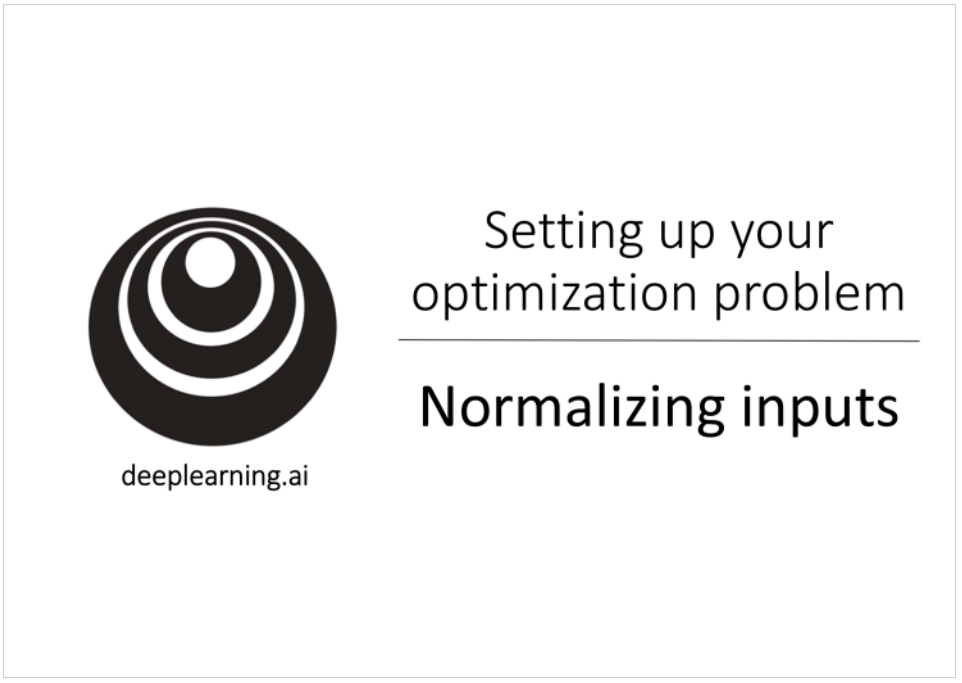
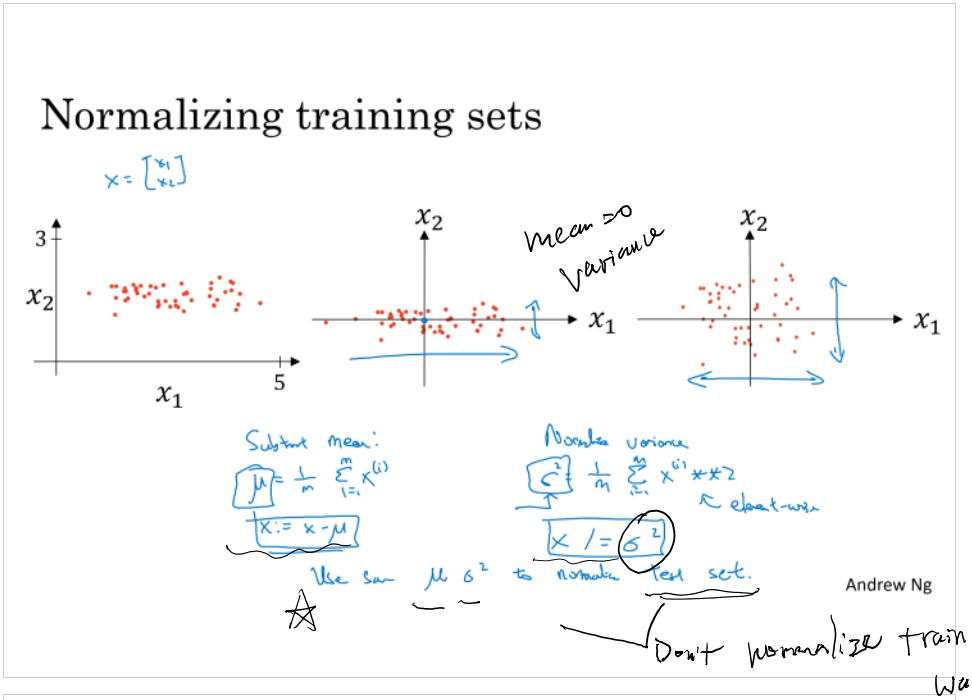
Normalizing input

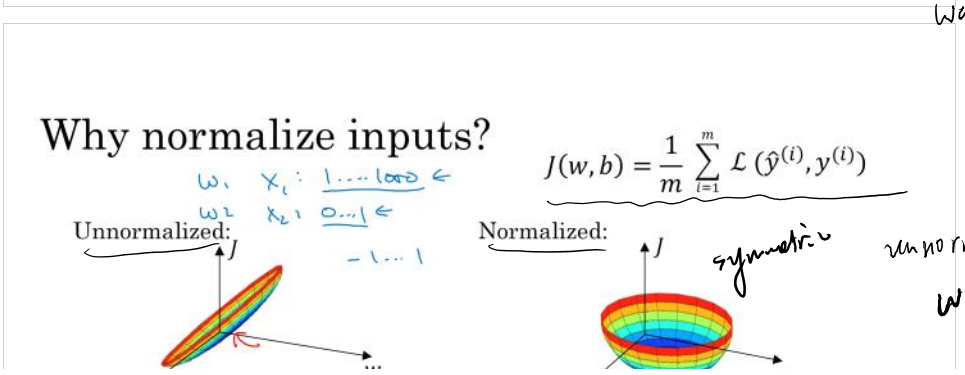
Friday, June 1, 2018 2:35 PM

normalize input



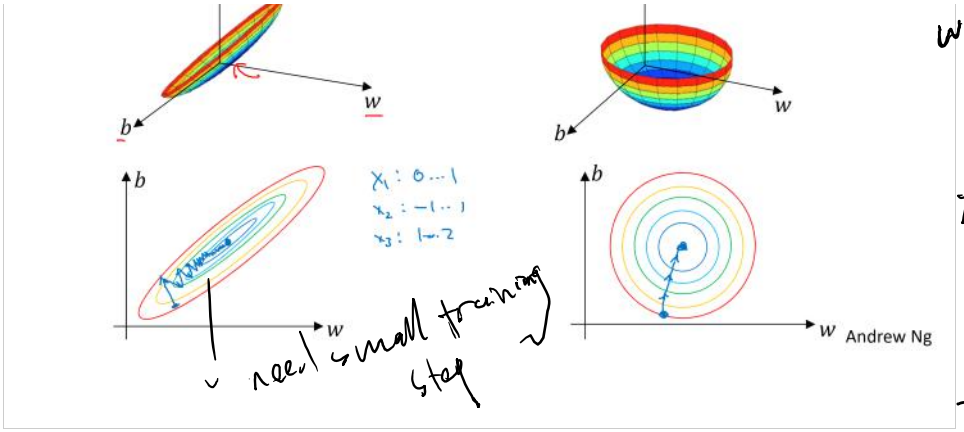


Don't normalize training & test set differently
Want same transformation



symmetric

unnormalize + feature of different scales
w_1, w_2 different value
elongated bowl



w_1, w_2 different value

elongated bowl

In pratice, w high dim

this two-dim plot shows the intuition

if feature similar scale, won't be important but this won't do any harm

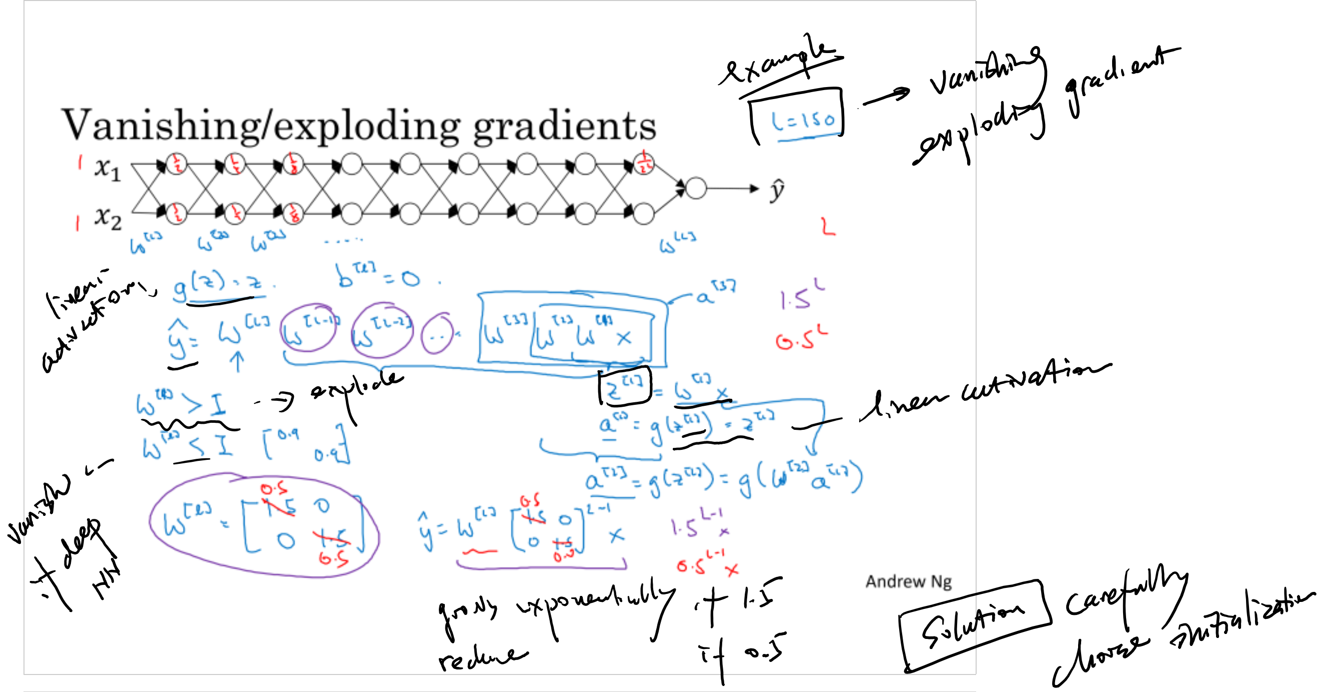
vanish exploding gradient

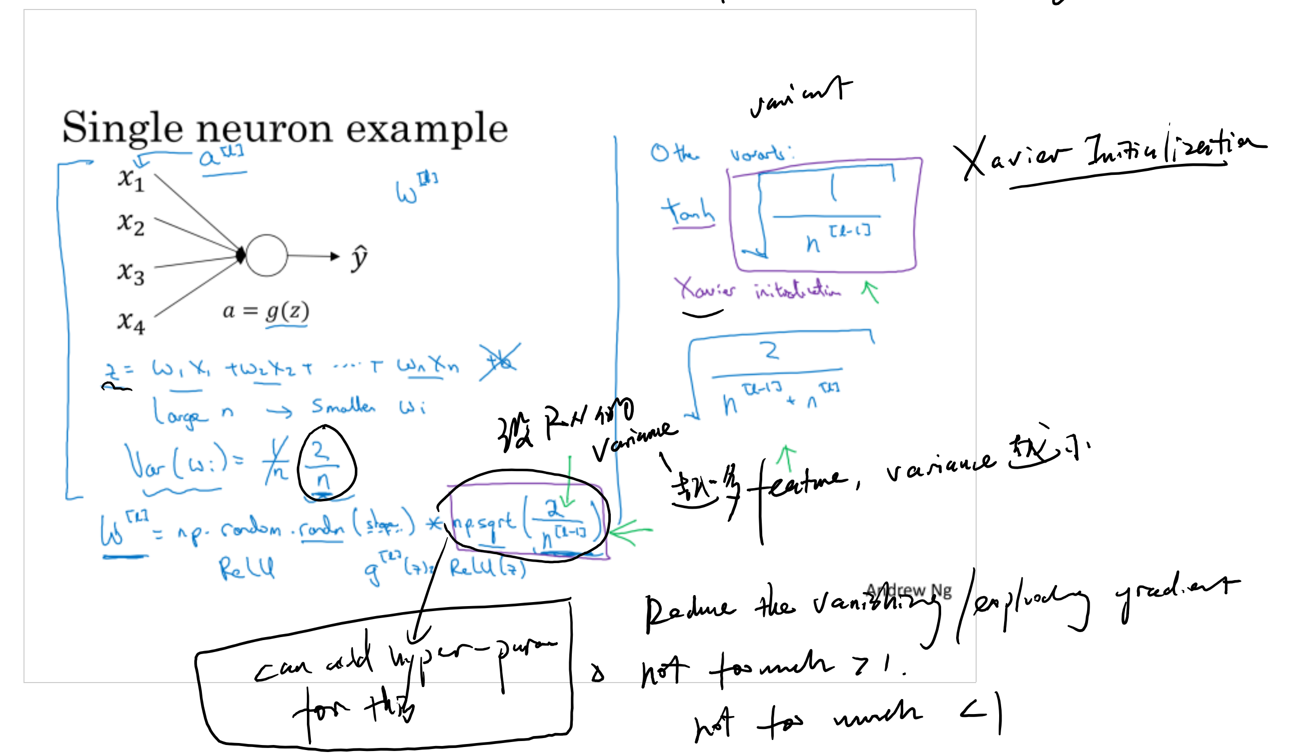
one big problem

deeplearning.ai

Setting up your optimization problem

Vanishing/exploding gradients





numerical approx

deeplearning.ai

Setting up your optimization problem

Numerical approximation of gradients

Checking your derivative computation

$f(\theta) = \theta^3$

$\theta \in \mathbb{R}$.

$g(\theta) = \frac{d}{d\theta} f(\theta) = f'(\theta)$

$g(\theta) = 3\theta^2$.

$g(\theta) = 3 \cdot (1)^2 = 3$ when $\theta = 1$

$\frac{dJ}{db}$

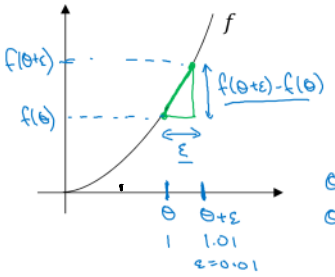
$\frac{f(\theta+\varepsilon) - f(\theta)}{\varepsilon} \approx g(\theta)$

$\frac{(1.01)^3 - 1^3}{0.01} = 3.0301 \approx 3$

0.0301

3.1

3.2

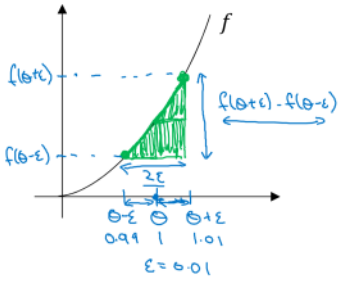


$\theta = 1$

$\theta + \varepsilon = 1.01$

Andrew Ng

Checking your derivative computation



$f(\theta) = \theta^3$

numerical approximation

Just calculus.

$\frac{f(\theta+\varepsilon) - f(\theta-\varepsilon)}{2\varepsilon} \approx g(\theta)$

$\frac{(1.01)^3 - (0.99)^3}{2(0.01)} = 3.0001 \approx 3$

$g(\theta) = 3\theta^2 = 3$

Approx error: 0.0001

(previous slide: 3.0301. error: 0.03)

$f'(\theta) = \lim_{\varepsilon \to 0} \frac{f(\theta+\varepsilon) - f(\theta-\varepsilon)}{2\varepsilon}$ $O(\varepsilon^2)$ 0.01 0.0001

$\frac{f(\theta+\varepsilon) - f(\theta)}{\varepsilon}$ error: $O(\varepsilon)$ 0.01

Two sided.

for gradient checking

We would rather use the two-sided difference

Andrew Ng

gradient check

deeplearning.ai

Setting up your optimization problem

Gradient Checking

Gradient check for a neural network

Take $W^{[1]}, b^{[1]}, \dots, W^{[L]}, b^{[L]}$ and reshape into a big vector θ.

concatenate

$J(W^{[1]}, b^{[1]}, \dots, W^{[L]}, b^{[L]}) = J(\theta)$

Take $dW^{[1]}, db^{[1]}, \dots, dW^{[L]}, db^{[L]}$ and reshape into a big vector $d\theta$.

concatenate

Is $d\theta$ the gradient of $J(\theta)$?

Andrew Ng

Gradient checking (Grad check)

$J(\theta) = J(\theta_1, \theta_2, \theta_3, \dots)$

why ε

for each i:

$$\to d\theta_{approx}[i] = \frac{J(\theta_1, \theta_2, \dots, \theta_i + \varepsilon, \dots) - J(\theta_1, \theta_2, \dots, \theta_i - \varepsilon, \dots)}{2\varepsilon}$$

$$\approx d\theta[i] = \frac{\partial J}{\partial \theta_i}$$

Where to get? still the prog thing?

$d\theta_{approx} \overset{?}{\approx} d\theta$

How to define whether two vectors are close, Euclidean distance

Check

$$\frac{\| d\theta_{approx} - d\theta \|_2}{\| d\theta_{approx} \|_2 + \| d\theta \|_2} \approx 10^{-7} \text{ - great!} \leftarrow$$

10^{-5}

$\to 10^{-3}$ - worry. $\leftarrow$

$\varepsilon = 10^{-7}$

normalize it into a ratio.

$\to$ if big value need debug

Andrew Ng

Gradient checking implementation notes

- Don't use in training – only to debug

- If algorithm fails grad check, look at components to try to identify bug.

- Remember regularization.

- Doesn't work with dropout.

- Run at random initialization; perhaps again after some training.

Andrew Ng