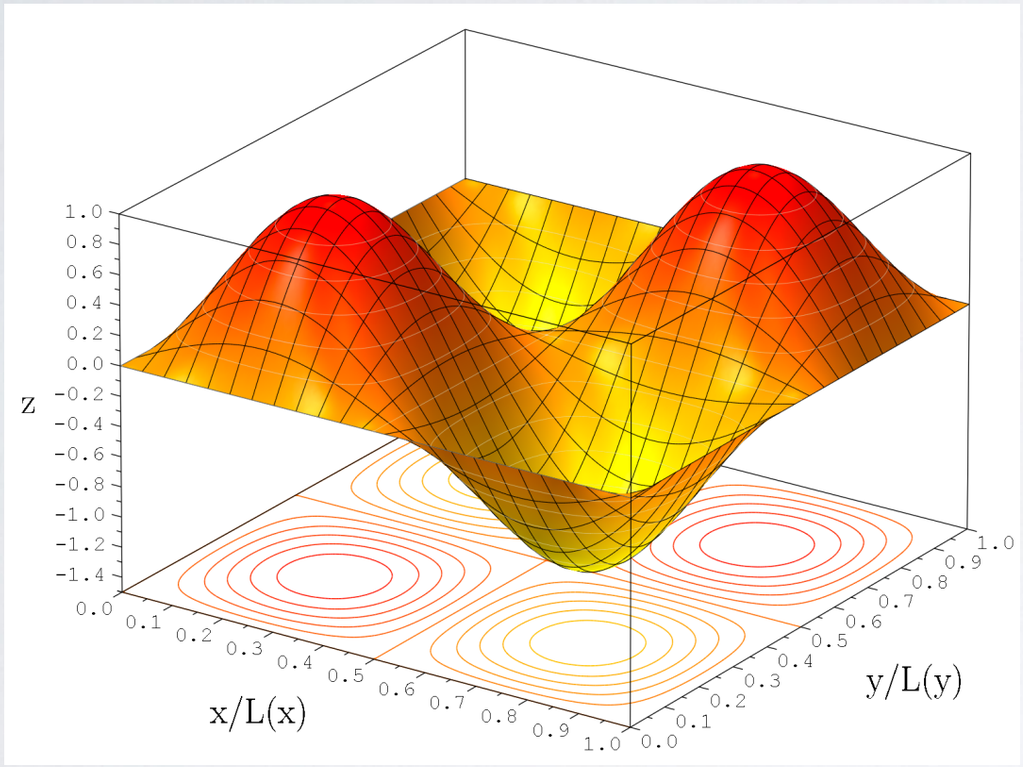
$$\nabla f(x^\star) = 0$$

First-order methods



Gradient descent

THE CALCULUS YO MOMMA TAUGHT YOU



$$f(x,y,z) : \mathbb{R}^n \to \mathbb{R}$$

$$\nabla f(x,y,z) = \partial_x f\,\mathbf{i} + \partial_y f\,\mathbf{j} + \partial_z f\,\mathbf{k}$$

What letter do we use for 4th variable?

What letter do we use for 1,000,000th variable?

DERIVATIVE

$$f : \mathbb{R}^n \to \mathbb{R}^m$$

Freshet derivative is a linear operator D with

$$\lim_{\|h\| \to 0} \frac{f(x+h) - f(x) - Dh}{\|h\|} = 0$$

A linear operator that **locally** behaves like f

In optimization: variables are usually **column vectors**

DERIVATIVE

In optimization: variables are usually **column vectors**

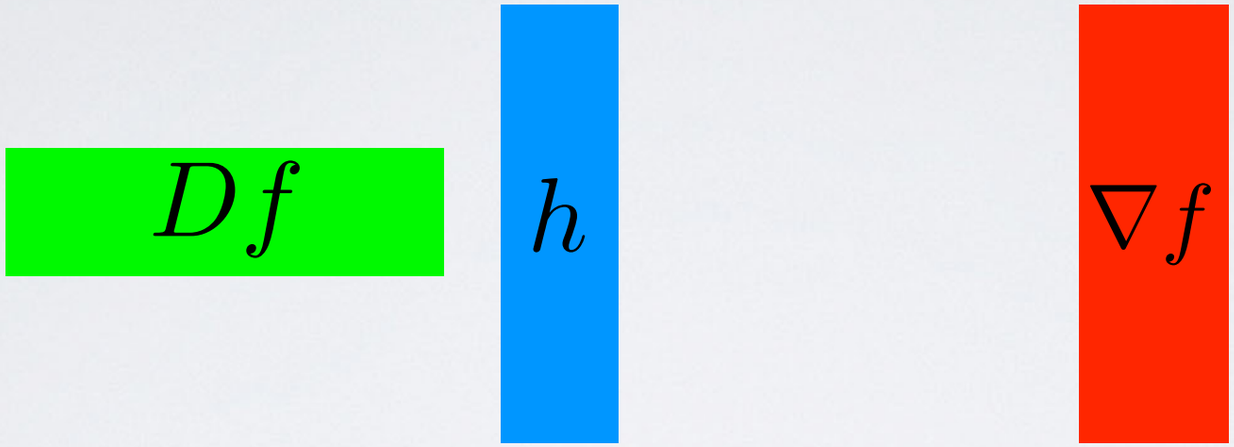
In this case the derivative is the **Jacobian**

$$Df = \begin{pmatrix} D_{x_1} f & D_{x_2} f & D_{x_3} f \end{pmatrix}$$

$$= \begin{pmatrix} D_{x_1} f_1 & D_{x_2} f_1 & D_{x_3} f_1 \\ D_{x_1} f_2 & D_{x_2} f_2 & D_{x_3} f_2 \\ D_{x_1} f_3 & D_{x_2} f_3 & D_{x_3} f_3 \end{pmatrix}$$

GRADIENT

In optimization: variables are usually **column vectors**



Definition: Gradient is the **adjoint** of the derivative

For scalar-valued functions

$$Df = \begin{pmatrix} \partial_{x_1} f, \partial_{x_2} f, \partial_{x_3} f \end{pmatrix} \qquad \nabla f = \begin{pmatrix} \partial_{x_1} f \\ \partial_{x_2} f \\ \partial_{x_3} f \end{pmatrix}$$

GRADIENT

In optimization: variables are usually **column vectors**

$$Df \qquad h \qquad \nabla f$$

Better Definition: The gradient of a scalar-valued function is a matrix of derivatives that is the **same shape** as the unknowns

example: matrix of unknowns

$$x = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix} \qquad \nabla f = \begin{pmatrix} \partial_{11} f & \partial_{12} f & \partial_{13} f \\ \partial_{21} f & \partial_{22} f & \partial_{23} f \\ \partial_{31} f & \partial_{32} f & \partial_{33} f \end{pmatrix}$$

WHAT'S THE GRADIENT?!?!

Function	gradient
$f(x) = Ax$	$\nabla f(x) = A^T$
$f(x) = \|x\|^2$	$\nabla f(x) = 2x$
$f(x) = \frac{1}{2} x^T A x$	$\nabla f(x) = Ax$

symmetric A

CHAIN RULE

$$h(x) = f \circ g(x) = f(g(x))$$

single-variable chain rule

$$\partial f(g(x)) = f'(g)g'(x)$$

multi-variable chain rule

$$Df(g(x)) = Df \circ Dg(x)$$

for gradients

$$\nabla f(g(x)) = \nabla g(x) \nabla f(g)$$

Swap

EXAMPLE: LEAST SQUARES

$$f(g(x)) = \frac{1}{2}\|Ax - b\|^2$$

$$f = \frac{1}{2}\|x\|^2 \qquad Ax - b = g$$

$$\nabla \qquad \nabla$$

$$x \qquad A^T$$

$$A^T(Ax - b)$$

$$\nabla g(x) \nabla f(g)$$

How would you find the gradient?

IMPORTANT RULES

If

$$h(x) = f(Ax)$$

Then

$$\nabla h(x) = A^T f'(Ax)$$

If

$$h(x) = f(xA)$$

Then

$$\nabla h(x) = f'(xA)A^T$$

Why?

$$g(x) = xA \qquad \nabla g \nabla f = \nabla f A^T$$

EXAMPLE: RIDGE REGRESSION

$$f(x) = \frac{\lambda}{2}\|x\|^2 + \frac{1}{2}\|Ax - b\|^2$$

Optimality Condition

$$\nabla f(x) = \lambda x + A^T(Ax - b) = 0$$

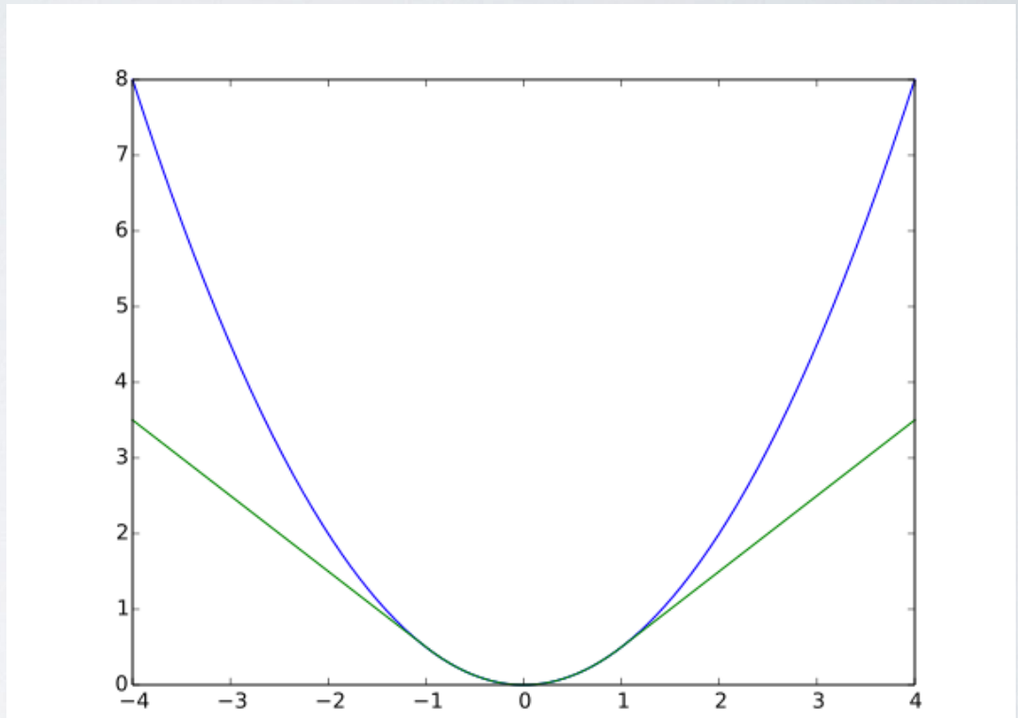
$$(A^T A + \lambda I)x = A^T b$$

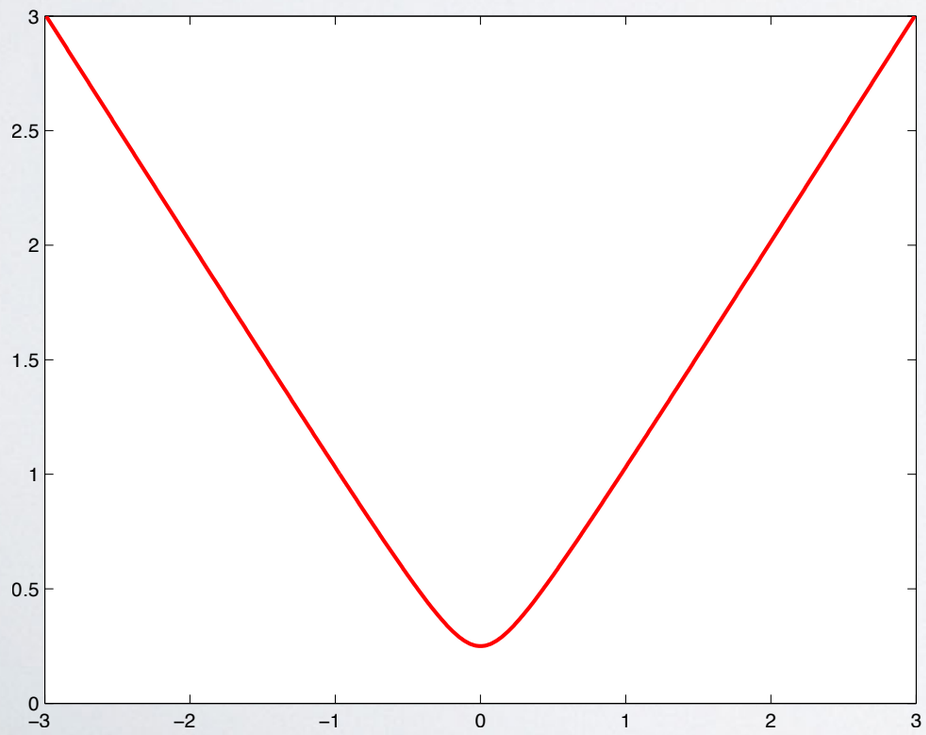
$$x^\star = (A^T A + \lambda I)^{-1} A^T b$$

REGULARIZED GRADIENTS

Huber regularization

$$|x| \approx h(x) = \begin{cases} \frac{1}{2}x^2, \text{ for } |x| \leq \delta \\ \delta(|a| - \frac{1}{2}\delta), \text{ otherwise} \end{cases}$$





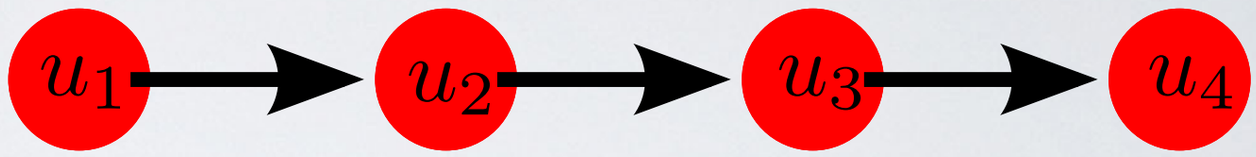
Hyperbolic regularization

$$|x| \approx \sqrt{x^2 + \epsilon^2}$$

FINITE DIFFERENCE OPERATOR

Neumann

$$\nabla u = \begin{pmatrix} u_1 - u_0 \\ u_2 - u_1 \\ u_3 - u_2 \end{pmatrix}$$



discrete gradient operator (finite difference)

$$\nabla = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

(negative) divergence

$$\nabla^T = \begin{bmatrix} -1 & 0 & 0 \\ -1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

GRADIENT OF TV

$$|\nabla x| + \frac{\mu}{2}\|x - f\|^2$$

Regularize L1

$$h(\nabla x) + \frac{\mu}{2}\|x - f\|^2$$

Differentiate

$$\nabla^T h'(\nabla x) + \mu(x - f)$$

$$h'(x) = \frac{x}{\sqrt{x^2 + \epsilon^2}}$$ divide component-wise

LOGISTIC REGRESSION

Data vectors

$$\{x_i\}$$

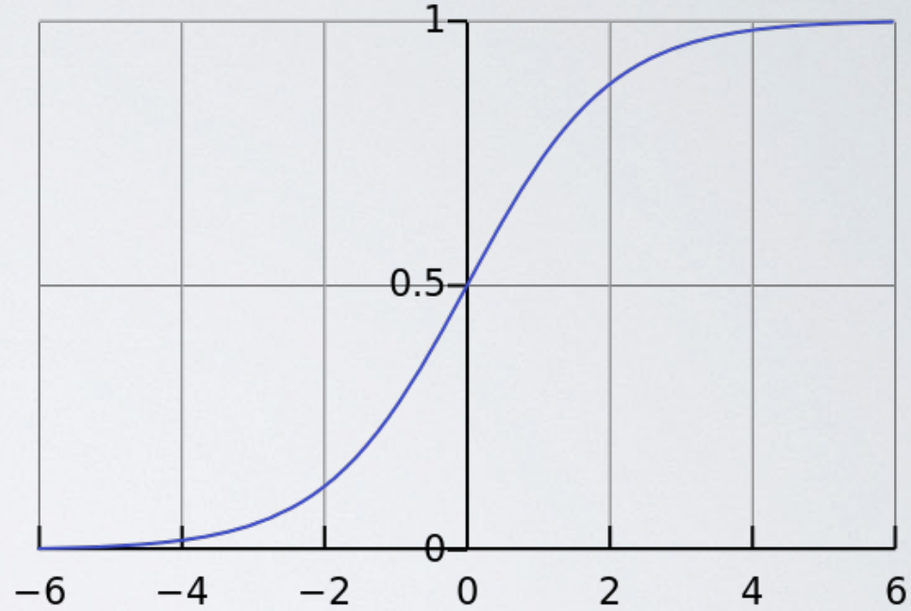
Labels

$$\{y_i\}$$

Model

$$p(y_i = 1|x_i) = \sigma(x_i w) = \frac{e^{x_i w}}{1 + e^{x_i w}}$$

$$\sigma =$$



Maximum Likelihood (MAP) Estimator

$$\text{maximize} \quad p(w|X, Y)$$

$$\underset{w}{\text{minimize}} \quad \sum_{k=1}^{m} \log(1 + \exp(-y_k \cdot x_k w))) = f_{lr}(YXw)$$

LOGISTIC REGRESSION

$$\underset{w}{\text{minimize}} \quad \sum_{k=1}^{m} \log(1 + \exp(-y_k \cdot x_k w))) = f_{lr}(YXw)$$

Break the problem down

$$f_{lr}(z) = \sum_{k=1}^{m} \log(1 + \exp(-z_k)) \qquad \hat{X} = YX$$

Coordinate separable

Linear operator

$$\underset{x}{\text{minimize}} \quad f_{lr}(\hat{X}w)$$

$$\nabla\{f_{lr}(\hat{X}w)\} = \hat{X}^T \nabla f_{lr}(\hat{X}w) \qquad \nabla f_{lr}(z) = \frac{-exp(-z)}{1 + exp(-z)}$$

NON-NEGATIVE MATRIX FACTORIZATION

$$\underset{X\geq 0, Y\geq 0}{\text{minimize}} \quad E(X,Y) = \frac{1}{2}\|XY - S\|^2$$

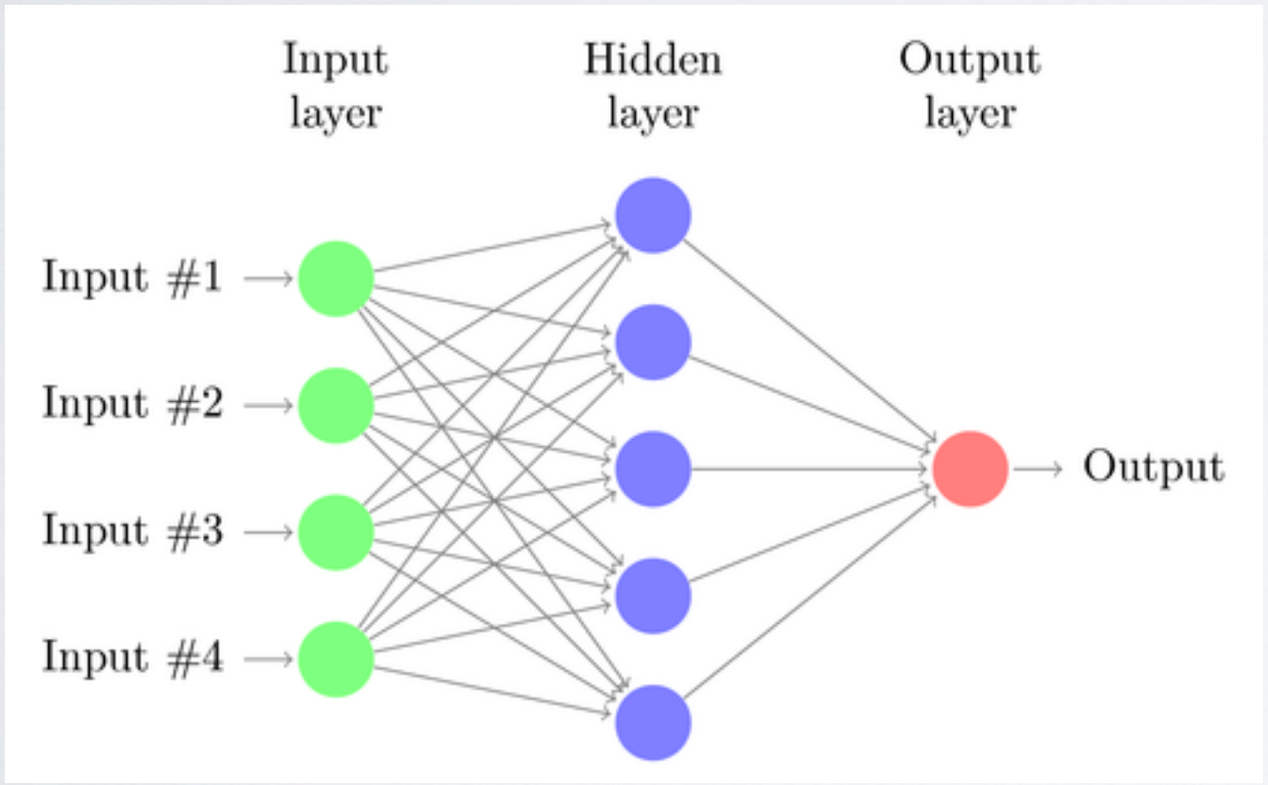
$$\partial_Y E = X^T(XY - S)$$

$$\partial_X E = (XY - S)Y^T$$

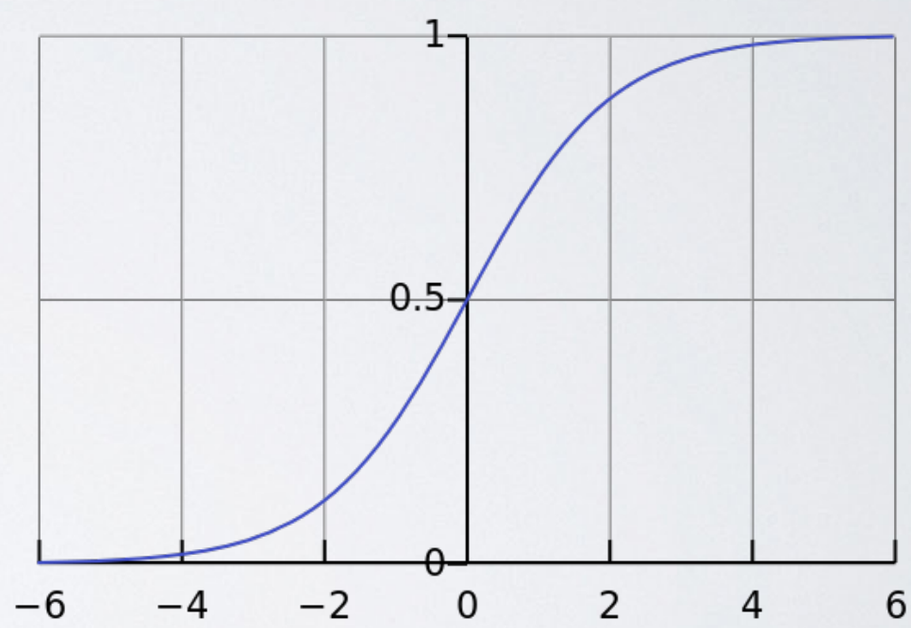
NEURAL NETS: LOGISTIC REGRESSION ON STEROIDS

$$\sigma(\sigma(\sigma(x_i W_1)W_2)W_3) = y_3$$

MLP = "Multilayer Perception"



$\sigma =$

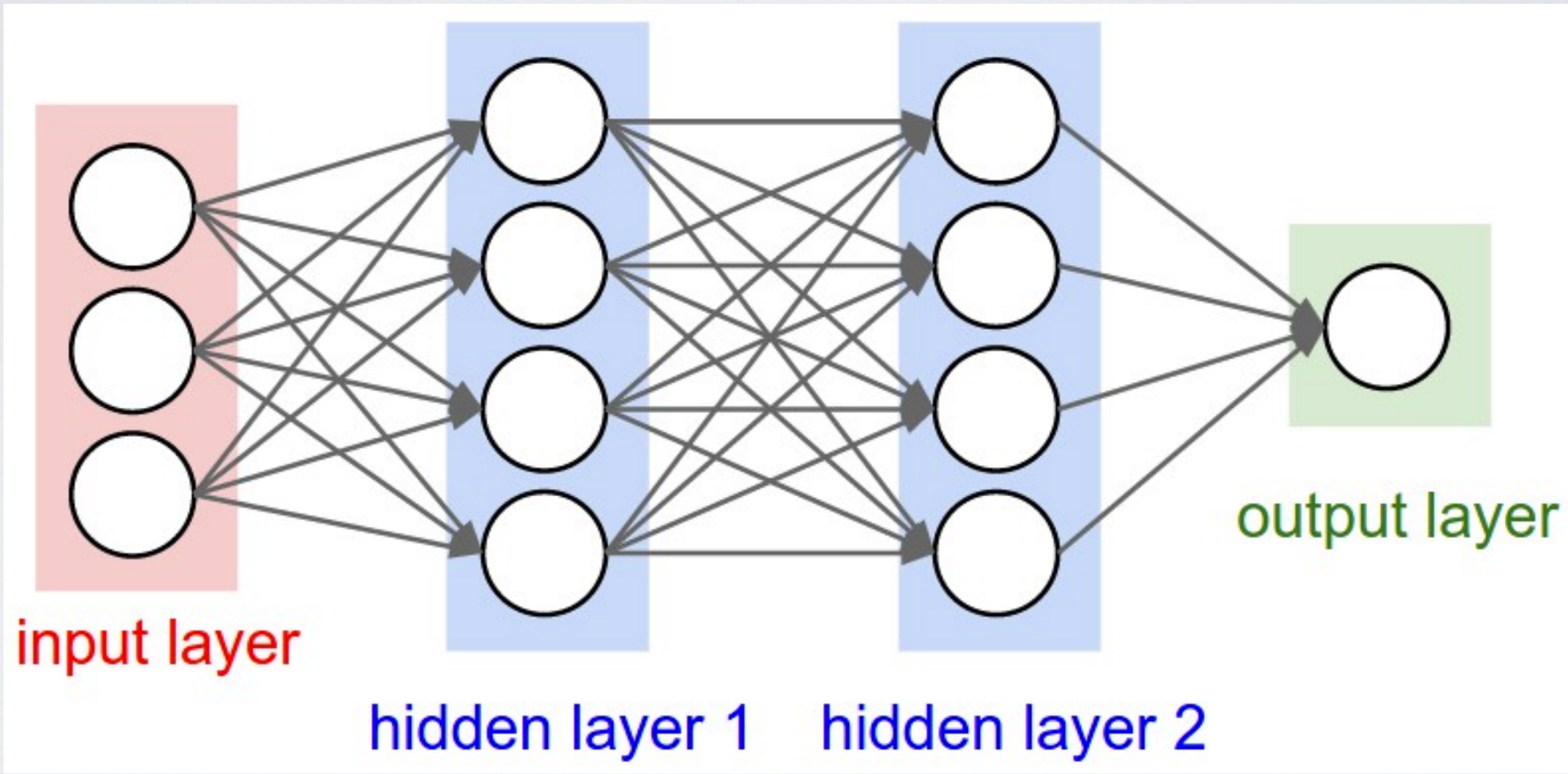


In neural net land, x is a row vector!

GRADIENT OF NET

$$\sigma(\sigma(\sigma(x_i W_1)W_2)W_3) = y_3$$

x_i



y_i

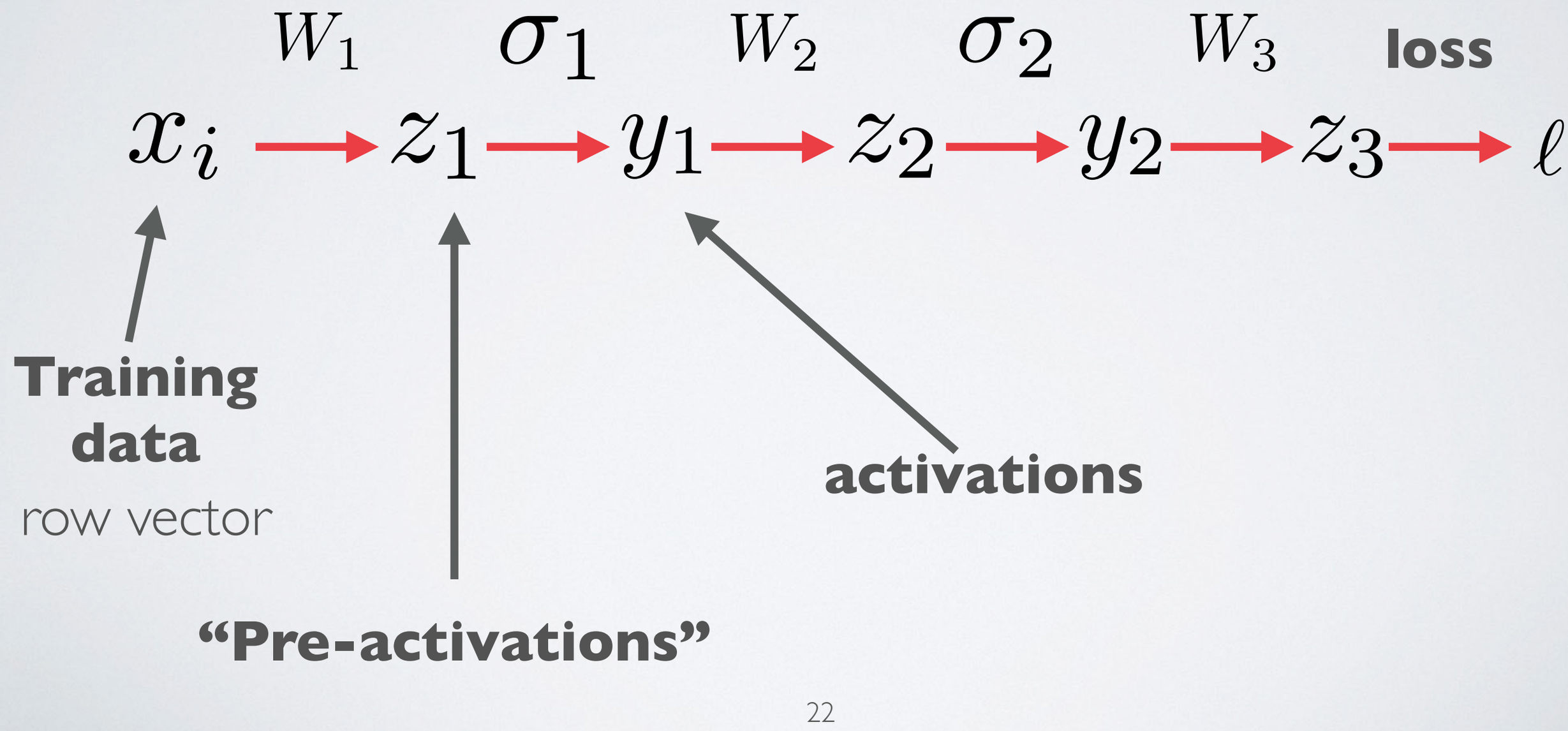
$\sigma(x_i W_1) = y_1$

$\sigma(y_1 W_2) = y_2$

$\sigma(y_2 W_3) = y_3$

STEP BY STEP

Break network apart into simple operations

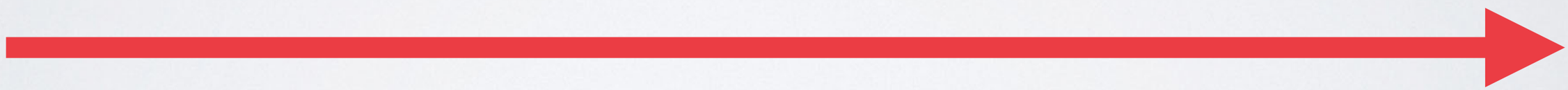


FORWARD PASS

Compute the intermediate values

W_1 σ_1 W_2 σ_2 W_3 **loss**

$x_i \longrightarrow z_1 \longrightarrow y_1 \longrightarrow z_2 \longrightarrow y_2 \longrightarrow z_3 \longrightarrow \ell$



Store the y's for later

BACKWARD PASS

Let's find derivative of loss with respect to x

$$\begin{array}{ccccccccccc} W_1 & & \sigma_1 & & W_2 & & \sigma_2 & & W_3 & & \textbf{loss} \\ x_i & \longrightarrow & z_1 & \longrightarrow & y_1 & \longrightarrow & z_2 & \longrightarrow & y_2 & \longrightarrow & z_3 \longrightarrow \ell \end{array}$$

derivative
with respect to this

$$\frac{\partial \ell}{\partial z_3}$$

BACKWARD PASS

Let's find derivative of loss with respect to x

$$\begin{array}{ccccccccccc} W_1 & & \sigma_1 & & W_2 & & \sigma_2 & & W_3 & & \textbf{loss} \\ x_i & \longrightarrow & z_1 & \longrightarrow & y_1 & \longrightarrow & z_2 & \longrightarrow & y_2 & \longrightarrow & z_3 & \longrightarrow & \ell \end{array}$$

derivative
with respect to this

$$W_3 \; \frac{\partial \ell}{\partial z_3}$$

BACKWARD PASS

Let's find derivative of loss with respect to x

$$\begin{array}{ccccccccccc} W_1 & & \sigma_1 & & W_2 & & \sigma_2 & & W_3 & & \textbf{loss} \\ x_i & \longrightarrow & z_1 & \longrightarrow & y_1 & \longrightarrow & z_2 & \longrightarrow & y_2 & \longrightarrow & z_3 & \longrightarrow & \ell \end{array}$$

derivative
with respect to this

$$\sigma_2' W_3 \frac{\partial \ell}{\partial z_3}$$

BACKWARD PASS

Let's find derivative of loss with respect to x

$$\begin{array}{ccccccccccc} & & W_1 & & \sigma_1 & & W_2 & & \sigma_2 & & W_3 & & \textbf{loss} \\ x_i & \longrightarrow & z_1 & \longrightarrow & y_1 & \longrightarrow & z_2 & \longrightarrow & y_2 & \longrightarrow & z_3 & \longrightarrow & \ell \end{array}$$

derivative
with respect to this

$$W_2\, \sigma_2'\, W_3\, \frac{\partial \ell}{\partial z_3}$$

BACKWARD PASS

Let's find derivative of loss with respect to x

$$\begin{array}{ccccccccccc} W_1 & & \sigma_1 & & W_2 & & \sigma_2 & & W_3 & & \textbf{loss} \\ x_i & \longrightarrow & z_1 & \longrightarrow & y_1 & \longrightarrow & z_2 & \longrightarrow & y_2 & \longrightarrow & z_3 & \longrightarrow & \ell \end{array}$$

derivative
with respect to this

$$\sigma_1' \, W_2 \, \sigma_2' \, W_3 \, \frac{\partial \ell}{\partial z_3}$$

BACKWARD PASS

Let's find derivative of loss with respect to x

$$\begin{array}{ccccccccccc} W_1 & & \sigma_1 & & W_2 & & \sigma_2 & & W_3 & & \textbf{loss} \\ x_i & \longrightarrow & z_1 & \longrightarrow & y_1 & \longrightarrow & z_2 & \longrightarrow & y_2 & \longrightarrow & z_3 & \longrightarrow & \ell \end{array}$$

derivative
with respect to this

$$W_1 \, \sigma_1' \, W_2 \, \sigma_2' \, W_3 \, \frac{\partial \ell}{\partial z_3}$$

GRADIENT VS DERIVATIVE

derivative

$$\frac{\partial \ell}{\partial x_1} = W_1 \, \sigma_1' \, W_2 \, \sigma_2' \, W_3 \, \frac{\partial \ell}{\partial z_3}$$

gradient

$$\nabla \ell_{x_1} = \nabla_{z_3} \ell \; W_3^\top \sigma_2' \; W_2^\top \sigma_1' \; W_1^\top$$

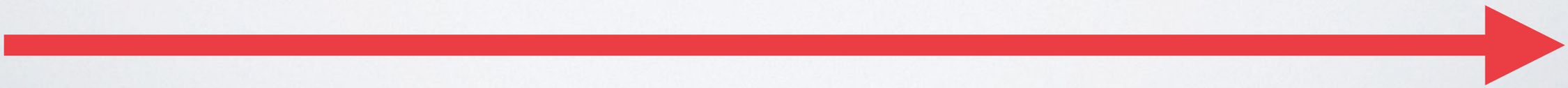
GRADIENT VS DERIVATIVE

derivative

$$\frac{\partial \ell}{\partial x_1} = W_1 \, \sigma_1' \, W_2 \, \sigma_2' \, W_3 \, \frac{\partial \ell}{\partial z_3}$$

gradient

$$\nabla \ell_{x_1} = \nabla_{z_3} \ell \; W_3^\top \sigma_2' \; W_2^\top \sigma_1' \; W_1^\top$$



“Backward pass”

SANITY CHECK

$$\nabla \ell_{x_1} = \nabla_{z_3} \ell \; W_3^\top \sigma_2' \; W_2^\top \sigma_1' \; W_1^\top$$

$\ell' = \nabla \ell$

W_1^T

What shape should the output be?

INSANITY CHECK: WHAT ABOUT WEIGHTS?

Let's find derivative of loss with respect to W_2

$$\begin{array}{ccccccccccc} & & W_1 & & \sigma_1 & & W_2 & & \sigma_2 & & W_3 & & \textbf{loss} \\ x_i & \longrightarrow & z_1 & \longrightarrow & y_1 & \longrightarrow & z_2 & \longrightarrow & y_2 & \longrightarrow & z_3 & \longrightarrow & \ell \end{array}$$

Use chain rule

$$\frac{\partial \ell}{\partial W_2} = \frac{\partial \ell}{\partial z_2} \frac{\partial z_2}{\partial W_2}$$

$$\nabla_{W_2} \ell = \nabla_{W_2} z_2 \nabla_{z_2} \ell$$

INSANITY CHECK: WHAT ABOUT WEIGHTS?

Let's find derivative of loss with respect to W_2

W_1 σ_1 W_2 σ_2 W_3 **loss**

$$x_i \rightarrow z_1 \rightarrow y_1 \rightarrow z_2 \rightarrow y_2 \rightarrow z_3 \rightarrow \ell$$

derivative
with respect to this

$$\sigma_2' \; W_3 \, \frac{\partial \ell}{\partial z_3}$$

INSANITY CHECK: WHAT ABOUT WEIGHTS?

Let's find **derivative** of loss with respect to W_2

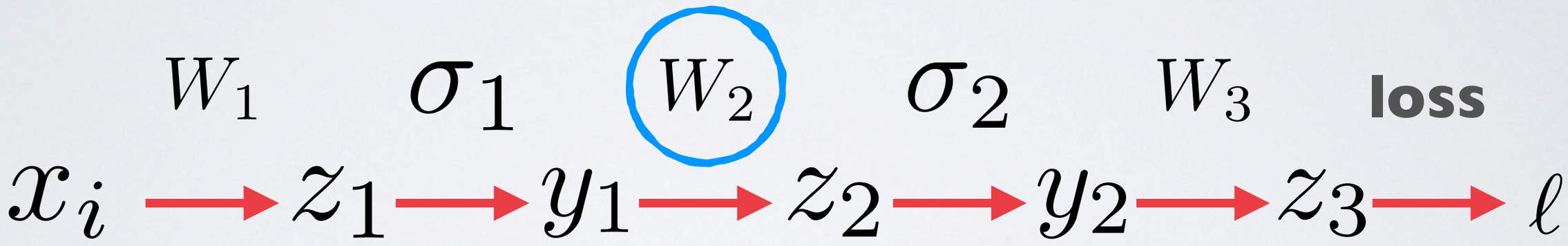
$$
\begin{array}{ccccccccccc}
 & & W_1 & & \sigma_1 & & W_2 & & \sigma_2 & & W_3 \\
x_i & \longrightarrow & z_1 & \longrightarrow & y_1 & \longrightarrow & z_2 & \longrightarrow & y_2 & \longrightarrow & z_3 \longrightarrow \ell
\end{array}
$$

loss

$$\sigma_2' \, W_3 \, \frac{\partial \ell}{\partial z_3} \, y_1 \leftarrow \frac{\partial z_2}{\partial W_2}$$

INSANITY CHECK: WHAT ABOUT WEIGHTS?

Let's find **gradient** of loss with respect to W_2



$$y_1^\top \nabla_{z_3} \ell \; W_3^\top \sigma_2'$$

COMPLETE BACKPROP

$$x_i \xrightarrow{W_1} z_1 \xrightarrow{\sigma_1} y_1 \xrightarrow{W_2} z_2 \xrightarrow{\sigma_2} y_2 \xrightarrow{W_3} z_3 \xrightarrow{\text{loss}} \ell$$

$$\nabla_{z_3} \ell$$

$$\nabla_{W_3} \ell = y_2^\top \nabla_{z_3} \ell$$

$$\nabla_{z_2} \ell = \nabla_{z_3} \ell W_3^\top \sigma_2'$$

$$\nabla_{W_2} \ell = y_1^\top \nabla_{z_2} \ell$$

$$\nabla_{z_1} \ell = \nabla_{z_3} \ell W_3^\top \sigma_2' W_2^\top \sigma_1'$$

$$\nabla_{W_1} \ell = x_i^\top \nabla_{z_1} \ell$$

This seem unnecessarily complicated…

REMEMBER THESE (SIMPLE) RULES?

If

$$h(x) = f(Ax)$$

Then

$$\nabla h(x) = A^T f'(Ax)$$

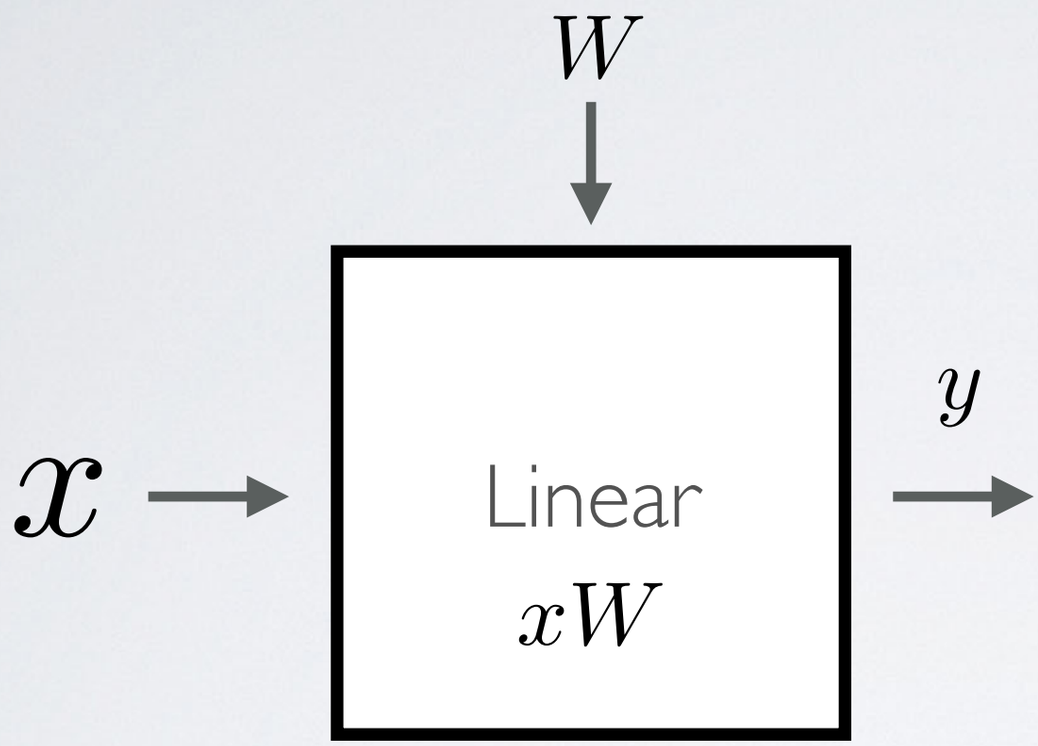
If

$$h(x) = f(xA)$$

Then

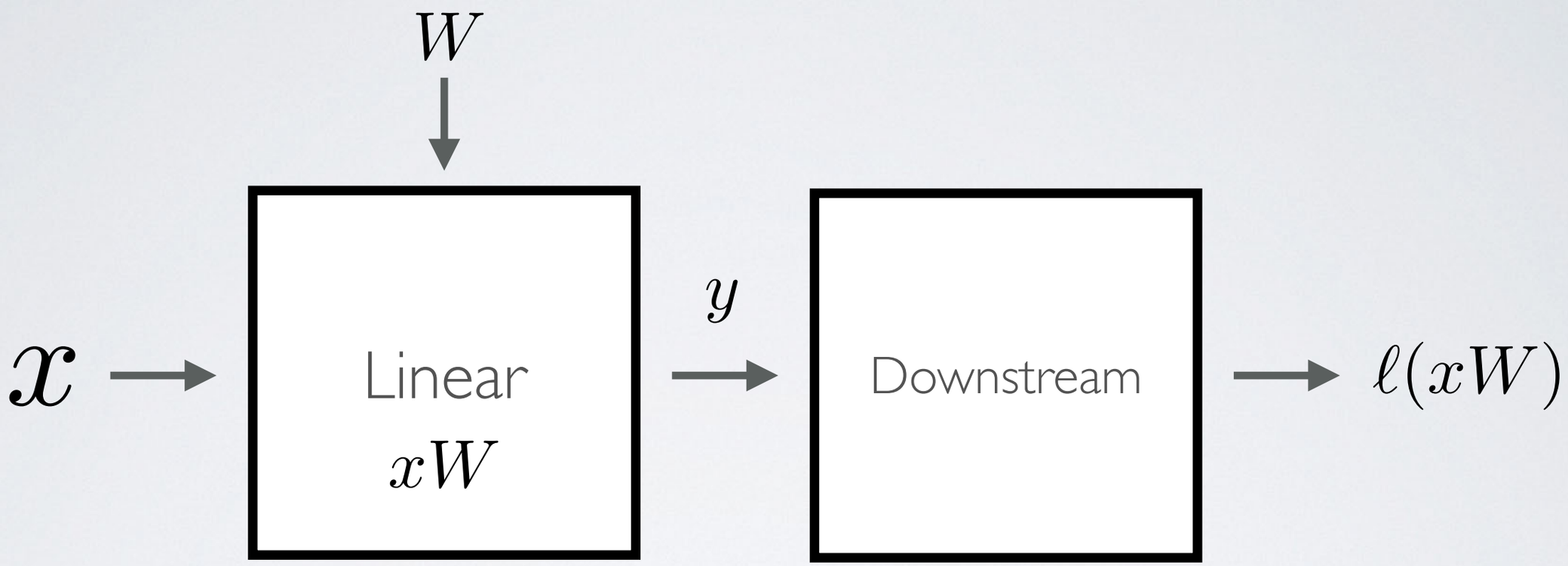
$$\nabla h(x) = f'(xA)A^T$$

AUTOGRAD




```
def forward(x, W):
    return x@W
```

AUTOGRAD

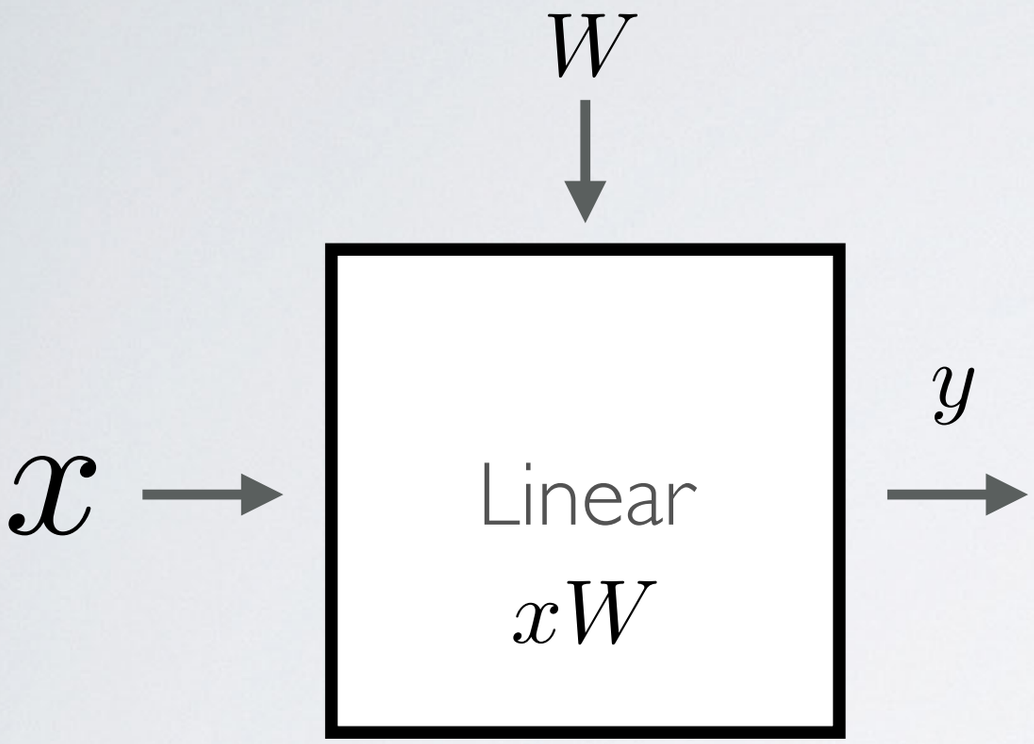


Gradient with respect to input?

$$\nabla_x \ell(xW) = \ell'(xW)W^T$$

$$\nabla_W \ell(xW) = x^T \ell'(xW)$$

AUTOGRAD




```
def forward(x, W):
    return  x@W
```

```
def backward(x, W, grad_ly):
    grad_lx = grad_ly@W.T
    grad_lW = x.T@grad_ly
    return  grad_lx, grad_lW
```

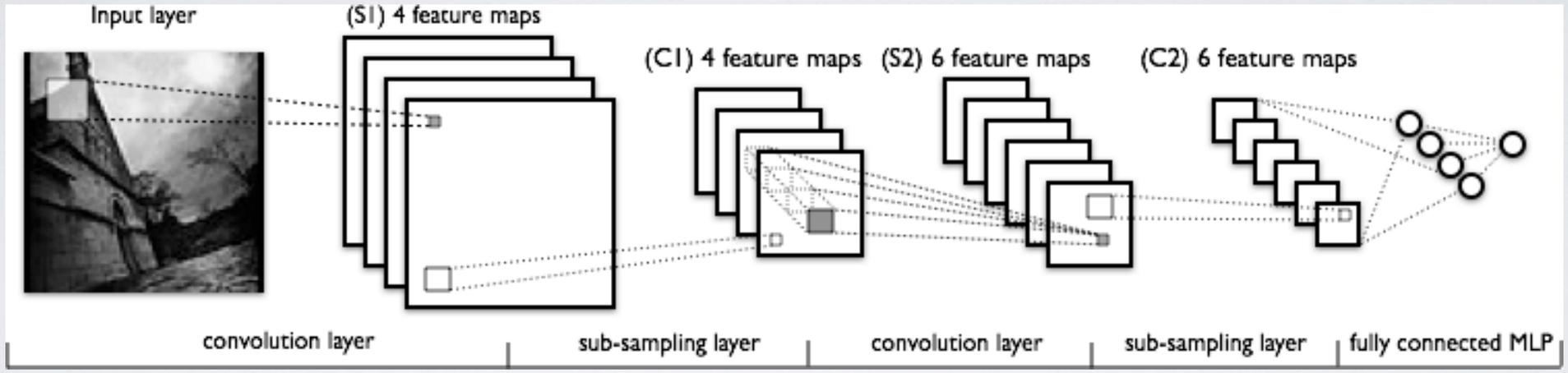
$$\nabla_x \ell(xW) = \ell'(xW)W^T$$

$$\nabla_W \ell(xW) = x^T \ell'(xW)$$

CONVOLUTIONAL NETS

Convolution-based template matching

CONVOLUTIONAL NETS

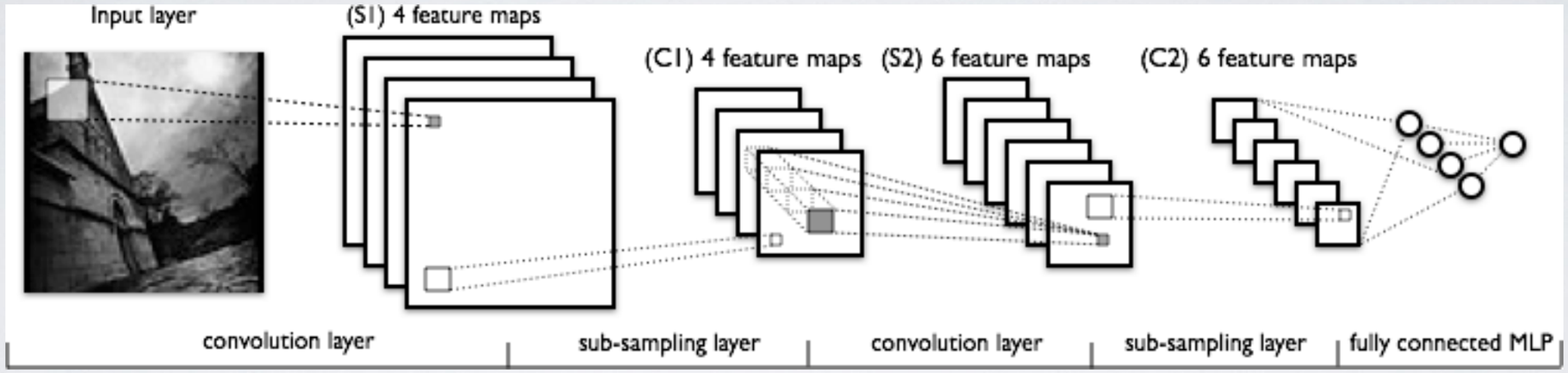


Regular-old MLP: $y_1 = \sigma(x_i W_1)$

Fancy-schmance convnet : $y_1 = \sigma(x_i K_1) = \sigma(x_i * k_1)$

conv matrix **stencil**

CONVOLUTIONAL NETS

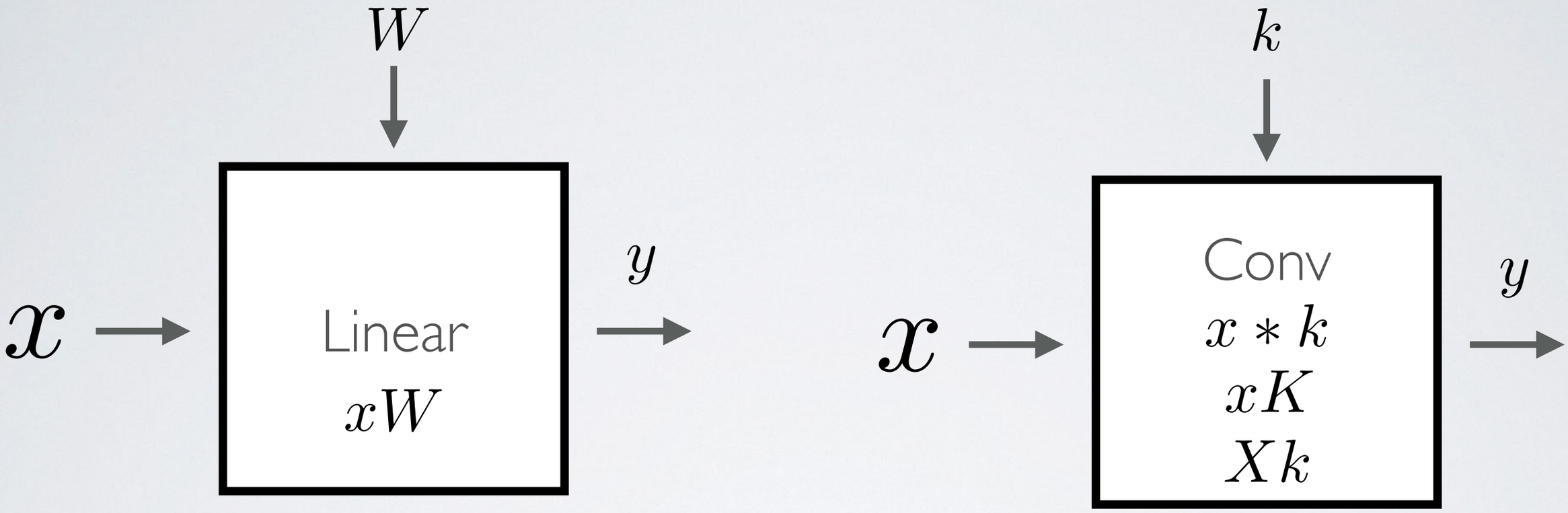


$$y = \sigma(\sigma(\sigma(x_i K_1)K_2)K_3)$$

$$\nabla_{z_1}\ell = \nabla_{z_3}\ell K_3^\top \sigma_2' K_2^\top \sigma_1'$$

Adjoint of convolution

AUTOGRAD



$$\nabla_x \ell(xW) = \ell'(xW)W^T$$

$$\nabla_W \ell(xW) = x^T \ell'(xW)$$

$$\nabla_x \ell(xK) = \ell'(xK)K^T$$

$$\nabla_W \ell(Xk) = X^T \ell'(Xk)$$

CONVOLUTIONAL NETS

$$\nabla_x \sigma_i(xK_i) = \mathrm{diag}(\sigma_i')K_i^T$$

Adjoint of convolution

Stencil

$\mathcal{F}k$

Diagonal

$K = \mathcal{F}^H D \mathcal{F}$

Adjoint

???

CONVOLUTIONAL NETS

$$\nabla_x \sigma_i(xK_i) = \mathrm{diag}(\sigma_i')K_i^T$$

Adjoint of convolution

Stencil

$\mathcal{F}k$

Diagonal

$K = \mathcal{F}^H D \mathcal{F}$

Adjoint

$K = \mathcal{F}^H \bar{D} \mathcal{F}$

???

CONVOLUTIONAL NETS

$$\nabla_x \sigma_i(xK_i) = \mathrm{diag}(\sigma_i')K_i^T$$

Adjoint of convolution

Stencil

$\mathcal{F}k$

Diagonal

$K = \mathcal{F}^H D \mathcal{F}$

Adjoint

$K = \mathcal{F}^H \bar{D} \mathcal{F}$

$\mathcal{F}\mathrm{flip}(k)$

Stencil