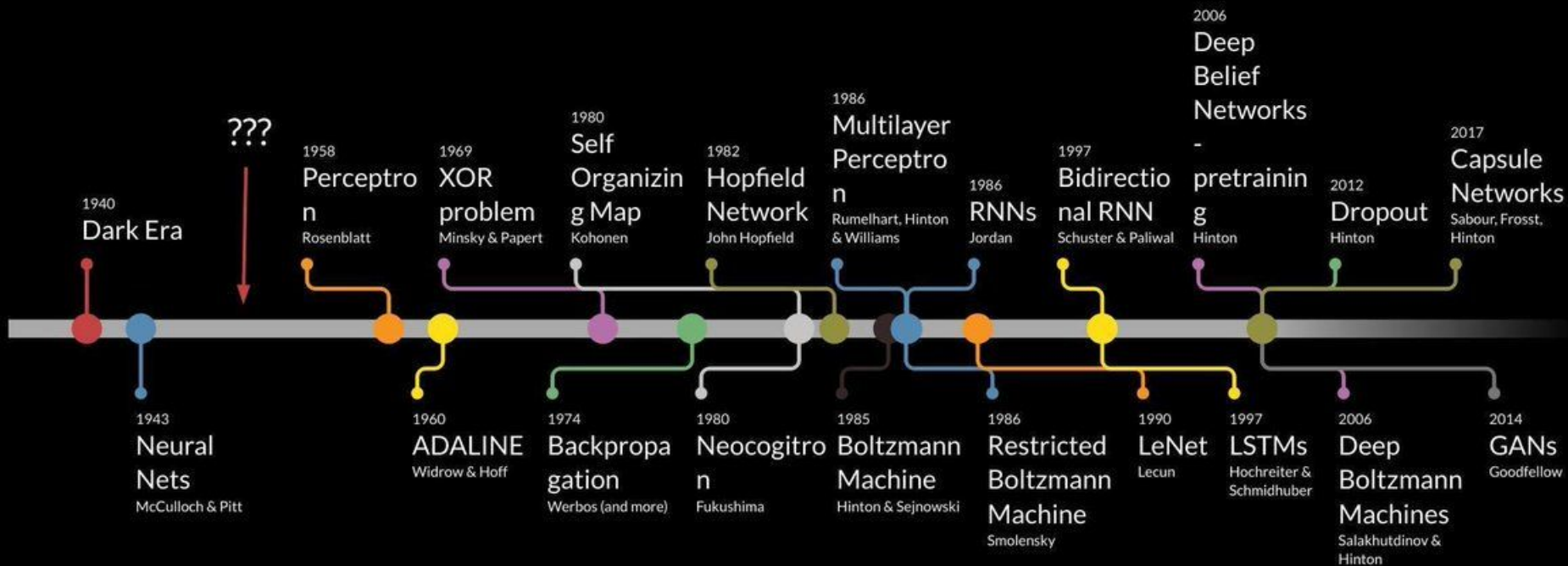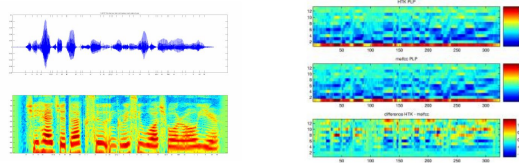# Introduction to Deep Learning

Outline

1. Neural Networks in different areas. Historical overview.

2. Backpropagation.

3. Activation functions.

4. Playground.

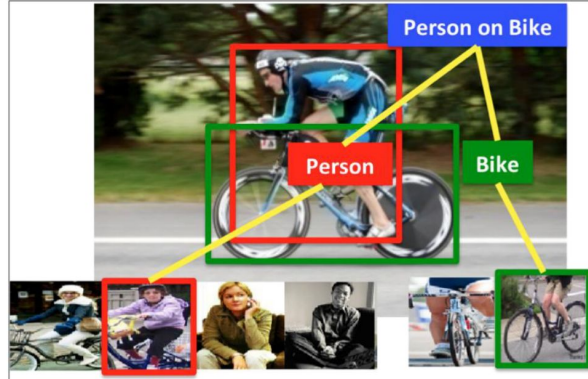5. Neural network from scratch.

Deep Learning Timeline

Made by Favio Vázquez

Real world problems

Audio Features

Spectrogram

MFCC

- Object detection
- Action classification
- Image captioning
- …

person
hammer
flower pot
power drill

Person on Bike

Person

Bike

"man in black shirt is playing guitar."

$$X \longrightarrow Wx + b \longrightarrow P(y)$$
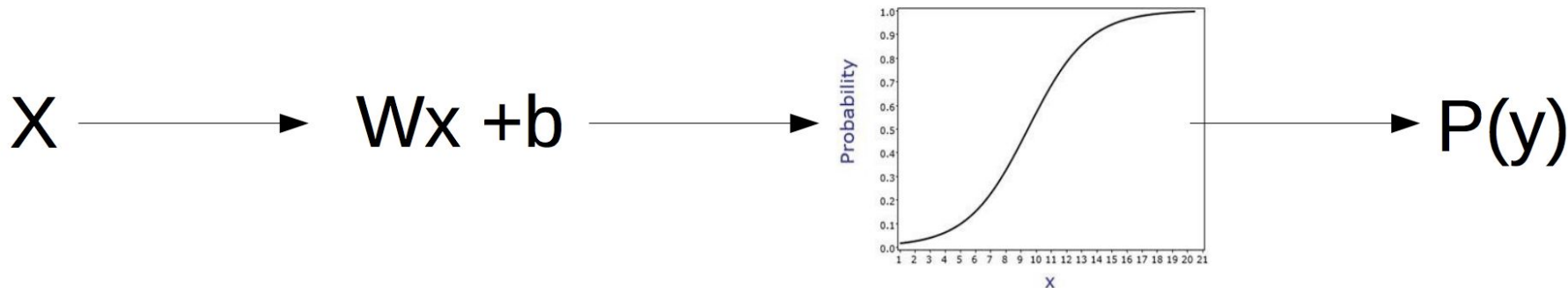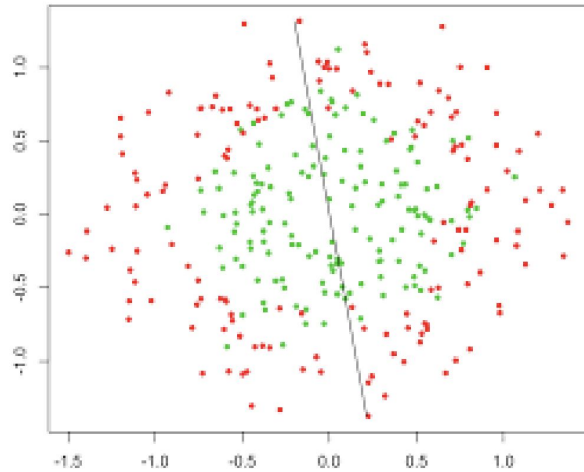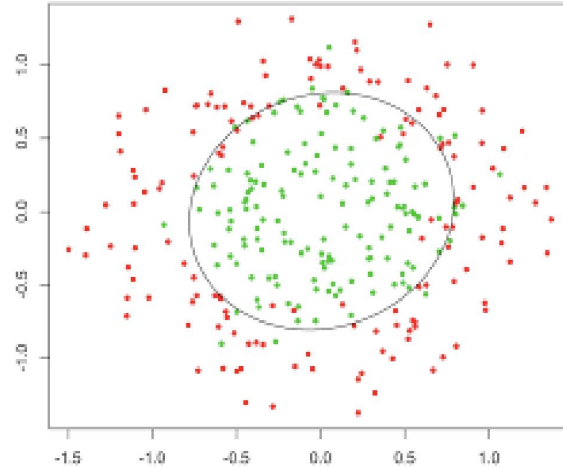
$$P(y|x) = \sigma(w \cdot x + b)$$

$$L = -\sum_i y_i \log P(y|x_i) + (1 - y_i) \log(1 - P(y|x_i))$$
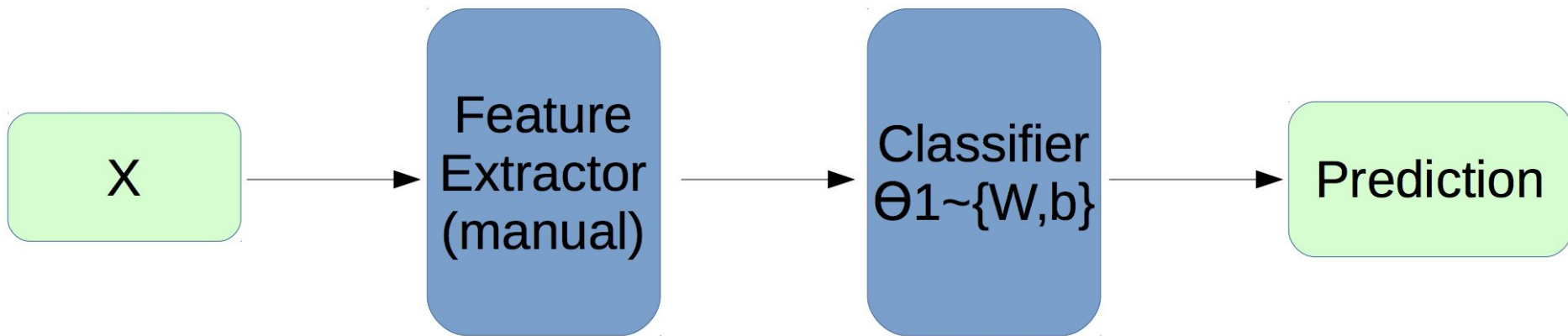
Problem: nonlinear dependencies

What we have

What we want

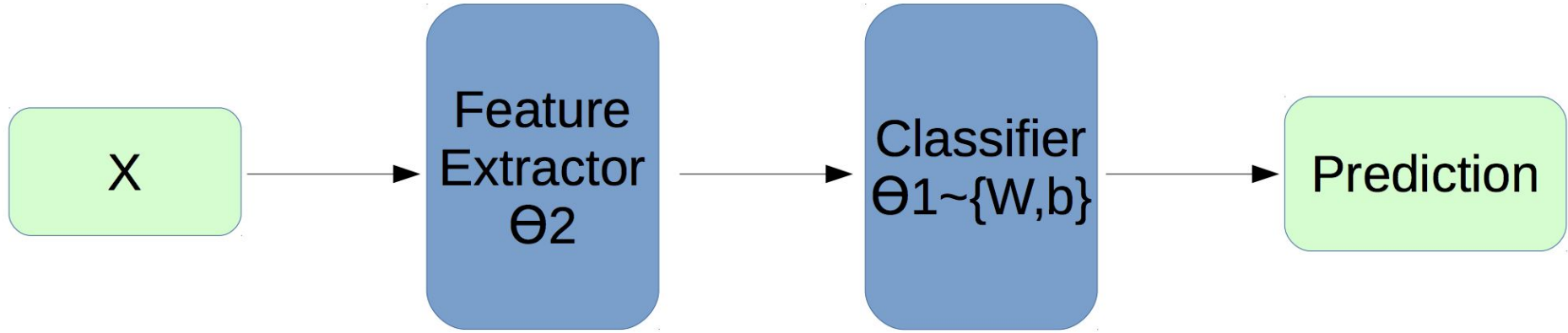Logistic regression (generally, linear model) need feature engineering to show good results.

And feature engineering is an *art*.

```
  ┌─────────┐        ┌──────────────┐        ┌──────────────┐        ┌──────────────┐
  │         │        │   Feature    │        │  Classifier  │        │              │
  │    X    │ ─────► │  Extractor   │ ─────► │  Θ1~{W,b}    │ ─────► │  Prediction  │
  │         │        │  (manual)    │        │              │        │              │
  └─────────┘        └──────────────┘        └──────────────┘        └──────────────┘
```

Handcrafted features, generated by experts.

NN pipeline

X → Feature Extractor $\Theta 2$ → Classifier $\Theta 1 \sim \{W, b\}$ → Prediction

Automatically extracted features.
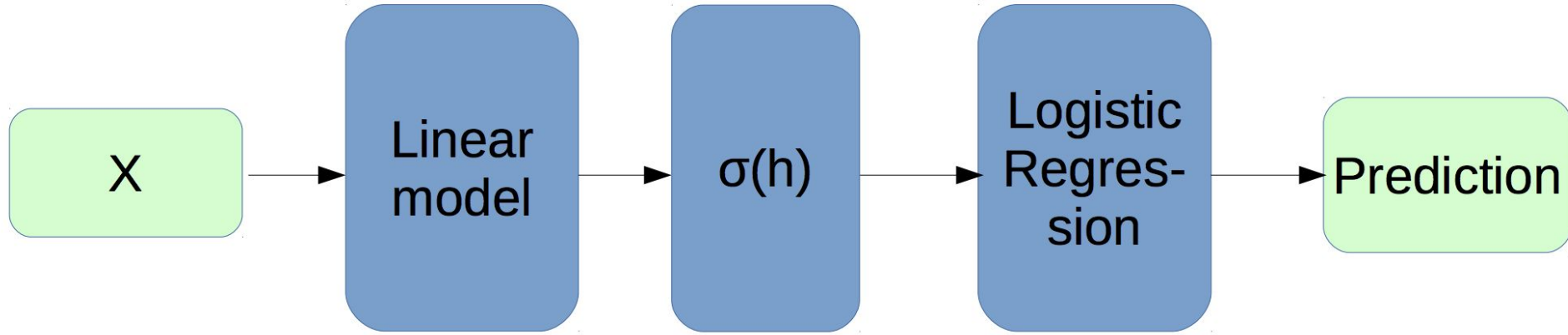
E.g. two logistic regressions one after another.

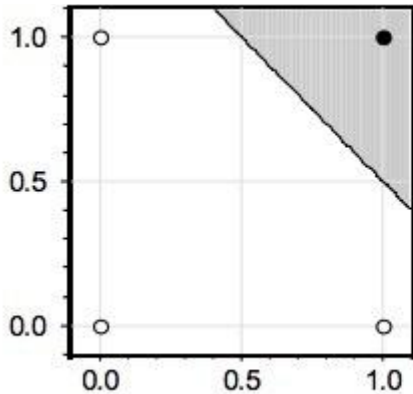X → Linear model → σ(h) → Logistic Regression → Prediction
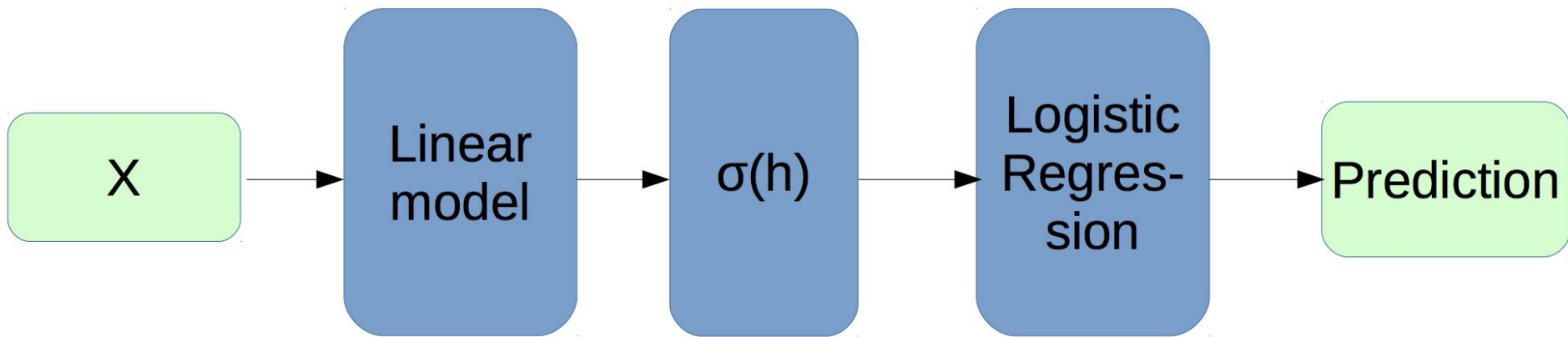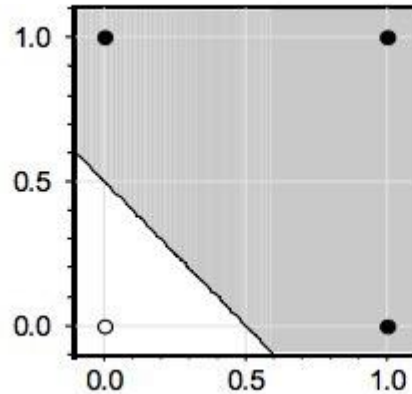
Actually, it's a neural network.

XOR problem
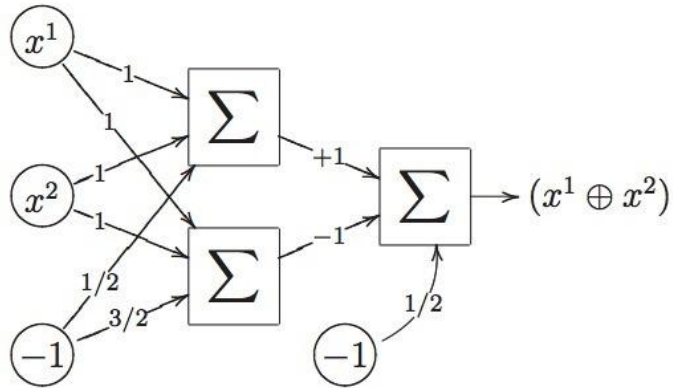
This 2-layer NN (on the left) implements XOR with only x1 and x2 features.
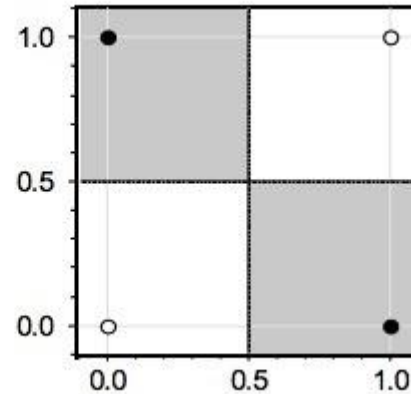
1-layer NN also can succeed, but only with extra feature x1*x2.
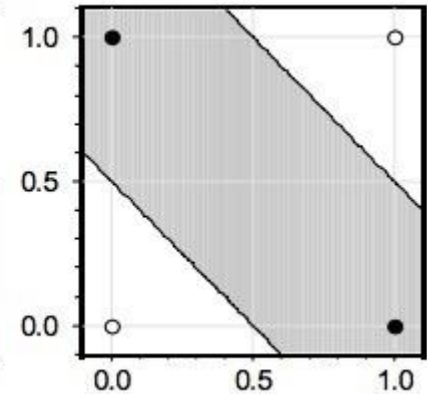
AND   OR   XOR(with x1*x2)   XOR

Activation functions: nonlinearities

$$f(a) = \frac{1}{1 + e^a}$$

$$f(a) = \tanh(a)$$

$$f(a) = \max(0, a)$$

$$f(a) = \log(1 + e^a)$$
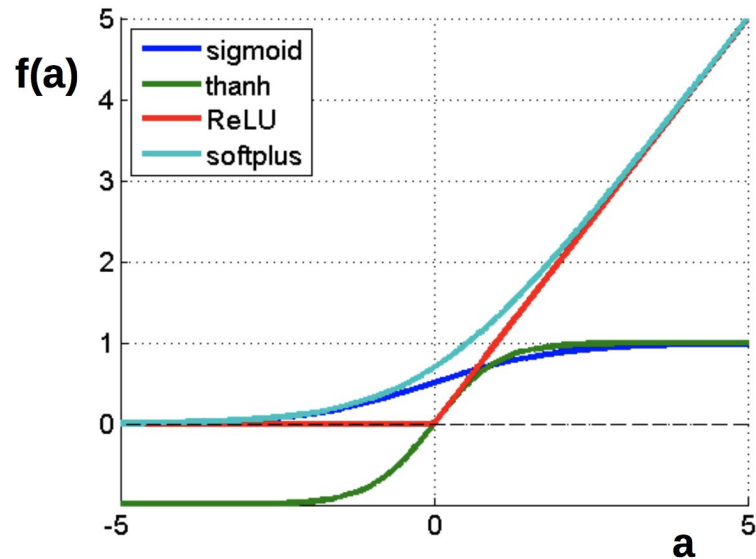
Some generally accepted terms

- Layer – a building block for NNs :
  - Dense layer: $f(x) = Wx+b$
  - Nonlinearity layer: $f(x) = \sigma(x)$
  - Input layer, output layer
  - A few more we will cover later
- Activation function – function applied to layer output
  - Sigmoid
  - tanh
  - ReLU
  - Any other function to get nonlinear intermediate signal in NN
- Backpropagation – a fancy word for "chain rule"

Dense layer
with a nonlinearity
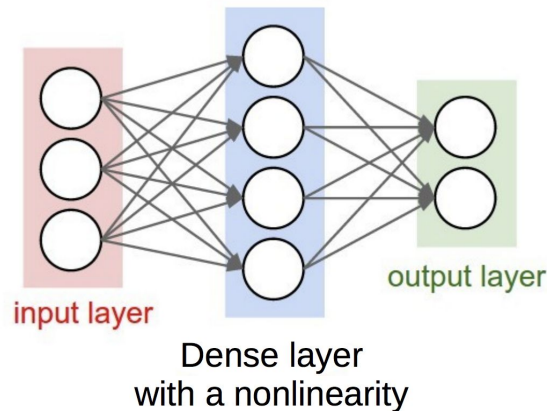
"Train it via backprop!"

Actually, it can be deeper

input layer

hidden layer 1   hidden layer 2

output layer

Much deeper...

15
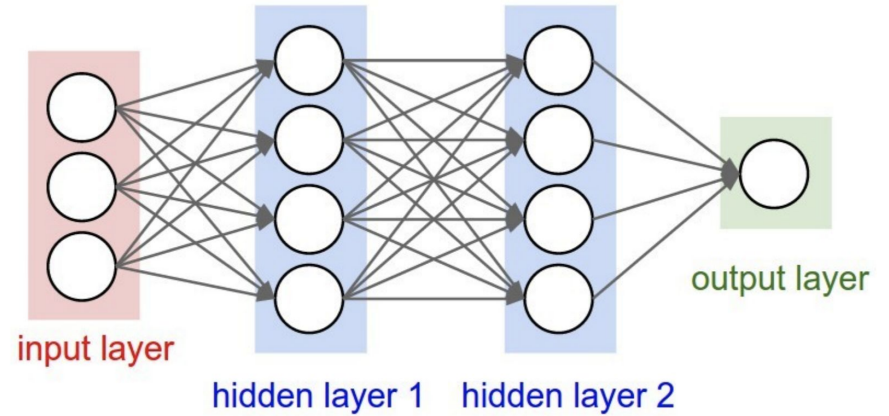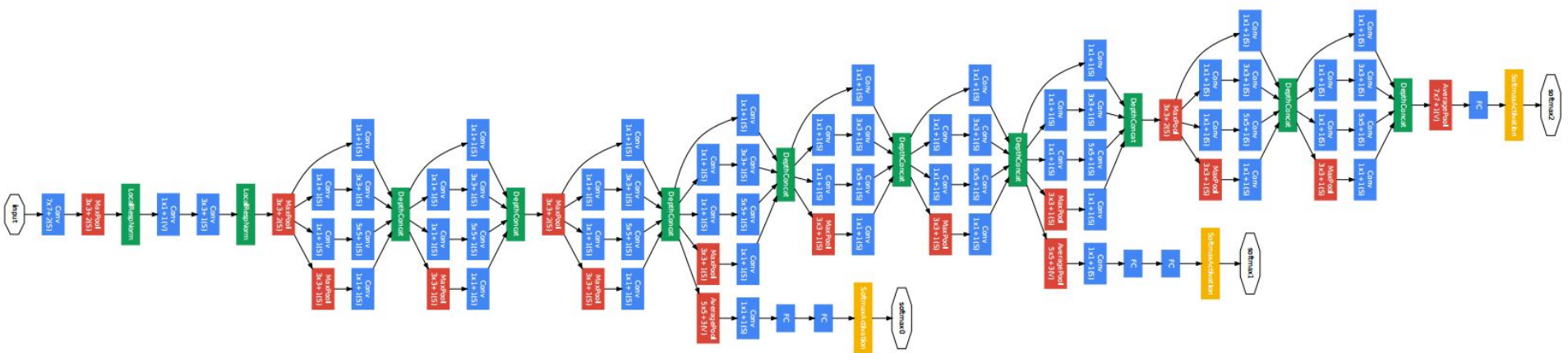
Much deeper...
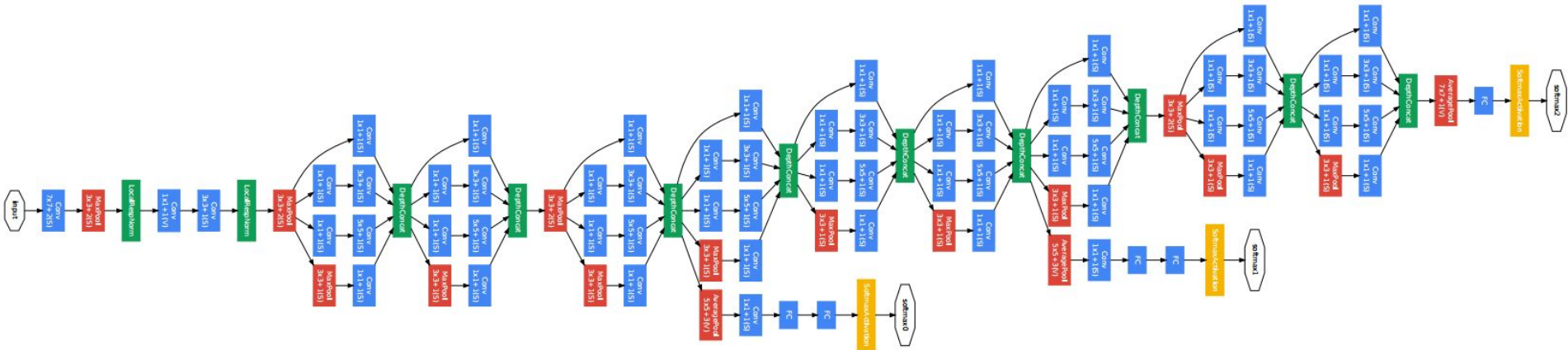
How to train it?

Backpropagation and chain rule

Chain rule is just simple math:

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z} \frac{\partial z}{\partial x}$$

Backprop is just way to
use it in NN training.

source: http://cs231n.github.io

Backpropagation example

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

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

source: http://cs231n.github.io

Backpropagation example

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

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

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

source: http://cs231n.github.io

Backpropagation example

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

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

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

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

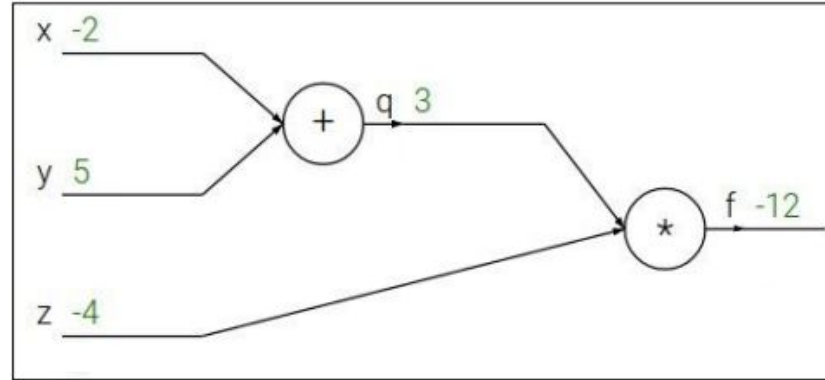source: http://cs231n.github.io

Backpropagation example

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

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

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

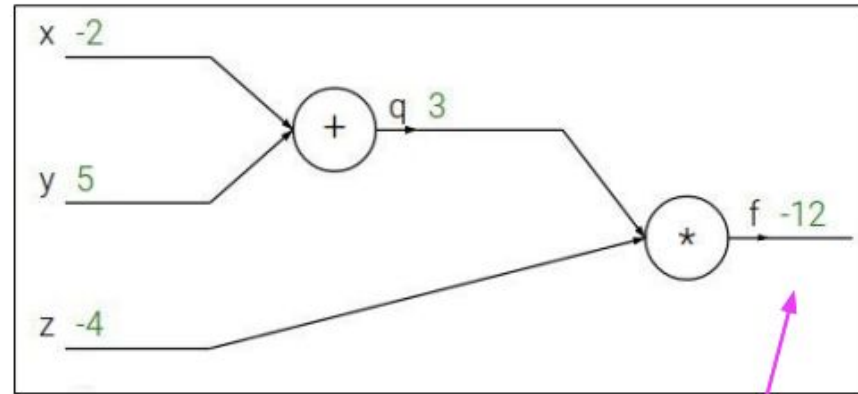$$\frac{\partial f}{\partial f}$$

Backpropagation example

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

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

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

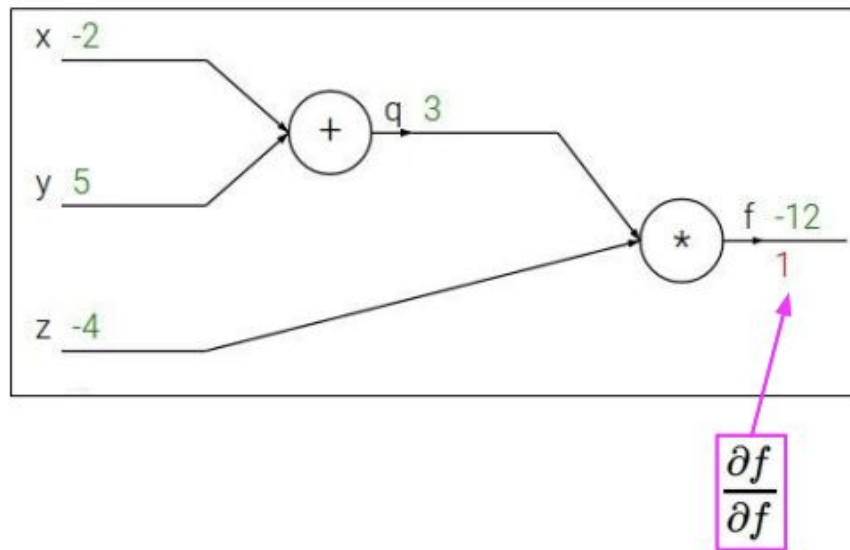$$\frac{\partial f}{\partial z}$$

Backpropagation example

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

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

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

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

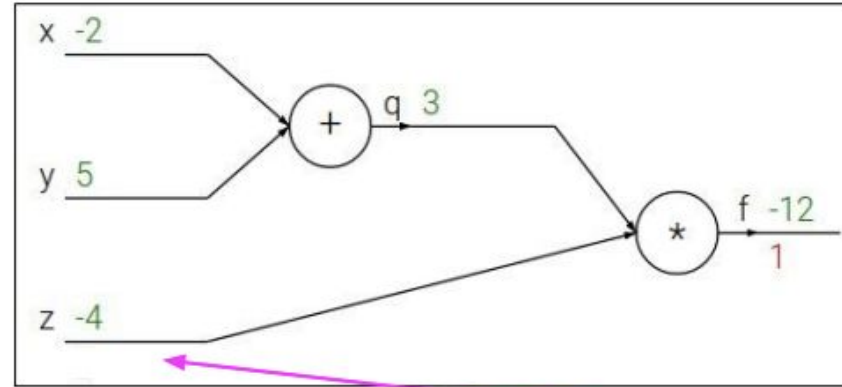source: http://cs231n.github.io

Backpropagation example

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

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

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

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

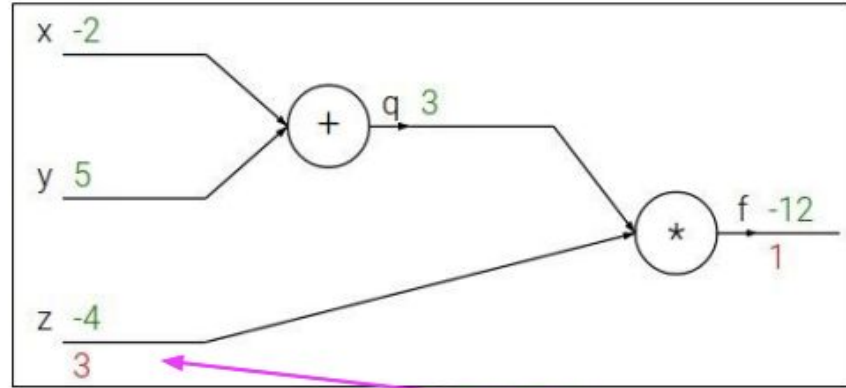source: http://cs231n.github.io

Backpropagation example

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

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

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

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

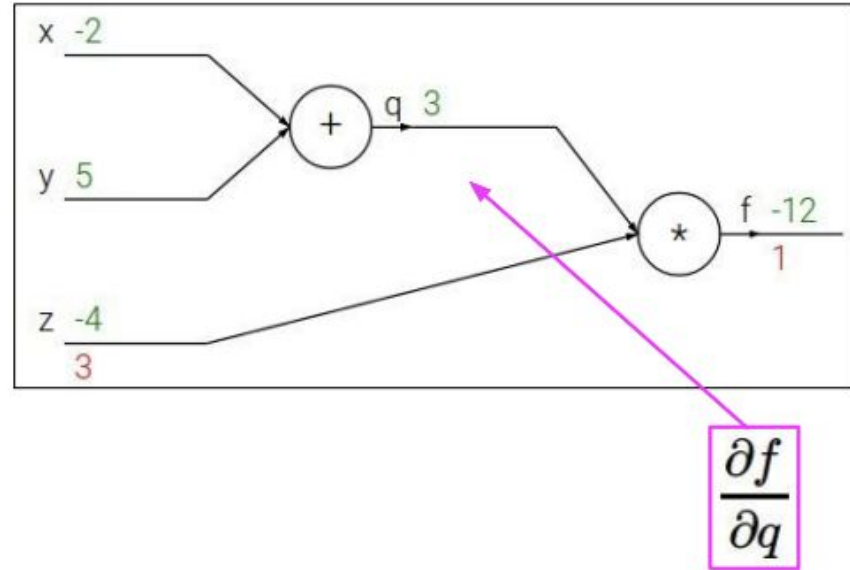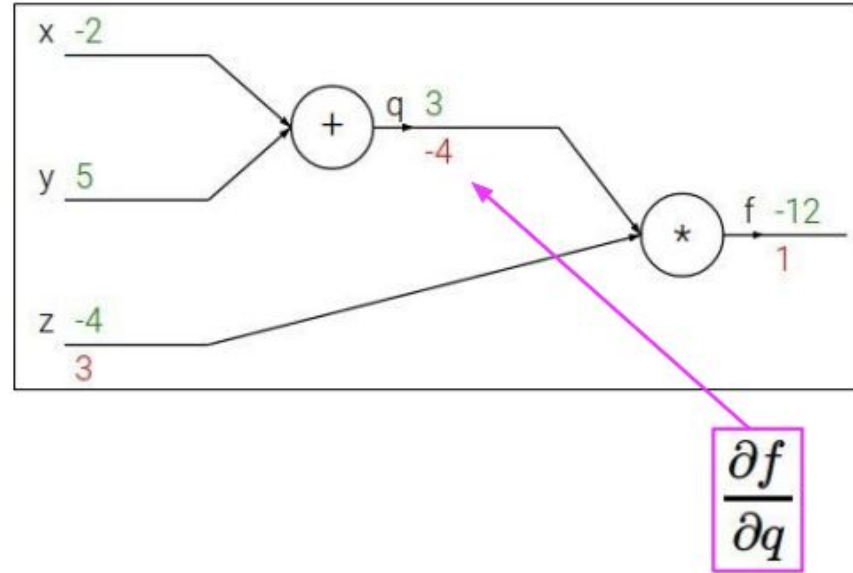source: http://cs231n.github.io

Backpropagation example

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

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

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

$$\frac{\partial f}{\partial y}$$
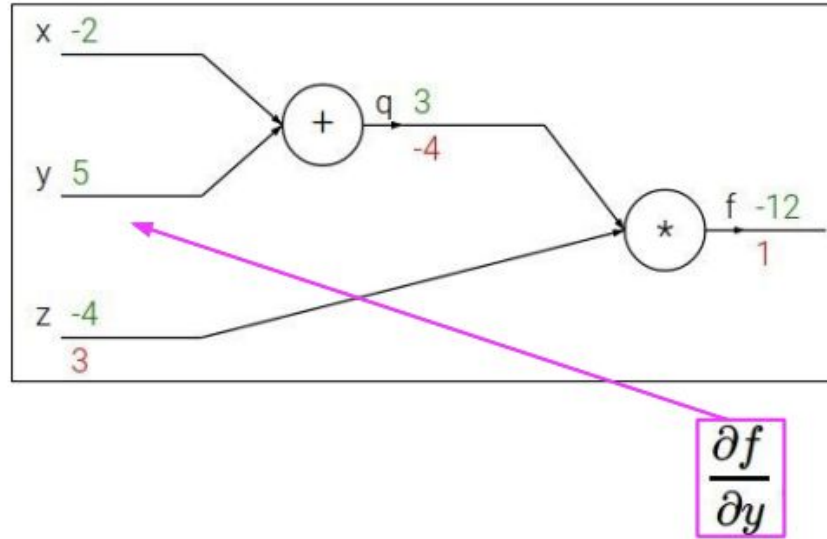
source: http://cs231n.github.io

Backpropagation example

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

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

$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$

$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$

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

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

Chain rule:

$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial y}$
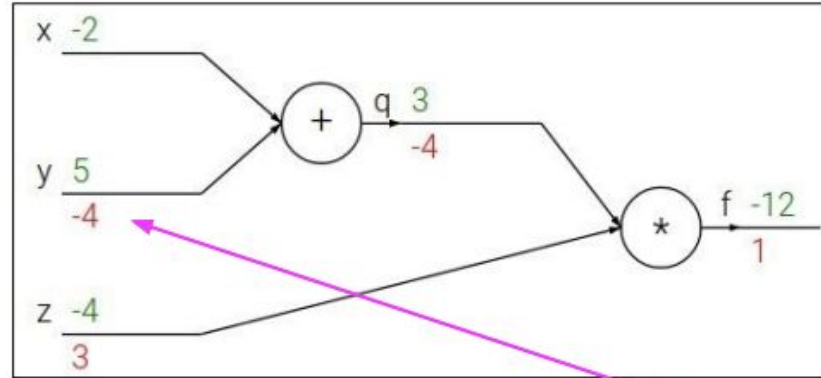
Backpropagation example

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

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

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

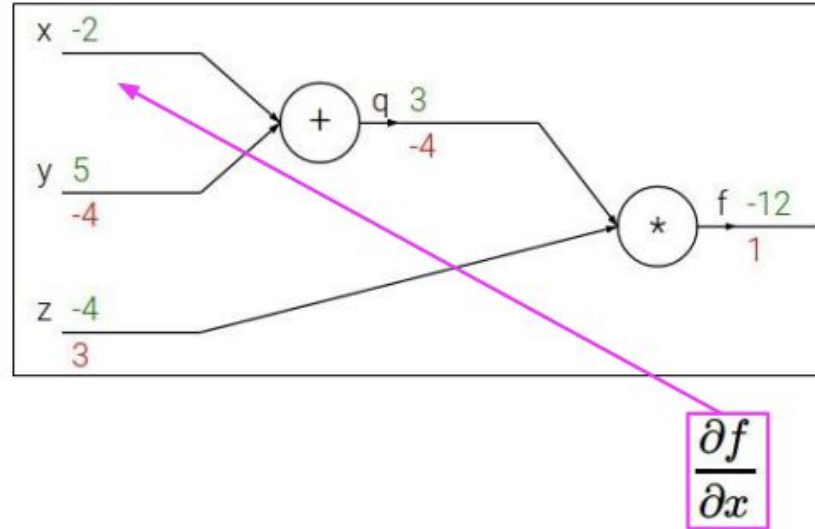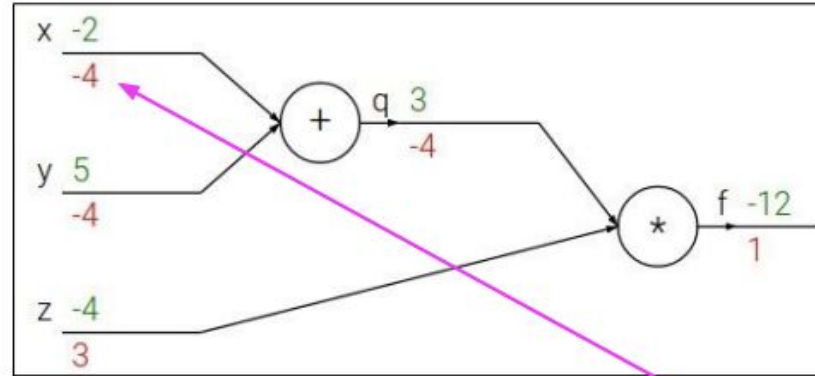source: http://cs231n.github.io

Backpropagation example

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

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

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

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

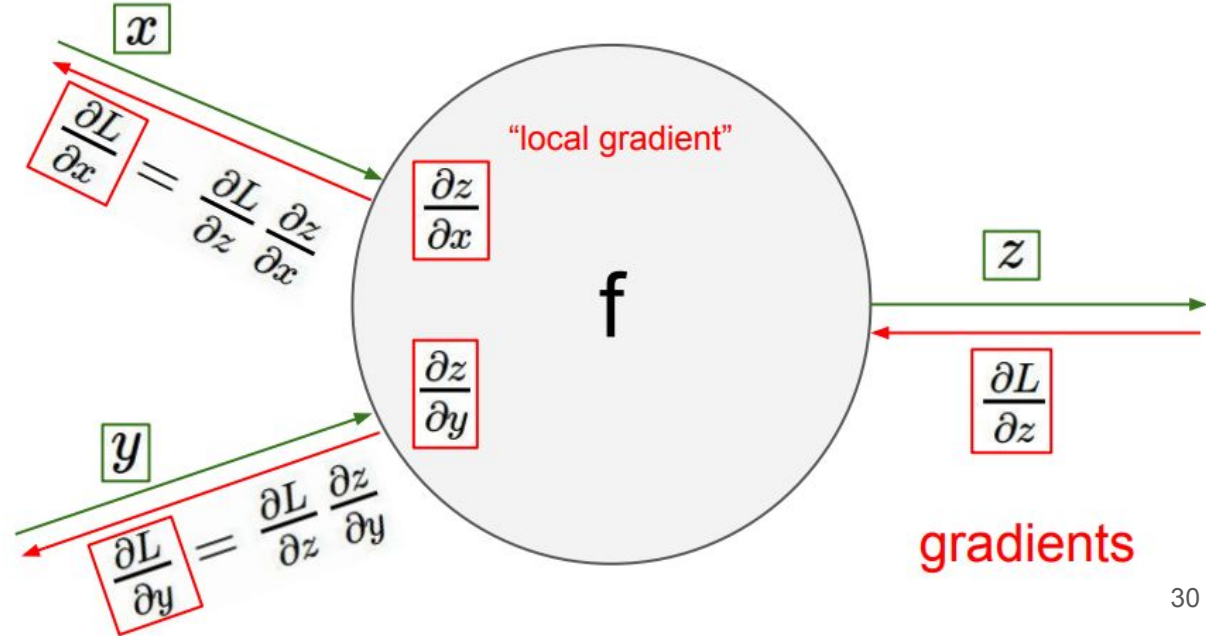$$\frac{\partial f}{\partial x}$$

Chain rule:

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

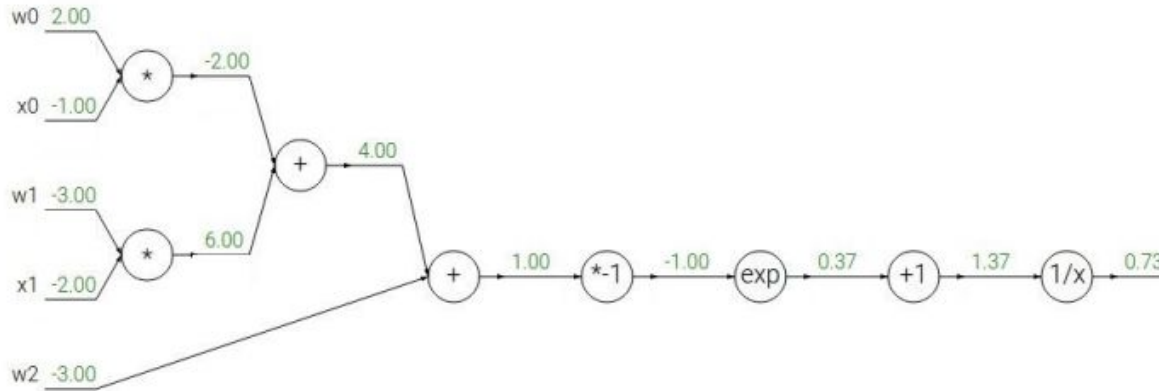# Backpropagation and chain rule

Chain rule is just simple math:

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial z}\frac{\partial z}{\partial x}$$
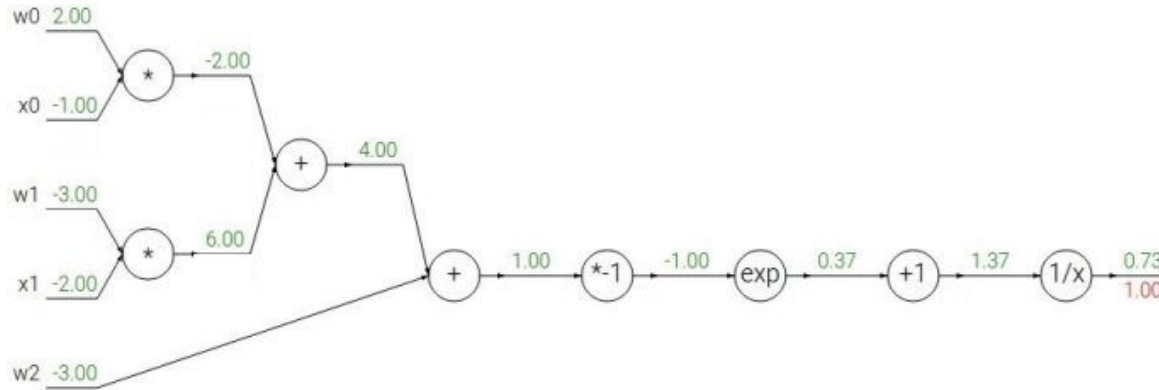
Backprop is just way to use it in NN training.

source: http://cs231n.github.io

Another example:

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

source: http://cs231n.github.io

Another example: $f(w,x) = \dfrac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$

$$f(x) = e^x \qquad \rightarrow \qquad \frac{df}{dx} = e^x$$

$$f_a(x) = ax \qquad \rightarrow \qquad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \qquad \rightarrow \qquad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \qquad \rightarrow \qquad \frac{df}{dx} = 1$$

source: http://cs231n.github.io

Backpropagation example

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

$$f(x) = e^x \qquad \rightarrow \qquad \frac{df}{dx} = e^x$$
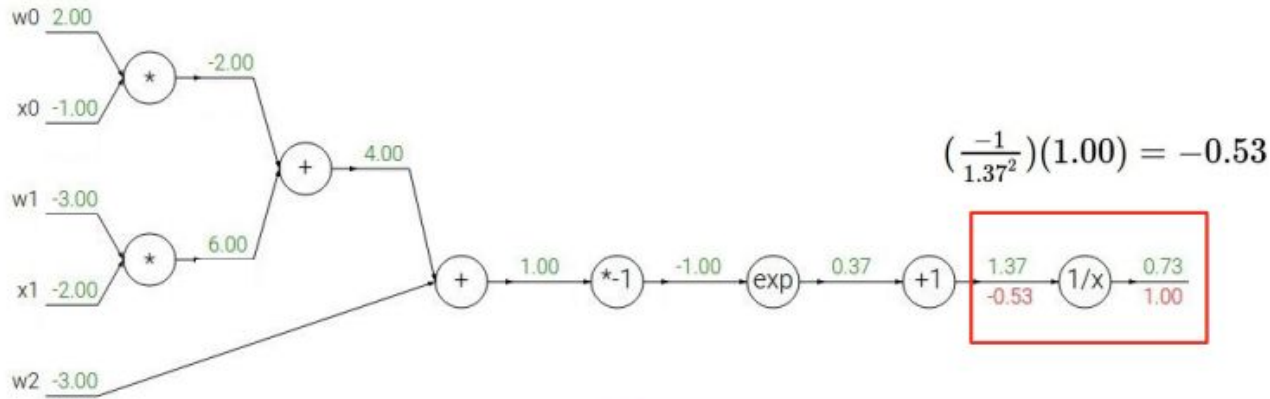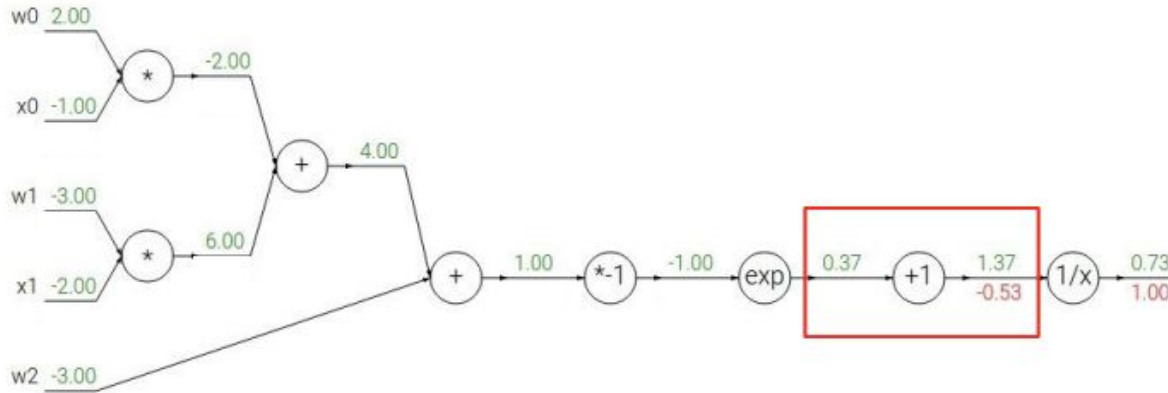
$$f_a(x) = ax \qquad \rightarrow \qquad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \qquad \rightarrow \qquad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \qquad \rightarrow \qquad \frac{df}{dx} = 1$$

source: http://cs231n.github.io

Another example:

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

$$\left(\frac{-1}{1.37^2}\right)(1.00) = -0.53$$

$$f(x) = e^x \qquad \rightarrow \qquad \frac{df}{dx} = e^x \qquad \Big| \qquad f(x) = \frac{1}{x} \qquad \rightarrow \qquad \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \qquad \rightarrow \qquad \frac{df}{dx} = a \qquad \Big| \qquad f_c(x) = c + x \qquad \rightarrow \qquad \frac{df}{dx} = 1$$

source: http://cs231n.github.io

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

$$f(x) = e^x \qquad \rightarrow \qquad \frac{df}{dx} = e^x \qquad \Bigg| \qquad f(x) = \frac{1}{x} \qquad \rightarrow \qquad \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \qquad \rightarrow \qquad \frac{df}{dx} = a \qquad \Bigg| \qquad f_c(x) = c + x \qquad \rightarrow \qquad \frac{df}{dx} = 1$$

source: http://cs231n.github.io

35

Backpropagation example

Another example: $\quad f(w, x) = \dfrac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}}$

$$(1)(-0.53) = -0.53$$

$$f(x) = e^x \qquad \rightarrow \qquad \frac{df}{dx} = e^x \qquad \Bigg| \qquad f(x) = \frac{1}{x} \qquad \rightarrow \qquad \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \qquad \rightarrow \qquad \frac{df}{dx} = a \qquad \Bigg| \qquad f_c(x) = c + x \qquad \rightarrow \qquad \frac{df}{dx} = 1$$

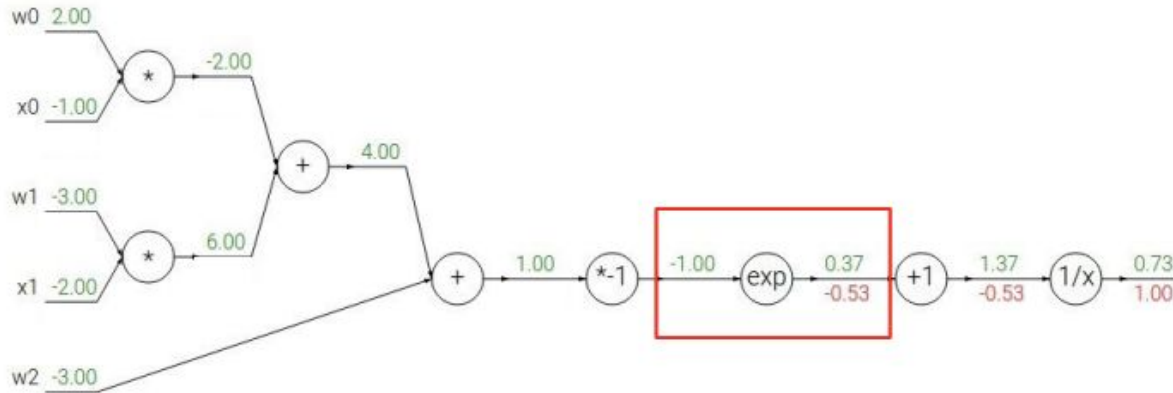source: http://cs231n.github.io

Backpropagation example

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

$$f(x) = e^x \qquad \rightarrow \qquad \frac{df}{dx} = e^x$$

$$f_a(x) = ax \qquad \rightarrow \qquad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \qquad \rightarrow \qquad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \qquad \rightarrow \qquad \frac{df}{dx} = 1$$

source: http://cs231n.github.io

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

$$(e^{-1})(-0.53) = -0.20$$

$$f(x) = e^x \qquad \rightarrow \qquad \frac{df}{dx} = e^x \qquad\bigg|\qquad f(x) = \frac{1}{x} \qquad \rightarrow \qquad \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \qquad \rightarrow \qquad \frac{df}{dx} = a \qquad\bigg|\qquad f_c(x) = c + x \qquad \rightarrow \qquad \frac{df}{dx} = 1$$

source: http://cs231n.github.io

Backpropagation example

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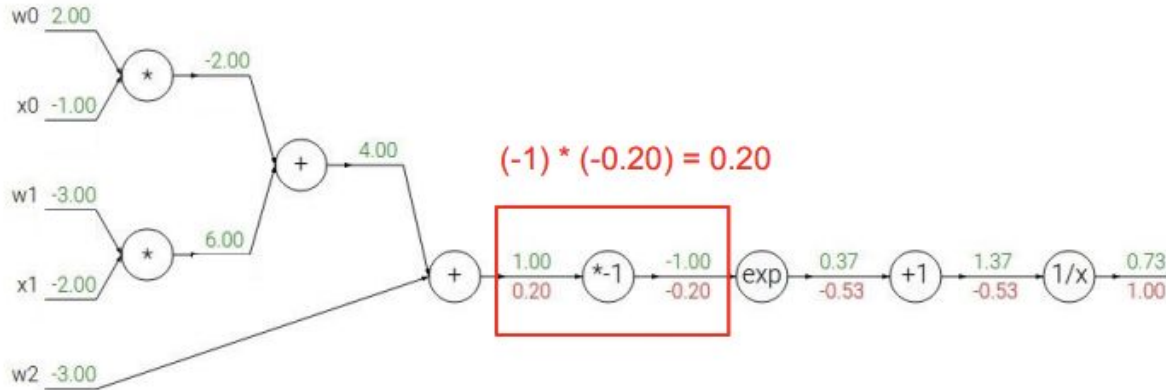

$$f(x) = e^x \qquad \rightarrow \qquad \frac{df}{dx} = e^x$$

$$f_a(x) = ax \qquad \rightarrow \qquad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \qquad \rightarrow \qquad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \qquad \rightarrow \qquad \frac{df}{dx} = 1$$

source: http://cs231n.github.io

Backpropagation example

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

w0 2.00
x0 -1.00
w1 -3.00
x1 -2.00
w2 -3.00

(-1) * (-0.20) = 0.20

$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

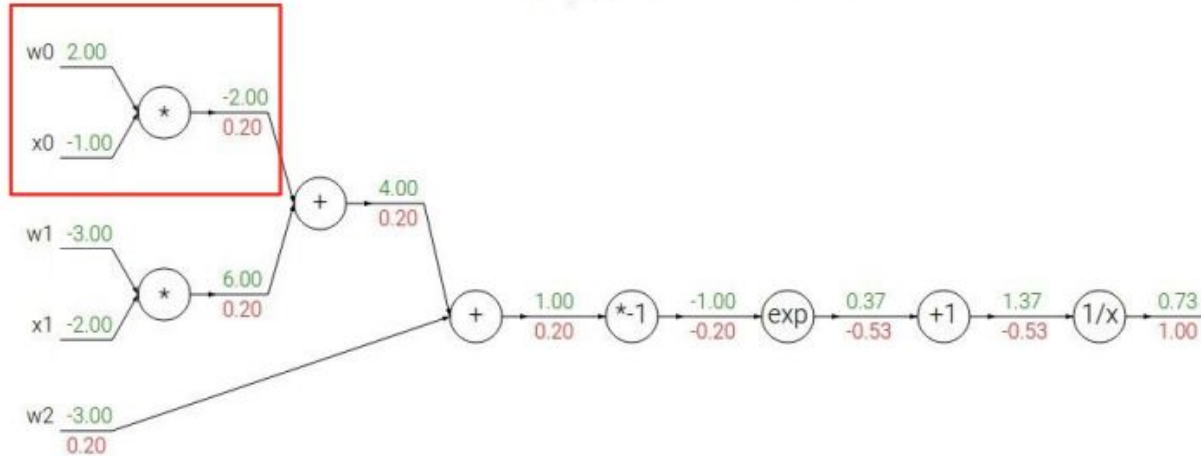source: http://cs231n.github.io

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

$f(x) = e^x$ $\rightarrow$ $\dfrac{df}{dx} = e^x$ $\bigg|$ $f(x) = \dfrac{1}{x}$ $\rightarrow$ $\dfrac{df}{dx} = -1/x^2$

$f_a(x) = ax$ $\rightarrow$ $\dfrac{df}{dx} = a$ $f_c(x) = c + x$ $\rightarrow$ $\dfrac{df}{dx} = 1$
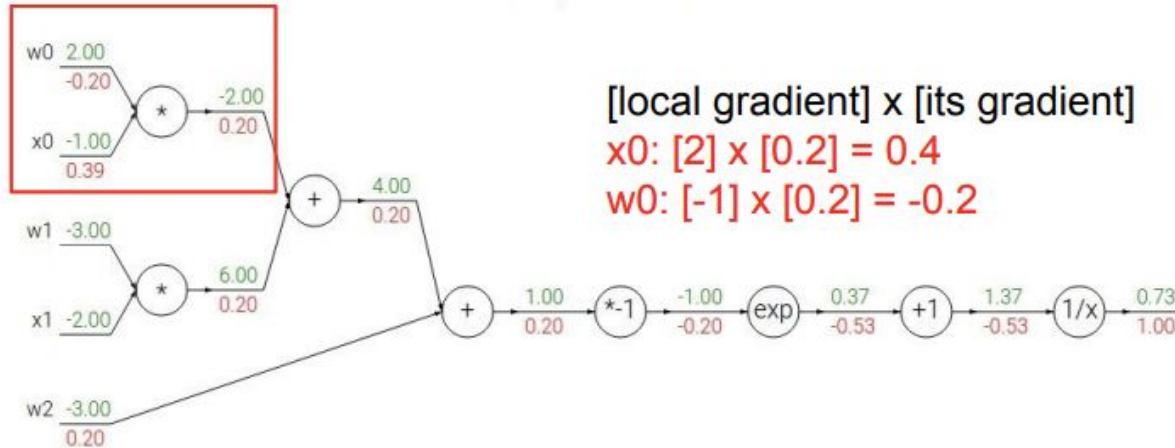
41

source: http://cs231n.github.io

Backpropagation example

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

[local gradient] x [its gradient]
[1] x [0.2] = 0.2
[1] x [0.2] = 0.2 (both inputs!)

$f(x) = e^x \qquad \rightarrow \qquad \dfrac{df}{dx} = e^x$

$f_a(x) = ax \qquad \rightarrow \qquad \dfrac{df}{dx} = a$

$f(x) = \dfrac{1}{x} \qquad \rightarrow \qquad \dfrac{df}{dx} = -1/x^2$

$f_c(x) = c + x \qquad \rightarrow \qquad \dfrac{df}{dx} = 1$

source: http://cs231n.github.io

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

$$f(x) = e^x \qquad \rightarrow \qquad \frac{df}{dx} = e^x \quad \bigg| \quad f(x) = \frac{1}{x} \qquad \rightarrow \qquad \frac{df}{dx} = -1/x^2$$

$$f_a(x) = ax \qquad \rightarrow \qquad \frac{df}{dx} = a \qquad \quad f_c(x) = c + x \qquad \rightarrow \qquad \frac{df}{dx} = 1$$

43

source: http://cs231n.github.io

Backpropagation example

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

[local gradient] x [its gradient]
x0: [2] x [0.2] = 0.4
w0: [-1] x [0.2] = -0.2

$$f(x) = e^x \qquad \rightarrow \qquad \frac{df}{dx} = e^x$$

$$f_a(x) = ax \qquad \rightarrow \qquad \frac{df}{dx} = a$$

$$f(x) = \frac{1}{x} \qquad \rightarrow \qquad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \qquad \rightarrow \qquad \frac{df}{dx} = 1$$

44

Backpropagation example

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

$$\boxed{\sigma(x) = \frac{1}{1 + e^{-x}}}$$ sigmoid function

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1 + e^{-x})^2} = \left(\frac{1 + e^{-x} - 1}{1 + e^{-x}}\right)\left(\frac{1}{1 + e^{-x}}\right) = (1 - \sigma(x))\,\sigma(x)$$

(0.73) * (1 - 0.73) = 0.2

source: http://cs231n.github.io
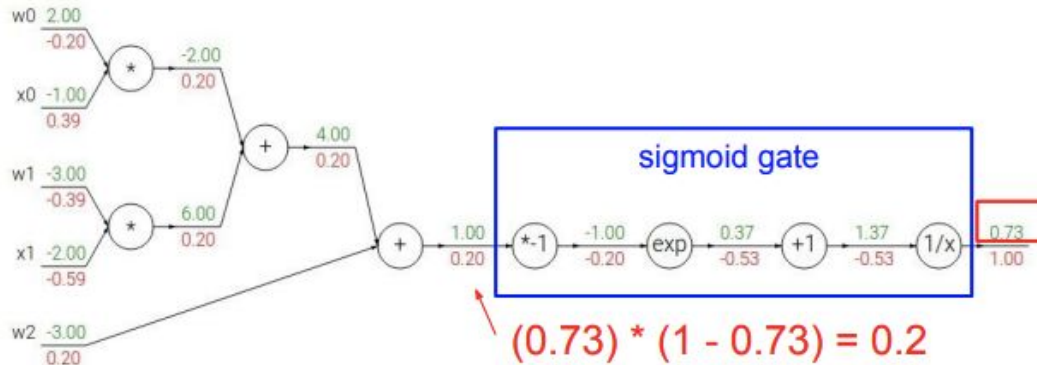
Stochastic gradient descent (and variations)
is used to optimize NN parameters.

$$x_{t+1} = x_t - \text{learning rate} \cdot dx$$

source: http://cs231n.github.io/neural-networks-3/

$$f(a) = \frac{1}{1 + e^a}$$

$$f(a) = \tanh(a)$$
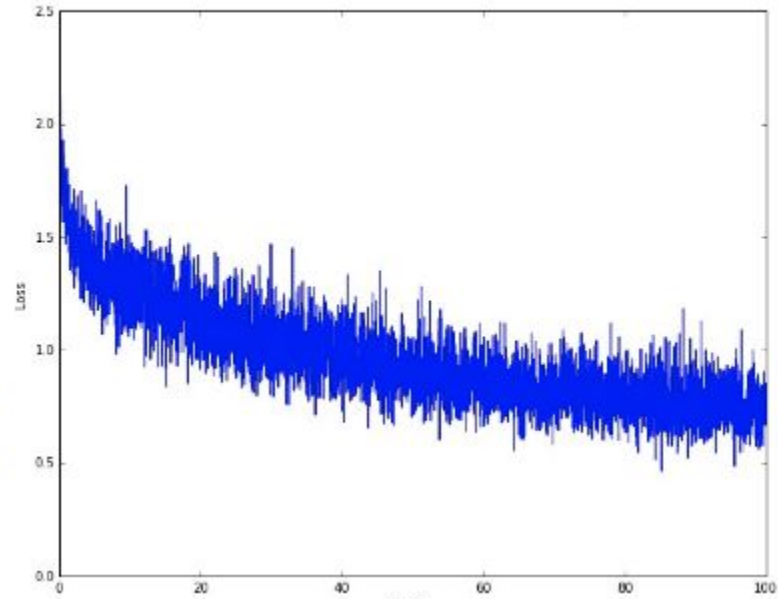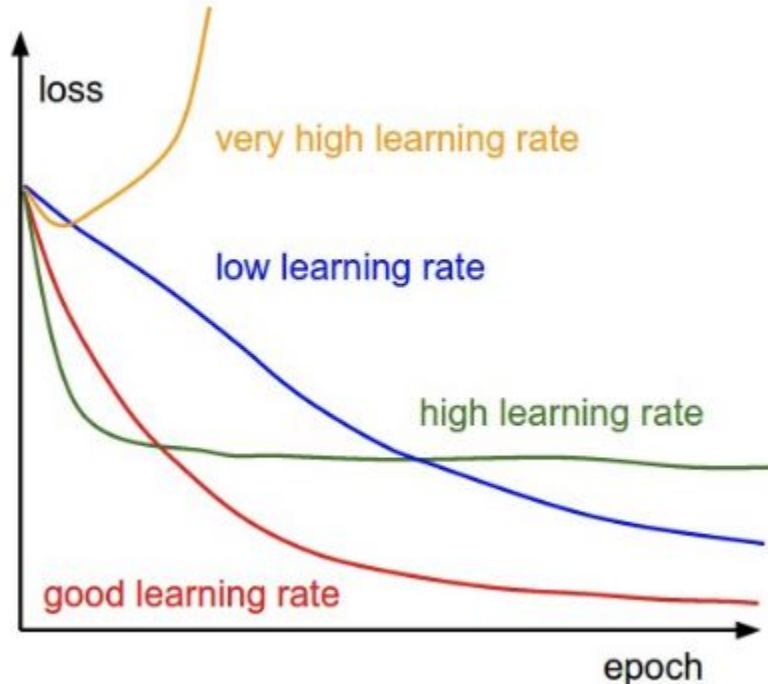
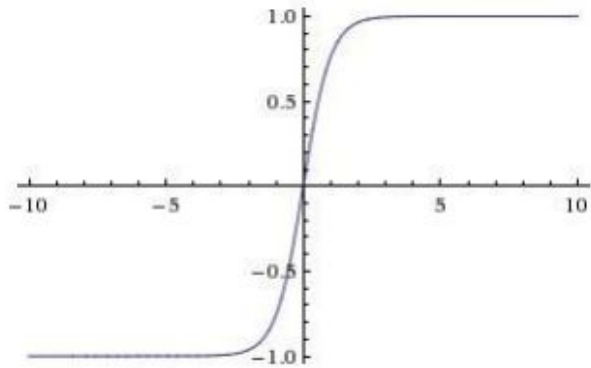$$f(a) = \max(0, a)$$

$$f(a) = \log(1 + e^a)$$
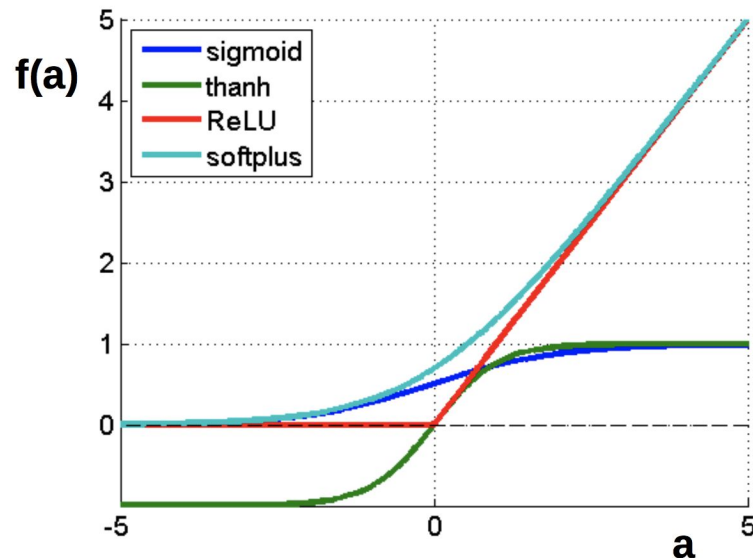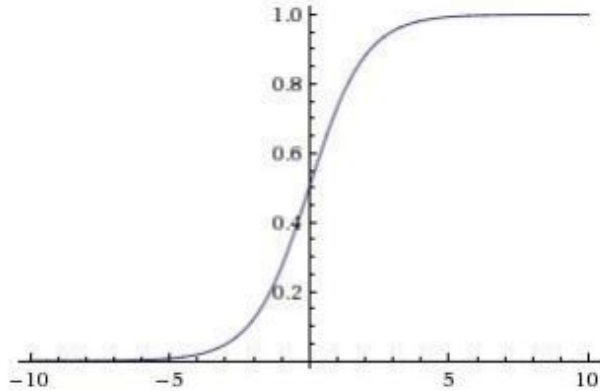
Sigmoid

$$f(a) = \frac{1}{1 + e^a}$$

- Squashes numbers to range [0,1]
- Historically popular since they have nice interpretation as a saturating "firing rate" of a neuron

3 problems:

1. Saturated neurons "kill" the gradients
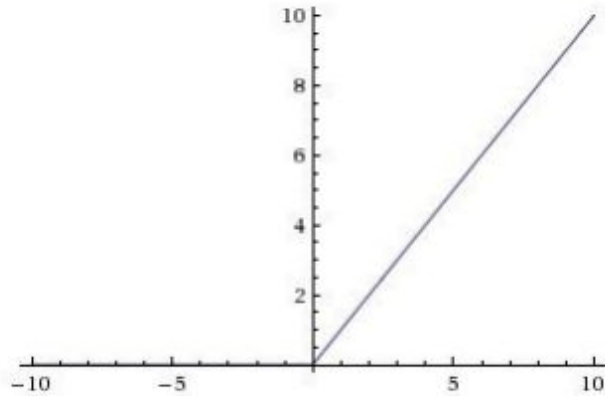2. Sigmoid outputs are not zero-centered
3. exp() is a bit compute expensive

48

tanh(x)

- Squashes numbers to range [-1,1]
- zero centered (nice)
- still kills gradients when saturated :(
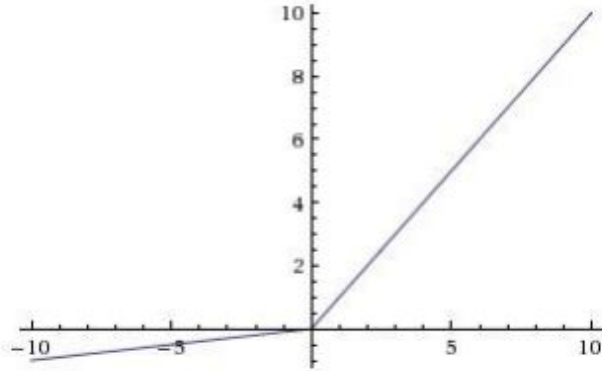
$$f(a) = \tanh(a)$$

ReLU
(Rectified Linear Unit)

$$f(a) = \max(0, a)$$

- Does not saturate (in +region)
- Very computationally efficient
- Converges much faster than sigmoid/tanh in practice (e.g. 6x)

- Not zero-centered output
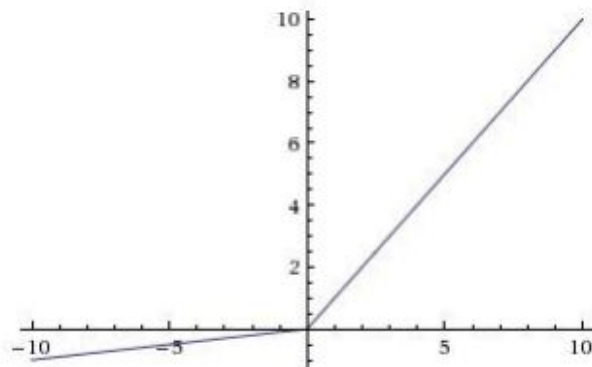- An annoyance:

hint: what is the gradient when x < 0?

- Does not saturate
- Computationally efficient
- Converges much faster than sigmoid/tanh in practice! (e.g. 6x)
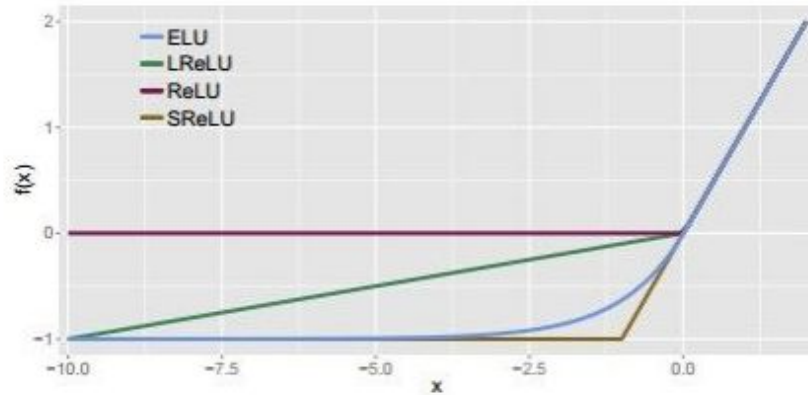- **will not "die".**

Leaky ReLU

$$f(x) = \max(0.01x, x)$$

- Does not saturate
- Computationally efficient
- Converges much faster than sigmoid/tanh in practice! (e.g. 6x)
- **will not "die".**

Leaky ReLU

$$f(x) = \max(0.01x, x)$$

Parametric Rectifier (PReLU)

$$f(x) = \max(\alpha x, x)$$

backprop into \alpha (parameter)

Exponential Linear Units (ELU)

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha\,(\exp(x) - 1) & \text{if } x \leq 0 \end{cases}$$

- All benefits of ReLU
- Does not die
- Closer to zero mean outputs

- Computation requires exp()

Activation functions: sum up

- Use ReLU as baseline approach
- Be careful with the learning rates
- Try out Leaky ReLU or ELU
- Try out tanh but do not expect much from it
- Do not use Sigmoid

Practice time: interactive playground

https://playground.tensorflow.org/
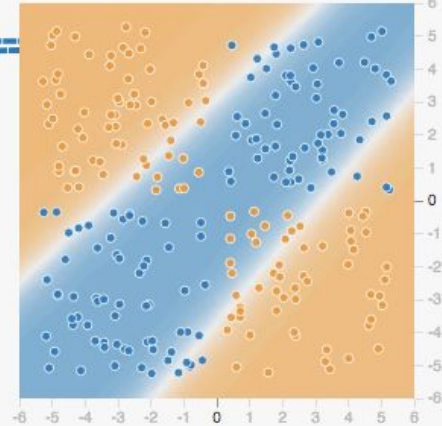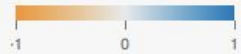
That's all. Time to build some NN.