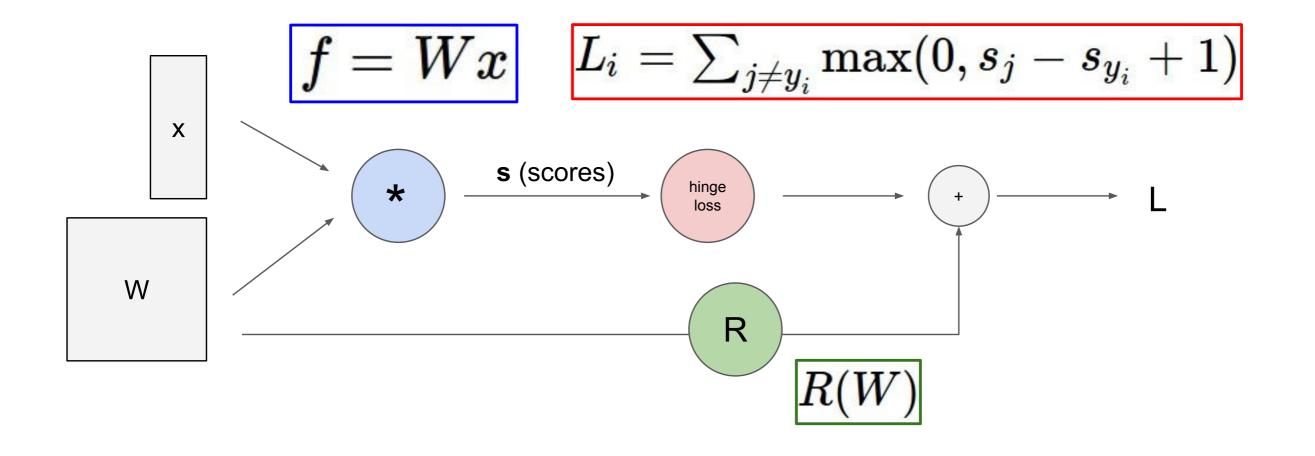
# Training Neural Networks

Slides adapted from http://cs231n.stanford.edu/

Gerard Pons-Moll

#### **Computational graphs**

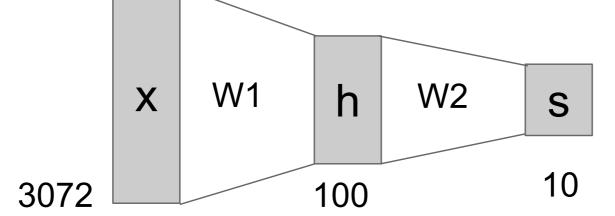


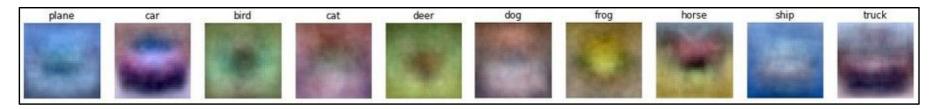
### **Neural Networks**

Linear score function:

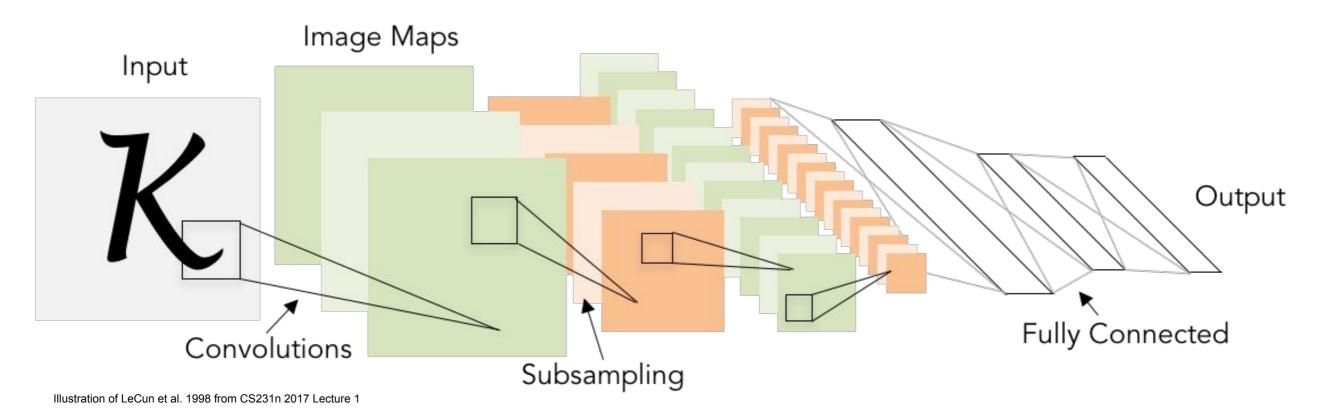
2-layer Neural Network

$$egin{aligned} f &= Wx \ f &= W_2 \max(0, W_1 x) \end{aligned}$$

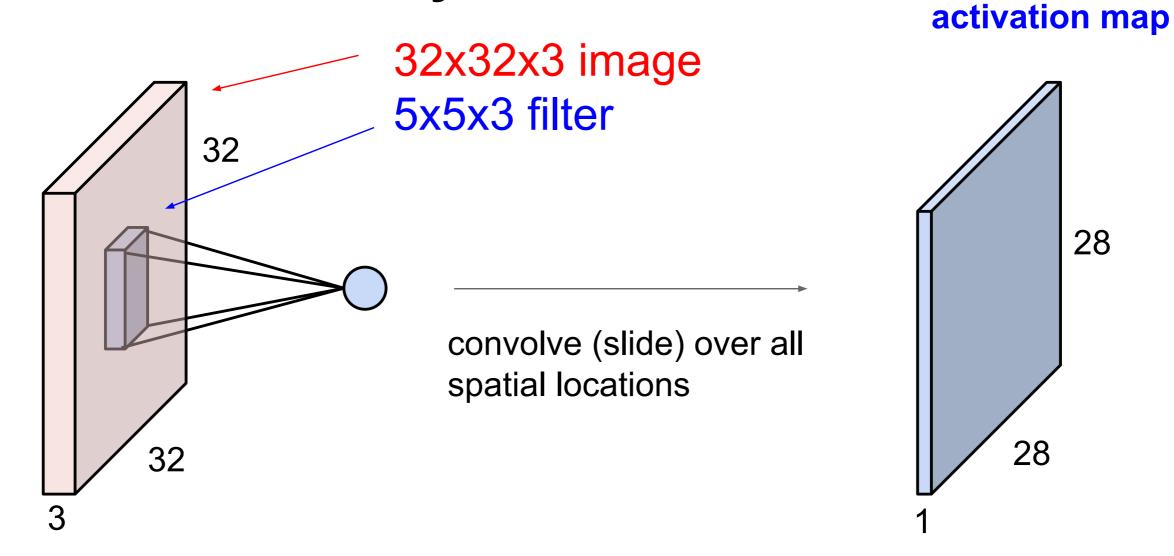




#### **Convolutional Neural Networks**



#### **Convolutional Layer**

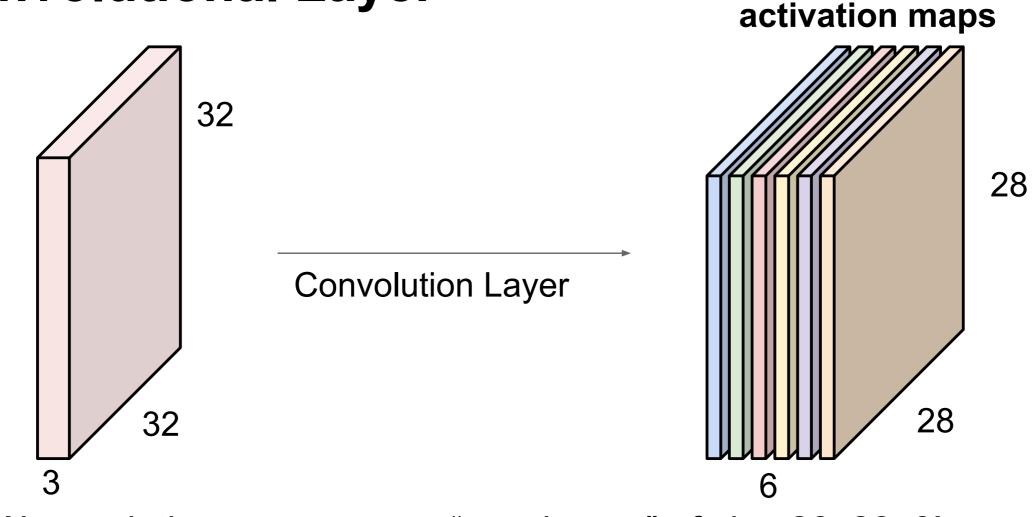


28

28

#### Where we are now... Convolutional Layer

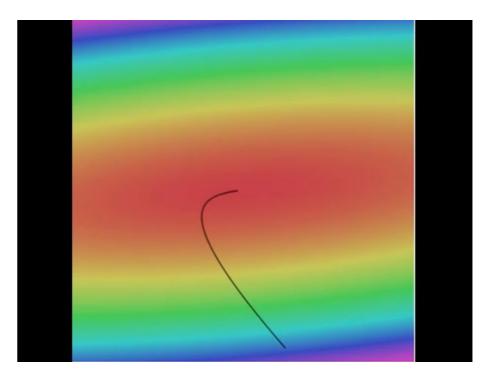
For example, if we had 6 5x5 filters, we'll get 6 separate activation maps:



We stack these up to get a "new image" of size 28x28x6!

#### Learning network parameters through optimization





# Vanilla Gradient Descent

while True:

Landscape image is <u>CC0 1.0</u> public domain Walking man image is <u>CC0 1.0</u> public domain weights\_grad = evaluate\_gradient(loss\_fun, data, weights)
weights += - step\_size \* weights\_grad # perform parameter update

# Mini-batch SGD

Loop:

- 1. Sample a batch of data
- 2. **Forward** prop it through the graph (network), get loss
- 3. Backprop to calculate the gradients
- 4. Update the parameters using the gradient

# Next: Training Neural Networks

# Overview

#### 1. One time setup

activation functions, preprocessing, weight initialization, regularization, gradient checking

#### 2. Training dynamics

babysitting the learning process,

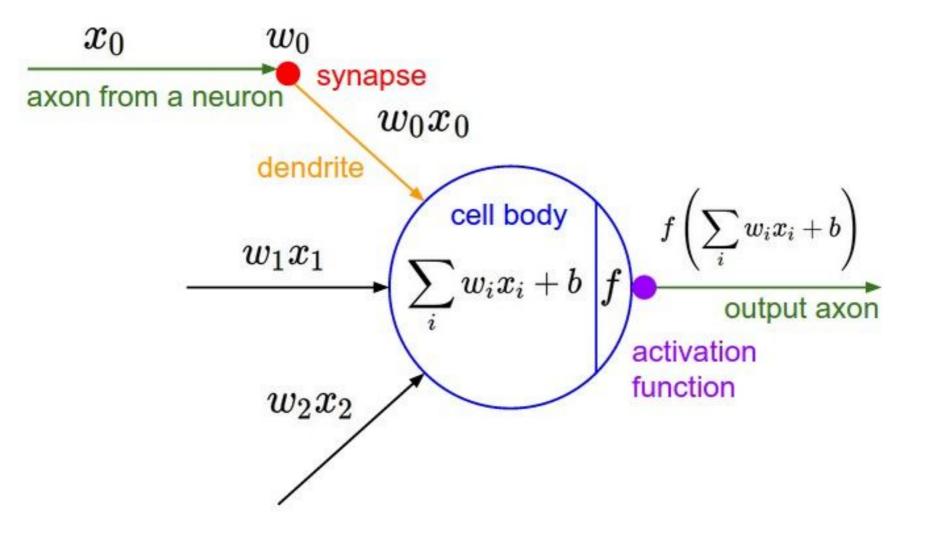
parameter updates, hyperparameter optimization

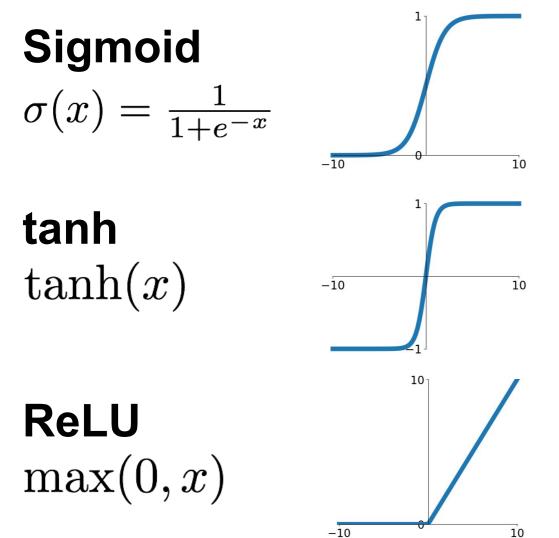
#### 3. Evaluation

model ensembles

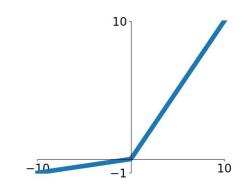
# Part 1

- Activation Functions
- Data Preprocessing
- Weight Initialization
- Batch Normalization
- Babysitting the Learning Process
- Hyperparameter Optimization

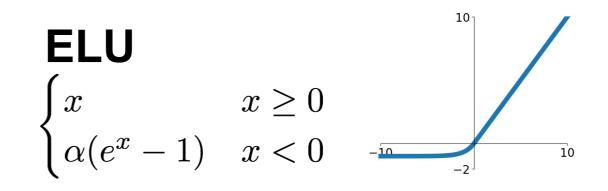


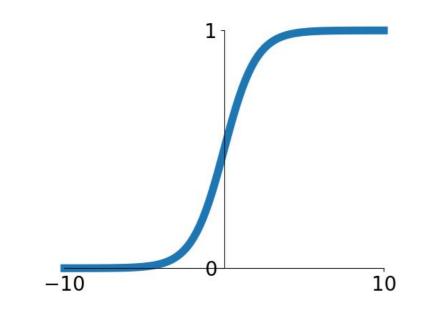


# Leaky ReLU $\max(0.1x, x)$



 $\begin{array}{l} \textbf{Maxout} \\ \max(w_1^T x + b_1, w_2^T x + b_2) \end{array}$ 

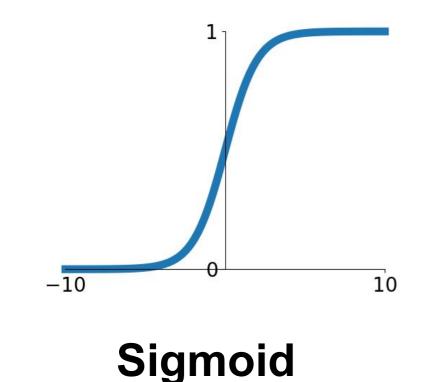




Sigmoid

$$\sigma(x) = 1/(1+e^{-x})$$

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

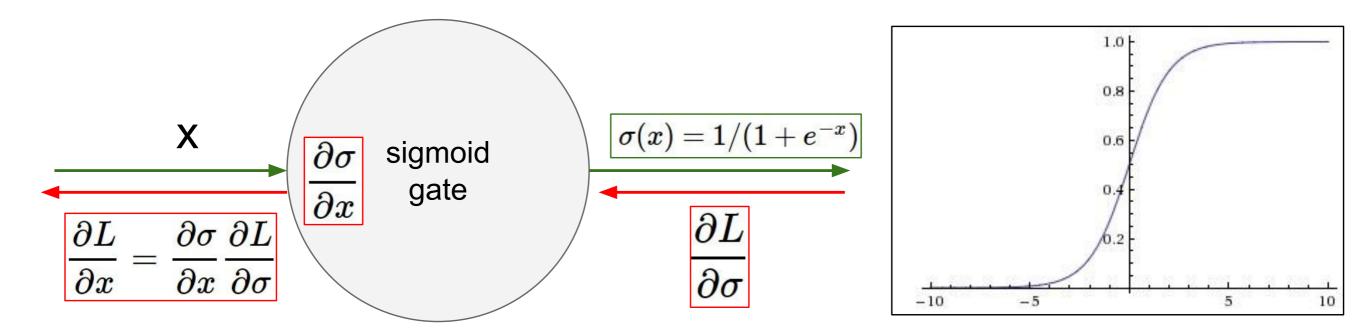


$$\sigma(x)=1/(1+e^{-x})$$

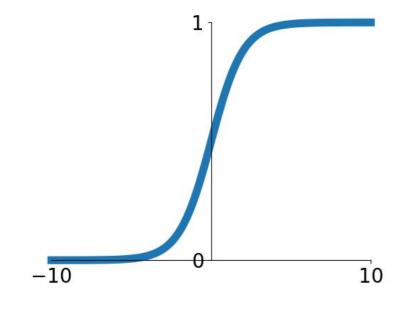
- Squashes numbers to range [0,1]
- Historically popular since they have nice interpretation as a saturating "firing rate" of a neuron

3 problems:

1. Saturated neurons "kill" the gradients



What happens when x = -10? What happens when x = 0? What happens when x = 10?



Sigmoid

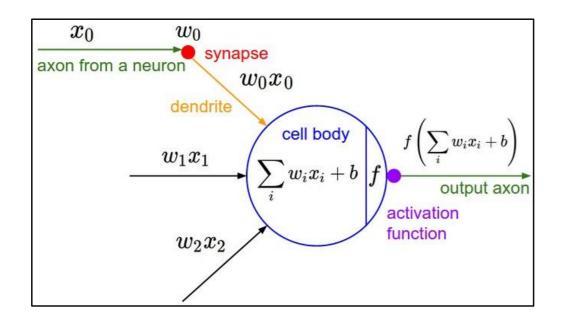
 $\sigma(x) = 1/(1+e^{-x})$ 

- 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

Consider what happens when the input to a neuron (x) is always positive:



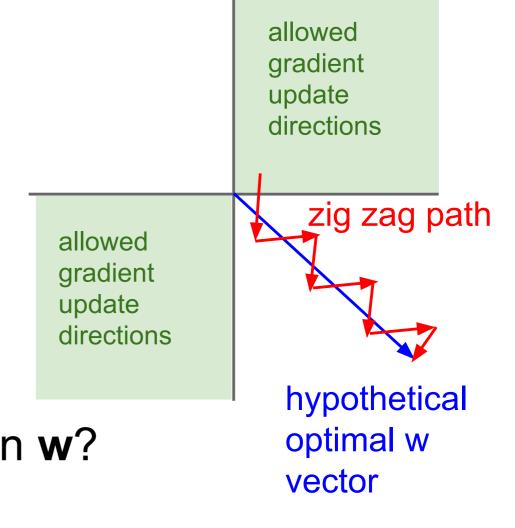
 $f\left(\sum_{i} w_i x_i + b\right)$ 

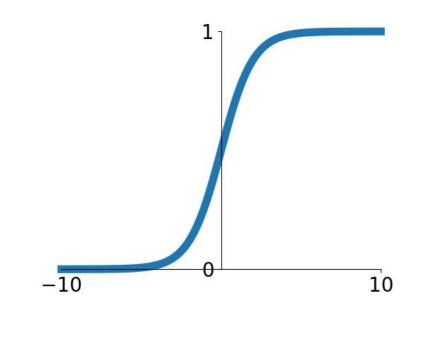
What can we say about the gradients on **w**?

Consider what happens when the input to a neuron is always positive...

$$f\left(\sum_{i} w_{i}x_{i} + b\right)$$

What can we say about the gradients on w? Always all positive or all negative :( (this is also why you want zero-mean data!)





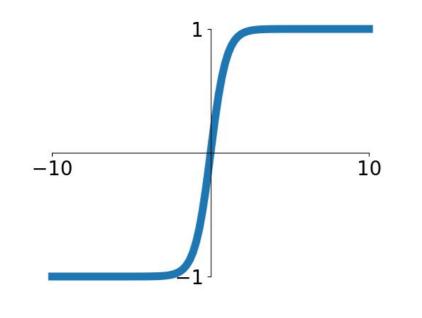
Sigmoid

 $\sigma(x) = 1/(1 + e^{-x})$ 

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

3 problems:

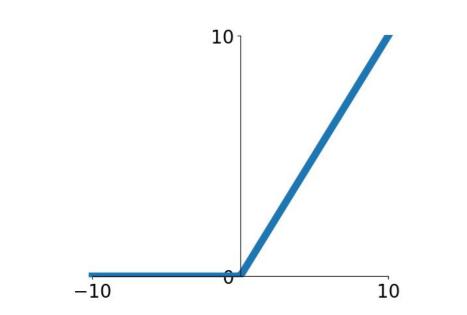
- 1. Saturated neurons "kill" the gradients
- 2. Sigmoid outputs are not zero-centered
- 3. exp() is a bit compute expensive



tanh(x)

- Squashes numbers to range [-1,1]
- zero centered (nice)
- still kills gradients when saturated :(

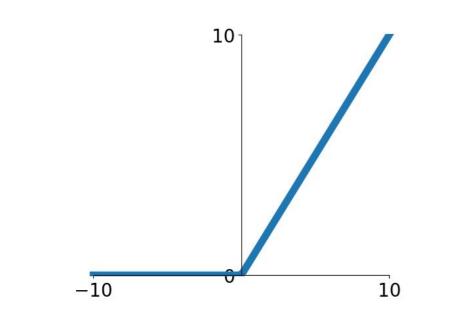
[LeCun et al., 1991]



#### **ReLU** (Rectified Linear Unit)

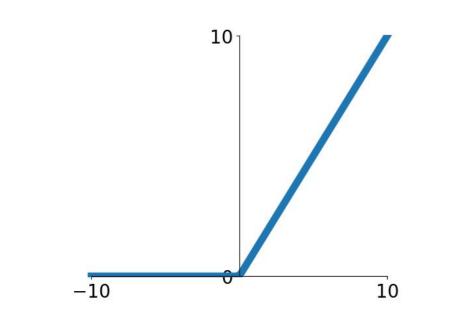
- Computes f(x) = max(0,x)
- Does not saturate (in +region)
- Very computationally efficient
- Converges much faster than sigmoid/tanh in practice (e.g. 6x)
- Actually more biologically plausible than sigmoid

[Krizhevsky et al., 2012]



**ReLU** (Rectified Linear Unit)

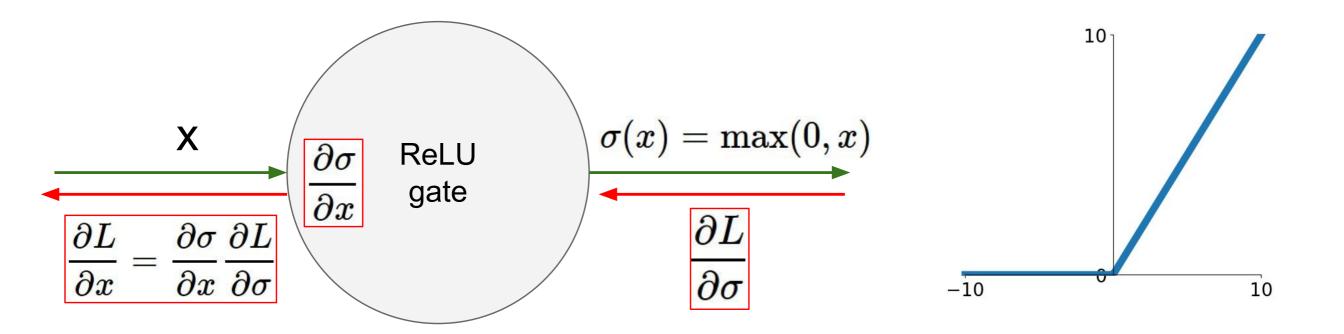
- Computes f(x) = max(0,x)
- Does not saturate (in +region)
- Very computationally efficient
- Converges much faster than sigmoid/tanh in practice (e.g. 6x)
- Actually more biologically plausible than sigmoid
- Not zero-centered output



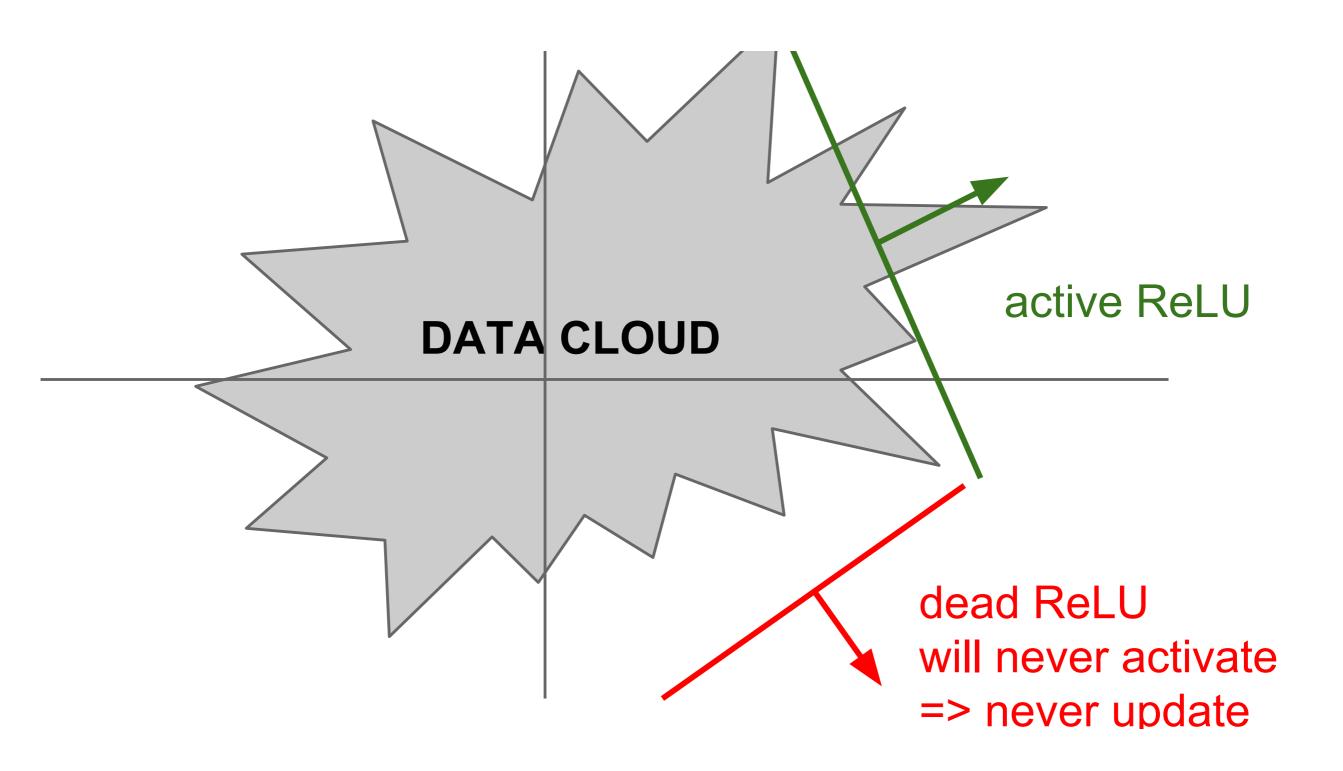
**ReLU** (Rectified Linear Unit)

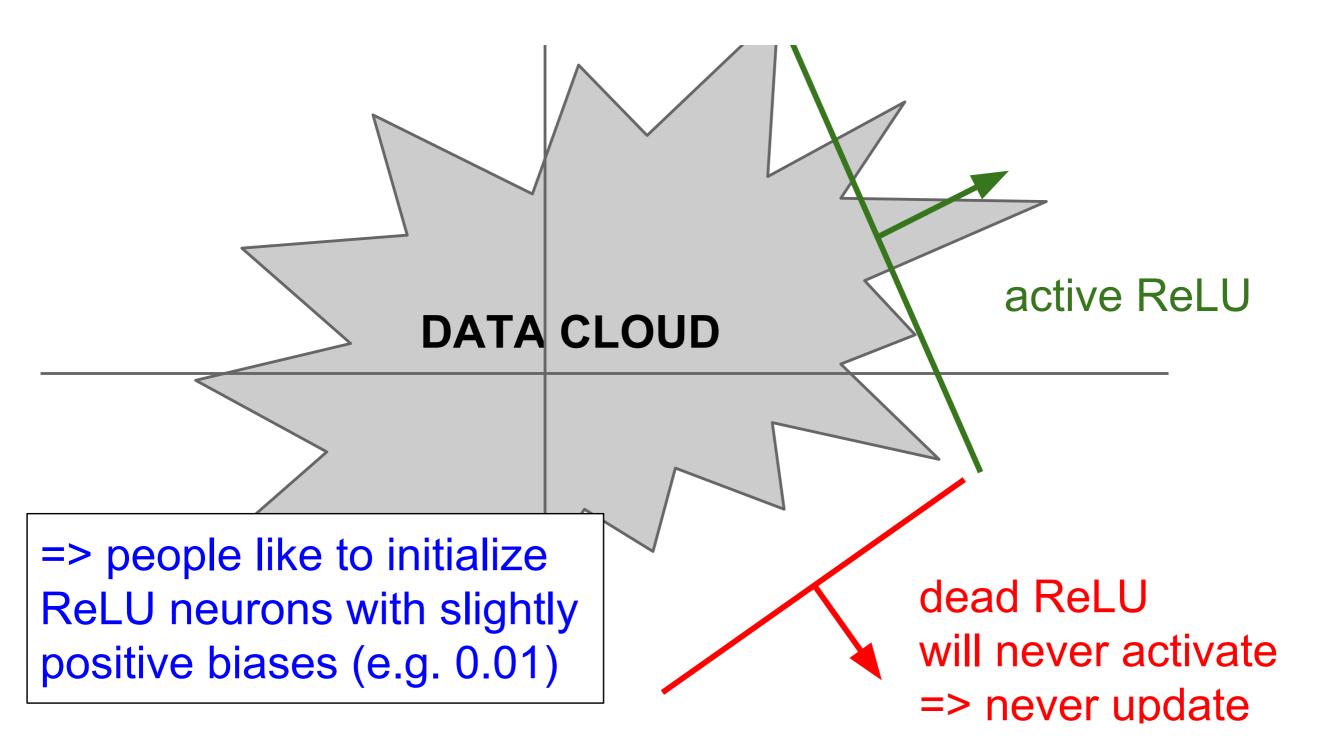
- Computes f(x) = max(0,x)
- Does not saturate (in +region)
- Very computationally efficient
- Converges much faster than sigmoid/tanh in practice (e.g. 6x)
- Actually more biologically plausible than sigmoid
- Not zero-centered output
- An annoyance:

hint: what is the gradient when x < 0?



What happens when x = -10? What happens when x = 0? What happens when x = 10?





10

[Mass et al., 2013] [He et al., 2015]

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

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

10

10

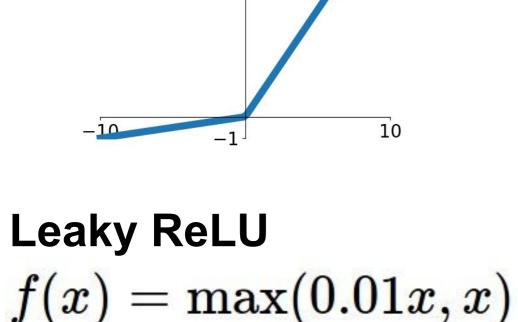
[Mass et al., 2013] [He et al., 2015]

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

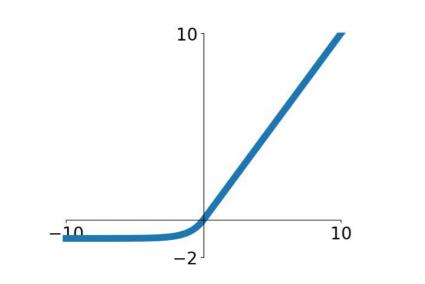
#### **Parametric Rectifier (PReLU)**

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

backprop into \alpha / (parameter)



#### **Exponential Linear Units (ELU)**



- All benefits of ReLU
- Closer to zero mean outputs
- Negative saturation regime compared with Leaky ReLU adds some robustness to noise

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

- Computation requires exp()

## Maxout "Neuron"

- Does not have the basic form of dot product -> nonlinearity
- Generalizes ReLU and Leaky ReLU
- Linear Regime! Does not saturate! Does not die!

$$\max(w_1^Tx+b_1,w_2^Tx+b_2)$$

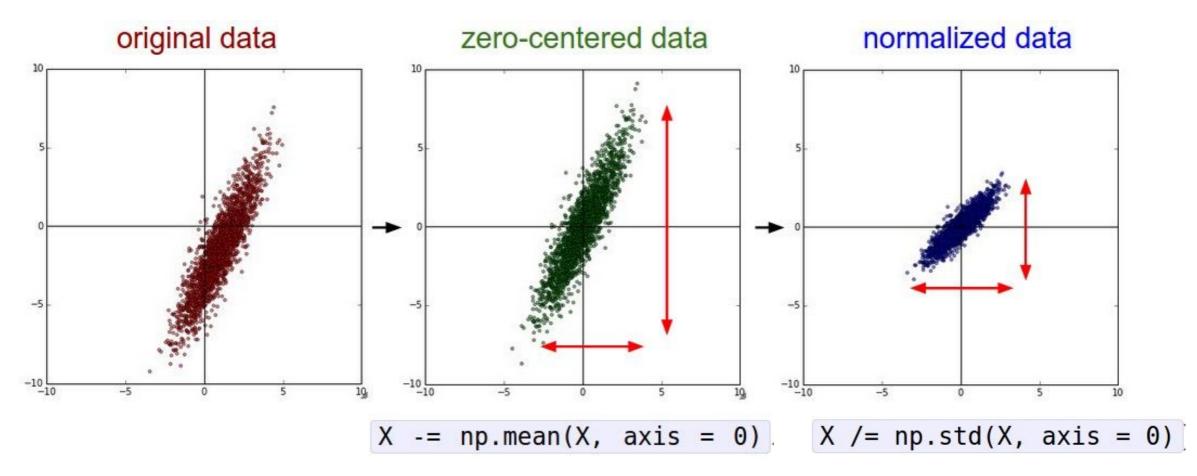
Problem: doubles the number of parameters/neuron :(

**TLDR: In practice:** 

- Use ReLU. Be careful with your learning rates
- Try out Leaky ReLU / Maxout / ELU
- Try out tanh but don't expect much
- Don't use sigmoid

# Data Preprocessing

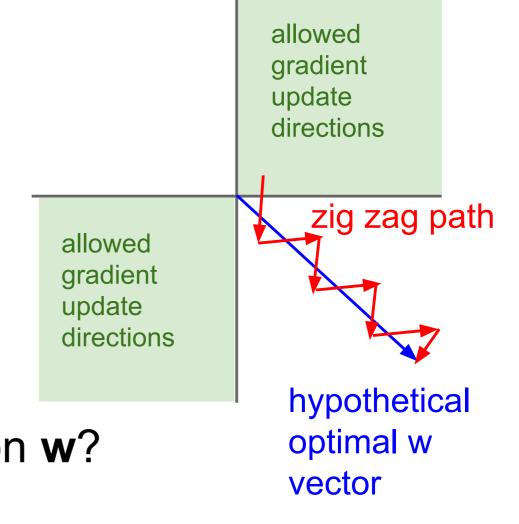
## Step 1: Preprocess the data



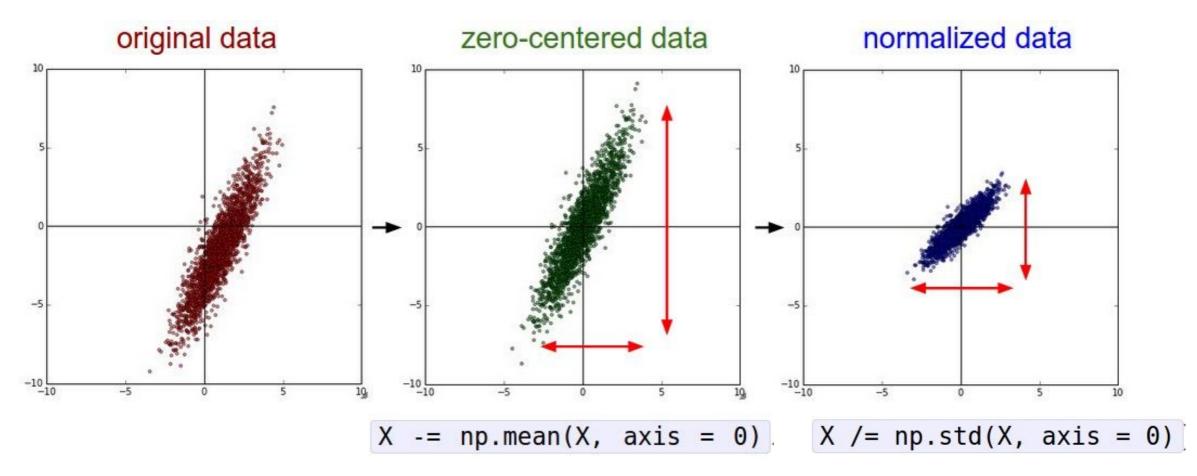
(Assume X [NxD] is data matrix, each example in a row) Remember: Consider what happens when the input to a neuron is always positive...

$$f\left(\sum_{i} w_{i}x_{i} + b\right)$$

What can we say about the gradients on w? Always all positive or all negative :( (this is also why you want zero-mean data!)



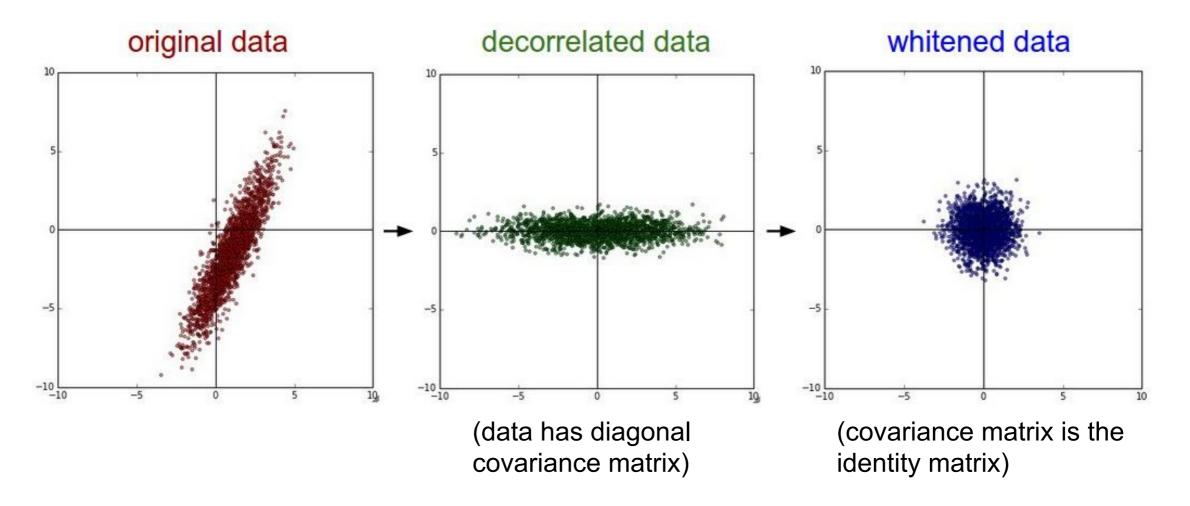
### Step 1: Preprocess the data



(Assume X [NxD] is data matrix, each example in a row)

#### Step 1: Preprocess the data

#### In practice, you may also see PCA and Whitening of the data



#### TLDR: In practice for Images: center only

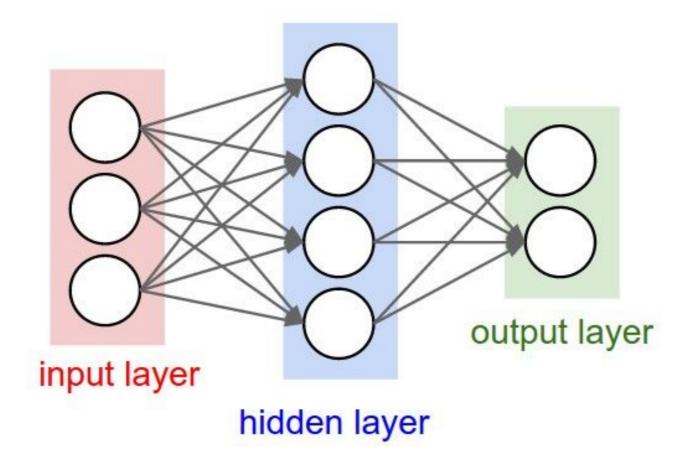
e.g. consider CIFAR-10 example with [32,32,3] images

- Subtract the mean image (e.g. AlexNet) (mean image = [32,32,3] array)
- Subtract per-channel mean (e.g. VGGNet) (mean along each channel = 3 numbers)

Not common to normalize variance, to do PCA or whitening

## Weight Initialization

- Q: what happens when W=constant init is used?



#### - First idea: Small random numbers

(gaussian with zero mean and 1e-2 standard deviation)

W = 0.01\* np.random.randn(D,H)

#### - First idea: Small random numbers

(gaussian with zero mean and 1e-2 standard deviation)

W = 0.01\* np.random.randn(D,H)

Works ~okay for small networks, but problems with deeper networks.

Lets look at some activation statistics

E.g. 10-layer net with 500 neurons on each layer, using tanh non-linearities, and initializing as described in last slide.

```
# assume some unit gaussian 10-D input data
D = np.random.randn(1000, 500)
hidden_layer_sizes = [500]*10
nonlinearities = ['tanh']*len(hidden layer sizes)
```

```
act = {'relu':lambda x:np.maximum(0,x), 'tanh':lambda x:np.tanh(x)}
Hs = {}
for i in xrange(len(hidden_layer_sizes)):
    X = D if i == 0 else Hs[i-1] # input at this layer
    fan_in = X.shape[1]
    fan_out = hidden_layer_sizes[i]
    W = np.random.randn(fan_in, fan_out) * 0.01 # layer initialization
```

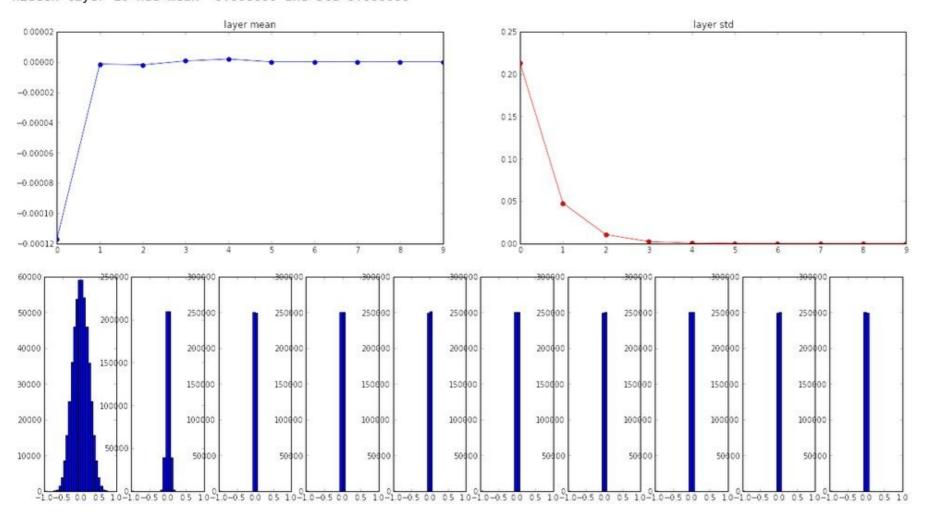
```
H = np.dot(X, W) # matrix multiply
H = act[nonlinearities[i]](H) # nonlinearity
Hs[i] = H # cache result on this layer
```

```
# look at distributions at each layer
print 'input layer had mean %f and std %f' % (np.mean(D), np.std(D))
layer_means = [np.mean(H) for i,H in Hs.iteritems()]
layer_stds = [np.std(H) for i,H in Hs.iteritems()]
for i,H in Hs.iteritems():
    print 'hidden layer %d had mean %f and std %f' % (i+1, layer_means[i], layer_stds[i])
```

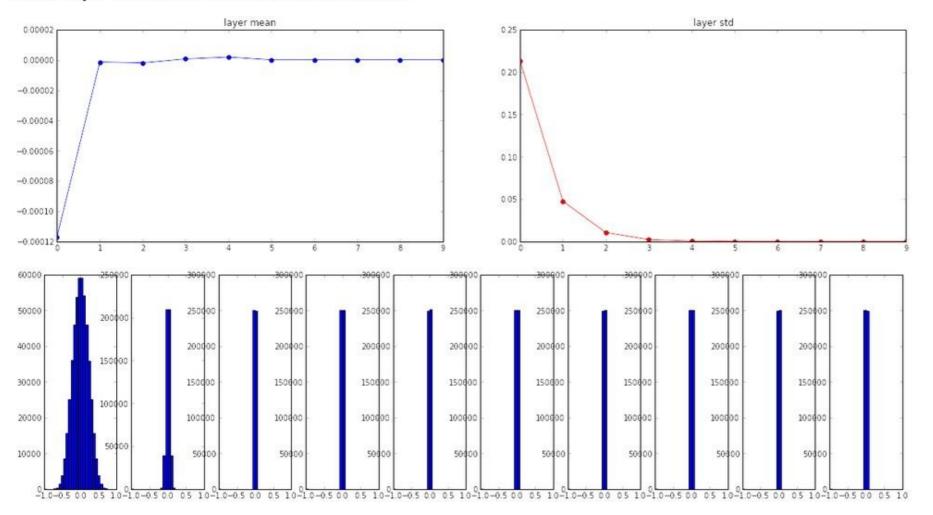
```
# plot the means and standard deviations
plt.figure()
plt.subplot(121)
plt.plot(Hs.keys(), layer_means, 'ob-')
plt.title('layer mean')
plt.subplot(122)
plt.plot(Hs.keys(), layer_stds, 'or-')
plt.title('layer std')
```

```
# plot the raw distributions
plt.figure()
for i,H in Hs.iteritems():
    plt.subplot(1,len(Hs),i+1)
    plt.hist(H.ravel(), 30, range=(-1,1))
```

input layer had mean 0.000927 and std 0.998388 hidden layer 1 had mean -0.000117 and std 0.213081 hidden layer 2 had mean -0.000001 and std 0.047551 hidden layer 3 had mean -0.000002 and std 0.010630 hidden layer 4 had mean 0.000001 and std 0.002378 hidden layer 5 had mean 0.000002 and std 0.000532 hidden layer 6 had mean -0.000000 and std 0.000119 hidden layer 7 had mean 0.000000 and std 0.000026 hidden layer 8 had mean -0.000000 and std 0.000006 hidden layer 9 had mean 0.000000 and std 0.000001 hidden layer 10 had mean -0.000000 and std 0.000000



input layer had mean 0.000927 and std 0.998388 hidden layer 1 had mean -0.000117 and std 0.213081 hidden layer 2 had mean -0.000001 and std 0.047551 hidden layer 3 had mean -0.000002 and std 0.010630 hidden layer 4 had mean 0.000001 and std 0.002378 hidden layer 5 had mean 0.000002 and std 0.000532 hidden layer 6 had mean -0.000000 and std 0.000119 hidden layer 7 had mean 0.000000 and std 0.000026 hidden layer 8 had mean -0.000000 and std 0.000006 hidden layer 9 had mean 0.000000 and std 0.000001 hidden layer 10 had mean -0.000000 and std 0.000000



# All activations become zero!

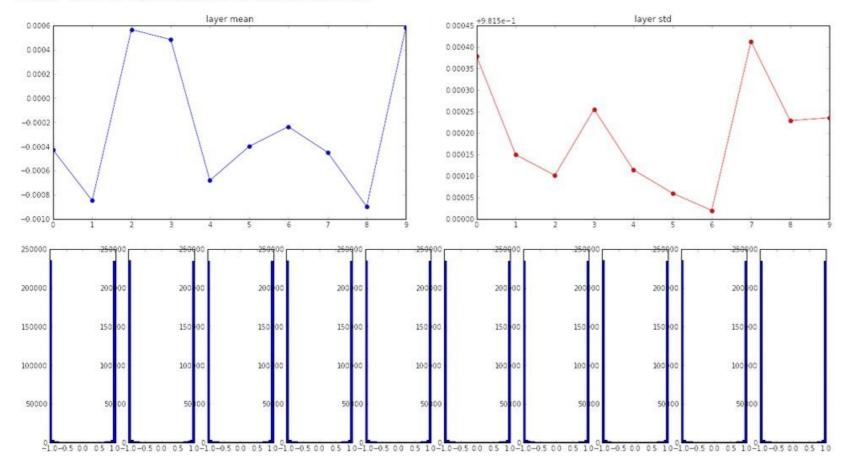
Q: think about the backward pass. What do the gradients look like?

Hint: think about backward pass for a W\*X gate.

#### W = np.random.randn(fan in, fan out) \* 1.0 # layer initialization

input layer had mean 0.001800 and std 1.001311 hidden layer 1 had mean -0.000430 and std 0.981879 hidden layer 2 had mean -0.000849 and std 0.981649 hidden layer 3 had mean 0.000566 and std 0.981601 hidden layer 4 had mean 0.000483 and std 0.981755 hidden layer 5 had mean -0.000682 and std 0.981614 hidden layer 6 had mean -0.000401 and std 0.981560 hidden layer 7 had mean -0.000237 and std 0.981520 hidden layer 8 had mean -0.000448 and std 0.981913 hidden layer 9 had mean -0.000899 and std 0.981728 hidden layer 10 had mean 0.000584 and std 0.981736

## \*1.0 instead of \*0.01



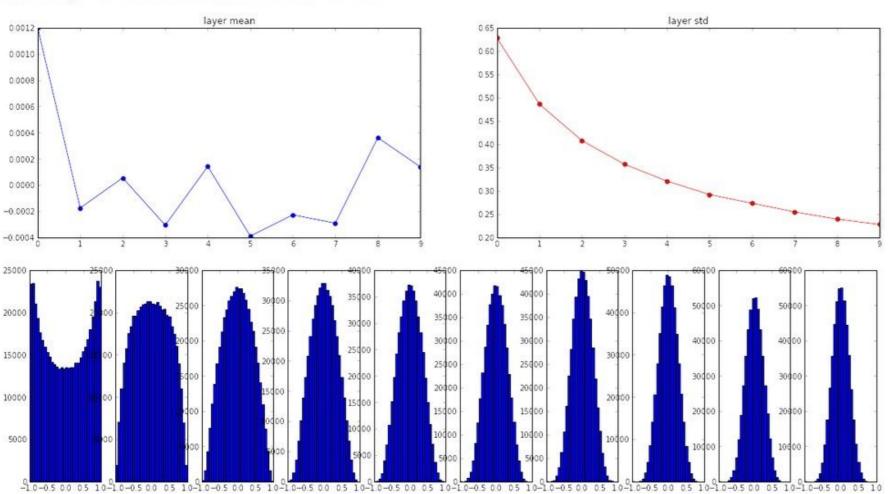
Almost all neurons completely saturated, either -1 and 1. Gradients will be all zero. W = np.random.randn(fan\_in, fan\_out) / np.sqrt(fan\_in) # layer initialization

#### "Xavier initialization" [Glorot et al., 2010]

#### **Reasonable initialization.**

(Mathematical derivation assumes linear activations)

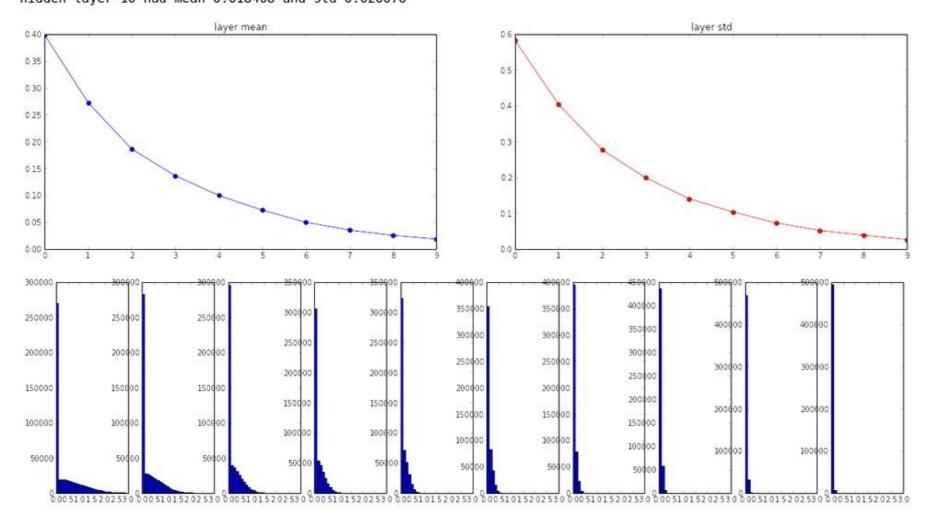
input layer had mean 0.001800 and std 1.001311 hidden layer 1 had mean 0.001198 and std 0.627953 hidden layer 2 had mean -0.000175 and std 0.486051 hidden layer 3 had mean 0.000055 and std 0.407723 hidden layer 4 had mean -0.000306 and std 0.357108 hidden layer 5 had mean 0.000142 and std 0.320917 hidden layer 6 had mean -0.000389 and std 0.292116 hidden layer 7 had mean -0.000228 and std 0.273387 hidden layer 8 had mean -0.000291 and std 0.254935 hidden layer 9 had mean 0.000361 and std 0.239266 hidden layer 10 had mean 0.000139 and std 0.228008



W = np.random.randn(fan\_in, fan\_out) / np.sqrt(fan\_in) # layer initialization

## but when using the ReLU nonlinearity it breaks.

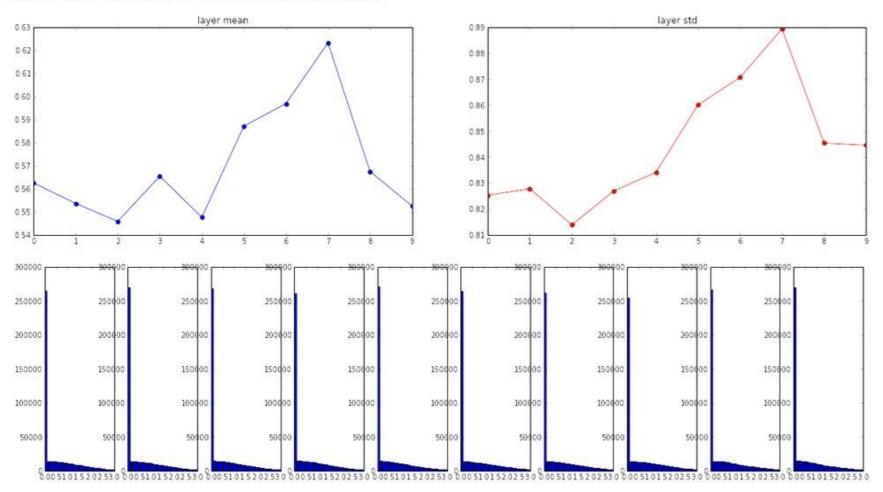
input layer had mean 0.000501 and std 0.999444 hidden layer 1 had mean 0.398623 and std 0.582273 hidden layer 2 had mean 0.272352 and std 0.403795 hidden layer 3 had mean 0.186076 and std 0.276912 hidden layer 4 had mean 0.136442 and std 0.198685 hidden layer 5 had mean 0.099568 and std 0.140299 hidden layer 6 had mean 0.072234 and std 0.103280 hidden layer 7 had mean 0.049775 and std 0.072748 hidden layer 8 had mean 0.035138 and std 0.051572 hidden layer 9 had mean 0.025404 and std 0.038583 hidden layer 10 had mean 0.018408 and std 0.026076



W = np.random.randn(fan\_in, fan\_out) / np.sqrt(2/fan\_in) # layer initialization

#### He et al., 2015 (note additional 2/)

input layer had mean 0.000501 and std 0.999444 hidden layer 1 had mean 0.562488 and std 0.825232 hidden layer 2 had mean 0.553614 and std 0.827835 hidden layer 3 had mean 0.545867 and std 0.813855 hidden layer 4 had mean 0.565396 and std 0.826902 hidden layer 5 had mean 0.547678 and std 0.834092 hidden layer 6 had mean 0.587103 and std 0.860035 hidden layer 7 had mean 0.596867 and std 0.870610 hidden layer 8 had mean 0.623214 and std 0.889348 hidden layer 9 had mean 0.567498 and std 0.845357 hidden layer 10 had mean 0.552531 and std 0.844523

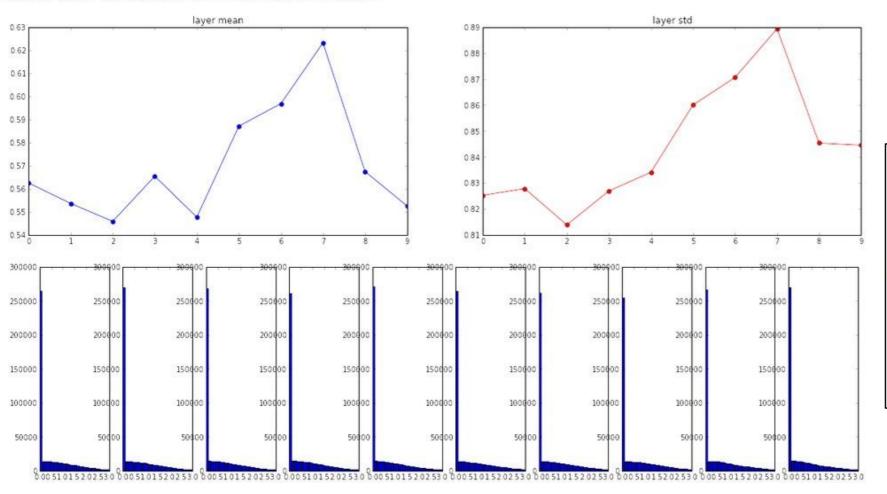


W = np.random.randn(fan\_in, fan\_out) / np.sqrt(2/fan\_in) # layer initialization

#### He et al., 2015 (note additional 2/)

 $\begin{array}{c} 0.95 \\ 0.96 \\ 0.85 \\ 0.85 \\ 0.86 \\ 0.86 \\ 0.86 \\ 0.86 \\ 0.75 \\ 0.75 \\ 0.75 \\ 0.75 \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ \hline \end{array} \right) \begin{array}{c} 0.95 \\ 0.86$ 

input layer had mean 0.000501 and std 0.999444 hidden layer 1 had mean 0.562488 and std 0.825232 hidden layer 2 had mean 0.553614 and std 0.827835 hidden layer 3 had mean 0.545867 and std 0.813855 hidden layer 4 had mean 0.565396 and std 0.826902 hidden layer 5 had mean 0.547678 and std 0.834092 hidden layer 6 had mean 0.587103 and std 0.860035 hidden layer 7 had mean 0.596867 and std 0.870610 hidden layer 8 had mean 0.623214 and std 0.889348 hidden layer 9 had mean 0.567498 and std 0.845357 hidden layer 10 had mean 0.552531 and std 0.844523



#### Proper initialization is an active area of research...

**Understanding the difficulty of training deep feedforward neural networks** by Glorot and Bengio, 2010

*Exact solutions to the nonlinear dynamics of learning in deep linear neural networks* by Saxe et al, 2013

*Random walk initialization for training very deep feedforward networks* by Sussillo and Abbott, 2014

Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification by He et al., 2015

Data-dependent Initializations of Convolutional Neural Networks by Krähenbühl et al., 2015

All you need is a good init, Mishkin and Matas, 2015

. . .

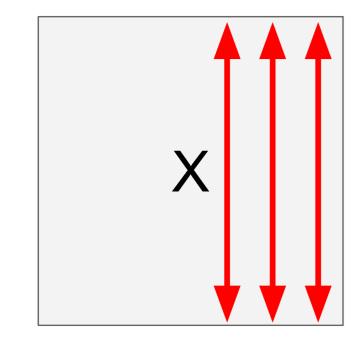
"you want zero-mean unit-variance activations? just make them so."

consider a batch of activations at some layer. To make each dimension zero-mean unit-variance, apply:

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathbb{E}[x^{(k)}]}{\sqrt{\operatorname{Var}[x^{(k)}]}}$$

this is a vanilla differentiable function...

"you want zero-mean unit-variance activations? just make them so."



Ν

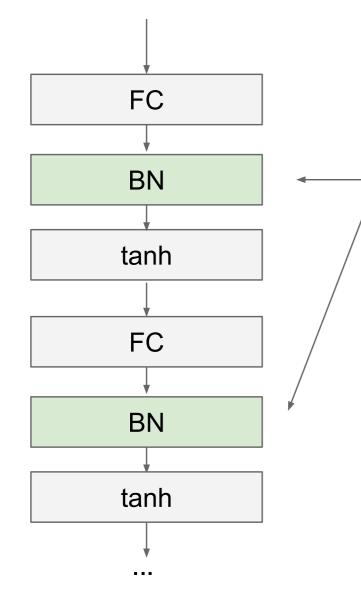
1. compute the empirical mean and variance independently for each dimension.

2. Normalize

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathbb{E}[x^{(k)}]}{\sqrt{\operatorname{Var}[x^{(k)}]}}$$

D

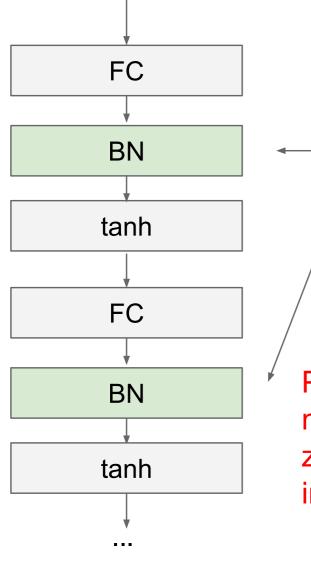
#### [loffe and Szegedy, 2015]



Usually inserted after Fully Connected or Convolutional layers, and before nonlinearity.

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathbb{E}[x^{(k)}]}{\sqrt{\operatorname{Var}[x^{(k)}]}}$$

#### [loffe and Szegedy, 2015]



Usually inserted after Fully Connected or Convolutional layers, and before nonlinearity.

Problem: do we necessarily want a zero-mean unit-variance input?

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathbb{E}[x^{(k)}]}{\sqrt{\operatorname{Var}[x^{(k)}]}}$$

Normalize:

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathbb{E}[x^{(k)}]}{\sqrt{\operatorname{Var}[x^{(k)}]}}$$

And then allow the network to squash the range if it wants to:

$$y^{(k)} = \gamma^{(k)} \widehat{x}^{(k)} + \beta^{(k)}$$

Note, the network can learn:  $\gamma^{(k)} = \sqrt{\text{Var}[x^{(k)}]}$   $\beta^{(k)} = \mathbb{E}[x^{(k)}]$ to recover the identity mapping.

**Input:** Values of x over a mini-batch:  $\mathcal{B} = \{x_{1...m}\}$ ; Parameters to be learned:  $\gamma, \beta$  **Output:**  $\{y_i = BN_{\gamma,\beta}(x_i)\}$   $\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^m x_i$  // mini-batch mean  $\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2$  // mini-batch variance  $\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$  // normalize  $y_i \leftarrow \gamma \hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i)$  // scale and shift

#### [loffe and Szegedy, 2015]

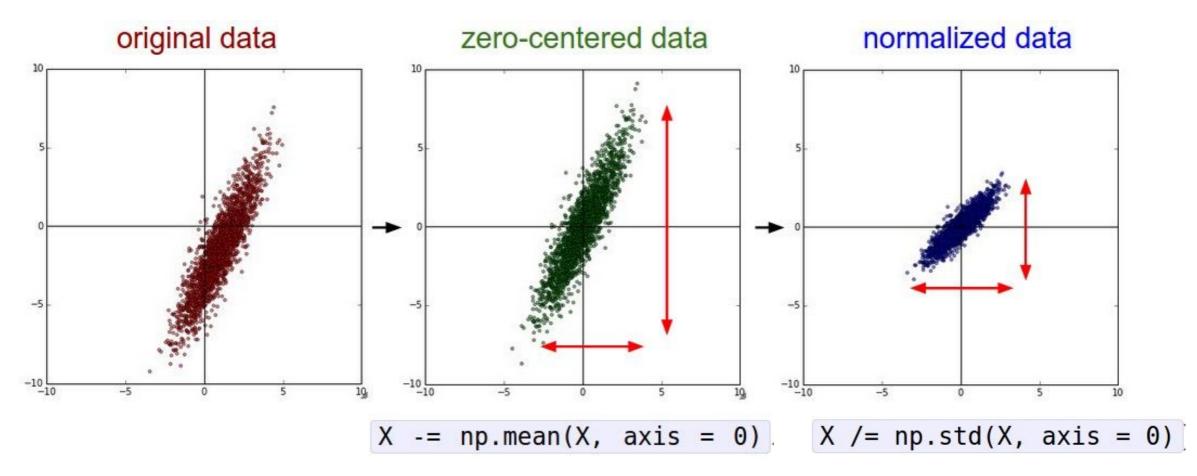
- Improves gradient flow through the network
- Allows higher learning rates
- Reduces the strong dependence on initialization
- Acts as a form of regularization in a funny way, and slightly reduces the need for dropout, maybe

#### [loffe and Szegedy, 2015]

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1m}\};$ Parameters to be learned: $\gamma, \beta$ Output: $\{y_i = BN_{\gamma,\beta}(x_i)\}$		Note: at test time BatchNorm layer functions differently:
$\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^{m} x_i$	// mini-batch mean	The mean/std are not computed based on the batch. Instead, a single fixed empirical mean of activations
$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2$	// mini-batch variance	during training is used.
$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$	// normalize	(e.g. can be estimated during training with running averages)
$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathbf{BN}_{\gamma,\beta}(x_i)$	// scale and shift	

## **Babysitting the Learning Process**

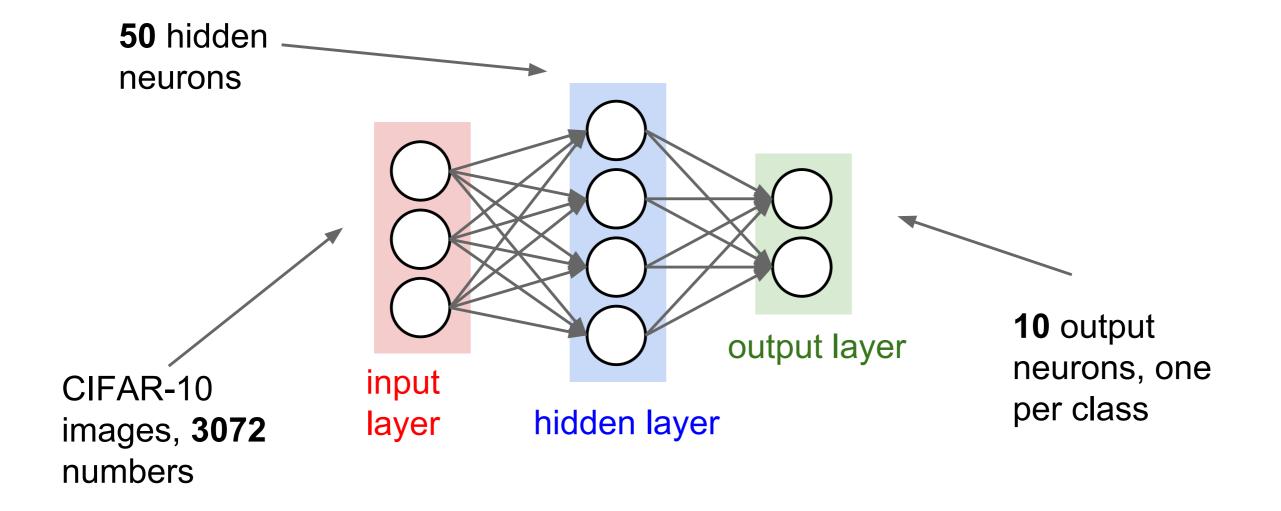
### Step 1: Preprocess the data



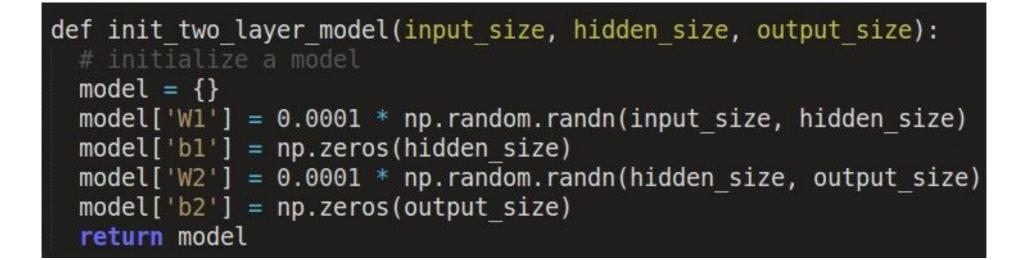
(Assume X [NxD] is data matrix, each example in a row)

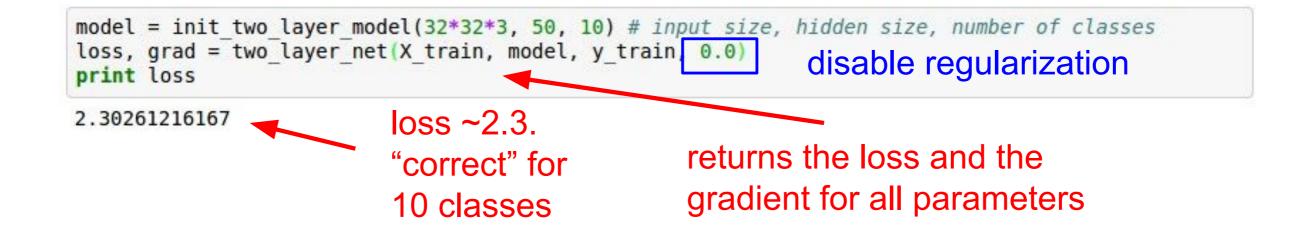
#### **Step 2: Choose the architecture:**

say we start with one hidden layer of 50 neurons:



#### Double check that the loss is reasonable:





#### Double check that the loss is reasonable:

```
def init_two_layer_model(input_size, hidden_size, output_size):
    # initialize a model
    model = {}
    model['W1'] = 0.0001 * np.random.randn(input_size, hidden_size)
    model['b1'] = np.zeros(hidden_size)
    model['W2'] = 0.0001 * np.random.randn(hidden_size, output_size)
    model['b2'] = np.zeros(output_size)
    return model
```

model = init\_two\_layer\_model(32\*32\*3, 50, 10) # input\_size, hidden size, number of classes
loss, grad = two\_layer\_net(X\_train, model, y\_train, 1e3) Crank up regularization
print loss

3.06859716482

loss went up, good. (sanity check)

**Tip**: Make sure that you can overfit very small portion of the training data

The above code:

- take the first 20 examples from CIFAR-10
- turn off regularization (reg = 0.0)
- use simple vanilla 'sgd'

**Tip**: Make sure that you can overfit very small portion of the training data

Very small loss, train accuracy 1.00, nice!

model = init two layer model(32\*32\*3, 50, 10) # input size, hidden size, number of classes trainer = ClassifierTrainer() X tiny = X train[:20] # take 20 examples y tiny = y train[:20] best model, stats = trainer.train(X tiny, y tiny, X tiny, y tiny, model, two layer net, num epochs=200, reg=0.0, update='sgd', learning rate decay=1, sample batches = False, learning rate=le-3, verbose=True) Finished epoch 1 / 200: cost 2.302603, train: 0.400000, val 0.400000, lr 1.000000e-03 Finished epoch 2 / 200: cost 2.302258, train: 0.450000, val 0.450000, lr 1.000000e-03 Finished epoch 3 / 200: cost 2.301849, train: 0.600000, val 0.600000, lr 1.000000e-03 Finished epoch 4 / 200: cost 2.301196, train: 0.650000, val 0.650000, lr 1.000000e-03 Finished epoch 5 / 200: cost 2.300044, train: 0.650000, val 0.650000, lr 1.000000e-03 Finished epoch 6 / 200: cost 2.297864, train: 0.550000, val 0.550000, lr 1.000000e-03 Finished epoch 7 / 200: cost 2.293595, train: 0.600000, val 0.600000, lr 1.000000e-03 Finished epoch 8 / 200: cost 2.285096, train: 0.550000, val 0.550000, lr 1.000000e-03 Finished epoch 9 / 200: cost 2.268094, train: 0.550000, val 0.550000, lr 1.000000e-03 Finished epoch 10 / 200: cost 2.234787, train: 0.500000, val 0.500000, lr 1.000000e-03 Finished epoch 11 / 200: cost 2.173187, train: 0.500000, val 0.500000, lr 1.000000e-03 Finished epoch 12 / 200: cost 2.076862, train: 0.500000, val 0.500000, lr 1.000000e-03 Finished epoch 13 / 200: cost 1.974090, train: 0.400000, val 0.400000, lr 1.000000e-03 Finished epoch 14 / 200: cost 1.895885, train: 0.400000, val 0.400000, lr 1.000000e-03 Finished epoch 15 / 200: cost 1.820876, train: 0.450000, val 0.450000, lr 1.000000e-03 Finished epoch 16 / 200: cost 1.737430, train: 0.450000, val 0.450000, lr 1.000000e-03 Finished epoch 17 / 200: cost 1.642356, train: 0.500000, val 0.500000, lr 1.000000e-03 Finished epoch 18 / 200: cost 1.535239, train: 0.600000, val 0.600000, lr 1.000000e-03 Finished epoch 19 / 200: cost 1.421527, train: 0.600000, val 0.600000, lr 1.000000e-03 Finished epoch 195 / 200: cost 0.002694, train: 1.000000, val 1.000000, lr 1.000000e-03 Finished epoch 196 / 200: cost 0.002674, train: 1.000000, val 1.000000, lr 1.000000e-03 Finished epoch 197 / 200: cost 0.002655, train: 1.000000, val 1.000000, lr 1.000000e-03 Finished epoch 198 / 200: cost 0.002635, train: 1.000000, val 1.000000, lr 1.000000e-03 Finished epoch 199 / 200: cost 0.002617, train: 1.000000, val 1.000000, lr 1.000000e-03 Finished epoch 200 / 200: cost 0.002597, train: 1.000000, val 1.000000, lr 1.000000e-03 finished optimization. best validation accuracy: 1.000000

Start with small regularization and find learning rate that makes the loss go down.

Start with small regularization and find learning rate that makes the loss go down.

model = init two layer model(32\*32\*3, 50, 10) # input size, hidden size, number of classes trainer = ClassifierTrainer() best model, stats = trainer.train(X train, y train, X val, y val, model, two layer net, num epochs=10, reg=0.000001, update='sgd', learning rate decay=1, sample batches = True, learning rate=le-6, verbose=True) Finished epoch 1 / 10: cost 2.302576, train: 0.080000, val 0.103000, lr 1.000000e-06 Finished epoch 2 / 10: cost 2.302582, train: 0.121000, val 0.124000, lr 1.000000e-06 Finished epoch 3 / 10: cost 2.302558, train: 0.119000, val 0.138000, lr 1.000000e-06 Finished epoch 4 / 10: cost 2.302519, train: 0.127000, val 0.151000, lr 1.000000e-06 Finished epoch 5 / 10: cost 2.302517, train: 0.158000, val 0.171000, lr 1.000000e-06 Finished epoch 6 / 10: cost 2.302518, train: 0.179000, val 0.172000, lr 1.000000e-06 Finished epoch 7 / 10: cost 2.302466, train: 0.180000, val 0.176000, lr 1.000000e-06 Finished epoch 8 / 10: cost 2.302452, train: 0.175000, val 0.185000, lr 1.000000e-06 Finished epoch 9 / 10: cost 2.302459, train: 0.206000, val 0.192000, lr 1.000000e-06 Finished epoch 10 / 10; cost 2.302420, train: 0.190000, val 0.192000, lr 1.000000e-06 finished optimization. best validation accuracy: 0.192000

#### Loss barely changing

Start with small regularization and find learning rate that makes the loss go down.

model = init two layer model(32\*32\*3, 50, 10) # input size, hidden size, number of classes trainer = ClassifierTrainer() best model, stats = trainer.train(X train, y train, X val, y val, model, two layer net, num epochs=10, reg=0.000001, update='sgd', learning rate decay=1, sample batches = True, learning rate=le-6, verbose=True) Finished epoch 1 / 10: cost 2.302576, train: 0.080000, val 0.103000, lr 1.000000e-06 Finished epoch 2 / 10: cost 2.302582, train: 0.121000, val 0.124000, lr 1.000000e-06 Finished epoch 3 / 10: cost 2.302558, train: 0.119000, val 0.138000, lr 1.000000e-06 Finished epoch 4 / 10: cost 2.302519, train: 0.127000, val 0.151000, lr 1.000000e-06 Finished epoch 5 / 10: cost 2.302517, train: 0.158000, val 0.171000, lr 1.000000e-06 Finished epoch 6 / 10: cost 2.302518, train: 0.179000, val 0.172000, lr 1.000000e-06 Finished epoch 7 / 10: cost 2.302466, train: 0.180000, val 0.176000, lr 1.000000e-06 Finished epoch 8 / 10: cost 2.302452, train: 0.175000, val 0.185000, lr 1.000000e-06 Finished epoch 9 / 10: cost 2.302459, train: 0.206000, val 0.192000, lr 1.000000e-06 Finished epoch 10 / 10; cost 2.302420, train: 0.190000, val 0.192000, lr 1.000000e-06 finished optimization. best validation accuracy: 0.192000

Loss barely changing: Learning rate is probably too low

loss not going down: learning rate too low

Start with small regularization and find learning rate that makes the loss go down.

loss not going down: learning rate too low

model = init two layer model(32\*32\*3, 50, 10) # input size, hidden size, number of classes trainer = ClassifierTrainer() best model, stats = trainer.train(X train, y train, X val, y val, model, two layer net, num epochs=10, reg=0.000001, update='sgd', learning rate decay=1, sample batches = True, learning rate=le-6, verbose=True) Finished epoch 1 / 10: cost 2.302576, train: 0.080000, val 0.103000, lr 1.000000e-06 Finished epoch 2 / 10: cost 2.302582, train: 0.121000, val 0.124000, lr 1.000000e-06 Finished epoch 3 / 10: cost 2.302558, train: 0.119000, val 0.138000, lr 1.000000e-06 Finished epoch 4 / 10: cost 2.302519, train: 0.127000, val 0.151000, lr 1.000000e-06 Finished epoch 5 / 10: cost 2.302517, train: 0.158000, val 0.171000, lr 1.000000e-06 Finished epoch 6 / 10: cost 2.302518, train: 0.179000, val 0.172000, lr 1.000000e-06 Finished epoch 7 / 10: cost 2.302466, train: 0.180000, val 0.176000, lr 1.000000e-06 Finished epoch 8 / 10: cost 2.302452, train: 0.175000, val 0.185000, lr 1.000000e-06 Finished epoch 9 / 10: cost 2.302459, train: 0.206000, val 0.192000, lr 1.000000e-06 Finished epoch 10 / 10; cost 2.302420, train: 0.190000, val 0.192000, lr 1.000000e-06 finished optimization. best validation accuracy: 0.192000

### Loss barely changing: Learning rate is probably too low

Notice train/val accuracy goes to 20% though, what's up with that? (remember this is softmax)

Start with small regularization and find learning rate that makes the loss go down. Now let's try learning rate 1e6.

**loss not going down:** learning rate too low Lets try to train now...

Start with small regularization and find learning rate that makes the loss go down.

```
model = init two layer model(32*32*3, 50, 10) # input size, hidden size, number of classes
trainer = ClassifierTrainer()
best model, stats = trainer.train(X train, y train, X val, y val,
                                  model, two layer net,
                                  num epochs=10, reg=0.000001,
                                  update='sgd', learning rate decay=1,
                                  sample batches = True,
                                  learning rate=le6, verbose=True)
/home/karpathy/cs231n/code/cs231n/classifiers/neural net.py:50: RuntimeWarning: divide by zero en
countered in log
  data loss = -np.sum(np.log(probs[range(N), y])) / N
/home/karpathy/cs231n/code/cs231n/classifiers/neural net.py:48: RuntimeWarning: invalid value enc
ountered in subtract
  probs = np.exp(scores - np.max(scores, axis=1, keepdims=True))
Finished epoch 1 / 10: cost nan, train: 0.091000, val 0.087000, lr 1.000000e+06
Finished epoch 2 / 10: cost nan, train: 0.095000, val 0.087000, lr 1.000000e+06
Finished epoch 3 / 10: cost nan, train: 0.100000, val 0.087000, lr 1.000000e+06
```

loss not going down: learning rate too low loss exploding: learning rate too high cost: NaN almost always means high learning rate... Lets try to train now...

Start with small regularization and find learning rate that makes the loss go down.

Finished epoch 1 / 10: cost 2.186654, train: 0.308000, val 0.306000, lr 3.000000e-03 Finished epoch 2 / 10: cost 2.176230, train: 0.330000, val 0.350000, lr 3.000000e-03 Finished epoch 3 / 10: cost 1.942257, train: 0.376000, val 0.352000, lr 3.000000e-03 Finished epoch 4 / 10: cost 1.827868, train: 0.329000, val 0.310000, lr 3.000000e-03 Finished epoch 5 / 10: cost inf, train: 0.128000, val 0.128000, lr 3.000000e-03 Finished epoch 6 / 10: cost inf, train: 0.144000, val 0.147000, lr 3.000000e-03

3e-3 is still too high. Cost explodes....

### loss not going down: learning rate too low loss exploding: learning rate too high

=> Rough range for learning rate we should be cross-validating is somewhere [1e-3 ... 1e-5]

# Hyperparameter Optimization

## **Cross-validation strategy**

#### coarse -> fine cross-validation in stages

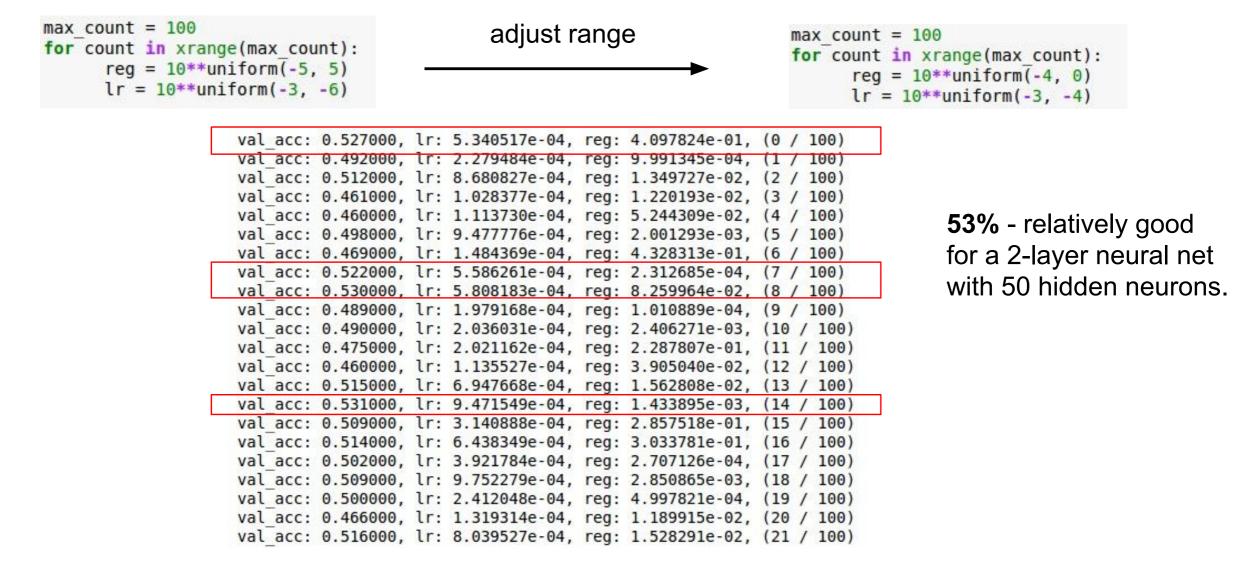
**First stage**: only a few epochs to get rough idea of what params work **Second stage**: longer running time, finer search ... (repeat as necessary)

Tip for detecting explosions in the solver: If the cost is ever > 3 \* original cost, break out early

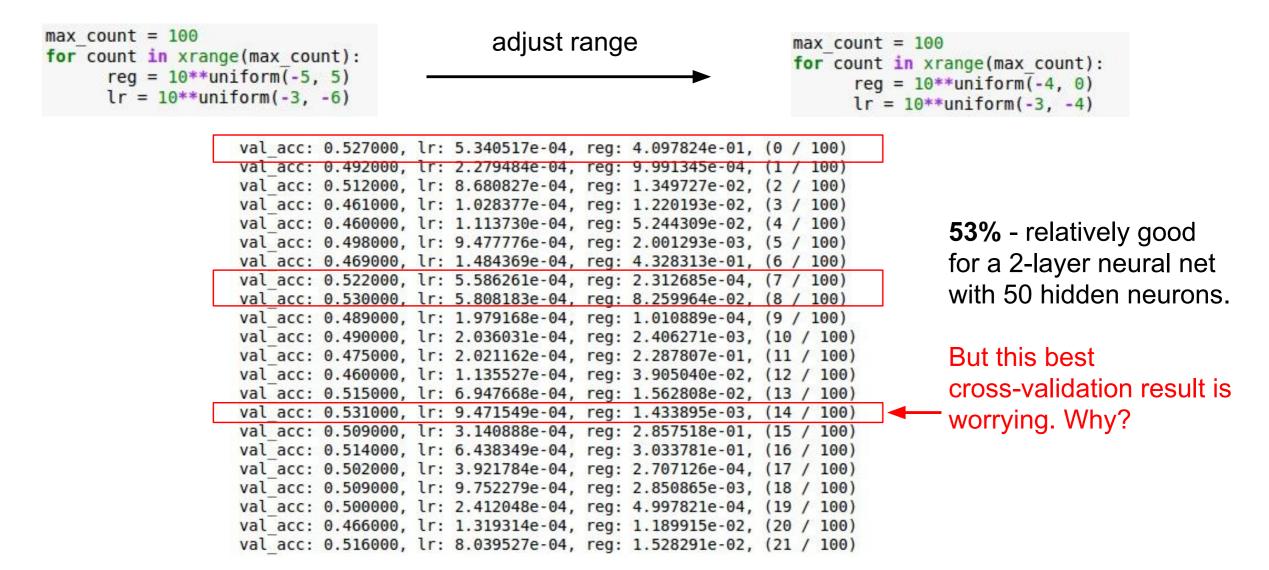
#### For example: run coarse search for 5 epochs

<pre>trainer = ClassifierTrainer(</pre>	(32*32*3, 50, 10) # input size, hidden size, number of classes
<pre>val_acc: 0.214000, lr val_acc: 0.208000, lr val_acc: 0.196000, lr val_acc: 0.079000, lr val_acc: 0.223000, lr val_acc: 0.441000, lr val_acc: 0.241000, lr val_acc: 0.482000, lr val_acc: 0.079000, lr</pre>	<pre>: 1.405206e-04, reg: 4.793564e-01, (1 / 100) : 7.231888e-06, reg: 2.321281e-04, (2 / 100) : 2.119571e-06, reg: 8.011857e+01, (3 / 100) : 1.551131e-05, reg: 4.374936e-05, (4 / 100) : 1.753300e-05, reg: 1.200424e+03, (5 / 100) : 4.215128e-05, reg: 4.196174e+01, (6 / 100) : 1.750259e-04, reg: 2.110807e-04, (7 / 100) : 6.749231e-05, reg: 4.226413e+01, (8 / 100) : 4.296863e-04, reg: 6.642555e-01, (9 / 100) : 5.401602e-06, reg: 1.599828e+04, (10 / 100) : 1.618508e-06, reg: 4.925252e-01, (11 / 100)</pre>

### Now run finer search...



### Now run finer search...



## Random Search vs. Grid Search

Random Search for Hyper-Parameter Optimization Bergstra and Bengio, 2012

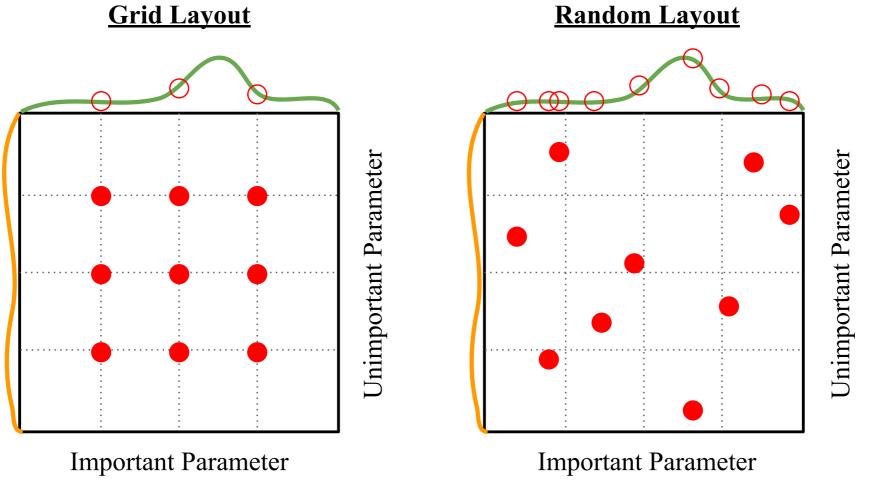
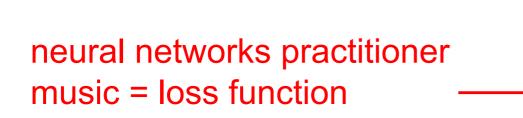


Illustration of Bergstra et al., 2012 by Shayne Longpre, copyright CS231n 2017

# Hyperparameters to play with:

- network architecture
- learning rate, its decay schedule, update type
- regularization (L2/Dropout strength)



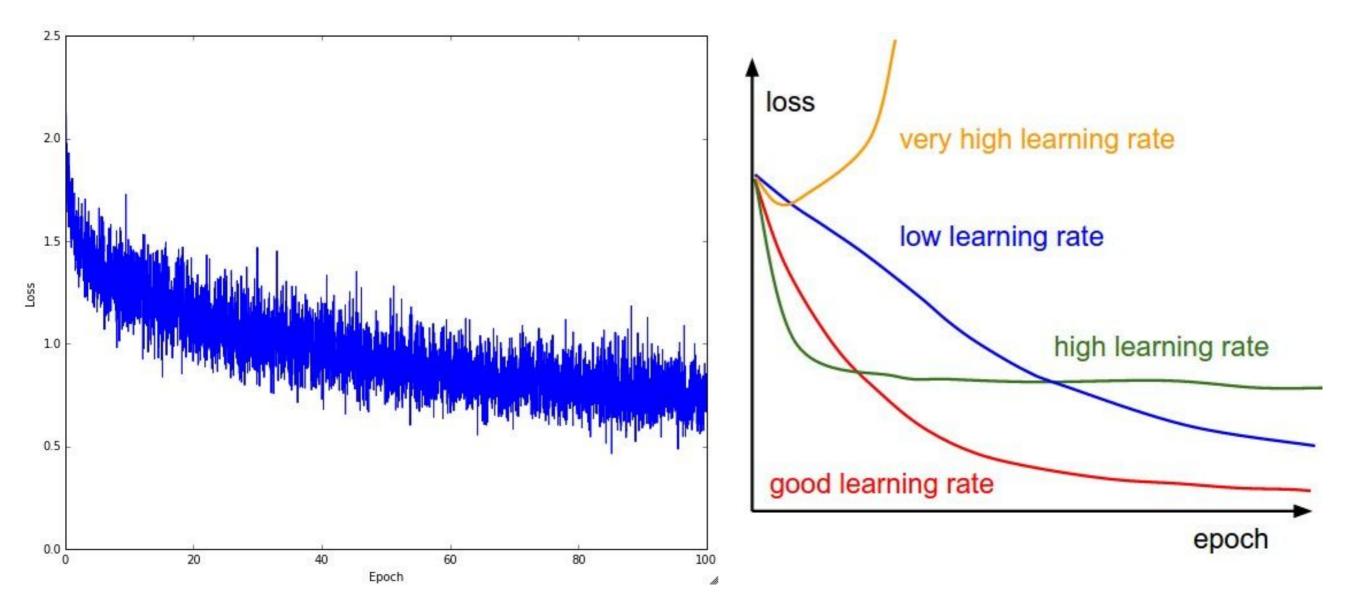


This image by Paolo Guereta is licensed under CC-BY 2.0

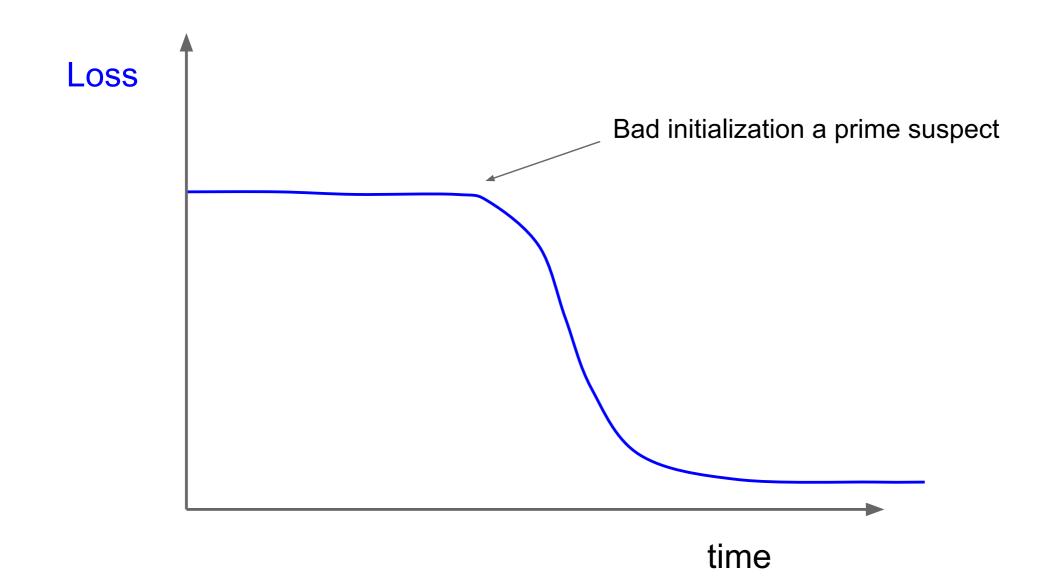
#### Cross-validation "command center"

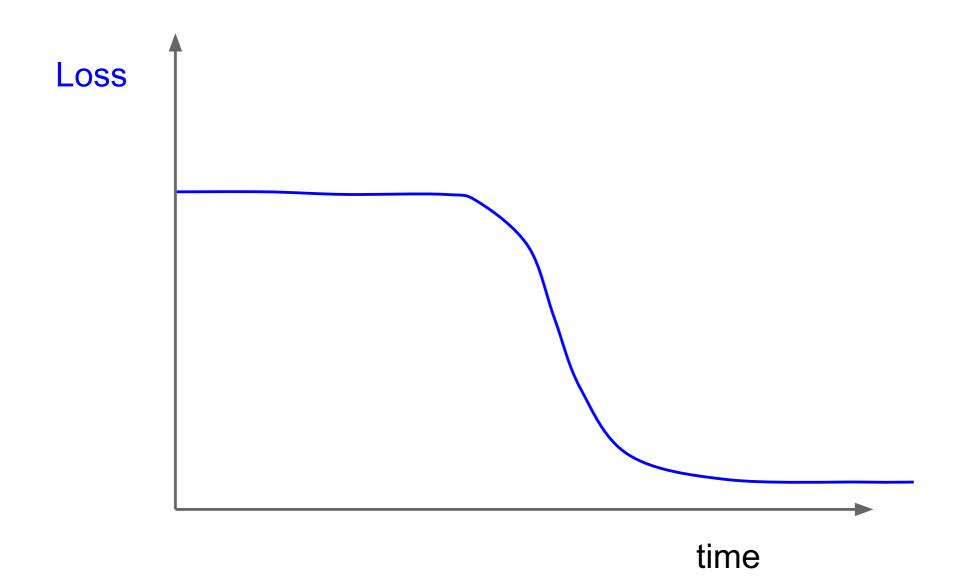
Incluses Incer

and and a second	Sector 1	Server 3	factor 4	Server 1	Service C	16 sheet	Berter 1	Service 2	Sector 11	10	Bucher II	Jacobre II
And Annual International Inter	Annual State -	Last growth, 2019-20- transferrer, 2019-20- transferrer, 2019-20- transferrer, 2019-2019 View 2019-2019-2019	International Control of Control	and the second s	Annual Control of Cont	Left underen 2017-Det Destandenen 2017-Det underen 2017-Det Dies 4004/24/00	And	And Andreas (1921-21) And Andreas (1921-21) Manual La Callin State 17756-52700	Annual Contract Statistics		Internation International Inter-De- monstrate Annual International Inter-De- International Inter-De- De- De- De- De- De- De- De- De- De-	Land and Lan
Markey (1997) A. Standard A. Standards, A. S. Maraganith and C. Taking exclusion adverti- tion, "Annual conduction adverti- ing annual conduction and and provide advertising and article advertising and advertising and article address and advertising and article address and advertising and article address and advertising and advertising advertising advecting adv	<ol> <li>Apple (1997) 11.</li> <li>Apple (1997) 14.</li> </ol>	**end of the statement of the statement of the statement of the statement of the statement of the statement of the statement of the statement of the "statement of the s	C. Scott, et al., Statute e	<ul> <li>South respectively. In: Proceedings of the Proceedings of</li></ul>	<ul> <li>Apple 1997 Oct.</li> <li>App</li></ul>	Apple 100 <sup>10</sup> S. Sono M., Sono M. S. Sanami M. Sana M. Sanati S. Sanami M. Sanati S. Sanati S. Sanati S. Sanati S. Sanati S. S. Sanati S. S	<ul> <li>Apart (1997) - R.</li> <li>Apart (1997) - R.</li> <li>Apart (1997) - Apart (1997)</li> </ul>	[1] April 10(1) - 3. Provide statements in the second statements from 1. Control resulting (1017) from result (1017) - 1.5. Provide statements (1017) Provide statements (1017) Provide statements (1017)	Algorit (1997) S. State (1997) S. State (1997) S. State (1997) S. State (1997) State (1997) State (1997) S. State (1997) State (1997) State (1997) State (1997) State (1997) State (1997) State (1997) State (1997) State (1997) State (1997) State (1997) State (1997) State (1997) State (1997) State (1997) State (1997) State (1997) State (1997) State (1997		<ul> <li>Appart (Project Oc)</li> <li>Appart (Project Oc)&lt;</li></ul>	1. "Point resident in the second s
(a) Annual Annua Annual Annual Annua Annual Annual Annu	Provide the standard of the Standard St	Dens, Jacob Jonefs, Oh, Peagemerts Control, "processory," Control Networks, Control Control Parally, Internet, Vol. Parally, Internet, Vol. Science, 1971 (Internet, Control of Science, 1971) (Internet, Control of Science, 2017) (Internet, Control of Science, 2017).	Construction of a site of the second state	Part according to the Tangent's line form a state of the task of the second the second second second second second second second s	Construction of the construction of the construction of the construction of the construction of the construction of the construction of the con	Constanting and Constanting	Construction and a link. Assessment Clark, Application Clarks, Clark, Construction Clarks, Clarks, environment, and Clarks, environment, and the Second Clarks, Construction, Control Construction, 2010.	Press, and exactly starts, etc., "sequently frame, and exactly and the sequence frame, and an analysis of the sequence frame, and the second start starts of the frame, and the second starts of the second starts of the frame, and the second starts of the second starts of the frame, and the second starts of the second starts of the frame, and the second starts of the second starts of the frame, and the second starts of the second starts of the frame, and the second starts of the second starts of the frame, and the second starts of the second starts of the frame, and the second starts of the second starts of the frame, and the second starts of the second starts of the frame, and the second start of the second starts of the second starts of the frame, and the second starts of the second starts of the second starts of the frame, and the second starts of the second starts of the second starts of the frame, and the second starts of the second starts of the second starts of the frame, and the second starts of	Participation, and a fair, fragments Party A. Santastari ("Santa") Party A. Santastari ("Santa") Party A. Santastari ("Santa") Party (Constraint Act, Santastari (") Party (Constraint Act, Santastari (")) Party (Constraint Act, Santastari ("		Production confusion, respective transfer descention, confusion and production and confusion and confusion and confusion and con	Frank, Janis, Carlos, M., San, S., Linnel, C., Sannan, C. Cardo, S. San, S. S. Sanna, S. S. Sand, S. Sanda, S. Sandar, S. S. Davids, and S. Sandar, S.
<ol> <li>Nobical's C. Status (Name and Status (Name (Name (Name (Name (Na</li></ol>	Prof. "Sealantad", & Sealant Johann, Sealant de Bernert 7. Seala - Constraint de Verninger Berninger, Constraint de Verninger Berninger, Statut 7. Sealant (Sealant de Sealant Frankriger, Sealant de Sealant	(1) Network (1) Network (Series (Series (Series))) Network (Series) (Series (Series)) Network (Series) (Series (Series)) Network (Series) (Series) (Series) Network (Series) (Series	<ol> <li>Statistical's T. Science Jonane (1999), disconsistent Science - Manageria (1999) Science (1999) (1999) Science (1999) (1999) Science (1999) (1999)</li> </ol>	<ul> <li>Constant L, Nolaki and Sul, Nolaki A, J.</li> <li>Constant A, S.</li> <li>Constant A, Solaki A, Solaki A, Solaki Patronic Indian, Patronic Indian, Patronic Indian,</li> </ul>	[10] "Number of C. Second Schwarz and an analysis" in the constraints of the Constraint Second and the Constraints Schwarz and Constraints. Nature Schwarz and Schwarz Schwarz Schwarz and Schwarz	[10] Nobel S. C. Scherold, C. S. Scherold, "Advanced particle and strategy and strategy of the strategy of	(c), "Numerica", C. Number, Andreas, and an Antonio, " Control, Andreas, and an Antonio, " Sciences, and and an Antonio and Antonio, Sciences, and Proceedings of the Antonio, Sciences, and Proceedings of the Antonio, Sciences, and Proceedings of the Antonio, Sciences, and Antonio, Sciences, and Antonio, Sciences, Antonio, Ant	[10] Numerica's C. Norma, Name and Annual America Structure of Control of Control Structure of Control of Control Structure and Control of Control Descention Structure of Control Descention Structure Descention Structure of Control Descention Structure of Control Descention Structure of Control Descention Structure Descention Structure of Control Descention Structure of Control Descention Structure of Control Descention Stru	(1) Number of States and American States (New York, States and States (New York, States (New York, States)), Number of States (New York, States), Number of States), Number of States (New York, States), Number of States)		M. "Solicitor", S. Schultz, Americany, " Solicitor, Schultz, and Schultz, Schultz	<ol> <li>Newsyngh, S. Santar, Daris, and A. Alm Contact, Status, and A. Alm Contact, Status, New York, New York, New York, N. Constant, 1997, 4</li> </ol>
Sector of England (Sector 2014) sector of Congress (Sector 2014) sector of Congress (Sector 2014) sector of Congress (Sector 2014) sector of Congress (Sector 2014) Sector 2014) Sector 2014 Sector 2014 Sect	<ul> <li>And contracting the search description of the search of the</li></ul>	Construction Construction International Construction International Construction Con- tension Construction Con- tension Construction Construction Construction Construction Con- tension Construction Construction Comparison, Construction Compared Comparison, Construction Compared Compared, Construction, Compared,	Consequences of an annual consequences in an annual consequence of the consequences in an annual consequence of the consequences in a second consequence of the consequence of the consequence of the consequence of the second consequence of the consequence of the consequence of the consequence of the consequence of the consequence of the consequence of the consequence of the consequence of the consequence of the consequence of the consequence of	Comparison of the second se	Construction (Construction) Construction (Co	Construction of the second sec	Compare and party for each pro- traction of party for each pro- party compared wave pro- traction of the party of the party for and party for the party of the party for party for the party of the party for the party for the party for the party for the party of the party of the party for the party of	Construction of the second sec	<ul> <li>Conservery and Provide Conservery International Conservery (Conservery), Con- Paration and Conservery (Conserver), Con- Paration Conserver, Conserver, Conserver, Conservery, Conserver, Conser</li></ul>		Consequences, and an annual consequences of the second second second second second second second, descendent of a methyle second, descendent of a methyle second, descendent of a methyle second second second second beams, second second second second second beams, second second second second second beams, second se	Construction of the second
an an advantation and reasons and there a p	the are been and a single to be a si	Lan pro Architelli, U.S. monte tracto geto de pro- 12. 462	tan ar anna 11.000 ana 11.000 12.000	Sum and States and States Sum for Second and 127,822	tion pro account 11% which there part to pup 121.12%	the product of the second test	Line per antier 1.000 Line per antier 1000	Line per Artelulli. 200 molet trate per of dis 27 878	Line and Arrive Links Second Galaxy and Second 21, 612		tone per anno (n. 16) anno (n. 16) 12 per	Law an anna 1915 main that an 1916 1819
-	-			- Umerdenie inner			12		-			
nerðar 14 na annar mir-ta-	Author 19	Service 16	Sectors 17	States 11	Service TR	Secritor 10 mm cont Let solutio (01-00	P	angler 22 San and 21	Automatica and a second	weather 34 manufacture (manufacture (manufac	aucher 25 Los aucher 25 - Co-	Forters 28 : Entry and the company
month of Arman month of Arman m	The antise of the second second second second methods (2000) (Second Sec	Loss of the second seco	Construction (2015) The International Construction International Construction (Construction) (Construction) (Construction) (Construction), Con- Version, Construction, Construc- tion, Neural Construction, Con- Neural, State (Construction, Con- Neural, State (Construction), Con- Neural, Construction, Con- Neural, Con- Neura	Linear second ingenits works and the Second	An experience (2023-10) setting the conversion ment of balance (a) and (a) balance (b) and (b) and (b) (b) and (b) and (b) and (b) and (b) (b) and (b) and (b) and (b) and (b) (b) and (b) and (b) and (b) and (b) and (b) (b) and (b) and (b) and (b) and (b) and (b) (b) and (b) and (b) and (b) and (b) and (b) and (b) and (b) (b) and (b) and (	International Contents International Contents Network Contents Network Contents International Cont	Landa La Rein anche La Rein La Construction Constructio	Annual State (Constraints) and a second state and a second state (Constraints)	The American State of State and State of State and State of State and State of State	A second	An annual and a second	Construction (CD-C)- terreturns (CD-C)- terreturns a cross much de K-Vel Loss Ministries (CD-C)- terreturns (CD-C)- terreturns (CD-C)- lister, etc., crossent (CD-C)- terreturns (CD-C)- lister, construction (CD-C)- lister, construction (CD-C)- lister, construction (CD-C)- terreturns (CD-C)- terretu
Holdson, and L. Hu, Dapart Mark, Annual T, Huert Mark, Annual Y, M. Mark, Annual Y, M. Mark, Annual Y, M. Mark, Annual Y, M. Mark, Mark Mark, Mark, Mark Mark, Mark M, L. Nobel, Joseff M. Nobel, 2017. U.	Provide start - Start Report Transf., Special - Transf. Provide Start - Start Provide Start	Seed, State Science, S. Sill, Fragmer's District, Space-Science, 1999 Science, Science, Science, Science, Science, Science, Science, Science, Science, Science, Science, Science, Science, Science, Science, 2007 Science, 2007 Scien	Frank, Jacob House, L. K., Franken, C. Sold, National Street, Weil, Son Statistics, C. S. Sondo, Contractive Activity, Parallel and Antipartic Activity (National Activity), and Sold Statistics, Statistics, Sold Street, Sold Street, Statistics, Sold Street, Sold Statistics, Statistics, Sold Street, Statistics, Statistics, Sold Street, Statistics, Sold Street, Statistics, Statistics, Statistics, Sold Street, Statistics, Statistics, S	Area, Janie Kolefs, D., Yagnerf, D.M., Die and Y. (2017). Weinger, A. 2017 March 1997 (2017). A 2017 March 1997 (2017). A 2017 March 1997 (2017). A 2017 March 1997 March 2017 March 1997 March 2017 March 2017 March 2017 March 2017 March 2017 March 2017.	Terra, and confr. ch., "Family," Terra, and confr. (Scott, Terra, and conference of the terra, and the second of the terra, "New confr. 1, "New confr. Terra, "New confr. Terra	Proc. Lands, and A. Lin, "Property Transf, Transmission," (2014), "Annu personal state of the Sector Annual State of the Sector Annual State, Sciences of the Proc. Management, Sci. Theorem, 2014 The Management, Sci. Theorem, 2014), The Management, Sci. Theorem, 2014 (2014), "New York," (2014)	<ul> <li>Anna Sana, Sandy Sill, Anamady Tanag, Sanapatory Datas, Tanag, Sanapatory Datas, Tanag, Sanapatory S. Patha Anamatani and Sanapatory Sanapato</li></ul>	Presi, and staffs (H., Stagnaff Statut, Antonio S., David Statut, A., Janefer M., Fand, A., Janefer M., Statut, Statut, J., Statut, Statut, S.	Proc. Later, and L. M., Sapandi Containing C., Spectra M., Sapandi M., 2014, 2014, 2017, 10 Proc., 2017, 10 Proc., 2017, 10 Proc., 2017, 2017, 2017, 2017, 2017, Nano, 2017, 2017, 2017, 2017, 2017, 2017, Nano, 2017, 2	Constrainty (1): New Joint, and C., Jin, Yuang M., Constrainty (1): An and the constrainty (1): An and the Data of the constraint (1): New Order and Annual Constrainty (1): And Constrainty (1): A second second second second (1): A second (1): A second second second (1): A second (1): A second (1): A second second (1): A second	Proceeding of the proceeding o	March Jann, Scathy Jan, S. Charl, Santasanari I, Salar Marchard, S. 2000 Str. National Conversion and Arti- National Conversion and Article Marchard Strength Scatter Scattering
Valida A. Statistical and the second seco	Control of America Processor Sector Sector, Sector Sector,	<ul> <li>Proc Complexity on Complexity of Complexity on Complexit</li></ul>	Provide a Comparison of Comparison Rectifier Density Density of Comparison Density of Co	2. Status - Construction Percentage Description (2007) - 5 Status Annual (2007) - 5 Discontration (2007) - 5	Control of the second s	Construction of the second sec	<ul> <li>Construction of the second seco</li></ul>	Contract Contraction on Contract Benefician Contract, A. Contraction Contract, A. Contraction Contract, A. Million, Contraction, Contract, March 19, 1999 (Contract, Contract, March 19, 1999) (Contract, Contract, March 19, 1999) (Contract, March 19, 1999) Service on Contract, March 19, 1999) Service on Contract, March 19, 1999	Construction of the second sec	Constraints, and a second seco	Comparison of the second secon	Product of sources and sources
mode and the first second second mode and the second second second mode and the second second second mode and the second second second second from any mode second	Particular is high class many and and an analytic heat provide the analytic heat provide the heat provide the and the provide the and the provide the and the provide the analytic the analytic the	Paral (127) 147 (147, 148, 73, 148, 73, 148, 148, 147, 147, 148, 148, 148, 148, 148, 148, 148, 148	Passel, and S. W., Sharki S., Sakari S., Sanato and S. Ser, Tachani S., Sanato and S. Sakari Sang Pasharenalist Lett. Camputer Sendulat 2014 South Tests 2015, 17 Aug. 21, 460	Despit, and one chapter protect transit way is and, inclusion reasoning, reasoning to a reasoning product of a reasoning to a persistent to a set to and the set of the set to and the set to a set to a set of the set o	Parent, seath, set, "happy cancel framed, and the set, and and the parent probability of the frame probability of the fra	Cardy, etc. <sup>2</sup> , no. (Supple Line Construction) and the Construction comparison and an inter- tion processing and the Con- tent processing of the Con- tent Construction and Construction (Con- tent Construction) and Construction (Con- tent Construction) and Construction (Con- tent Construction) and Construction (Con- tent Construction) (Construction) (Construction) (Construction) (Construction) (Construction) (Construction) (Construction) (Construction) (Construction) (Constructio	Design and the "support forwards Consists and the "support forwards" Consists and the support of the support	<ul> <li>Constraints and and constraints</li> <li>Constraints and and constraints</li> <li>Constraints</li> <li>Constraints</li></ul>	b) Anishi (provi cardy francustry) Prospany, from period 21, 20, 20 that per second 27, 20, 3 has per second 27, 20, 3 has per second and 10, 20 59, 803.	<ul> <li>Transformer and a set of the product of the comparison of the set of the set of the theory and and set of the set and an and and set of the set and the set of the set of the set of the set of the set of the set of the set of the set of</li></ul>	David, and Case, Marris Loke, D., Greense and Card, Sciller T., David, C. & Card, M. & Card, S. & San and Sciller Control Sciences Service Control Society Tests (2017) 101 (2017) 2017 (2017) 2017 2017 (2017)	Parent, and C. and Anaper- Parents, and C. and A. Smole Parents, Annual Annual Con- training and Annual Control (Any Physical Annual Control (Any Physical Annual Con- stance), and a specific and a specific and a spec
								111	- hanne	-	- hereiter	
etrier IJ	scriter 11	Secolar 01	Secolar 18	Sorbar 21	berber 31	H. 1620	Jenciar 34	sociar 35	soften 36	stpker 37	stoker 38	sorier N
ner annen erstelle Frieden in erstelle Frieden in erstelle an erstelle fra Annel erstelle fra	And appendix division birth the training spectra is a firm over Wildowind in Neuropeint Sciences of the second science of the second	And another intervals of the second s	Last annexes distants Distant and annexes plants in a constraint plants in the last plant with last plant with last plants of the last plants of t	and approve restricts Transition of Annual profits in the two main institutions in the constitution in the co	And Antonio International Antonio International Antonio Antoni	Lang and the second sec	And Antonio States	acchar 33 internet and the access	Lar andre dette Lara andre dette Lara andre dette Lara andre dette Lara andre dette Lara andre dette	Lan where some	en verstene annen en verstene annen en verstenen en verstenen en verstenenen verstenenen verstenen verstenen versten	Let un an entre annue transmissioner south to date the weighter has weighter base states an base states and base stat
Teories, "Programming Last" (2), "Prod. and Second", "Al- song prod. Second "Prod. Second Second Second "Prod. Second Second Second "Prod. Second	<ul> <li>Tenter, Comp. Receiving parts?</li> <li>Tente parts, Beneficial Conf., Sci.</li> <li>Tente parts, Beneficial Conf.</li></ul>	<ul> <li>Porter, Maria Access Arrives, P. et al., No. 79(1), and head of the control of the period period and access to the No. 6 access to the control of the No. 6 access to the control of the Control of the control of the control of the control of the control of the Control of the control of the Control of the control of the No. 6 access to the control of the control of the No. 6 access to the control of the control of the control of the No. 6 access to the control of the control of the control of the No. 6 access to the control of the control</li></ul>	<ul> <li>Denter, "Parameter and the second stands".</li> <li>Denter and the second of the second stands of</li></ul>	<ul> <li>Perez's, "Deserving control points", 2015, "Perez's and transfer of site Perez's provide and provide the Perez's provide and provide the Perez's and provide the perez's and Perez's and provide the perez's and Perez's and provide the perez's and Perez's and perez's and perez's and Perez's and Perez's</li></ul>	Control C. Marcon, Proceeding and C. S. S. S. S	Processing Carls, The Processing Carls, The	<ul> <li>Control, These proceeding of the control of the control of the Network of the Network of the control of the Network of t</li></ul>	<ul> <li>Control, Comp. Recording, 201971</li> <li>Control and Comp. Phys. 11, 11</li> <li>Paragi provide and comp. Phys. 11, 11</li> <li>Paragi provide and provide and the physical strategy of the physica</li></ul>	Control, "Temp, Location, 2010", 11, 210, "New Joint Control of Language and provide the state of the Party Location of the State of the State of the Data Annual State of the State of the Party Location of the State of the	<ul> <li>Perers', "Output processing processing Sillin, "Search and compart?" All:</li> <li>Perers, processing recently, string, "Perers, processing recently, string, "Perers, processing, recently, string, "Perers, processing," and processing Science, "Perersong, processing," Programmer Science, "Perersong, processing," Proceedings, Pererson, processing, processing, pro- perty processing, processing, processing, pro- perty processing, processing, processing, pro- perty processing, processing, pro- perty processing, processing, processing, pro- perty processing, processing, pro- perty processing, processing, pro- perty pro- p</li></ul>	<ul> <li>Trantis, Telana, modera prateri</li></ul>	<ul> <li>Personal A., "Contrast, the contrast of a static strength of the contrast of a static strength of the contrast of a static strength of the contrast of the static strength of the strength of the static strength of the strength of the strength of the strength of the strengt</li></ul>
Mich C., Standardson, J., St., Standard, M., C., Shang, March. U., Washing, Standard Stan, Standardson, Y., Washing, Standard DJ, Theoperature J., Michael J., Marcalastica, editeration, "C Spaces of Systems and Academic Systems," <i>International Conference on Confer</i>	An or constraint and community of the constraint	An experiment in constant is a second second second second second is a second	And a second sec	Annual Control of the Annual Annua Annual Annual Annua	Construction of American State Section 1, Section 2, Section 2, Section 2, Section 1, Section 2, Section 2, Section 2, Section 2, Section 2, Section 2, Section 2, Section 2,	And a second sec	Contraction of Contract Contraction of Contract Contraction of Contraction Contraction of Contraction Contra	And a second sec	And the second s	Anna Anna Anna Anna Anna Anna Anna Anna Anna	<ul> <li>Service American Andrewski (Strandarder)</li> <li>Service American American</li></ul>	<ul> <li>Annie, Annekowski, and Annie Str., Annie (Anne, 1994), and Annie (Anne, 1994).</li> <li>Press, Annie (Anne, 1994), Press, Annie (Anne, 1994</li></ul>
Santang (2017) i colore, internet, orbert dereteriet sons i colorente internet antipation (2017) internet anti- antipation (2017) internet antipation antipation (2017) internet antipat	<ul> <li>Adolina, March Lofena</li> <li>Felenanda annas derenantes</li> <li>Statuta annas derenantes</li> <li>Statuta annas derenantes</li> <li>Statuta annas derenantes</li> <li>Statut and der Statut annas derenantes</li> </ul>	Augusticity (2004) is adverse by the second	<ul> <li>Anachine, 1997. A selection</li> <li>Antenna M., Salari, S., Salari, Y., Antenna M., Salari, S., Salari, P., Martinez, C. S., Salari, S. Salari, S. S. Narad, State and Salari, 2017. In</li> <li>Antenna Salari, Salari, 2017. In</li> <li>Antenna Salari, Salari, 2017. In</li> </ul>	Providence, 20197 - 1 defents Providence of provide distances of providence of the second distances of providence of the second distance of the Providence of the se	<ul> <li>Andreas Argent and the second s</li></ul>	Contraction (Contraction) Contraction (Contraction) Contraction (Contraction) Contraction (Contraction) Contraction (Contraction) Contraction (Contraction) Contraction (Contraction) Contraction (Contraction) Contraction (Contraction) Contraction (Contraction)	Provide strength of the second strength of the second seco	Automatical (2007) is altered. Transformatic strategy and the material of the state of the strategy and participation of the strategy and the participation of the strategy and the Property construction of the strategy of the Property construction of the strategy of the Property construction of the strategy of the strategy of the participation of the strategy of the strategy of the participation of the strategy of the s	Participation Control Control Protocol Control Control Control Control Control Control Control	<ul> <li>Constitution (2019)</li> <li>Construction (2019)&lt;</li></ul>	<ul> <li>Andreas (1997) Andreas (1997)</li> <li>Antering (1997) Andreas (1997)</li> </ul>	Constrainty and the statement of the statement of the statement and the statement of the statement and the statement of the s
<ul> <li>"Willing gave loop, "Automatic memory," and appropriate light start provide starting of the start international starting memory and starting starting approximation with light/ 2,124</li> </ul>	<ul> <li>Personal and Constraints, and Constraints, New Constraints, State Science, Scien</li></ul>	<ul> <li>Automotory (and Automatic Systems), State and Automatic Systems and Automat</li></ul>	<ul> <li>State and A. Saranati, Source and S. Saranati, S. Saranat</li></ul>	<ul> <li>Statistic approvals, A.S., Statistics, Strang gata associative value value gata associative value annual value approval value value approval value approvalue approval value approval value approval value approval v</li></ul>	Compared agent cards, "sciences," Compared States, and States "compared second-", it is compared second-", it is compa	Sealard (1977) and Personal Control Sealard (1977) and (1977) and (1978) and (1977) and (1977) and (1977) (1977) a	[25] Andrew Stern, Sandy Andrews, Pressing and Annual Article and Strain for Annual Article and Strain for Annual Article and Annual Annual Annual (1998). 2014;405	Compared Speed and A Second State Compared Systems and Second State State and Advanced State Speed States and State Speed States and States Speed Speed States and Speed States Speed Speed Speed Speed Speed Speed Speed Speed Speed Speed Speed Speed Speed Speed Speed Speed S	<ol> <li>Andres spectral states, "Andrease Program (2019) and an analysis of the same pair and an analysis of the same pair when shares the same pair when shares the same pair of the same states of the same states of the same states of the same states of the same states and the same states of</li></ol>	Present part with framework of the memory of the second of the test part and the second of the test part and the second of the memory of the second of the memory of the second of the memory of the second of the second of the second of the memory of the second of the second of the second of the second of the second	<ul> <li>Salina gano lacit, Nacionario, Presento e Salina (Secondario), S. S. Pasa gana accident, S. St. Salina faire science (Secondario), ante faire science (Secondario), anter science (Second</li></ul>	<ul> <li>Same and a set of require and a set of the</li></ul>
					-	-	1.1.1				1.4	
nahar 43 Tananan aslatan Martin Sanatan Martin Sanatan Martin Sanatan	Anthen 12 Trans John San galaxies mittan- singer and the work 12 MOP San anthen mit	Barber 42 rangi marki marki- rangi marki marki- rati marki marki marki 53.00 rangi markimma	walkin (1) ber genale mitjele procession walkin web 11,80% bas webstere	Million 41 with the second se	mictur 15 mictur 15 har province mittane province any province any province any har province more	ataliari 40 contaitori 40 contaitori etterneti anteriori etternetie menti 27.43/30 contaitori etternetie	water 17 ten genet state ten genet state ten see ten see ten see ten state ten see ten see	And their 11 which are an and the second se	Institut D tore and the second states when a second	Replace M. Constructions in the second secon	million Ti Contents and Contents and the re- charged and the re- perty at 1998	August March
*Amag (2007) 202 one of the discretion of (1). The second fit (2007) *Amag and the second f	1. April (1975) X. Description of the second state of the secon	<ol> <li>"Specify Control State of the Control of the Control</li></ol>	[1] "State of all states in the second states of	<ul> <li>Sense (1997) De-</li></ul>	<ul> <li>Annal Alexis Lin</li> <li>Annal Alexis Links, A. Stelland, M. S. Stelland, M. Stelland, S. Stelland, M. Stelland, S. Stelland, M. Stelland, M. Stelland, S. Stelland, M. Stelland, S. Stelland, M. Stelland, S. Stelland, S. Stelland, M. Stelland, S. Stelland, S. Stelland, M. Stelland, S. Stelland,</li></ul>	Arman Garden La Description of the second state of the second state of the second state of the second second state of the Description of the second state of the Description of the second state of the Description of the second second state Description of the second second state Description of the second second state Description of the second	<ul> <li>Anna (1997) L.</li> <li>Anna (1997) L.<td><ul> <li>Arme (Lipe", di);</li> <li>Arme (Lipe", di);&lt;</li></ul></td><td>Stand (Sight E) State of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state second state of the second state of the second state second state of the second state of the second state second state of the second state of the second state second state of the second state of the second state second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the</td><td><ul> <li>Annal, Shafe's He.</li> <li>Deriv (and Statement?)</li> <li>Annal, Shafe's He.</li> <li>Manage, Mathematical States of the states o</li></ul></td><td><ul> <li>Apack starts (A).</li> <li>Para (and presented) (L. Spanner).</li> <li>Parate presented (L. Spanner).</li> </ul></td><td><ul> <li>Annalyzing Collage</li> <li>Annalyzing Collag</li></ul></td></li></ul>	<ul> <li>Arme (Lipe", di);</li> <li>Arme (Lipe", di);&lt;</li></ul>	Stand (Sight E) State of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state second state of the second state of the second state second state of the second state of the second state second state of the second state of the second state second state of the second state of the second state second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the	<ul> <li>Annal, Shafe's He.</li> <li>Deriv (and Statement?)</li> <li>Annal, Shafe's He.</li> <li>Manage, Mathematical States of the states o</li></ul>	<ul> <li>Apack starts (A).</li> <li>Para (and presented) (L. Spanner).</li> <li>Parate presented (L. Spanner).</li> </ul>	<ul> <li>Annalyzing Collage</li> <li>Annalyzing Collag</li></ul>
Hold, Sandy, Scieff, S. 20, Nagarathy (1997). "Advances in Academic Scieffic and accounter of the Scieffic Scieffic Academic Scieffic S	Construction of the second sec	Prof. Inch. and Y. Hu, "Hands", "Color, "Anticastic "Color," (2017) "Note pro- training of anticipation of the "State", Anticipation (2017) "Anticipation of the state of the Distribution of the state of the state of the state of the Distribution of the state of the state of the state of the Distribution of the state of the state of the state of the Distribution of the state of the state of the state of the Distribution of the state of the state of the state of the Distribution of the state of the state of the state of the state of the Distribution of the state of the state of the state of the state of the Distribution of the state of the state of the state of the state of the Distribution of the state of the Distribution of the state of th	[14] J. S. B. (1997). HL, "Statement Weight, "galaxies," Samp, "State, pp., 17, 2017; S. S. S. S. S. S. S	Fred Back, eds., 10, Passan Nation, Sciences, Context Nation, Sciences, Context Nation, Sciences, Context Nation, National Sciences, National National Context National Conte	(1) Construction of the Construction of the Construction of the	Construction of the August States of August	and the start of the "cases" of the start of the start of the start	Control Locales, and Y. Hu, "Control of Control of American Control of Control Control of American Control of Control Control of Control of Control of Control Control of Control of Control of Control Control of Control of Control of Control of Control Control of Control of Co	Terri Land, and J. H., Danest, S. & Sang, S. & Sang, S. & Sang, F. & Sang, S. & Sang,	<ul> <li>Data America's Dis- based and a series and integrated based and a series of the series Provide and the series of the series of the Provide America and the Provide America</li></ul>	Part West, '19, '19, '19, '19, '19, '19, '19, '19	New York, N. S. Star, N. S. Star, N. S. Star, N. S. Star, S. S. Star, S. S. Star, N. S. Star, N. S. Star, S.
n, "Anderse", r. Anderse Jahren (Schermerse") Sont Distance (Schermerse")	The second	Part - Sustainable 1. National Product Company of Company - Product of Supports, One Company -	No. "Weldwart 1. Network Addar robrid (Lamona ", "Social Colorderia Ver-Chap-	nt, maanan i hore nava waa kinaan hala shaqada ba daga	<ol> <li>Stalande, S.</li> <li>Stalande, S.</li> <li>Stalander, S.</li> <li>Stalander, S. Stalander, S. Stalander, S. Stalander, S. Stalander, S. Stalander, S. Stalander, Stalander,</li></ol>	Part, Penderson - 2, Decemptarian autore (directory), Policy (constrained autore), pro-	ent, "enabling" (c), Sector station (control control of Transformation (control of the control o	Strate and Strategies, Control of Strategies, States and Strategies, Strate	Personal and a second s	Weit, Sant Disco, J., "Count proce- tic," Supported by Probabilitation offent, Scientific,"1	<ul> <li>State (1997) S. State (1997)</li> <li>State (1997) S.</li> <li>State (1997) S.</li> <li>State (1997) State (1997)</li> </ul>	Post Aug Denvis 1, State al., Senatorial 1, Post her states spage (1)

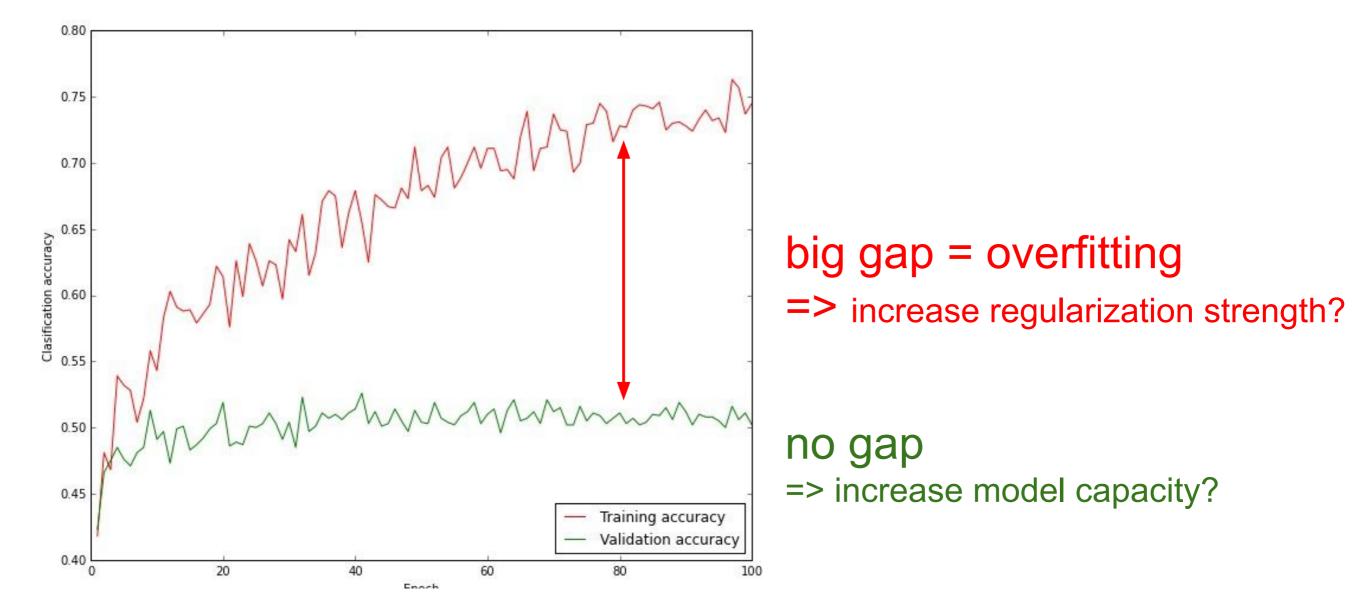


#### Monitor and visualize the loss curve









Track the ratio of weight updates / weight magnitudes:

```
# assume parameter vector W and its gradient vector dW
param_scale = np.linalg.norm(W.ravel())
update = -learning_rate*dW # simple SGD update
update_scale = np.linalg.norm(update.ravel())
W += update # the actual update
print update_scale / param_scale # want ~1e-3
```

ratio between the updates and values: ~ 0.0002 / 0.02 = 0.01 (about okay) want this to be somewhere around 0.001 or so

# Summary We looked in detail at:

# TLDRs

- Activation Functions (use ReLU)
- Data Preprocessing (images: subtract mean)
- Weight Initialization (use Xavier/He init)
- Batch Normalization (use)
- Babysitting the Learning process
- Hyperparameter Optimization (random sample hyperparams, in log space when appropriate)

# Next time: Training Neural Networks, Part 2

- Parameter update schemes
- Learning rate schedules
- Gradient checking
- Regularization (Dropout etc.)
- Evaluation (Ensembles etc.)
- Transfer learning / fine-tuning