

NEURAL NETWORKS

Deep learning = Deep neural networks =
neural networks

GMM/Gaussian fitting

- How many parameters are there in a 2x2 covariance matrix?
- How many data points do you need to estimate a 2v2 covariance matrix (at least)?

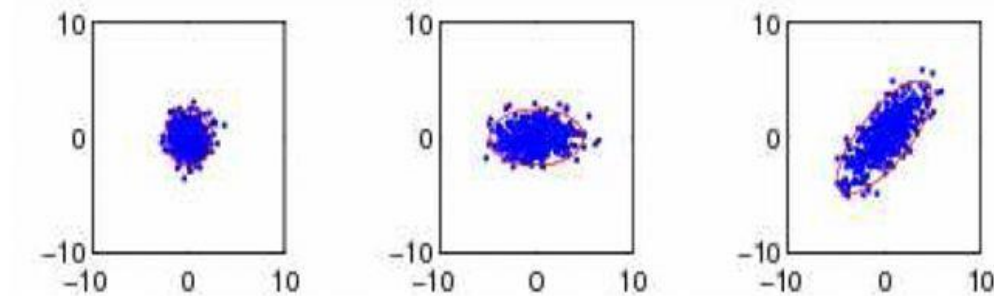
$$m_j = \frac{1}{N} \sum_n w_{n,j}$$

$$\vec{\mu}_j = \frac{\sum_n w_{n,j} \vec{x}_n}{\sum_n w_{n,j}}$$

$$\Sigma_j = \frac{\sum_n w_{n,j} (\vec{x}_n - \vec{\mu}_j)(\vec{x}_n - \vec{\mu}_j)^T}{\sum_n w_{n,j}}$$

Many forms of covariance matrix

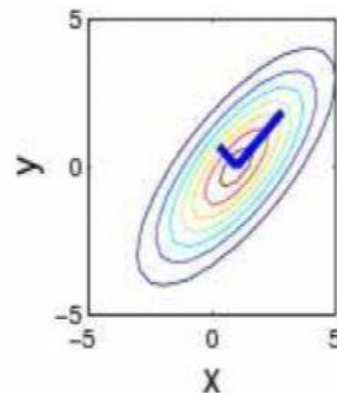
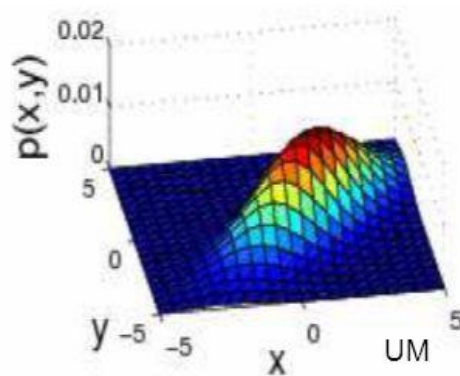
Spherical, diagonal, full covariance



$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$



Whitening and GMM fitting

- Spherical/diagonal covariance are less prone to overfitting (less parameters)
- Data are not always distributed like that
- Use **whitening** to help make them spherical/diagonal distributed
 - Still not quite true, but oh well

PCA as a feature normalization technique

- We said it's good to normalize features to $[0,1]$, $[-1,1]$, $N(0,1)$.
 - Normalize each dimension independently
- Can we do better?

Whitening (PCA)

- Find the project along the dimensions that has the highest variance in the data
- Let Σ be the covariance matrix. E is the matrix of eigenvectors, and D has eigenvalues along the diagonal. With eigen decomposition:
 - Note for covariance matrices, $E^t = E^{-1}$

$$\Sigma = EDE'$$

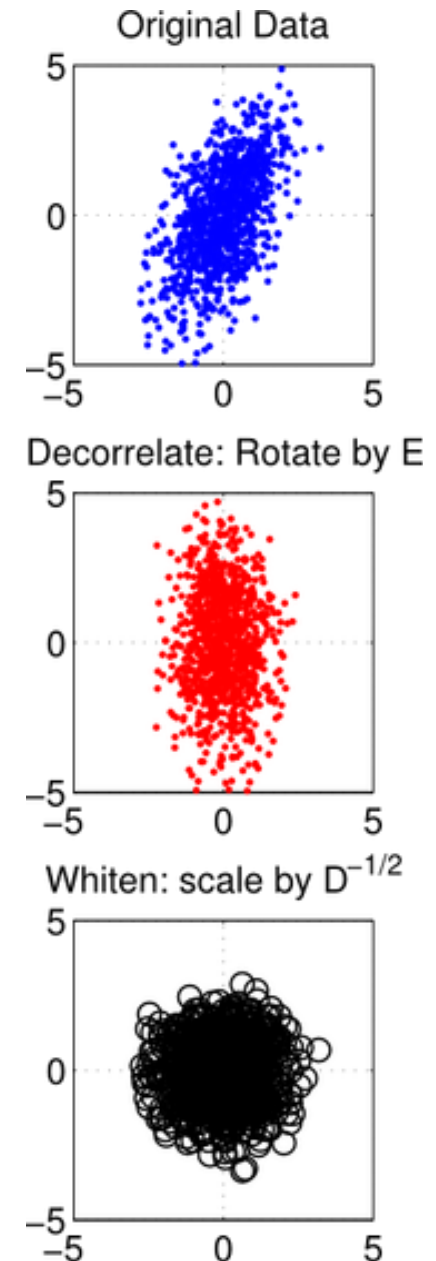
The diagram illustrates the equation $\Sigma = EDE'$ using matrix visualizations. On the left, matrix E is represented by a large square bracket containing three vertical grey rectangles, each labeled with a white 'e'. In the middle, matrix D is represented by a large square bracket containing a diagonal sequence of white 'd's. On the right, matrix E' is represented by a large square bracket containing three horizontal grey rectangles, each labeled with a white 'e'.

Whitening (PCA)

- Whitening decorrelates and scale

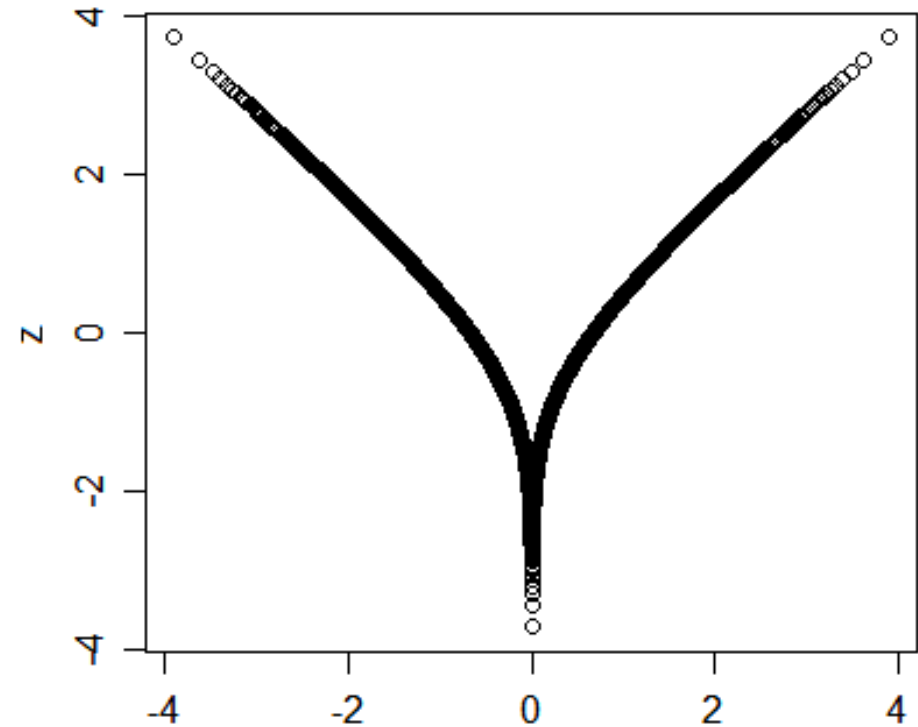
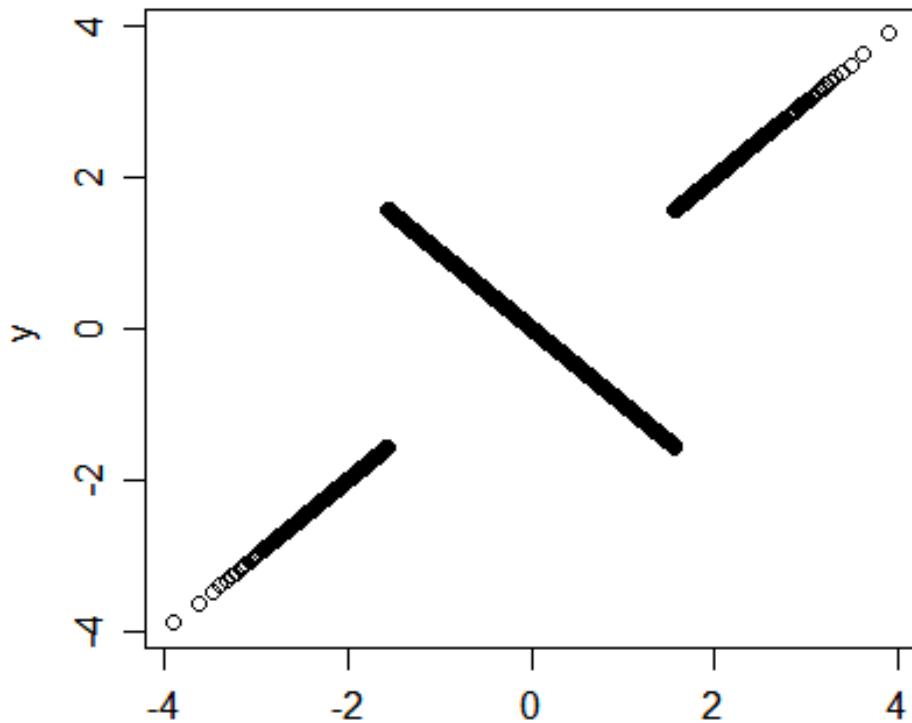
$$Y = D^{-1/2} E' X$$

- In homework we only use the decorrelates part (rotation)
- Some models prefer features to be of equal variance (SVMs, Neural networks)
- Scale according to the inverse of the variance.
- This decorrelates the features (on the global scale)
 - Correlations can still exist given class
 - Uncorrelated-ness does not imply independence
 - We usually assume so though
- It is often a good idea to **normalize the variance of each feature first before doing PCA dimensionality reduction**



Uncorrelated but dependence

- Below are example of variables with 0 correlation but definitely not independent



For multivariate normal distribution, uncorrelated implies independence



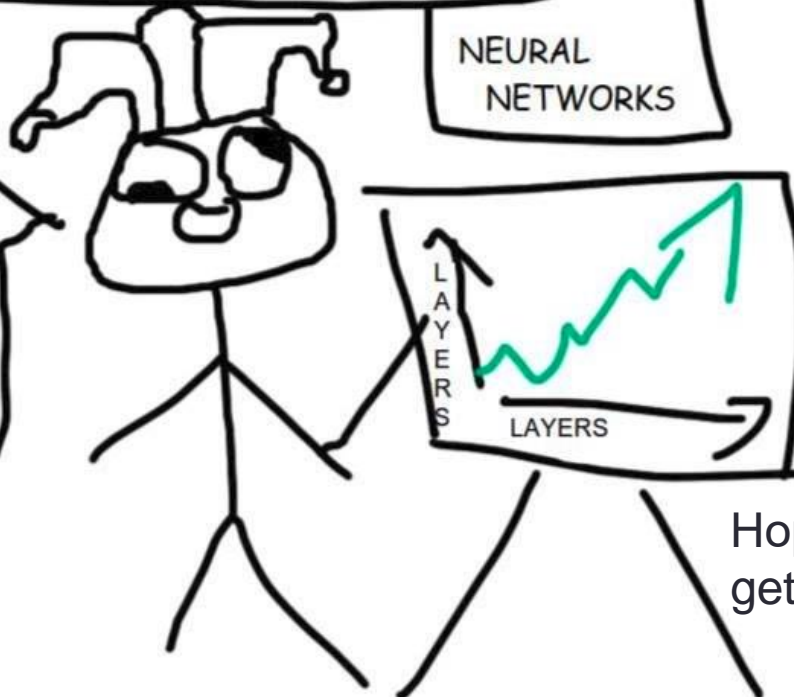
STATISTICAL LEARNING

Gentlemen, our learner overgeneralizes because the VC-Dimension of our Kernel is too high, Get some experts and minimize the structural risk in a new one. Rework our loss function, make the next kernel stable, unbiased and consider using a soft margin



NEURAL NETWORKS

STACK
MORE
LAYERS



Hopefully this won't be all you get from this class

DNNs (Deep Neural Networks)

- Why deep learning?
- Greatly improved performance in ASR and other tasks (Computer Vision, Robotics, Machine Translation, NLP, etc.)
- Surpassed human performance in many tasks

Task	Previous state-of-the-art	Deep learning (2012)	Deep learning (2019)
TIMIT	24.4%	20.0%	13.8%
Switchboard	23.6%	16.1%	5.0%
Google voice search	16.0%	12.3%	4.9%
MOOC (Thai)	38.7%		19.6%

Google's AlphaGo Defeats Chinese Go Master in Win for A.I.

[点击查看本文中文版](#)

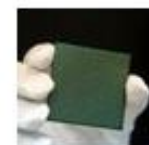
By PAUL MOZUR MAY 23, 2017



RELATED COVERAGE



A.I. Is
Repla



China
FEB. 3,



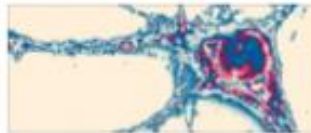
THE FU
The P



Master
Goog

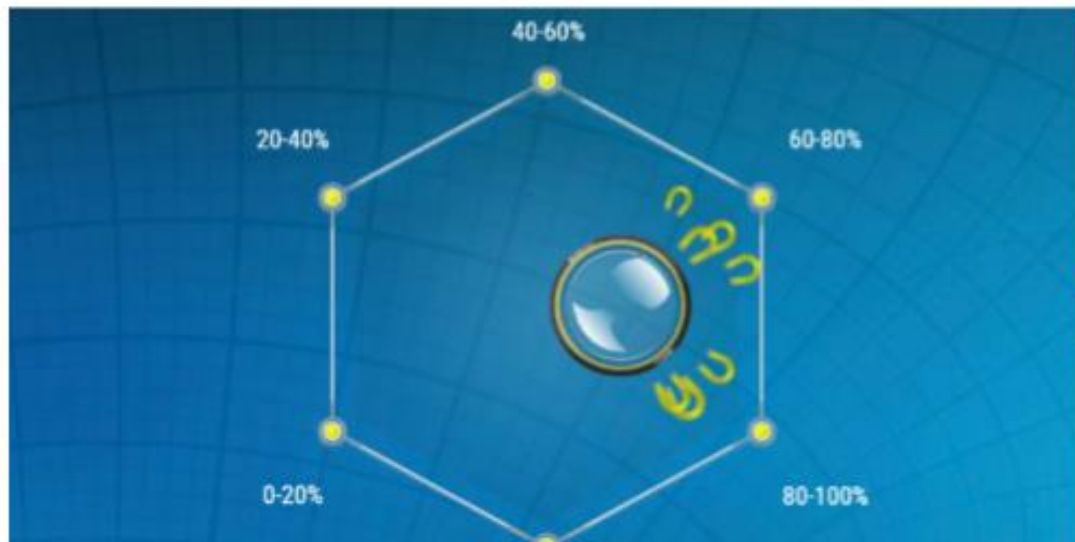
<https://www.nytimes.com/2017/05/23/business/google-deepmind-alphago-go-champion-defeat.html>

Artificial swarm intelligence diagnoses pneumonia better than individual computer or doctor



Hear from leading minds and find inspiration for your own research

by Fan Liu — September 27, 2018 0



Courtesy of Unanimous AI

✈ Bangkok to Tokyo

THB 4,030

BOOK NOW

✈ Bangkok to Hangzhou

THB 4,030

BOOK NOW

eZoiC

report this

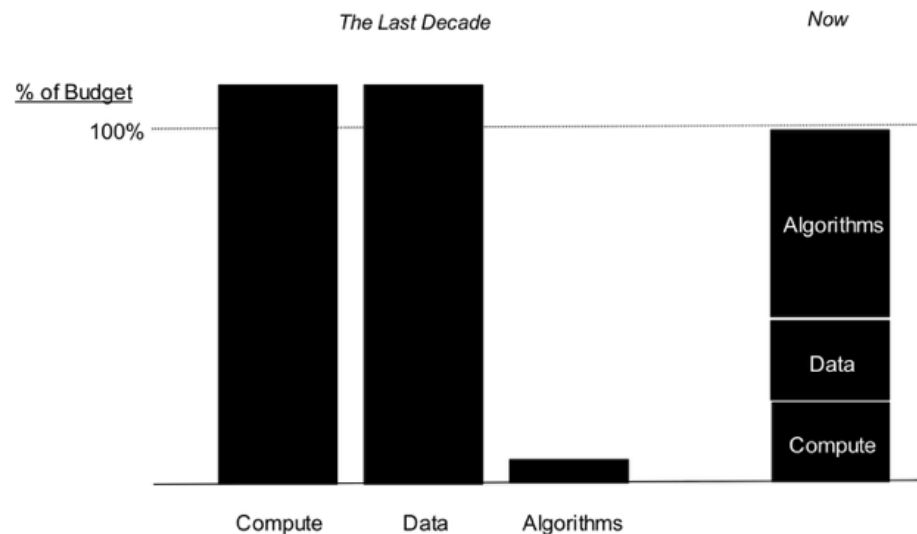
Popular Posts

Artificial swarm intelligence diagnoses pneumonia better than individual computer or

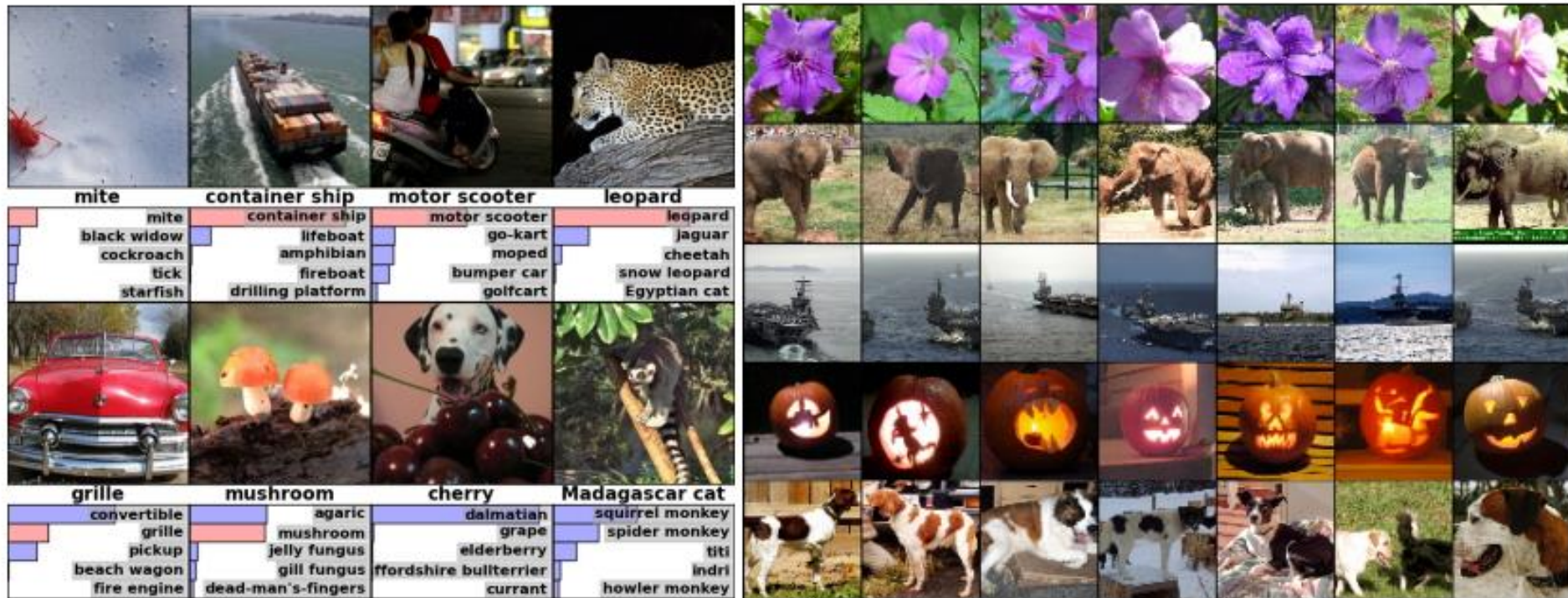
<https://www.stanforddaily.com/2018/09/27/artificial-swarm-intelligence-diagnoses-pneumonia-better-than-individual-computer-or-doctor/>

Why now

- Neural Networks has been around since 1990s
- **Big data** – DNN can take advantage of large amounts of data better than other models
- **GPU** – Enable training bigger models possible
- **Deep** – Easier to avoid bad local minima when the model is large



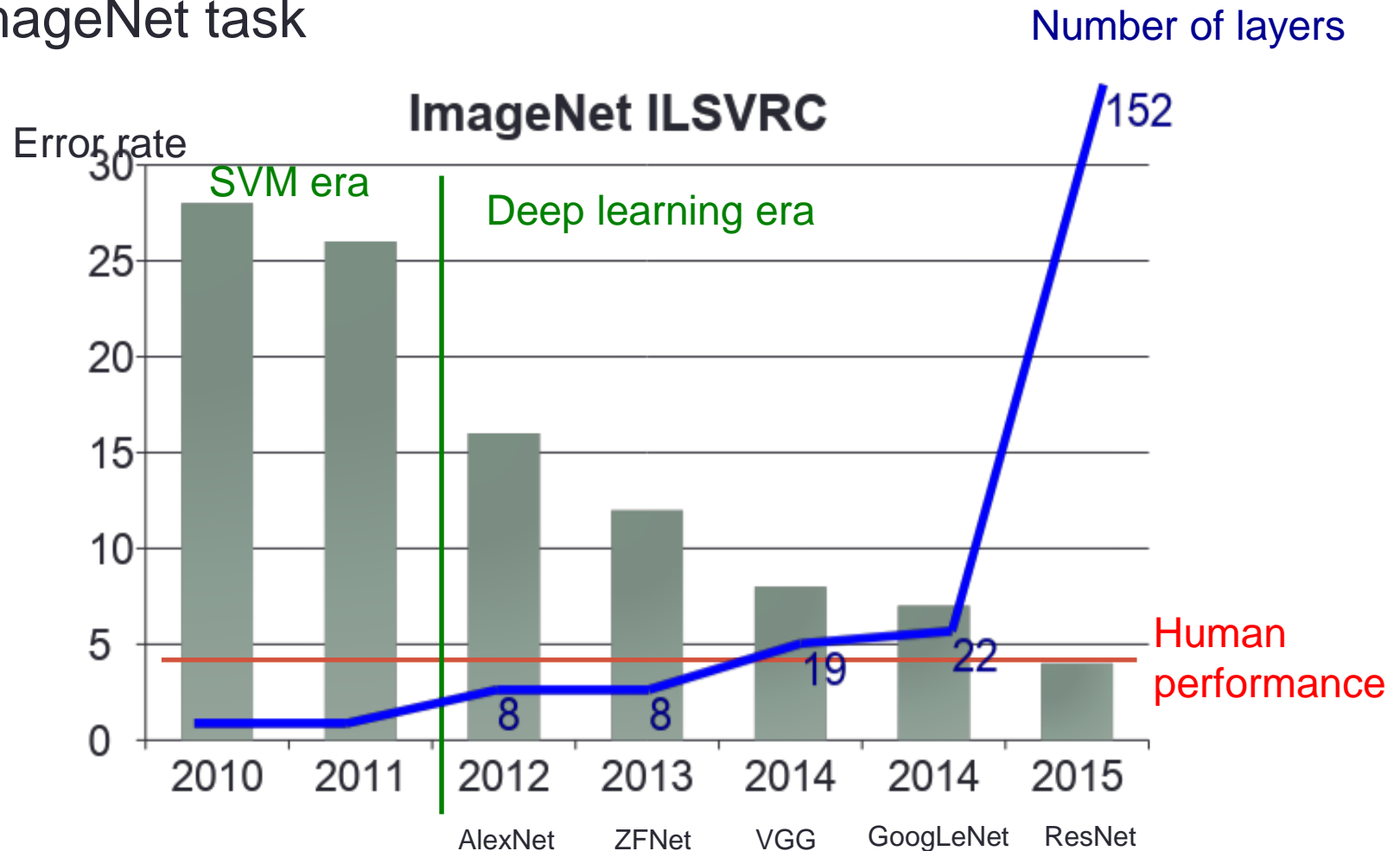
ImageNet - Object classification



Alex, Krizhevsky, Imagenet classification with deep convolutional neural networks, 2012

Wider and deeper networks

- ImageNet task





Statistics > Machine Learning

Dynamical Isometry and a Mean Field Theory of CNNs: How to Train 10,000-Layer Vanilla Convolutional Neural Networks

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel S. Schoenholz, Jeffrey Pennington

(Submitted on 14 Jun 2018)

In recent years, state-of-the-art methods in computer vision have utilized increasingly deep convolutional neural network architectures (CNNs), with some of the most successful models employing hundreds or even thousands of layers. A variety of pathologies such as vanishing/exploding gradients make training such deep networks challenging. While residual connections and batch normalization do enable training at these depths, it has remained unclear whether such specialized architecture designs are truly necessary to train deep CNNs. In this work, we demonstrate that it is possible to train vanilla CNNs with ten thousand layers or more simply by using an appropriate initialization scheme. We derive this initialization scheme theoretically by developing a mean field theory for signal propagation and by characterizing the conditions for dynamical isometry, the equilibration of singular values of the input-output Jacobian matrix. These conditions require that the convolution operator be an orthogonal transformation in the sense that it is norm-preserving. We present an algorithm for generating such random initial orthogonal convolution kernels and demonstrate empirically that they enable efficient training of extremely deep architectures.

Comments: ICML 2018 Conference Proceedings

Subjects: **Machine Learning (stat.ML)**; Machine Learning (cs.LG)Cite as: [arXiv:1806.05393](#) [[stat.ML](#)](or [arXiv:1806.05393v1](#) [[stat.ML](#)] for this version)

Submission history

From: Samuel Schoenholz [[view email](#)]

[v1] Thu, 14 Jun 2018 07:04:15 GMT (6734kb,D)

[Which authors of this paper are endorsers?](#) | [Disable MathJax](#) ([What is MathJax?](#))Link back to: [arXiv](#), [form interface](#), [contact](#).

Download:

- [PDF](#)
- [Other formats](#)

[\(license\)](#)

Current browse context:

stat.ML

[< prev](#) | [next >](#)[new](#) | [recent](#) | [1806](#)

Change to browse by:

[cs](#)[cs.LG](#)[stat](#)

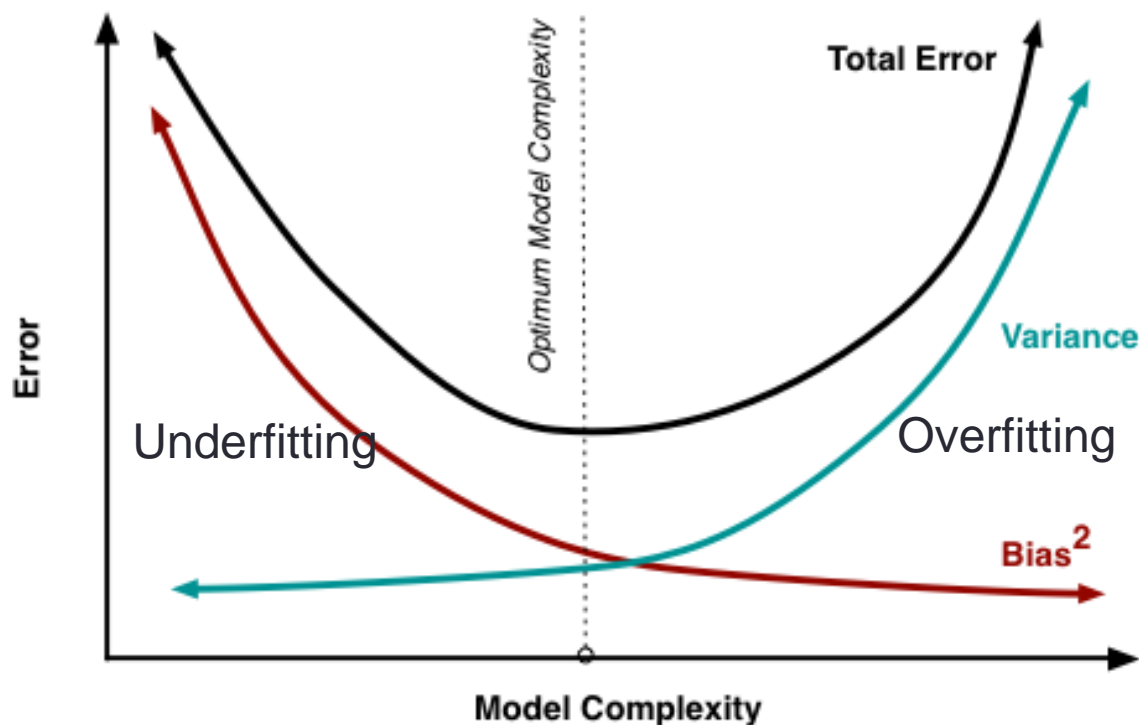
References & Citations

- [NASA ADS](#)

Bookmark ([what is this?](#))

Bias-Variance Underfitting-Overfitting

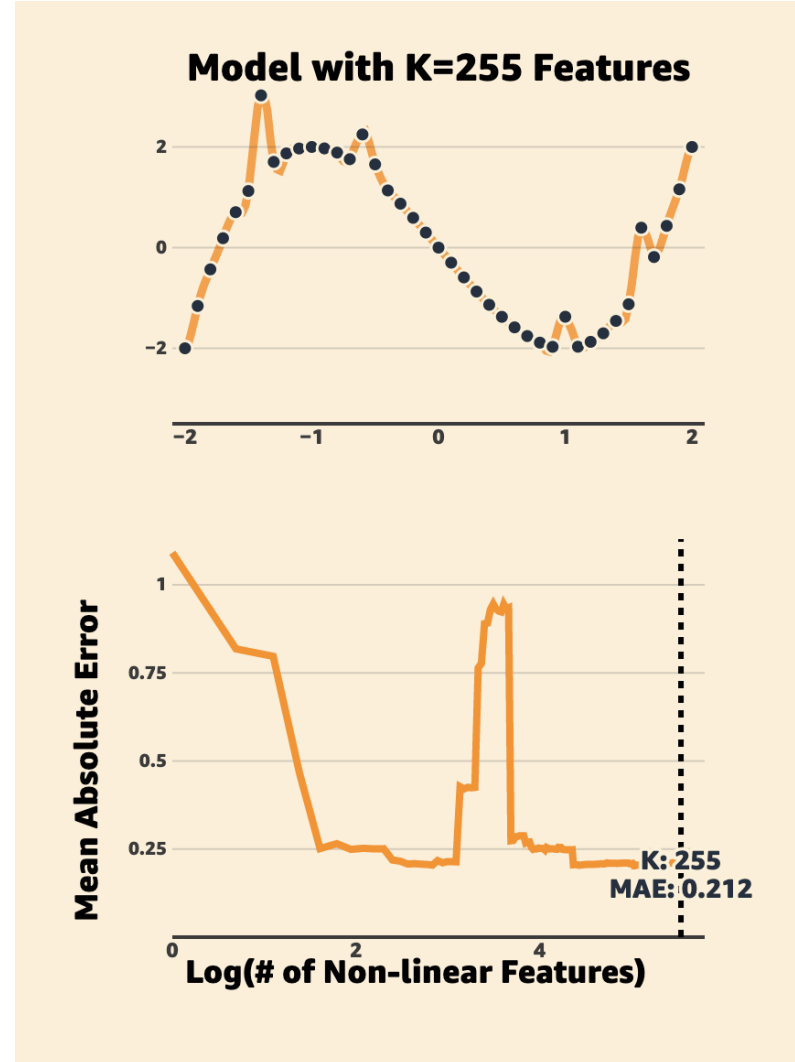
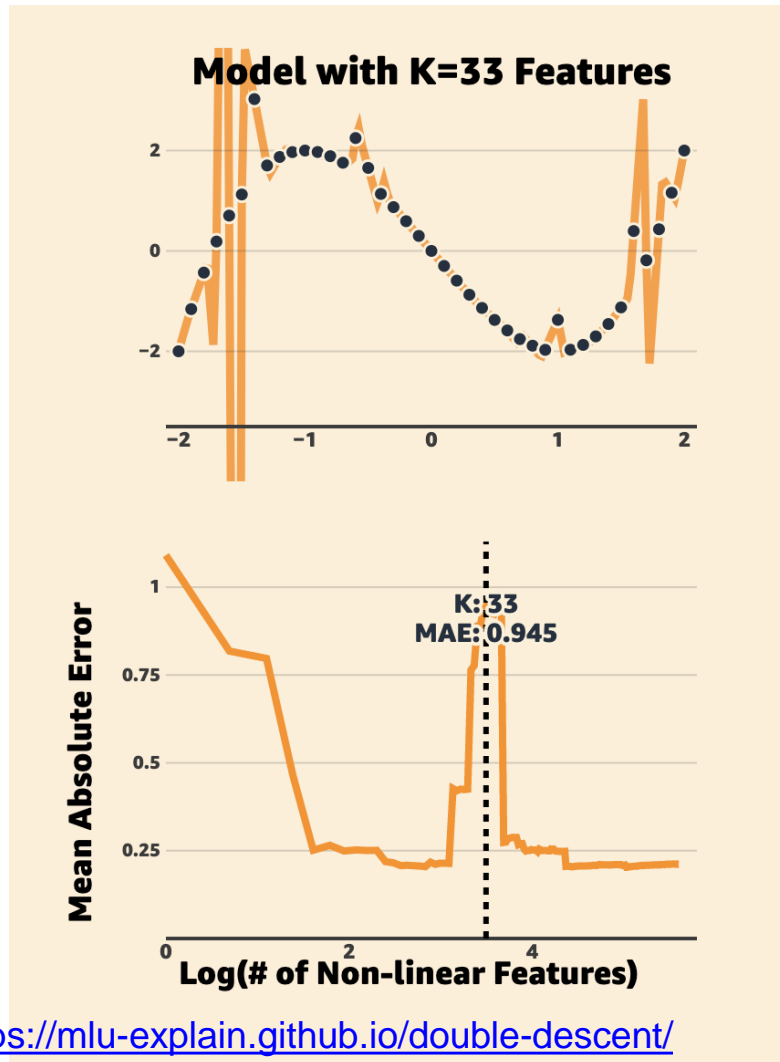
- Usually if you try to reduce the bias of your model, the variance will increase, and vice versa.
- Called the bias-variance trade-off



The double descent problem



The double descent problem

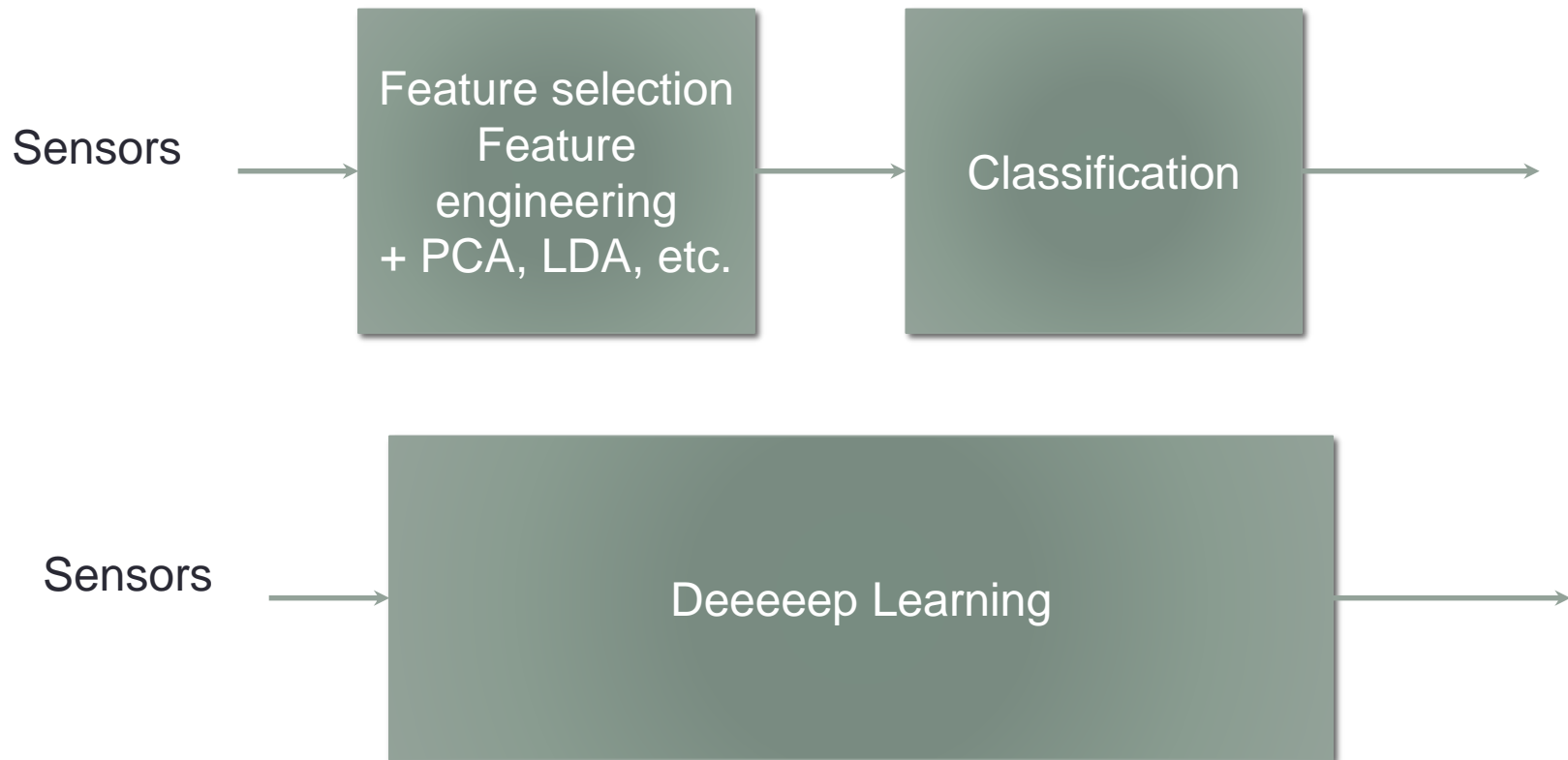


Inductive bias and ML

- The **inductive bias** of a learning algorithm is a set of assumptions that the algorithm used to generalize to new inputs.
http://www.cs.cmu.edu/~tom/pubs/NeedForBias_1980.pdf
- Your choice of model forces a certain type of behavior
 - Tells the model how to “overfit” the training data
- Putting inductive bias into deep learning model is easier than other models
 - ~~• Domain knowledge to construct features~~
 - Domain knowledge to encourage certain learning behaviors



Traditional VS Deep learning

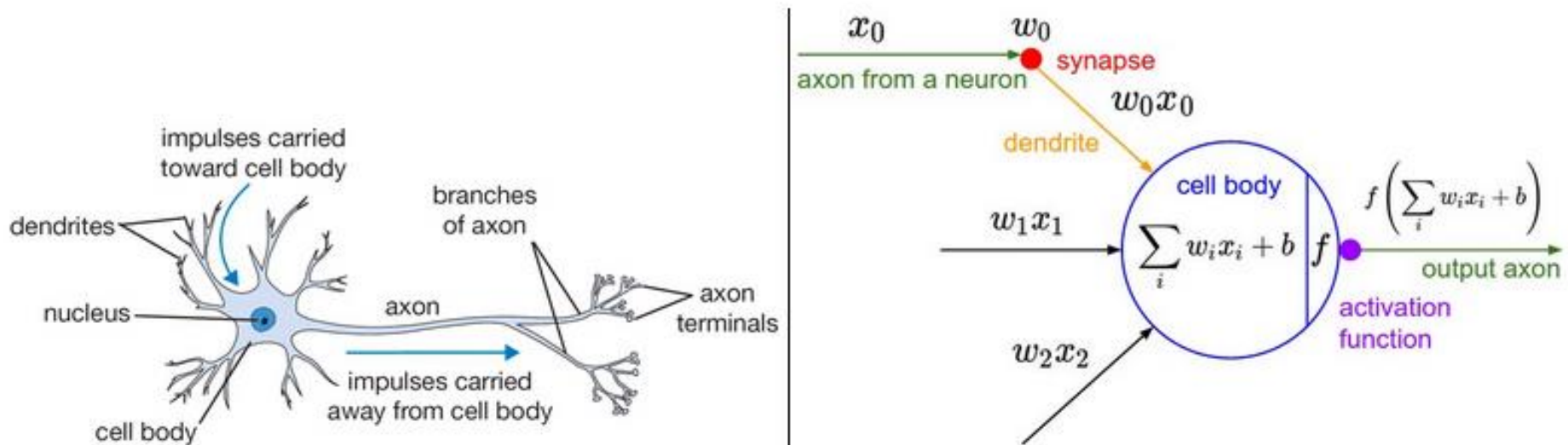


Neural networks

- Fully connected networks
 - Neuron
 - Non-linearity
 - Softmax layer
- DNN training
 - Loss function and regularization
 - SGD and backprop
 - Learning rate
 - Overfitting – dropout, batchnorm
- CNN, RNN, LSTM, GRU <- Next class

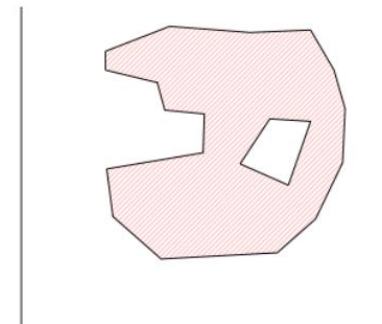
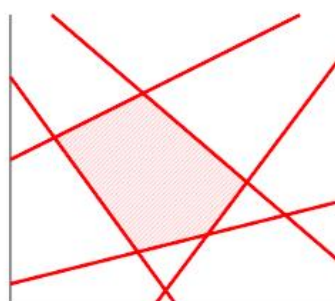
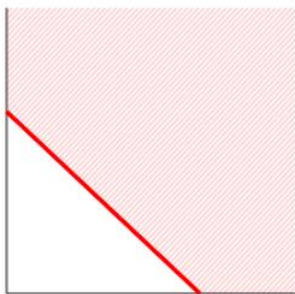
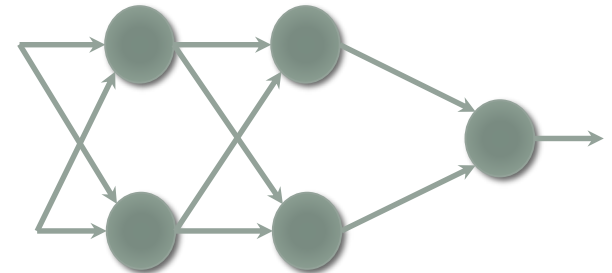
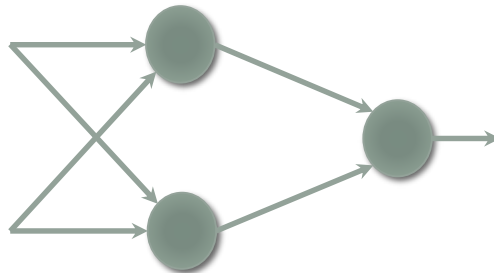
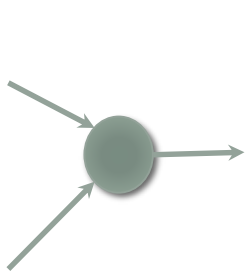
Fully connected networks

- Many names: feed forward networks or deep neural networks or multilayer perceptron or artificial neural networks
- Composed of multiple neurons



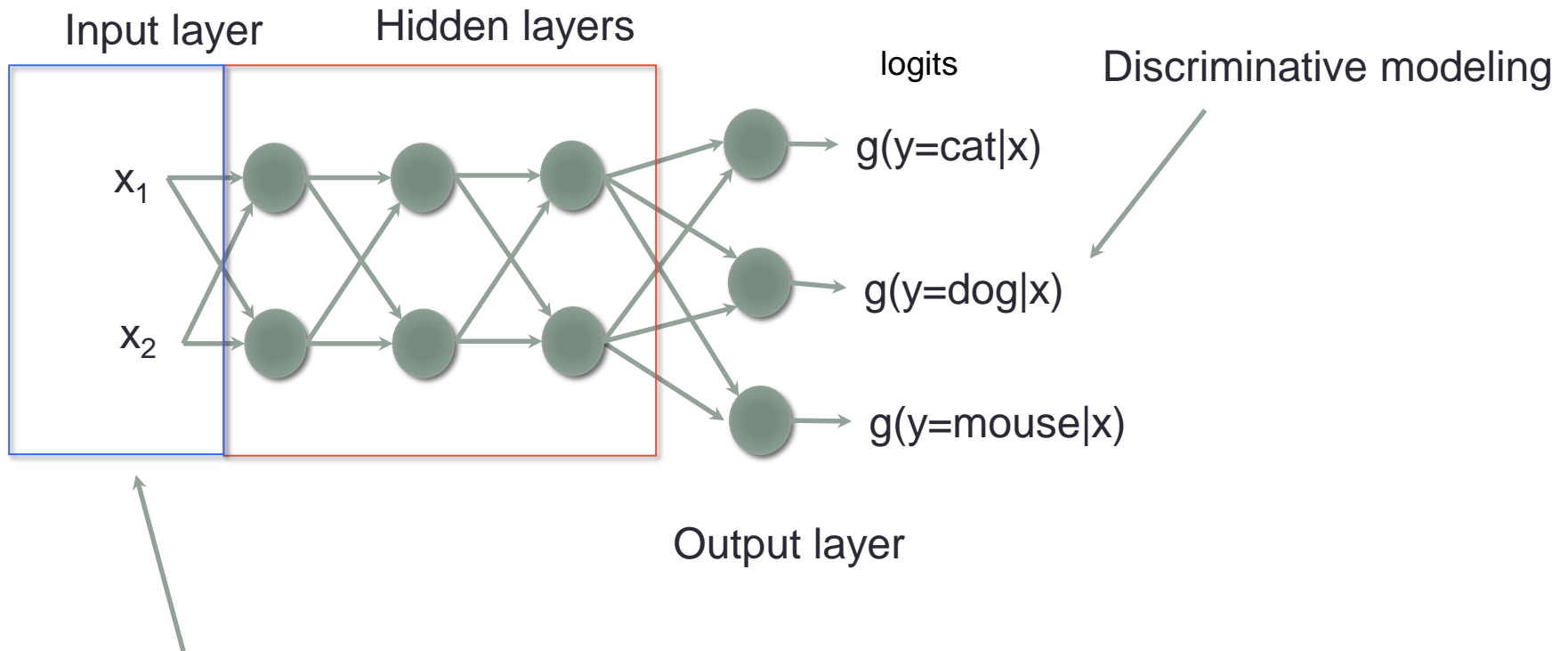
Combining neurons

- Each neuron splits the feature space with a hyperplane
- Stacking neuron creates more complicated decision boundaries
- More powerful but prone to overfitting



Terminology

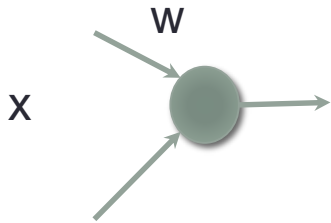
Deep in Deep neural networks means many hidden layers



Input should be scaled to have zero mean unit variance

Projections and Neural network weights

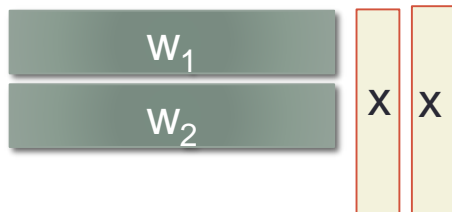
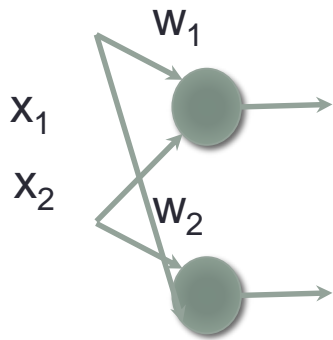
- $w^T x$



Projections and neural network weights

- $W^T[x_1, x_2]$

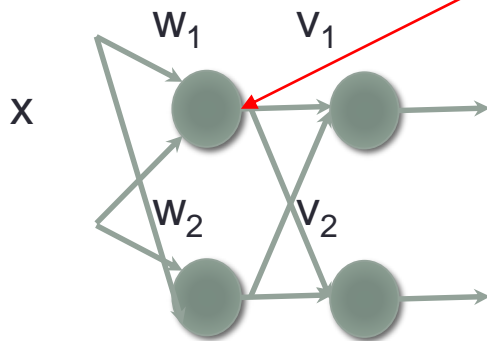
You can pack multiple inputs together to do one single matrix multiply



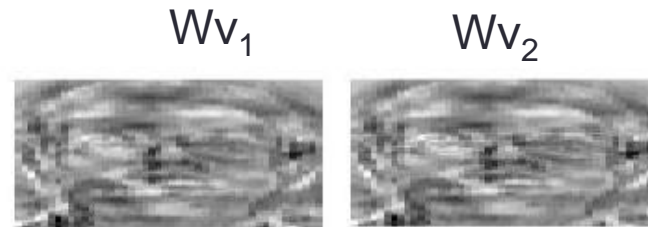
Neural network layer acts as nonlinear feature transform

- $W^T x$

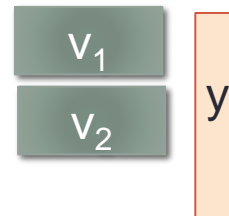
Without the nonlinearity the two matrices combine into one operation



$$\text{fisher projection} = V^T W^T x \\ = (WV)^T x$$

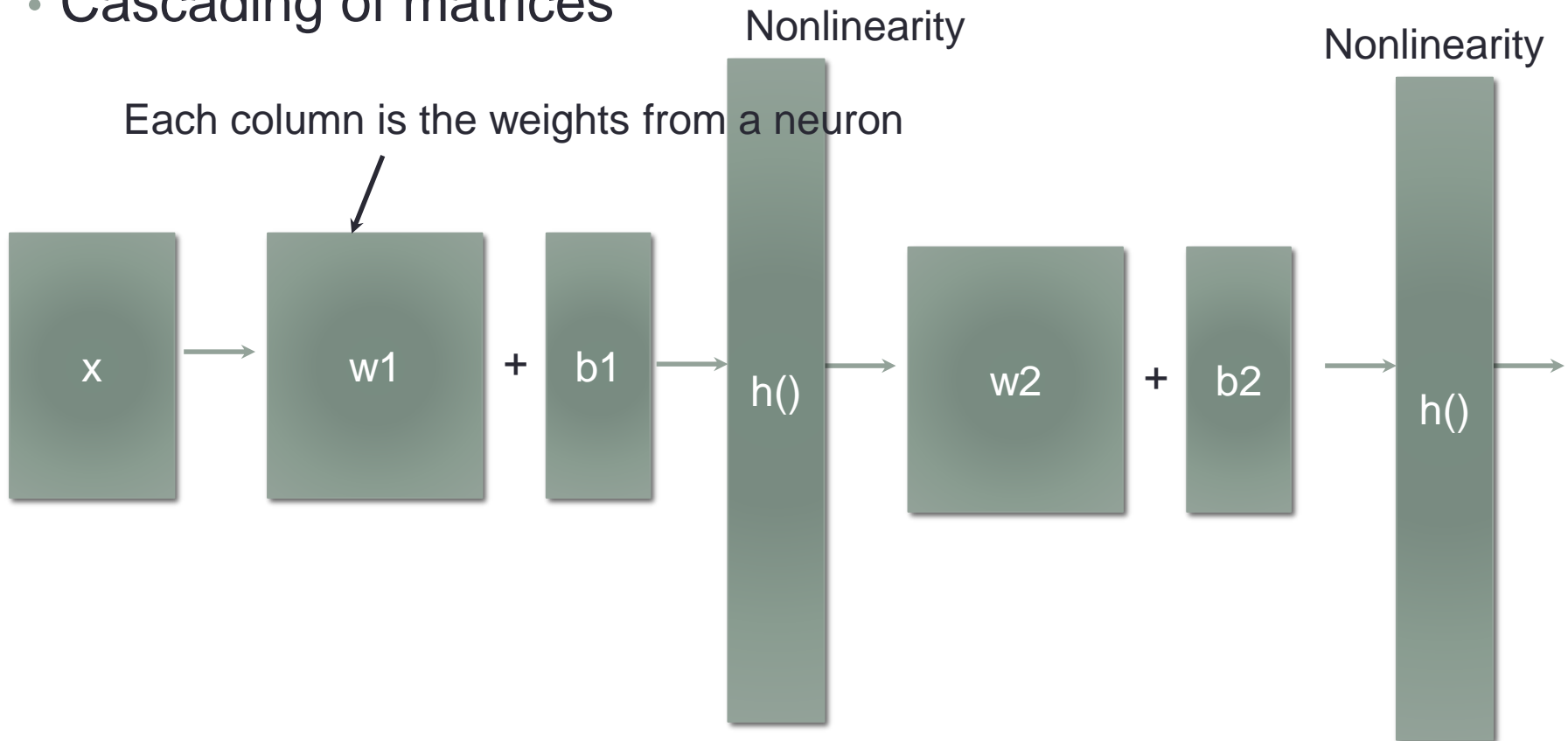


LDA projections



More linear algebra

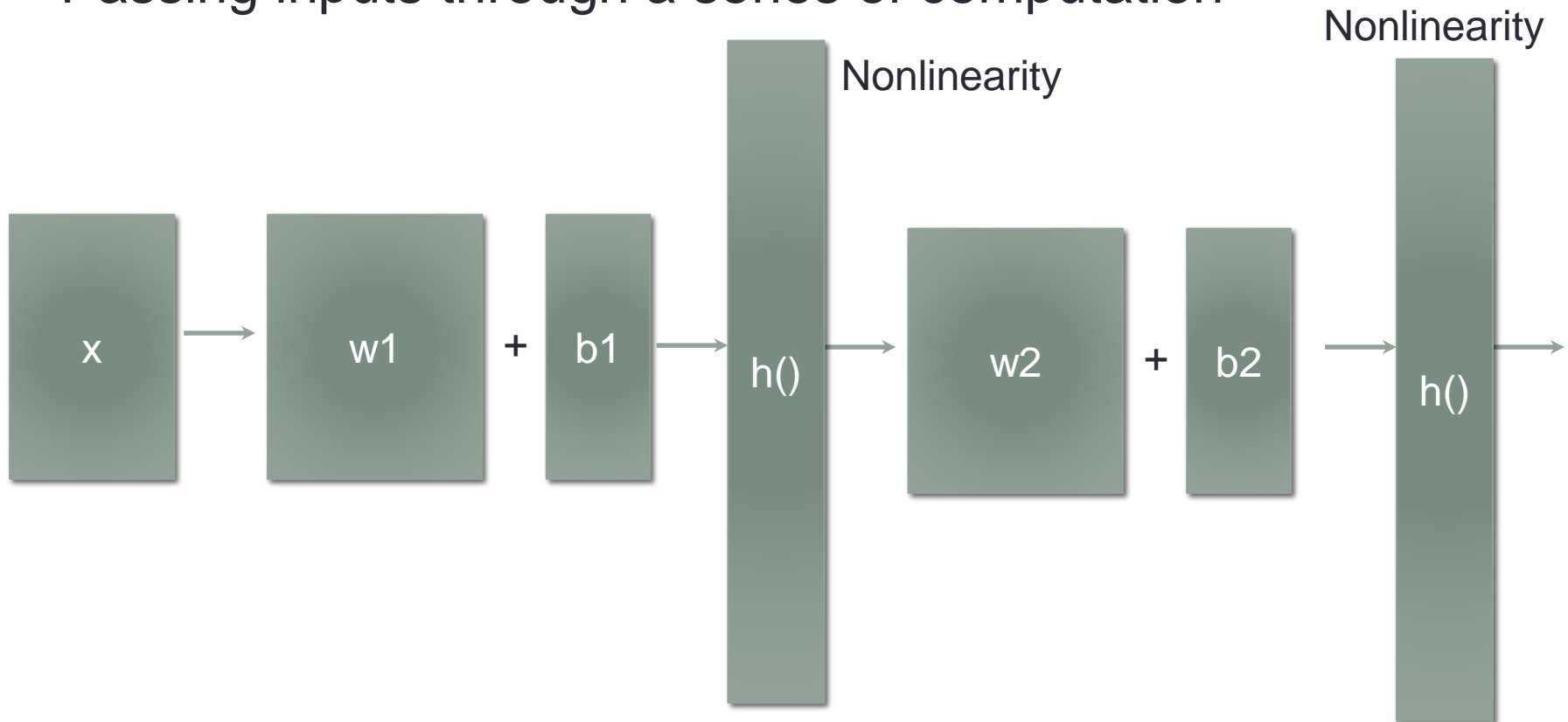
- Cascading of matrices



$$h(W_2^T h(W_1^T X + b_1) + b_2)$$

Computation graph

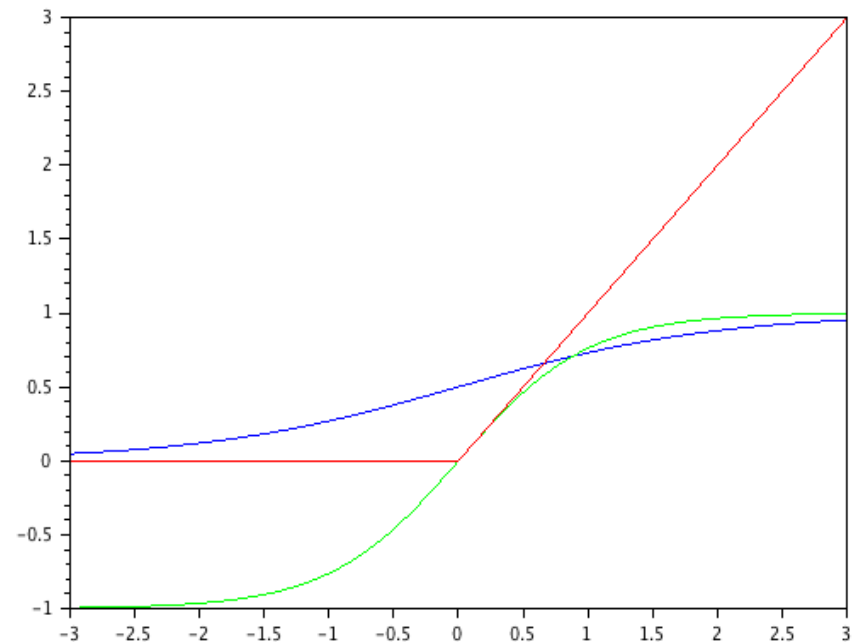
- Passing inputs through a series of computation



$$h(W_2^T h(W_1^T X + \mathbf{b}_1) + \mathbf{b}_2)$$

Non-linearity

- The Non-linearity is important in order to stack neurons
- Sigmoid or logistic function
- \tanh
- Rectified Linear Unit (ReLU)
 - LeakyReLU, ELU, PreLU
- Sigmoid Linear Units (SiLU)
 - Swish, Mish, GELU
- Most popular is ReLU and its variants



Non-linearity

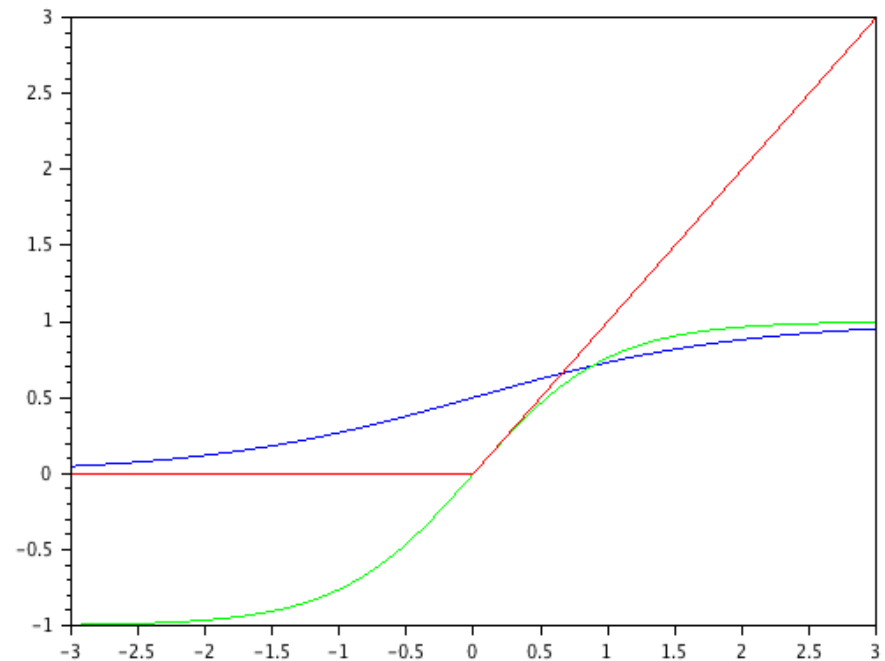
- Sigmoid $\frac{1}{1 + e^{-x}}$

- tanh

$$\tanh(x)$$

- Rectified Linear Unit (ReLU)

$$\max(0, x)$$



Swish (Sigmoid type)

Found through reinforcement learning to be the best **general** non-linearity

$$x \cdot \text{sig}(\beta x)$$

sig refers to a sigmoid function

Beta is a learnable parameter
or can be set to 1 for slightly
worse performance

Beta \rightarrow inf, then Swish \rightarrow ReLu

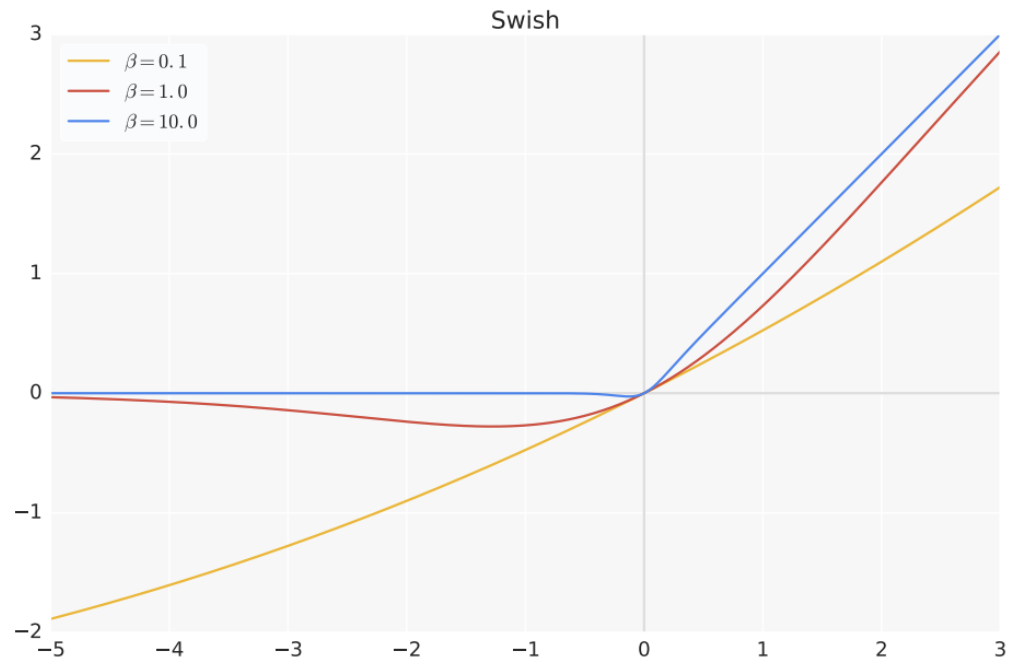


Figure 4: The Swish activation function.

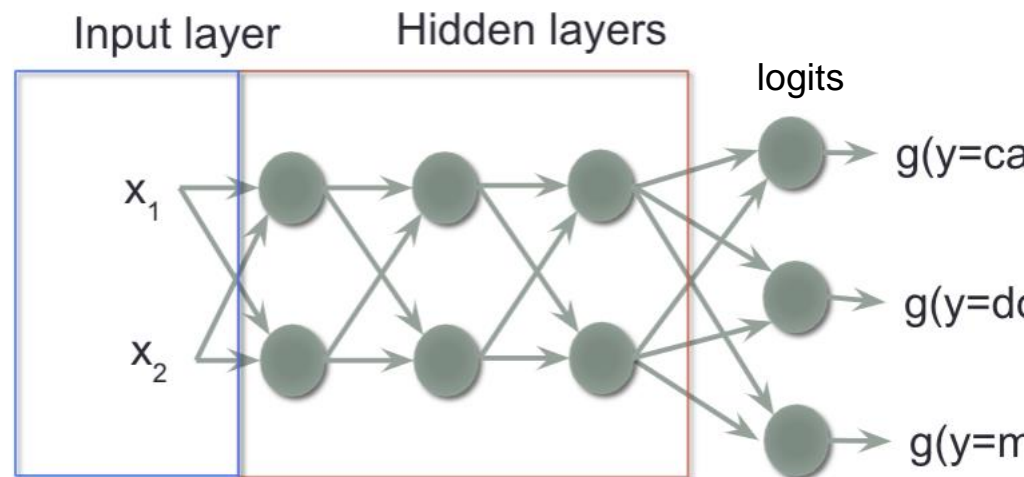
<https://arxiv.org/abs/1710.05941>

Proven theoretically to be optimal

[Expectation propagation: a probabilistic view of Deep Feed Forward Networks](#)

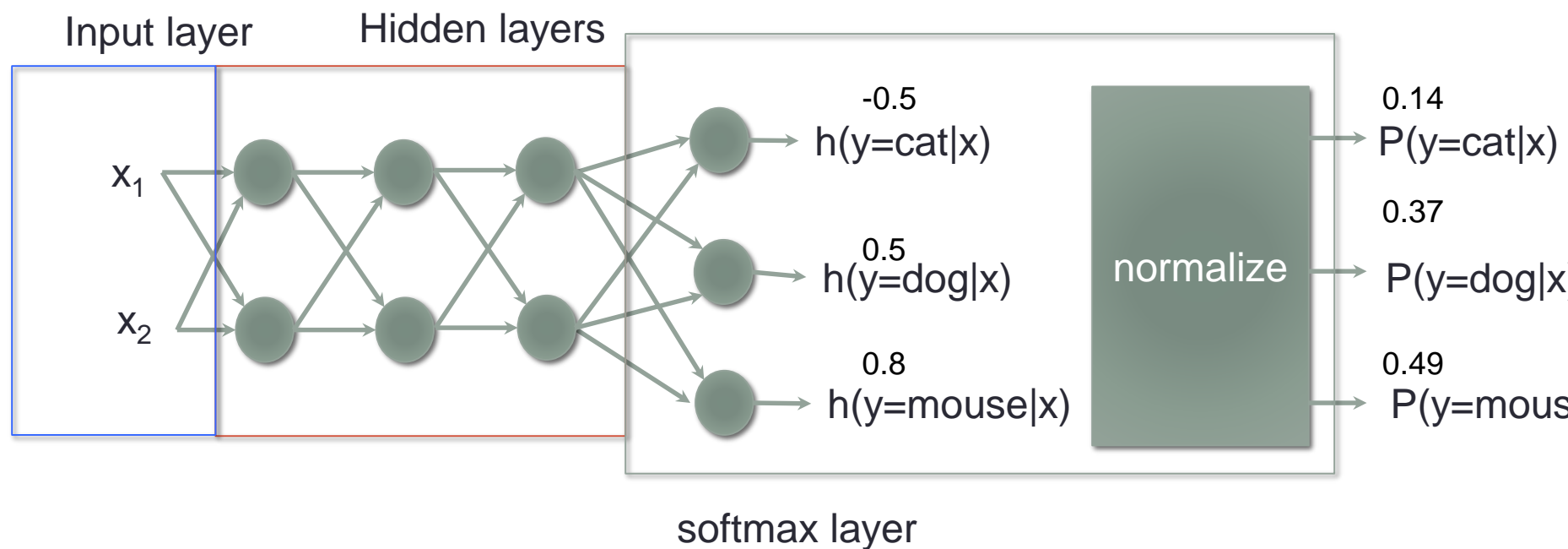
Output layer – Softmax layer

- We usually want the output to mimic a probability function ($0 \leq P \leq 1$, sums to 1)
- Current setup has no such constraint
- The current output should have highest value for the correct class.
 - Value can be positive or negative number
- Takes the exponent
- Add a normalization



Softmax layer

$$P(y = j|x) = \frac{e^{h(y=j|x)}}{\sum_y e^{h(y|x)}}$$

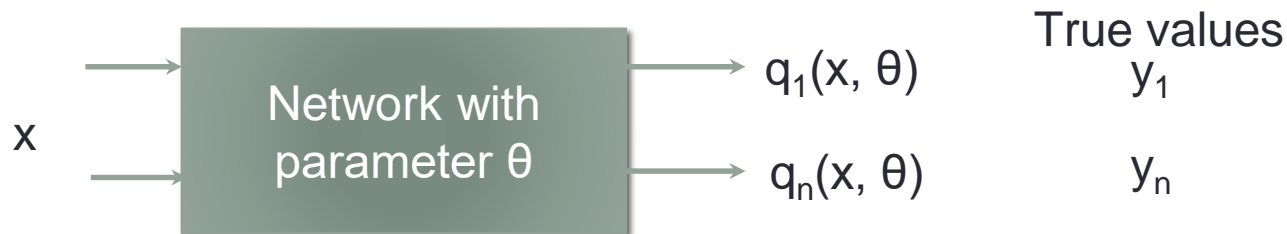


Neural networks

- Fully connected networks
 - Neuron
 - Non-linearity
 - Softmax layer
- DNN training
 - Loss function and regularization
 - SGD and backprop
 - Learning rate
 - Overfitting – dropout, batchnorm
- CNN, RNN, LSTM, GRU <- Next class

Objective function (Loss function)

- Can be any function that summarizes the performance into a single number
- Cross entropy
- Sum of squared errors



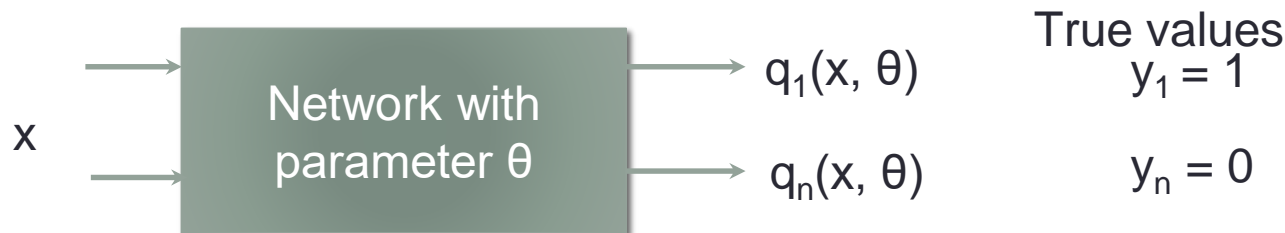
Cross entropy loss

- Used for softmax outputs (probabilities), or classification tasks

$$L = -\sum_n y_n \log q_n(x, \theta)$$

Note Eq is calculated on one sample
Sum over whole batch for total loss

- Where y_n is 1 if data x comes from class n
0 otherwise
- L only has the term from the correct class
- L is non negative with highest value when the output matches the true values, a “loss” function



Cross entropy loss & Logarithmic Loss (log loss)

- Minimizing the CE can be considered as the maximizing the log likelihood

$$L = -\sum_n y_n \log q_n(x, \theta)$$

- Where y_n is 1 if data x comes from class n
0 otherwise

- For binary class: $L(x_n) = \begin{cases} -\log(h(x_n)) & , \text{ if } y_n = 1 \\ -\log(1-h(x_n)) & , \text{ if } y_n = 0 \end{cases}$

$$L = [y_n \log(h(x_n))] + [(1 - y_n) \log(1 - h(x_n))]$$

Same as log likelihood of logistic regression

Negative in front because we are minimizing the loss vs maximizing the probability

$$p(y \mid x; \theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y}$$

Probabilistic view of Logistic Regression

- Let's assume, we'll classify as 1 with probability in accordance to the output of

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$P(y = 1 \mid x; \theta) = h_{\theta}(x)$$

$$P(y = 0 \mid x; \theta) = 1 - h_{\theta}(x)$$

or

$$p(y \mid x; \theta) = (h_{\theta}(x))^y (1 - h_{\theta}(x))^{1-y}$$

Other views of CE loss

Relationship between Entropy, CE, and KL Divergence

<https://www.youtube.com/watch?v=ErfnhcEV1O8>

A Short Introduction to Entropy, Cross-Entropy and KL-Divergence

Kullback-Leibler Divergence

Aurélien Geron, 2018

p = true distribution

q = predicted distribution

True Distribution (p)	1%	1%	4%	4%
00	$1/2^2 = 25\%$	$1/2^2 = 25\%$	$1/2^3 = 12.5\%$	$1/2^3 = 12.5\%$
10	$1/2^4 = 6.25\%$	$1/2^4 = 6.25\%$	$1/2^5 = 3.125\%$	$1/2^5 = 3.125\%$

KL Divergence:

$$D_{KL}(p || q) = H(p, q) - H(p)$$

8:22 / 10:40

Notes on CE loss

- If classes is ordered, it might not be ideal
 - Class 0: Perfect, Class 1: Good, Class 2: Average, Class 3: Bad
 - Soln: use Squared EMD loss <https://arxiv.org/abs/1611.05916>
- You can have better loss but worse accuracy
 - Soln: monitor accuracy as well as loss
- Leads to overconfidence
 - Soln: label smoothing, <https://paperswithcode.com/method/label-smoothing>, Calibration <https://arxiv.org/abs/1706.04599>

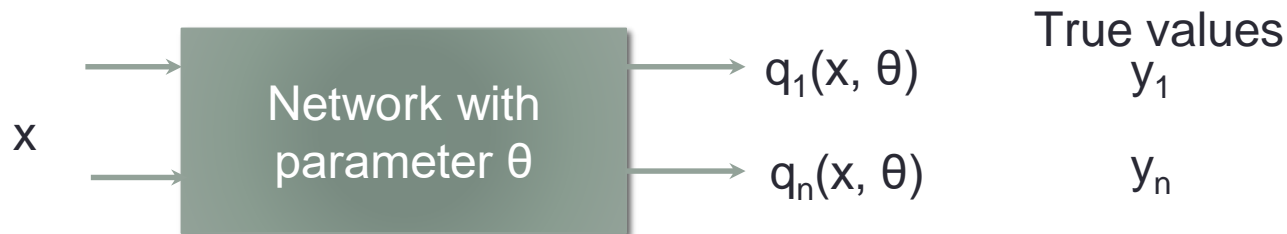
Sum of squared errors (MSE)

- Used for any real valued outputs such as regression

$$L = \frac{1}{2} \sum_n (y_n - q_n(x, \theta))^2$$

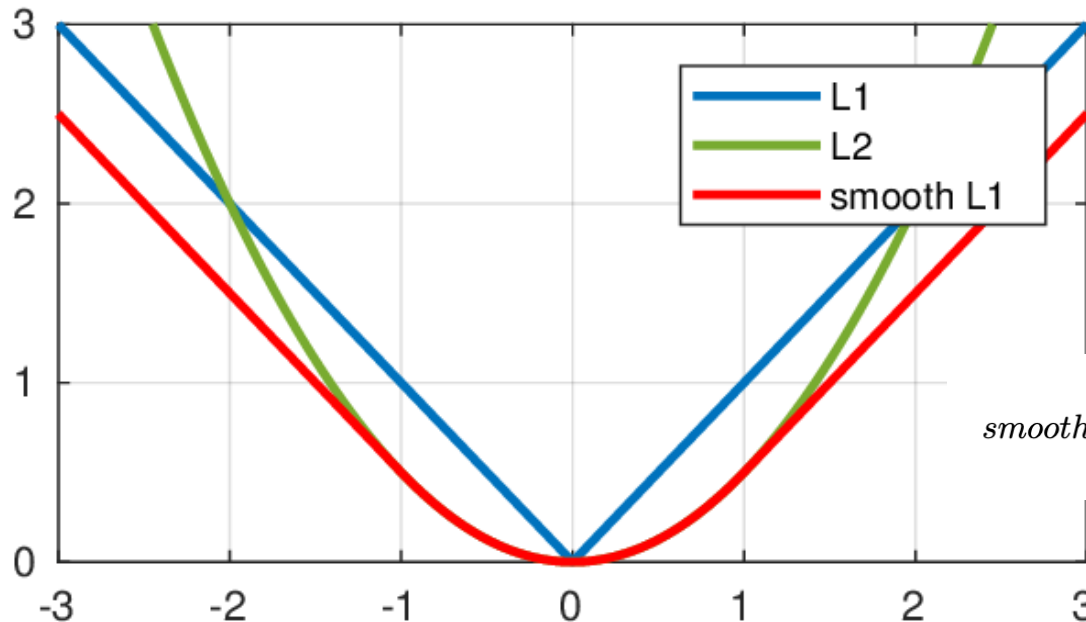
Note Eq is calculated on one sample
Sum over whole batch for total loss

- Nonnegative, the better the lower the loss



Notes on MSE loss

- L1 vs L2 loss
 - L1 robust to outliers
 - L2 easier to optimize (smooth gradient)
- Smoothed L1 combination of L1 and L2

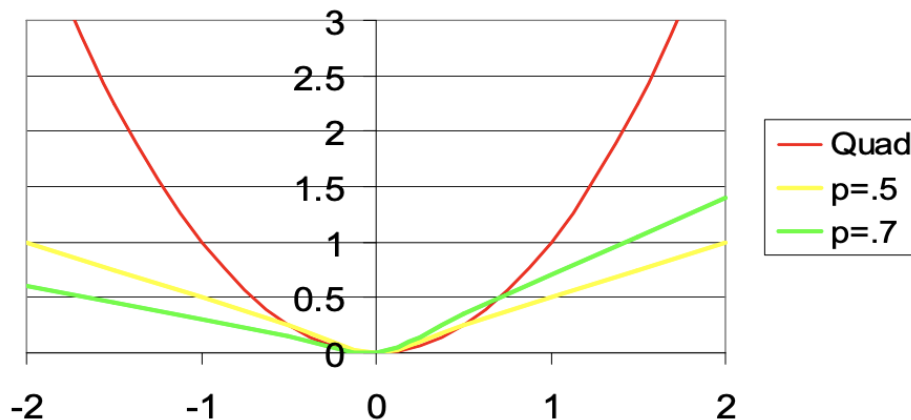


$$\text{smooth}_{L1}(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| < 1 \\ |x| - \frac{1}{2} & \text{otherwise} \end{cases}, \quad (4)$$

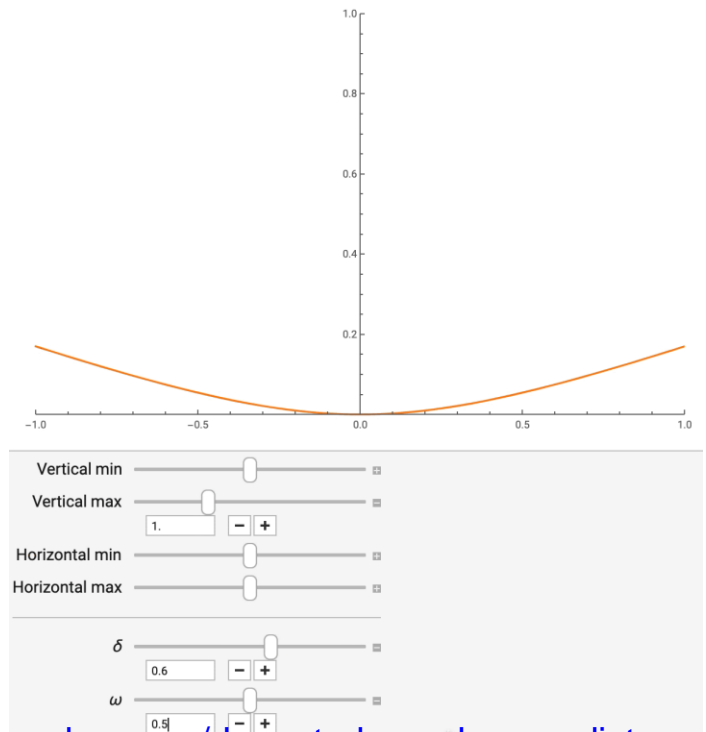
Notes on MSE loss

- Loss weights underestimate and overestimate to be the same
Soln: asymmetric loss (Quantile loss – L1, Huber loss – smoothed L1)

Absolute Loss vs. Quadratic Loss over errors



<https://www.bauer.uh.edu/rsusmel/phd/ec1-25.pdf>



<https://eng.uber.com/deepeta-how-uber-predicts-arrival-times/>

Regularization

There are two main approaches to regularize neural networks

- Explicit regularization
Deals with the loss function
- Implicit regularization
Deals with the network

Regularization in one slide

- What?
 - Regularization is a method to lower the model variance (and thereby increasing the model bias)
- Why?
 - Gives more generalizability (lower variance)
 - Better for lower amounts of data (reduce overfitting)
- How?
 - Introducing regularizing terms in the original loss function
 - Can be anything that make sense
 - $\mathbf{w}^T \mathbf{w} + C \sum \epsilon_i$
 - MAP estimate is MLE with regularization (the prior term)

Famous types of regularization

- L1 regularization: Regularizing term is a sum
 - $\mathbf{x}^T \mathbf{w} + C \sum |w_i|$
- L2 regularization: Regularizing term is a sum of squares
 - $\mathbf{x}^T \mathbf{w} + C \sum w_i^2$

Regularization in neural networks

L2

- We want to improve generalization somehow.
- Observation, models are better when the weights are spread out (no peaky weights).
 - Try to use every part of the model.
- Add a cost if we put some value to the weights
- Regularized loss = Original loss + $C \sum w^2$
- We sum the square of weights of the whole model
- C is a hyperparameter weighting the regularization term

Regularization in neural networks

L1

- We want to improve generalization somehow.
- Observation, models behave better when we force the weights to be sparse.
 - Sparse means many weights are zero or close to zero
 - Force the model to focus on only important parts
 - Less prone to noise
- Add a cost if we put some value to the weights
- Regularized loss = Original loss + $0.5 C \sum |w|$
- We sum the absolute weights of the whole model
- 0.5 is for prettiness when we take derivative
- C is a hyperparameter weighting the regularization term

Numerical example

Training data $x = [3, 2, 1]$ $y = 10$, regression task

Objective: $10 = w_1 * 3 + w_2 * 2 + w_3 * 1$ Find w_1, w_2, w_3

w_1	w_2	w_3	L1 loss	L2 loss
3	0.25	0.5	3.75	9.31
5	-2	-1	8	30
3.33	0	0	3.33	11.11
2.14	1.42	0.71	4.29	7.14

L1 does feature selection (makes most numbers 0)

L2 spreads the numbers (no 0)

L1 L2 regularization notes

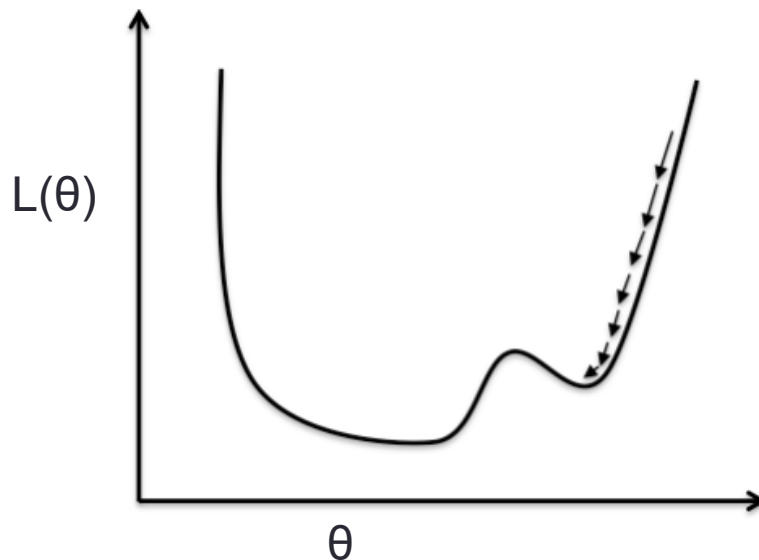
- Can use both at the same time
 - People claim L2 is superior
 - Weight decay can be considered as L2 regularization (under simple optimization techniques)
 - <https://arxiv.org/abs/1711.05101>
- I found them useless in practice for deep neural networks
 - Works when data is small (transfer learning)
- Other regularization methods exist (we will go over these later)

Neural networks

- Fully connected networks
 - Neuron
 - Non-linearity
 - Softmax layer
- DNN training
 - Loss function and regularization
 - SGD and backprop
 - Learning rate
 - Overfitting – dropout, batchnorm
- CNN, RNN, LSTM, GRU <- Next class

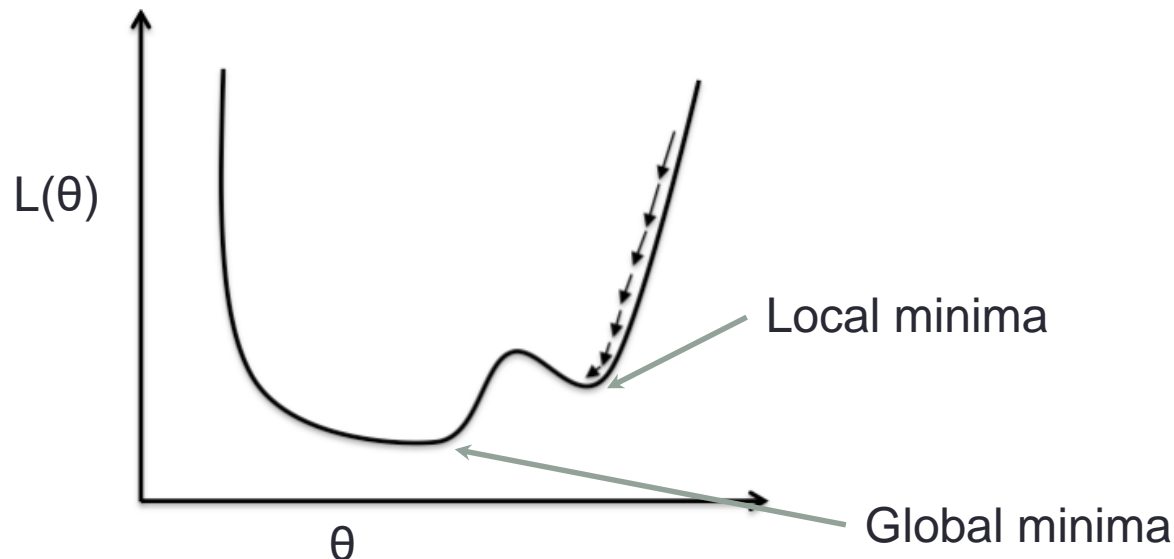
Minimization using gradient descent

- We want to minimize L with respect to θ (weights and biases)
 - Differentiate with respect to θ
 - Gradients passes through the network by Back Propagation



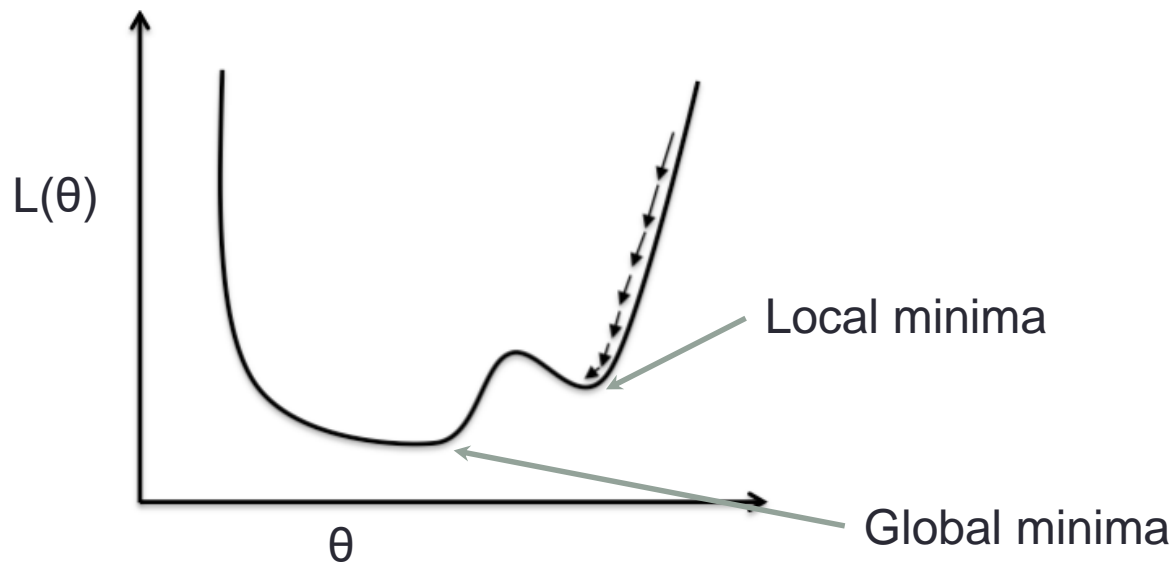
Deep vs Shallow

- The loss function of neural network is non-convex (and non-concave)
 - Local minimas can be avoided with convexity
 - Linear regression, SVM are convex optimization
 - Convexity gives easier training
 - Does not imply anything about the generalization of the model
 - The loss is optimized by the training set



Deep vs Shallow

- If deep, most local minimas are the global minima!
 - Always a way to lower the loss in the network with millions of parameters
 - Enough parameters to remember every training examples
 - Does not imply anything about generalization



Differentiating a neural network model

- We want to minimize loss by gradient descent
- A model is very complex and have many layers! How do we differentiate this!!?



Back propagation

- Forward pass
 - Pass the value of the input until the end of the network
- Backward pass
 - Compute the gradient starting from the end and passing down gradients using chain rule

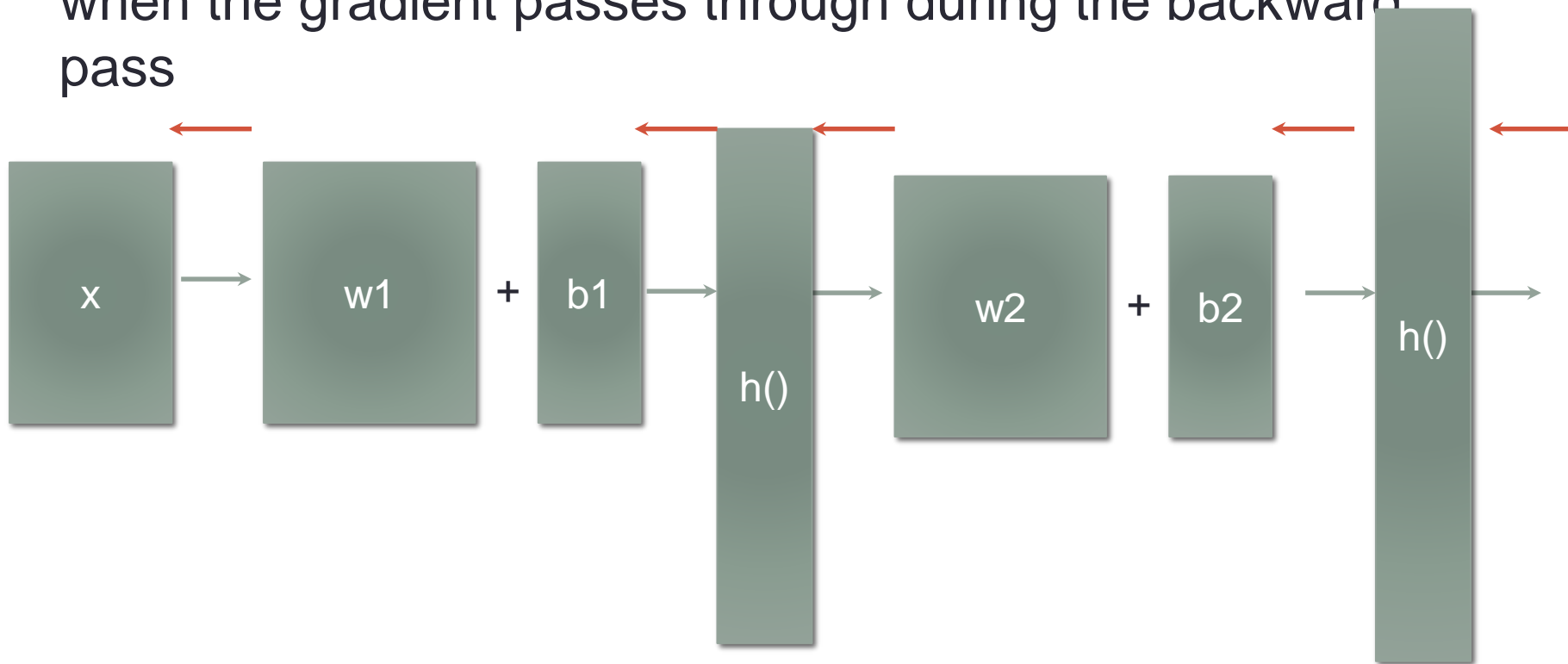
Examples to read

<https://alonalj.github.io/2016/12/10/What-is-Backpropagation/>

<https://mattmazur.com/2015/03/17/a-step-by-step-backpropagation-example/>

Backprop and computation graph

- We can also define what happens to a computing graph when the gradient passes through during the backward pass



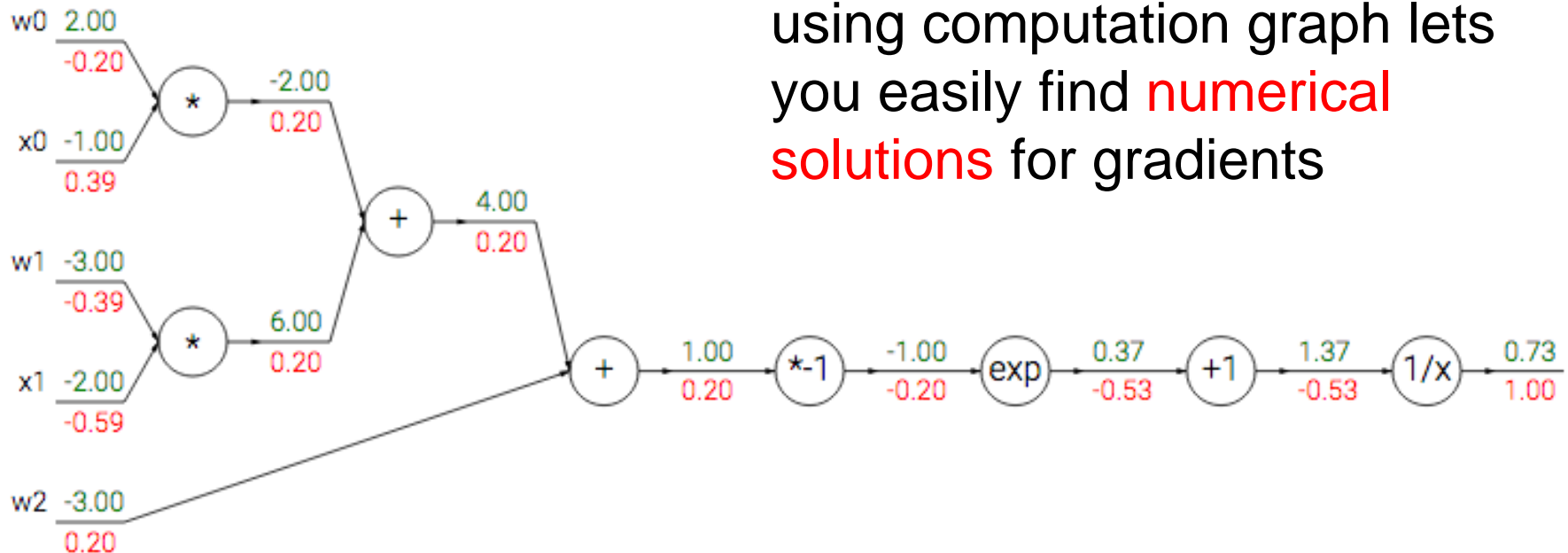
This lets us to build any neural networks without having to redo all the derivation as long as we define a forward and backward computation for the block.

Numerical gradient flow

- Let's find the gradient of

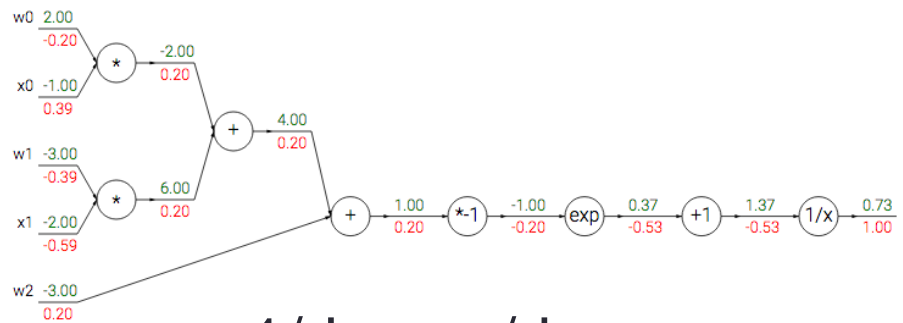
$$f(w, x) = \frac{1}{1 + e^{-(w_0x_0 + w_1x_1 + w_2)}}$$

Computation graph



Doing backprop (chain rule) by using computation graph lets you easily find **numerical solutions** for gradients

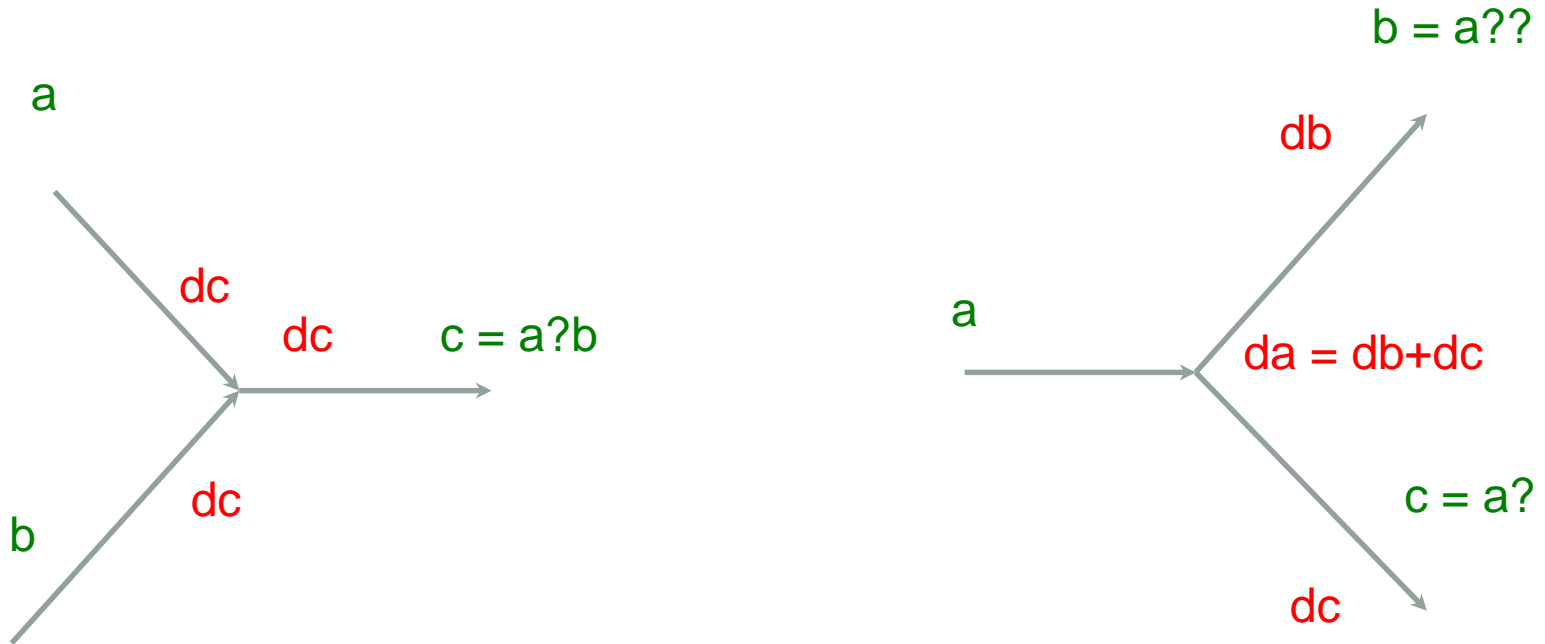
$$f(w, x) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}}$$



- $w = [0, -3, -3]$
- $x = [-1, -2]$
- $t_0 = w[0] * x[0]$
- $t_1 = w[1] * x[1]$
- $t_{01} = t_0 + t_1$
- $t_{012} = t_{01} + w[2]$
- $n_t = -t_{012}$
- $e = \exp(n_t)$
- $\text{denom} = e + 1$
- $f = 1/\text{denom}$
- $\text{ddenom} = -1/\text{denom}/\text{denom}$
- $\text{de} = 1 * \text{ddenom}$
- $\text{dn}_t = \exp(n_t) * \text{de}$
- $\text{dt}_{012} = -\text{dn}_t$
- $\text{dw}_2 = 1 * \text{dt}_{012}$
- $\text{dt}_{01} = 1 * \text{dt}_{012}$
- $\text{dt}_0 = 1 * \text{dt}_{01}$
- $\text{dt}_1 = 1 * \text{dt}_{01}$
- $\text{dw}_1 = x[1] \text{dt}_1$
- $\text{dx}_1 = w[1] \text{dt}_1$
- $\text{dw}_0 = x[0] \text{dt}_0; \text{dx}_0 = w[0] \text{dt}_0$

Perform backward pass in reverse order. No need to explicitly find overall derivative

Gradient flow at forks



Forward and backward pass acts differently at forks

Gradient and non-linearities

We can now talk about how good a non-linearity is by looking at the gradients.

We want

- Something that is **differentiable numerically**

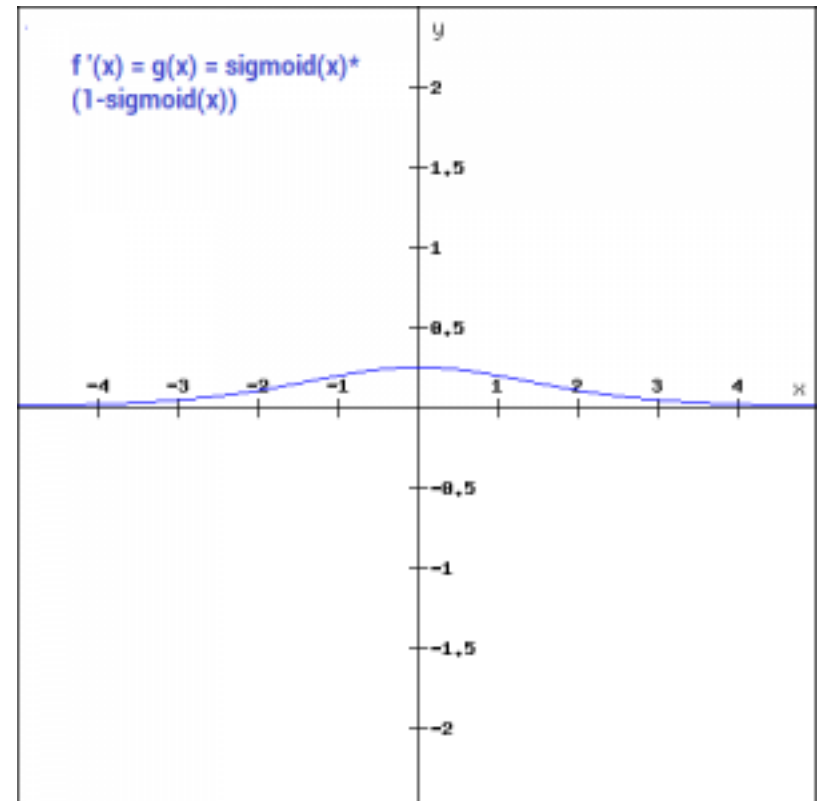
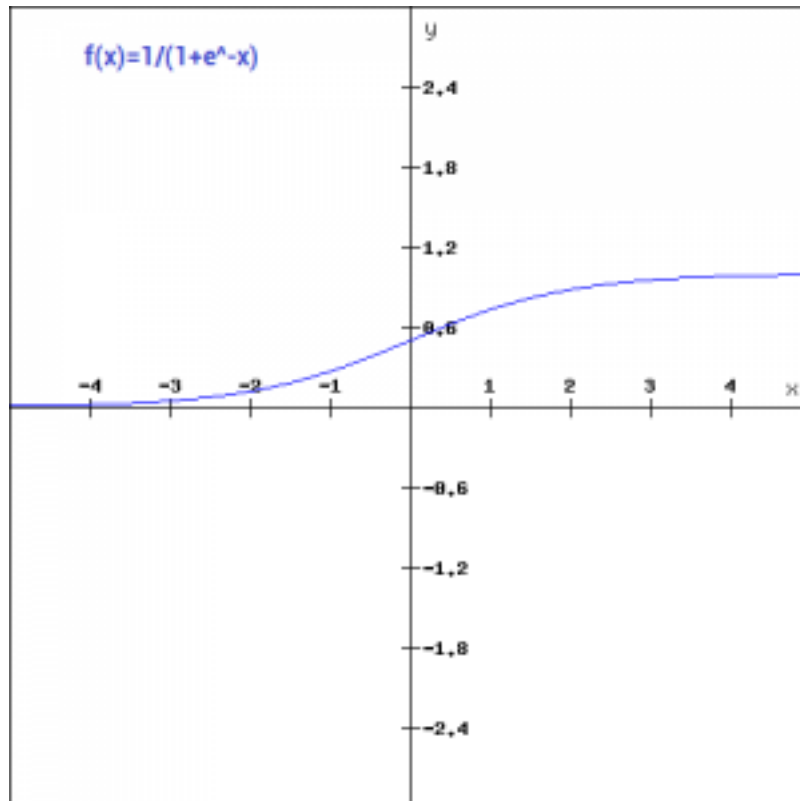
- Cheap to compute

- Big gradients at every point

Notes on non-linearity

- Sigmoid

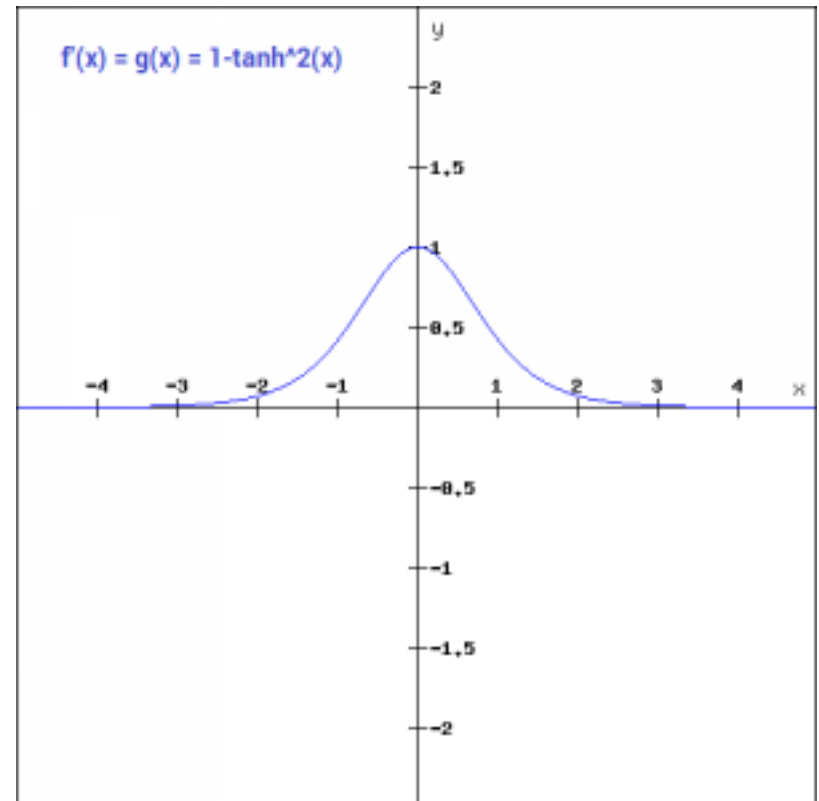
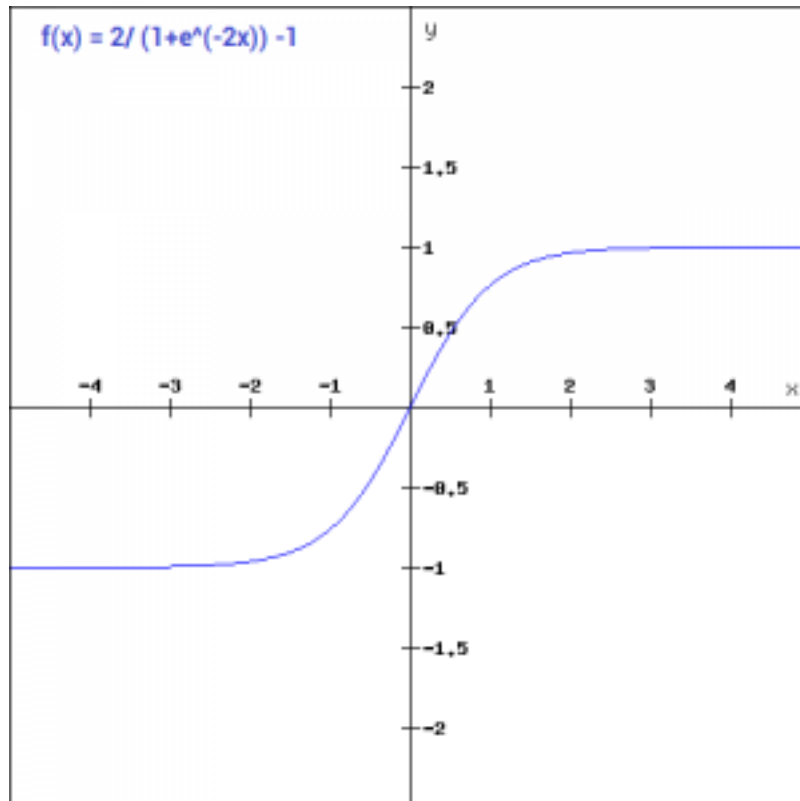
Models get stuck if fall go far away from 0. Output always positive



Notes on non-linearity

- Tanh

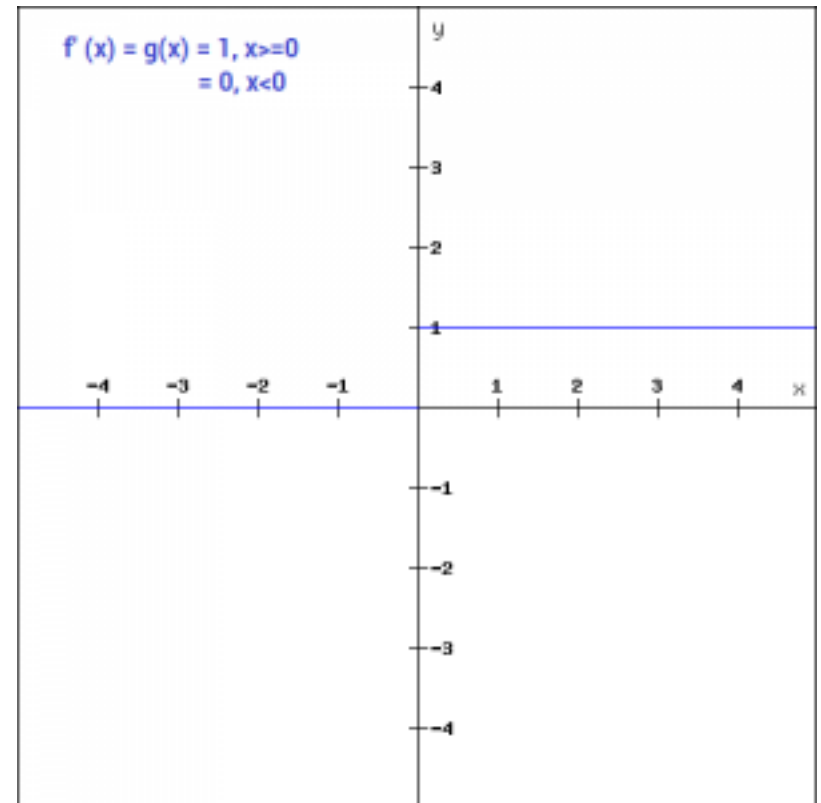
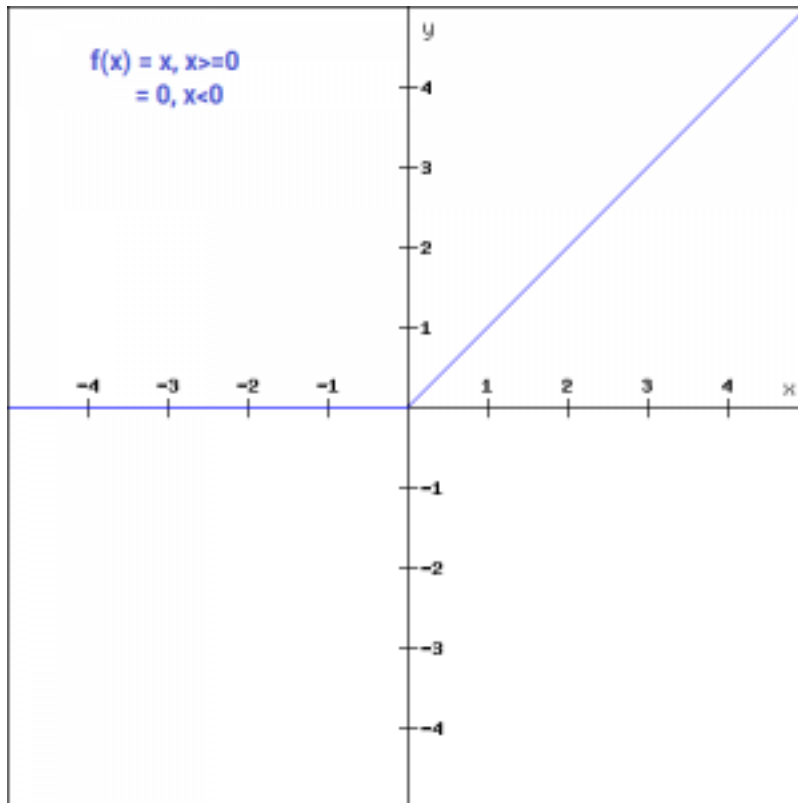
Output can be ± 1 . Models get stuck if far away from 0



Notes on non-linearity

- ReLU

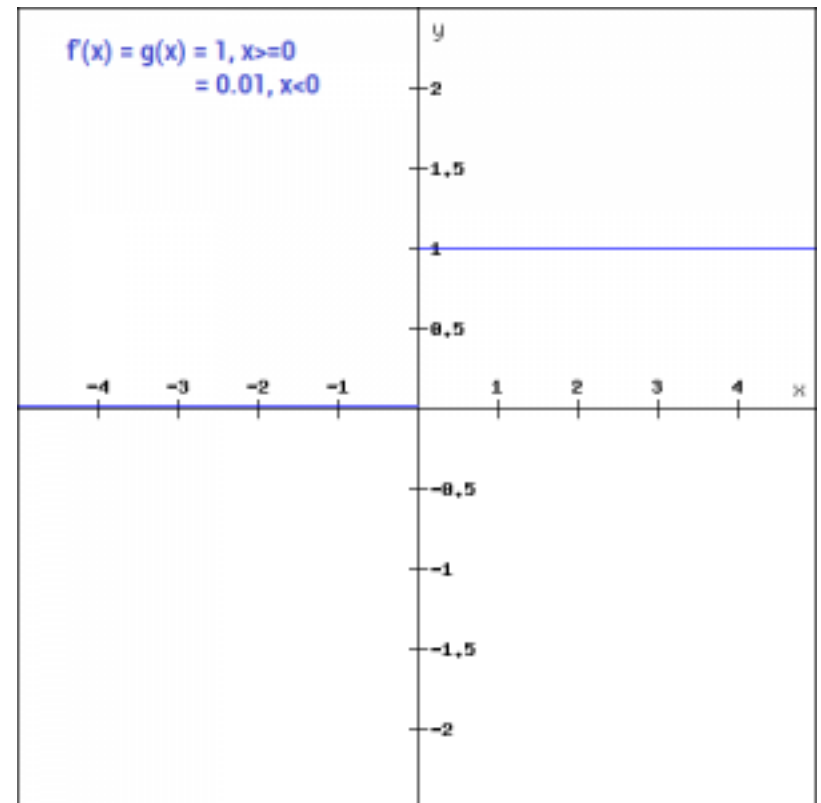
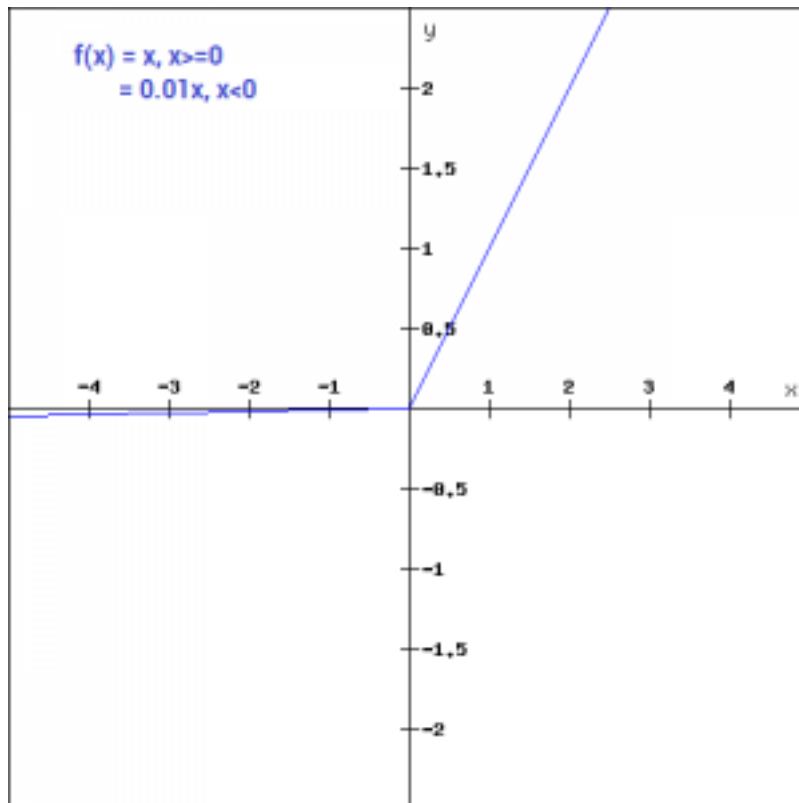
High gradient in positive. Fast compute. Gradient doesn't move in negative



Notes on non-linearity

- Leaky ReLU

Negative part now have some gradient. Small improvements depending on tasks



Notes on non-linearity

- Leaky ReLU
Fixed slope
- PreLU
Slope is learnable
Different layer can have different slope

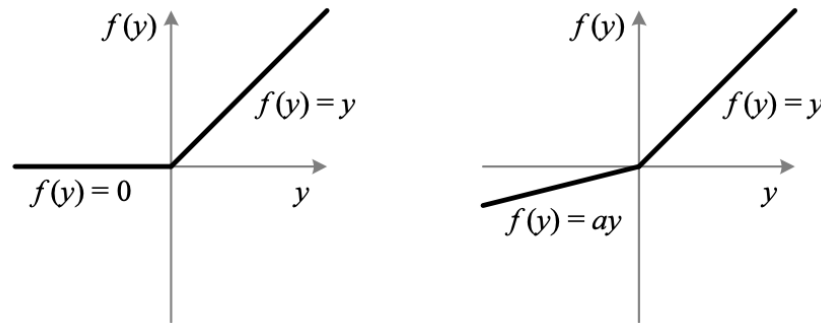


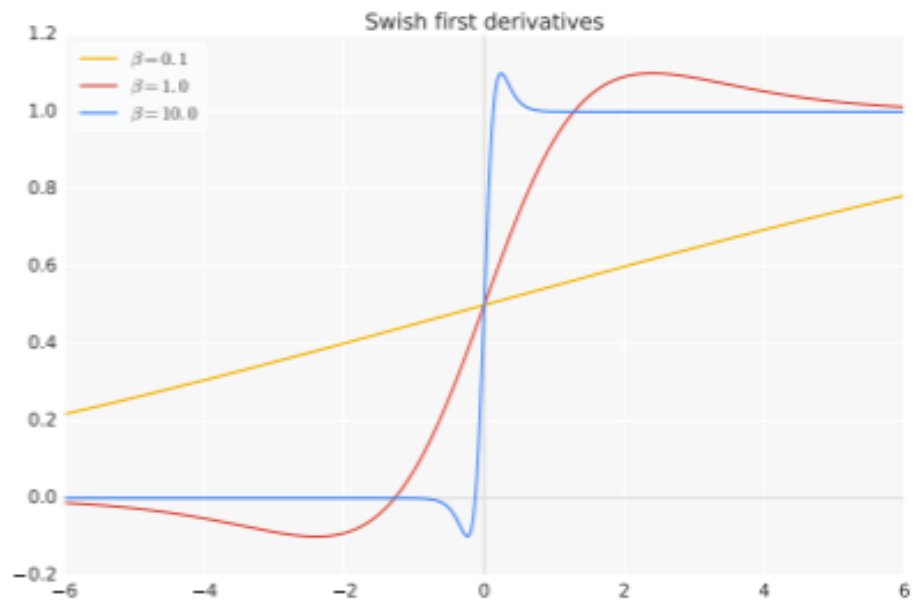
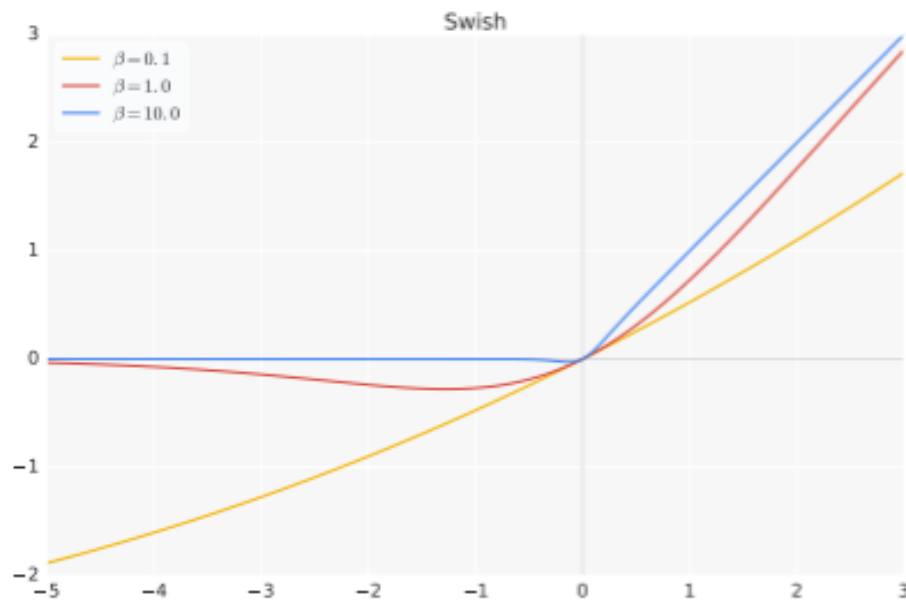
Figure 1. ReLU vs. PReLU. For PReLU, the coefficient of the negative part is not constant and is adaptively learned.

<https://paperswithcode.com/paper/delving-deep-into-rectifiers-surpassing-human>

Notes on non-linearity $x \cdot \text{sig}(\beta x)$

- Swish

Nonnegative everywhere. Not monotonic. Learnable Beta or set to 1



Notes on non-linearity

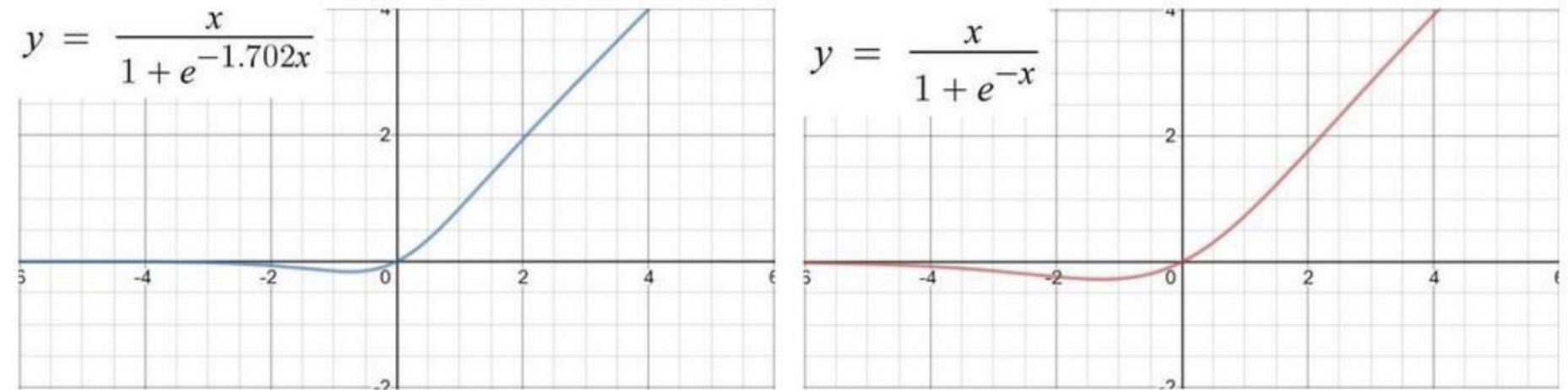
- Swish

Nonnegative everywhere. Not monotonic. Learnable Beta or set to 1

- Gaussian Error Linear Units (GELU)

Set beta to 1.702 (approximation of Gaussian distribution)

Trains faster somehow might be due to initialization strategy(?)



GELU (left) vs. Swish-1 (right): Image by Author

Notes on non-linearity

- Start with ReLU
 - No parameter to tune, many papers/settings uses this so it might work better with the matched settings
 - Try LeakyReLU or GELU afterwards

Initialization

- The starting point of your descent
- Important due to local minimas
- Not as important with large networks AND big data
- Now usually initialized randomly
 - One strategy (Xavier init)
$$\text{var}(w) = 2/(\text{fan_in} + \text{fan_out})$$
 - For ReLUs (He init)
$$\text{var}(w) = 2/(\text{fan_in})$$
- Or use a pre-trained network as initialization

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. 2010
<https://proceedings.mlr.press/v9/glorot10a.html>

Kaiming He, Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. 2015
<https://arxiv.org/abs/1502.01852>

Stochastic gradient descent (SGD)

- Consider you have one million training examples
 - Gradient descent computes the objective function of **all** samples, then decide direction of descent
 - Takes too long
 - SGD computes the objective function on **subsets** of samples
 - The subset should not be biased and properly randomized to ensure no correlation between samples
- The subset is called a mini-batch
- Size of the mini-batch determines the training speed and accuracy
 - Usually somewhere between 32-1024 samples per mini-batch
 - Generally, works well with default settings
- Definition: 1 batch vs 1 epoch

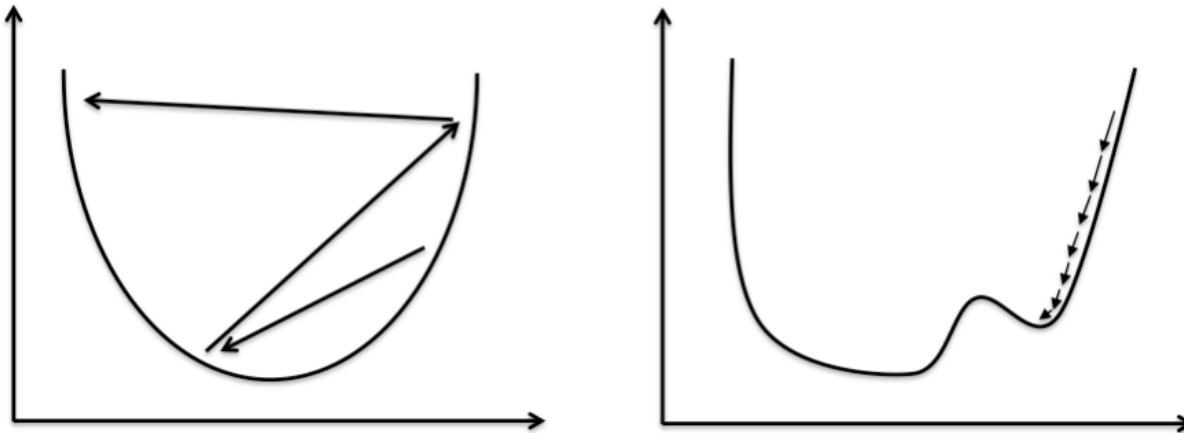
Self regularizing property of SGD

- SGD by its randomized nature does not overfit (as fast)
 - Considered as an implicit regularization (no change in the loss)

<https://cbmm.mit.edu/sites/default/files/publications/CBMM-Memo-067-v3.pdf>

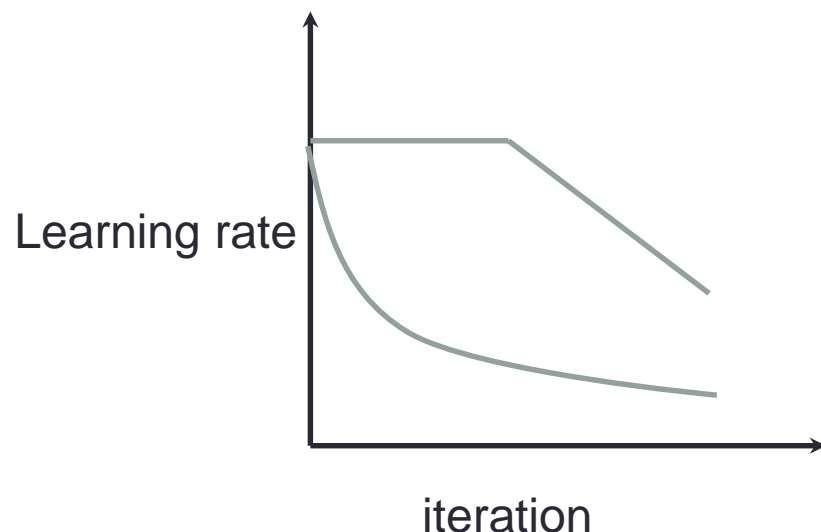
Learning rate

- How fast to go along the gradient direction is controlled by the learning rate
- Too large models diverge
- Too small the model get stuck in local minimas and takes too long to train



Learning rate scheduling

- Usually starts with a large learning rate then gets smaller later
- Depends on your task
- Automatic ways to adjust the learning rate : Adagrad, Adam, etc. (still need scheduling still)



Learning rate strategies (annealing)

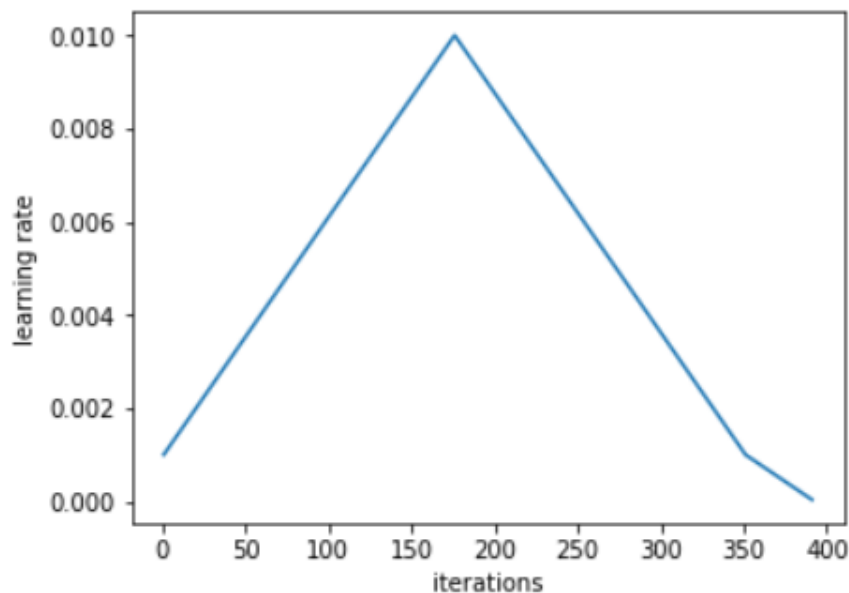
- Step decay: reduce learning rate by x after y epochs
- New bob method: half learning rate every time the validation error goes up. Only plausible in larger tasks
- Exponential decay: multiplies the learning rate by $\exp(-\text{rate} * \text{epoch number})$

Learning rate warm up

Initial point of the network can be at a bad spot.

Try not to go too fast - has a warm up period.

Useful for large datasets, or adaption (transfer learning)



Potentially leads to faster convergence and better accuracy

See links below for methods to select the shape of the triangle

<https://sgugger.github.io/the-1cycle-policy.html#the-1cycle-policy>

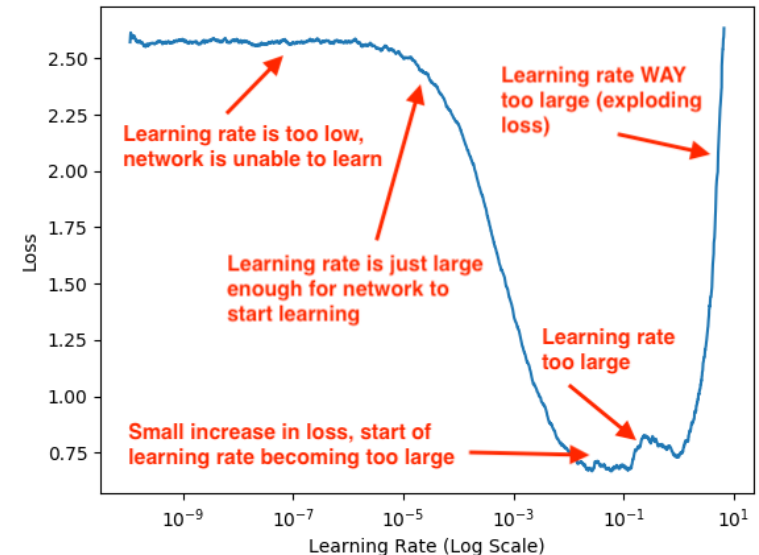
[Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour](#)

[Cyclical Learning Rates for Training Neural Networks](#)

Learning rate finder

- Define sweep range of LR
- Set $R = \min(\text{LR})$
- Until $R = \max(\text{LR})$
 - Train for 1 mini-batch
 - Exponentially increase R

Plot the loss after each mini-batch

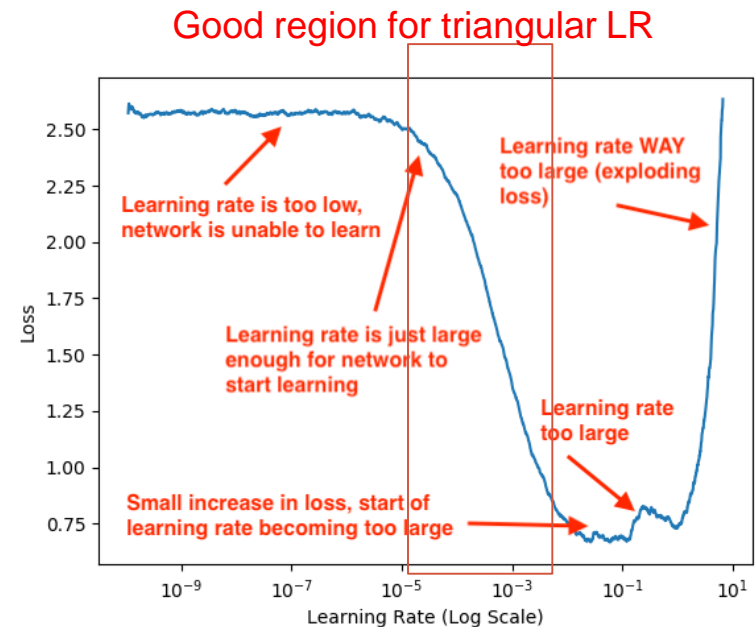


Learning rate finder

- Define sweep range of LR
- Set $R = \min(\text{LR})$
- Until $R = \max(\text{LR})$
 - Train for 1 mini-batch
 - Exponentially increase R

Plot the loss after each mini-batch

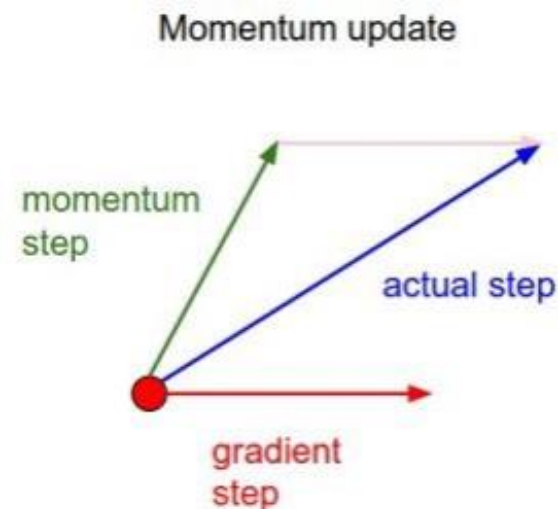
lr_find in many popular framework



Momentum

- Gradient descent can get stuck on small local minimas
 - Or slow down at saddle points
- Have concept of speed

$$\underbrace{V_t}_{\text{speed}} = \underbrace{\beta}_{\text{Momentum rate}} V_{t-1} + (1 - \beta) \underbrace{\nabla_w L(W, X, y)}_{\text{gradient}}$$
$$W = W - \underbrace{\alpha}_{\text{learning rate}} V_t$$



Nesterov Momentum

Old

$$V_t = \beta V_{t-1} + (1 - \beta) \nabla_w L(W, X, y)$$

$\nearrow W = W - \alpha V_t$

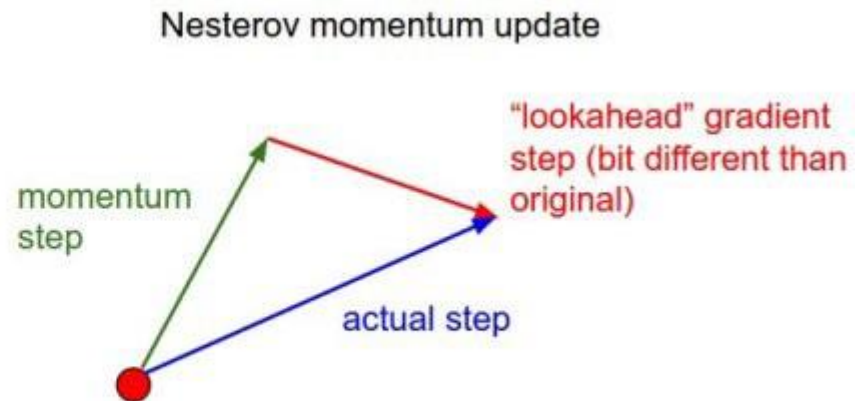
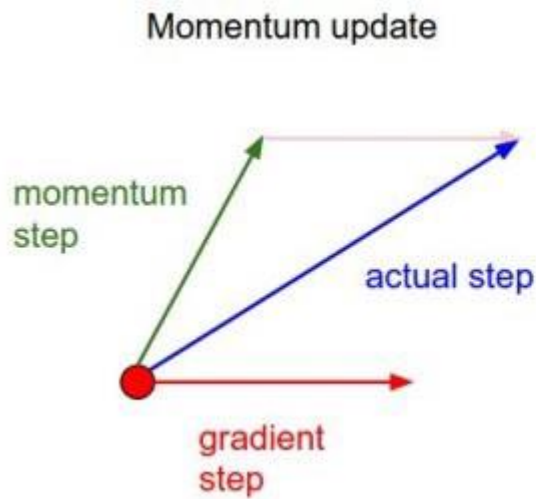
W is about to be at $W - \beta V$ due to momentum

- Momentum with look ahead.
 - Compute gradient as if we took an additional step

$$V_t = \beta V_{t-1} + \alpha \nabla_w L(\underbrace{W - \beta V_{t-1}}_{\text{lookahead}}, X, y)$$

$$W = W - V_t$$

gradient is computed as if we took a step



Adaptive learning rates

How to have the updates be different for different layers?

LR of each weight is scaled by the size of the gradient

Can we trust the moment estimates?

Decay running estimates of the gradient size/momentums

RMSProp (normalized by gradient + decay)

Adam (normalized by gradient + decay both grad and momentum)

AdamW (Adam but deal with weight decay differently)

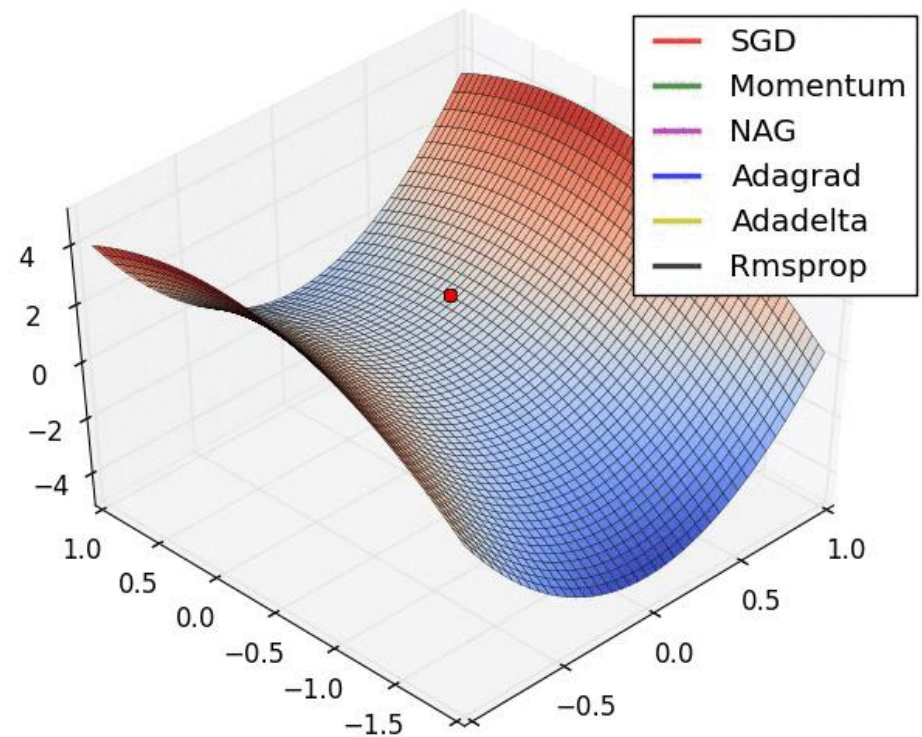
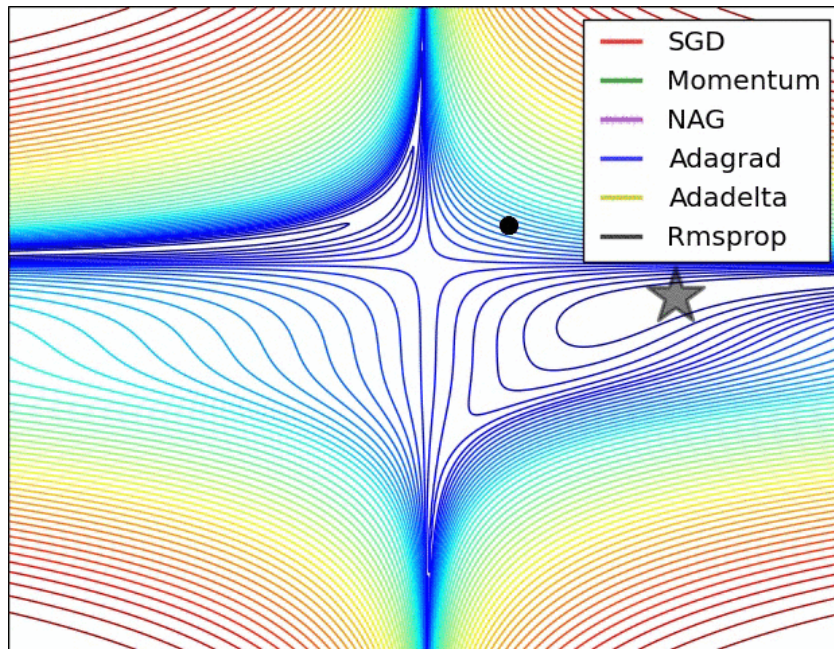
You still need to do learning rate scheduling

More details see

<http://ruder.io/optimizing-gradient-descent/index.html#whichoptimizertochoose>

<https://towardsdatascience.com/why-adamw-matters-736223f31b5d>

Optimization method and speed



Learning rate tricks

- At least decay the learning rate
 - Monitor validation set performance
- If the loss never goes down -> decrease the learning rate (by factor of 10)
- Start with ADAM or ADAMW. Also try RMSprop and SGD with Nesterov Momentum if you have time
- Learning rate is the most important hyperparameter that will affect your model performance

Learning rate and batch size

- Learning rate and batch size interacts with each other
 - Larger batch has low variance -> can use larger learning rate
 - Increase batch size by k , increase learning rate by k (can also scale the momentum etc, see papers)
 - <https://arxiv.org/abs/1706.02677>
 - <https://arxiv.org/abs/1711.00489>
- Set batch size first
 - largest to fit the GPU (trains faster, and get better accuracy)
- Then set your learning rate
 - Still need to tune. Above theory is just a guideline

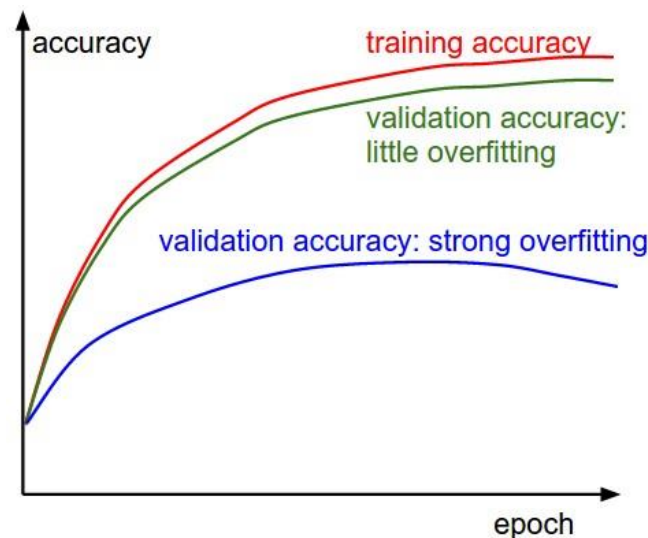


Small batchsize training

- Gradient accumulation
 - Accumulate gradients over multiple small batches before applying an parameter update
- Retune all learning rate hyperparamters

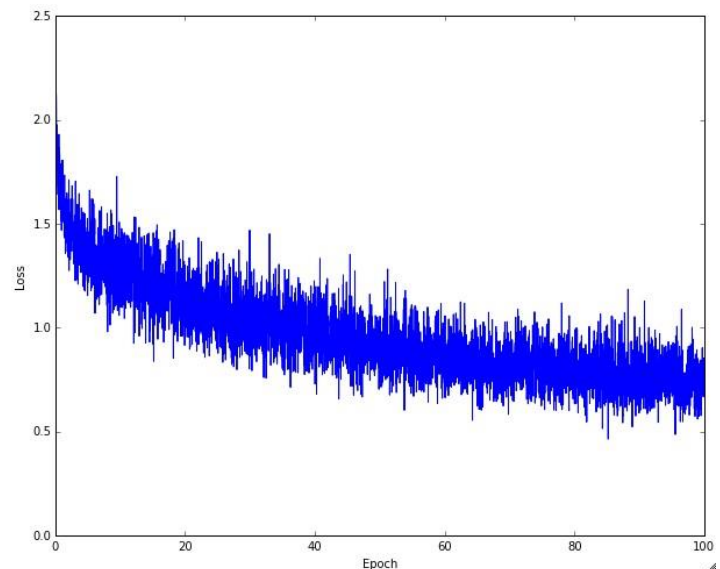
Overfitting

- You can keep doing back propagation forever!
- The training loss will always go down
- But it overfits
- Need to monitor performance on a held out set
- Stop or decrease learning rate when overfit happens



Monitoring performance

- Monitor performance on a dev/validation set
 - This is NOT the test set
- Can monitor many criteria
 - Loss function
 - Classification accuracy
- Sometimes these disagree
- Actual performance can be noisy, need to see the trend



Dropout

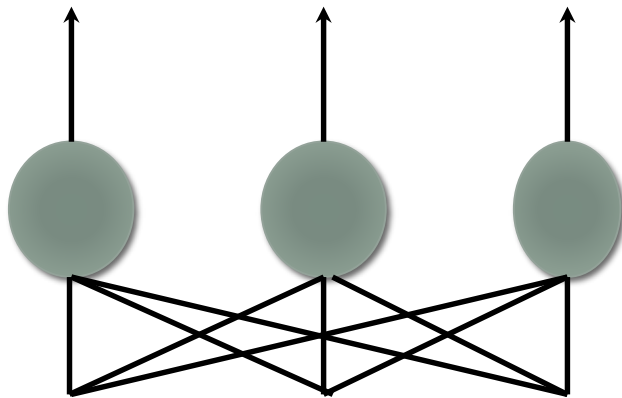
An **implicit regularization** technique for reducing overfitting

Randomly turn off different subset of neurons during training

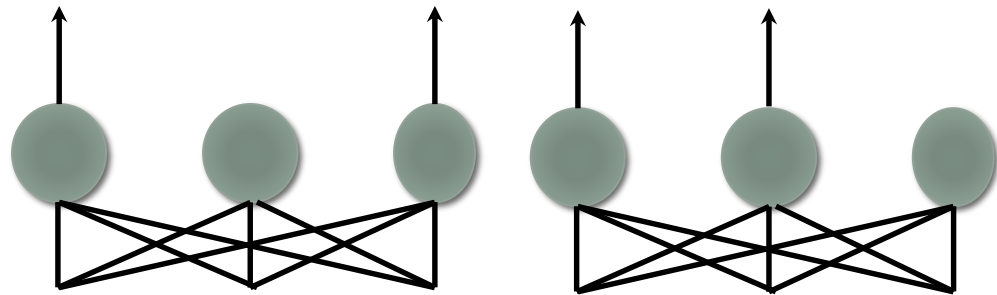
Network no longer depend on any particular neuron

Force the model to have redundancy – robust to any corruption in input data

A form of performing model averaging (ensemble of experts)



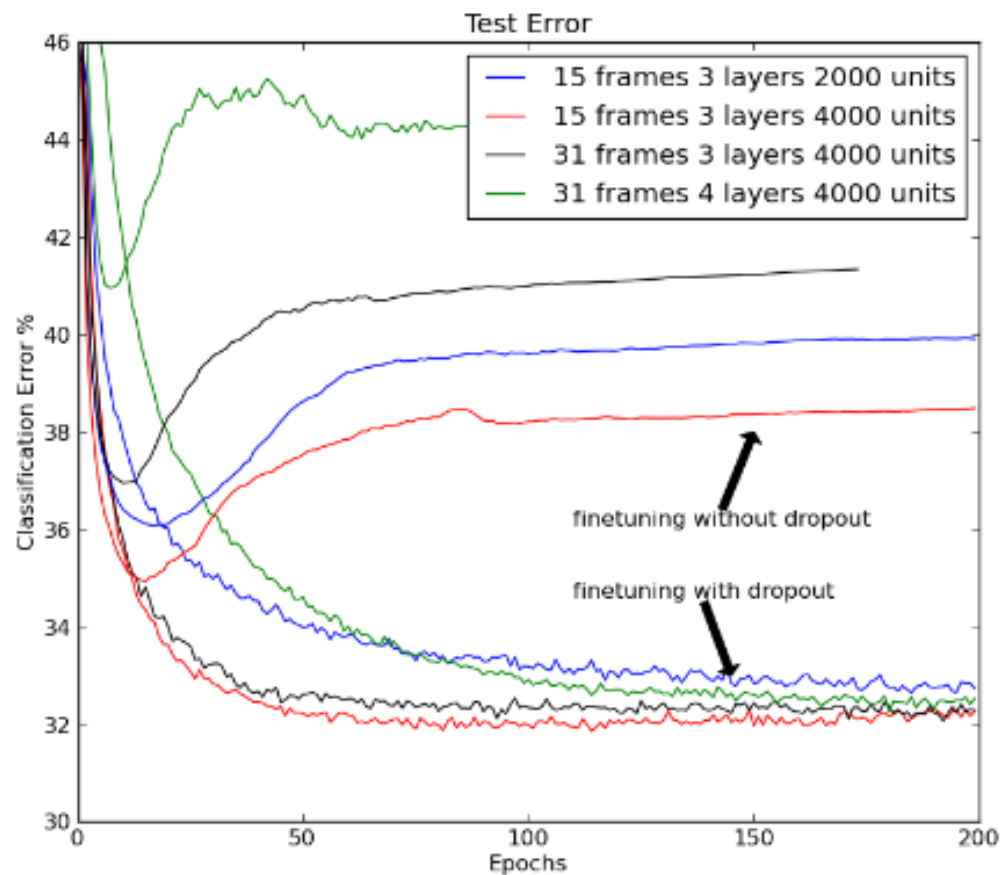
Model



Dropout rate of 0.33

Dropout on TIMIT

- A phoneme recognition task



Batch normalization

- Recent technique for (implicit) regularization
- **Normalize every mini-batch** at various batch norm layers to standard Gaussian (different from global normalization of the inputs)
- Place batch norm layers before non-linearities
- Faster training and better generalizations

For each mini-batch that goes through batch norm

1. Normalize by the mean and variance of the mini-batch for each dimension
2. Shift and scale by learnable parameters

Replaces dropout in some networks

$$\hat{x} = \frac{x - \mu_b}{\sigma_b}$$

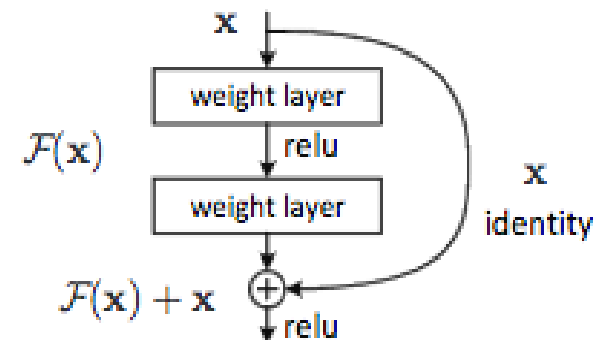
$$y = \alpha \hat{x} + \beta$$

<https://arxiv.org/abs/1502.03167>

Vanishing/Exploding gradient

- Backprop introduces many multiplications down chain
- The gradient value gets smaller and smaller
 - The deeper the network the smaller the gradient in the lower layers
 - Lower layers changes too slowly (or not at all)
 - Hard to train very deep networks (>6 layers)
- The opposite can also be true. The gradient explodes from repeated multiplication
 - Put a maximum value for the gradient (Gradient clipping)

- How to deal with this?
 - Residual connection



<https://arxiv.org/abs/1512.03385>

Tips to tune



- Feeling and experience $\neg \backslash(^{\circ}_o)/$
- Take numbers from papers
- Grid search
 - Heuristic search
 - Random search
 - Genetic Algorithm
- Picking the right type of model is more important than picking the right number of neurons
 - Inductive bias
- Overfit then regularize
- Tips for debugging (will make more sense next class)
 - <http://karpathy.github.io/2019/04/25/recipe/>

Neural networks

- Fully connected networks
 - Neuron
 - Non-linearity
 - Softmax layer
- DNN training
 - Loss function and regularization
 - SGD and backprop
 - Learning rate
 - Overfitting – dropout, batchnorm
- CNN, RNN, LSTM, GRU <- Next class

