Topics:

- Optimization

CS 4644-DL / 7643-A
ZSOLT KIRA

- **Assignment 1 – Due Friday!!!**
  - **DO NOT SEARCH FOR CODE!!!!**

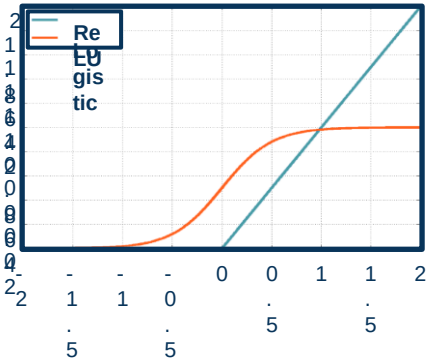
- **Assignment 2**
  - Implement convolutional neural networks

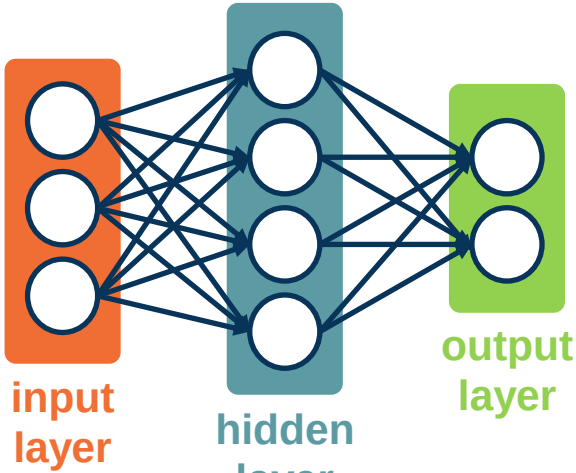
- **Piazza**: Start with public posts so that others can benefit!
  - Doesn't mean don't post!

- **Meta OH**: Data wrangling Friday 01/31 3pm ET
  - **OMSCS** Lessons (videos) linked as dropbox
  - **Full schedule and discussions on** https://ai-learning.org/

$$\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} & b_1 \\ w_{21} & w_{22} & \cdots & w_{2m} & b_2 \\ w_{31} & w_{32} & \cdots & w_{3m} & b_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \\ 1 \end{bmatrix}$$

$\boldsymbol{W}$ $\boldsymbol{x}$





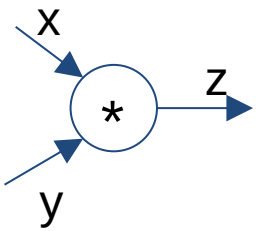
- Gradient Descent
- Compute gradients via chain rule
  - Backpropagation
  - Computation Graph + Automatic Differentiation

$$w_i = w_i - \alpha \frac{\partial L}{\partial w_i}$$

So Far

Slide Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

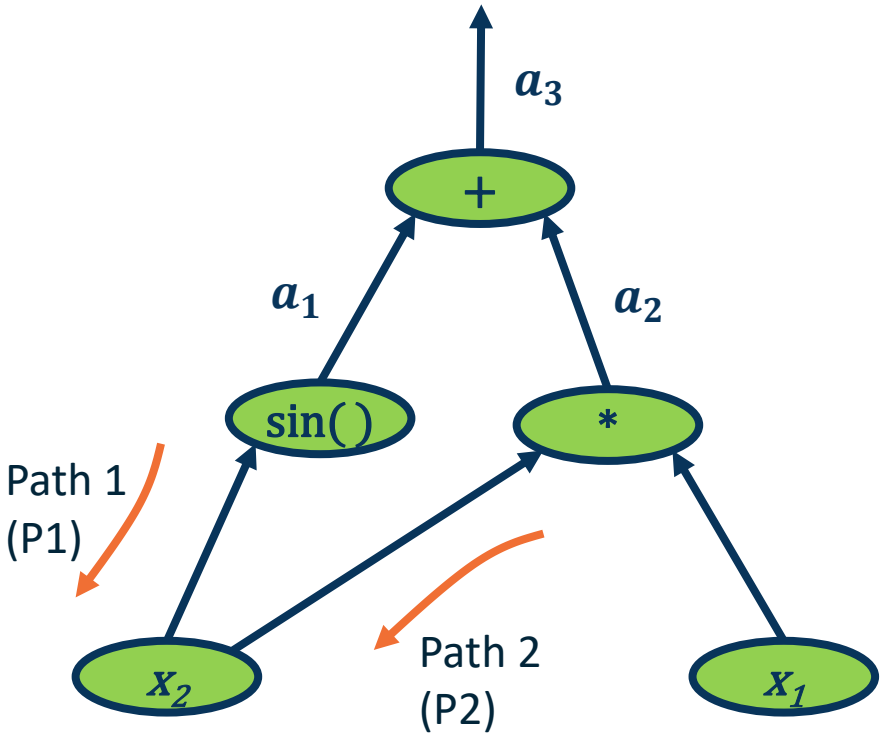
Modularized implementation: forward / backward API



(x,y,z are scalars)

```
class MultiplyGate(object):
    def forward(x,y):
        z = x*y
        self.x = x # must keep these around!
        self.y = y
        return z
    def backward(dz):
        dx = self.y * dz # [dz/dx * dL/dz]
        dy = self.x * dz # [dz/dy * dL/dz]
        return [dx, dy]
```

$$f(x_1, x_2) = x_1 x_2 + \sin(x_2)$$



$$\overline{a_3} = \frac{\partial f}{\partial a_3} = 1$$

$$\overline{a_1} = \frac{\partial f}{\partial a_1} = \frac{\partial f}{\partial a_3}\frac{\partial a_3}{\partial a_1} = \frac{\partial f}{\partial a_3}\frac{\partial (a_1+a_2)}{\partial a_1} = \frac{\partial f}{\partial a_3}\,1 = \overline{a_3}$$

$$\overline{a_2} = \frac{\partial f}{\partial a_2} = \frac{\partial f}{\partial a_3}\frac{\partial a_3}{\partial a_2} = \overline{a_3}$$

$$\overline{x_2^{P1}} = \frac{\partial f}{\partial a_1}\frac{\partial a_1}{\partial x_2} = \overline{a_1}\,\cos(x_2)$$

$$\overline{x_2^{P2}} = \frac{\partial f}{\partial a_2}\frac{\partial a_2}{\partial x_2} = \frac{\partial f}{\partial a_2}\frac{\partial (x1x2)}{\partial x_2} = \overline{a_2} x_1$$

Gradients from multiple paths **summed**

$$\overline{x_1} = \frac{\partial f}{\partial a_2}\frac{\partial a_2}{\partial x_1} = \overline{a_2} x_2$$

Example

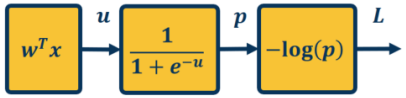
A graph is created on the fly

```
from torch.autograd import Variable

x = Variable(torch.randn(1, 20))
prev_h = Variable(torch.randn(1, 20))
W_h = Variable(torch.randn(20, 20))
W_x = Variable(torch.randn(20, 20))

i2h = torch.mm(W_x, x.t())
h2h = torch.mm(W_h, prev_h.t())
next_h = i2h + h2h
```

(Note above)



$L = 1$

$\bar{p} = \frac{\partial L}{\partial p} = -\frac{1}{p}$

where $p = \sigma(w^T x)$ and $\sigma(x) = \frac{1}{1+e^{-x}}$

$\bar{u} = \frac{\partial L}{\partial u} = \frac{\partial L}{\partial p}\frac{\partial p}{\partial u} = \bar{p}\,\sigma(1-\sigma)$

$\bar{w} = \frac{\partial L}{\partial w} = \frac{\partial L}{\partial u}\frac{\partial u}{\partial w} = \bar{u}x^T$

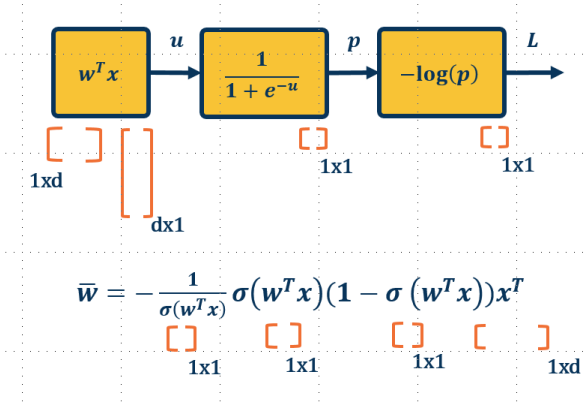
We can do this in a combined way to see all terms together:

$\bar{w} = \frac{\partial L}{\partial p}\frac{\partial p}{\partial u}\frac{\partial u}{\partial w} = -\frac{1}{\sigma(w^T x)}\sigma(w^T x)(1-\sigma(w^T x))x^T$

$= -\left(1-\sigma(w^T x)\right)x^T$

This effectively shows gradient flow along path from L to w

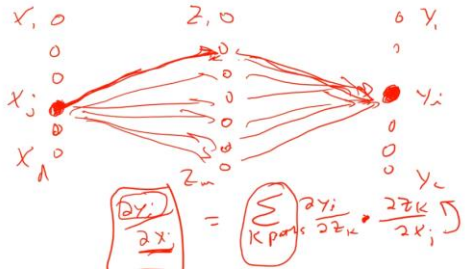
Computation Graph / Global View of Chain Rule



$$\bar{w} = -\frac{1}{\sigma(w^T x)}\sigma(w^T x)(1-\sigma(w^T x))x^T$$

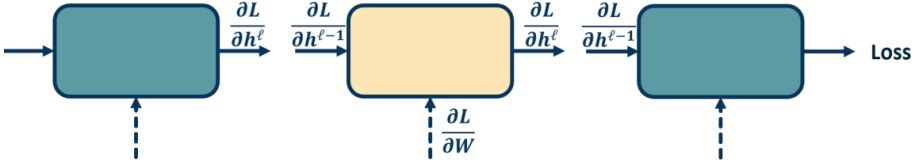
1x1 1x1 1x1 1xd

Computational / Tensor View



Graph View

- We want to to compute: $\left\{\frac{\partial L}{\partial h^{\ell-1}}, \frac{\partial L}{\partial W}\right\}$



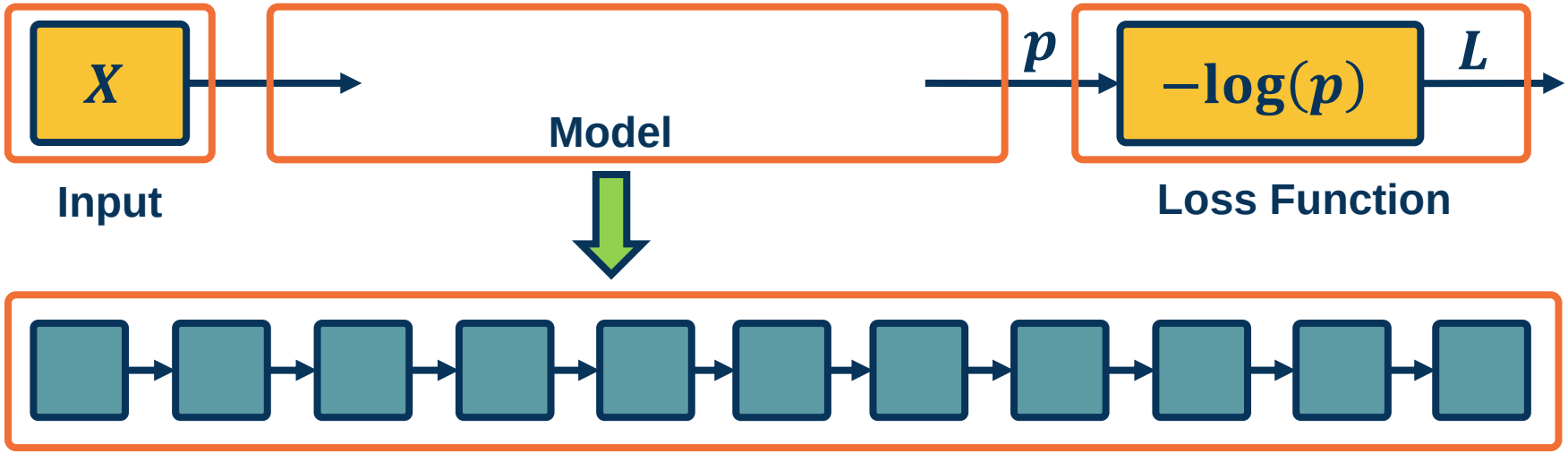
Backpropagation View (Recursive Algorithm)

Different Views of Equivalent Ideas

Georgia Tech

Backpropagation, and automatic differentiation, allows us to optimize **any** function composed of differentiable blocks

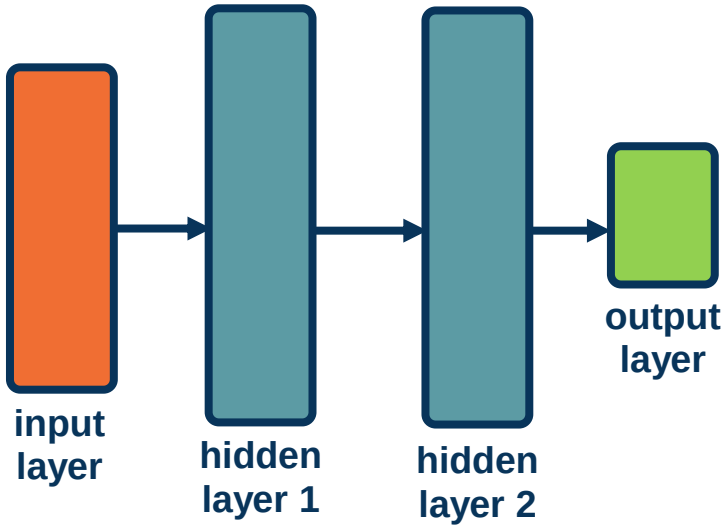
- **No need to modify** the learning algorithm!
- The complexity of the function is only limited by **computation and memory**



A network with two or more hidden layers is often considered a **deep** model

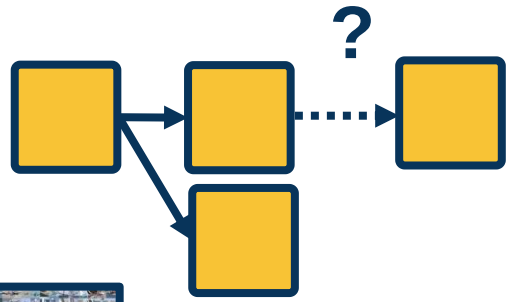
Depth is important:

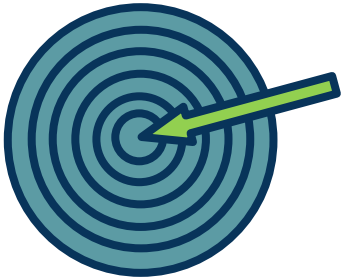
- Structure the model to represent an inherently compositional world
- Theoretical evidence that it leads to parameter efficiency
- Gentle dimensionality reduction (if done right)



There are still many design decisions that must be made:

- **Architecture**
- **Data Considerations**
- **Training and Optimization**
- **Machine Learning Considerations**





Local Minima

Machine Learning Considerations

The practice of machine learning is complex: For your particular application you have to **trade off** all of the considerations together

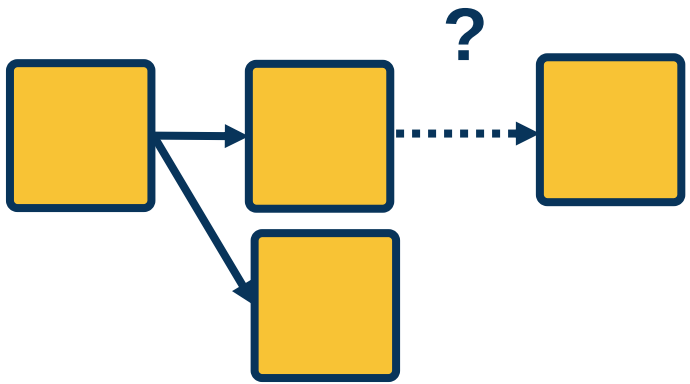
- Trade-off between **model capacity** (e.g. measured by # of parameters) and **amount of data**
- Adding **appropriate biases** based on knowledge of the domain

Georgia Tech

Architectural Considerations

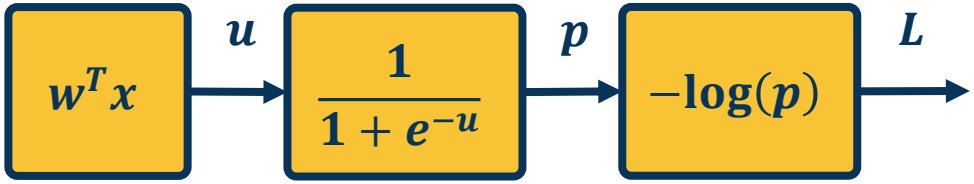
Determining what modules to use, and how to connect them is part of the **architectural design**

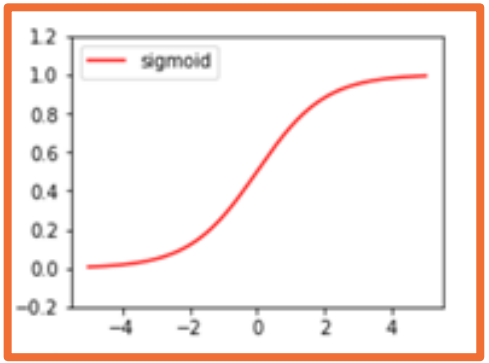
- Guided by the **type of data used** and its **characteristics**
  - Understanding your data is always the first step!
- **Lots of data types (modalities)** already have good architectures
  - Start with what others have discovered!
- **The flow of gradients** is one of the key principles to use when analyzing layers



- **Combination** of linear and non-linear layers
- Combination of **only** linear layers has same representational power as one linear layer
- **Non-linear layers** are crucial
  - Composition of non-linear layers **enables complex transformations of the data**

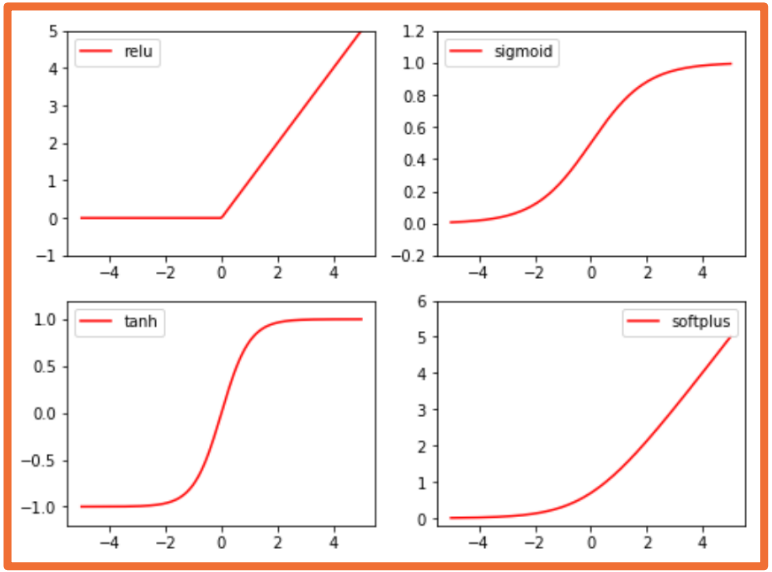
$$w_1^T(w_2^T(w_3^T x)) = w_4^T \mathbf{x}$$



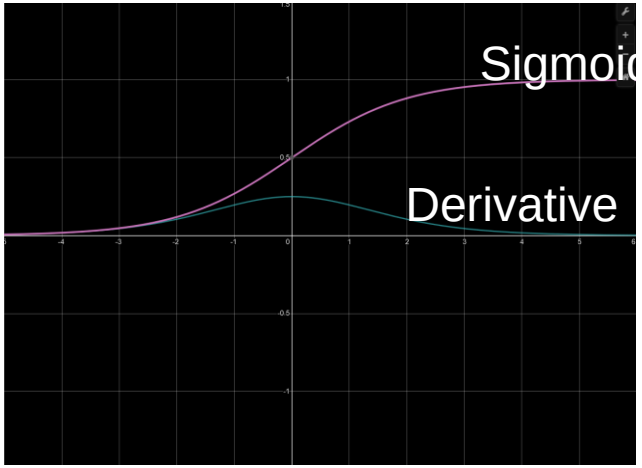


Several aspects that we can **analyze**:

- Min/Max
- Correspondence between input & output statistics
- **Gradients**
  - At initialization (e.g. small values)
  - At extremes
- Computational complexity

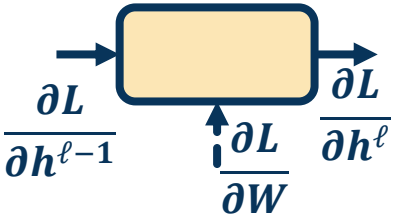


- **Min:** 0, **Max:** 1
- Output **always positive**
- Saturates at **both ends**
- **Gradients**
  - Vanishes at both end
  - Always positive
- **Computation: Exponential term**



$$h^{\ell} = \sigma\left(h^{\ell-1}\right)$$

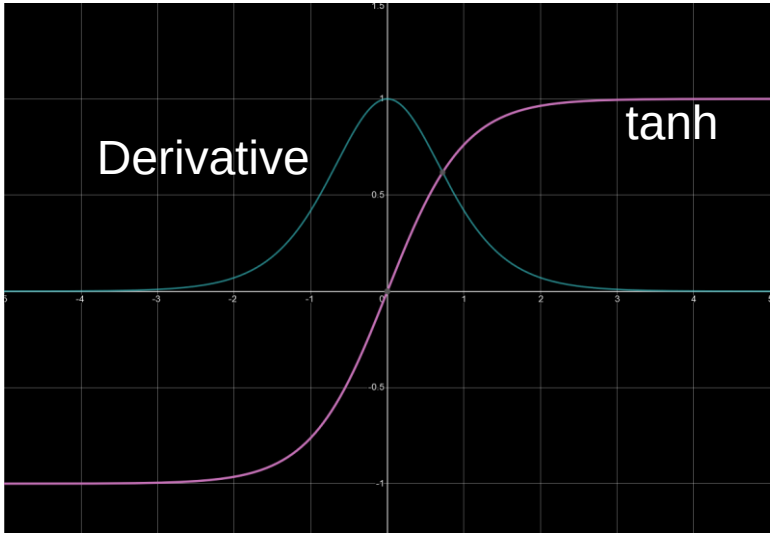
$$\sigma(x) = \frac{1}{1+e^{-x}}$$



$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial h^{\ell}} \frac{\partial h^{\ell}}{\partial W}$$

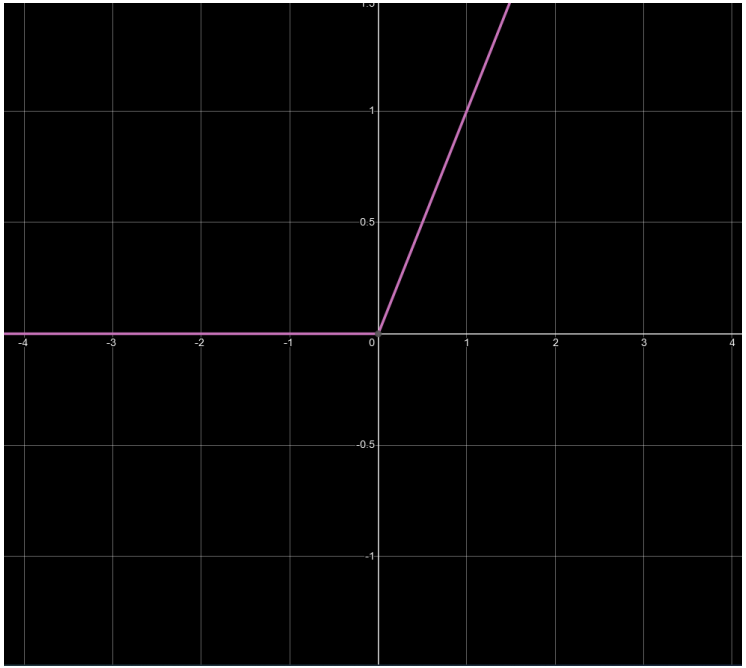
Sigmoid Function

- **Min:** -1, **Max:** 1
  - **Centered**
- Saturates at **both ends**
- **Gradients**
  - Vanishes at both end
  - Always positive
- **Still somewhat computationally heavy**



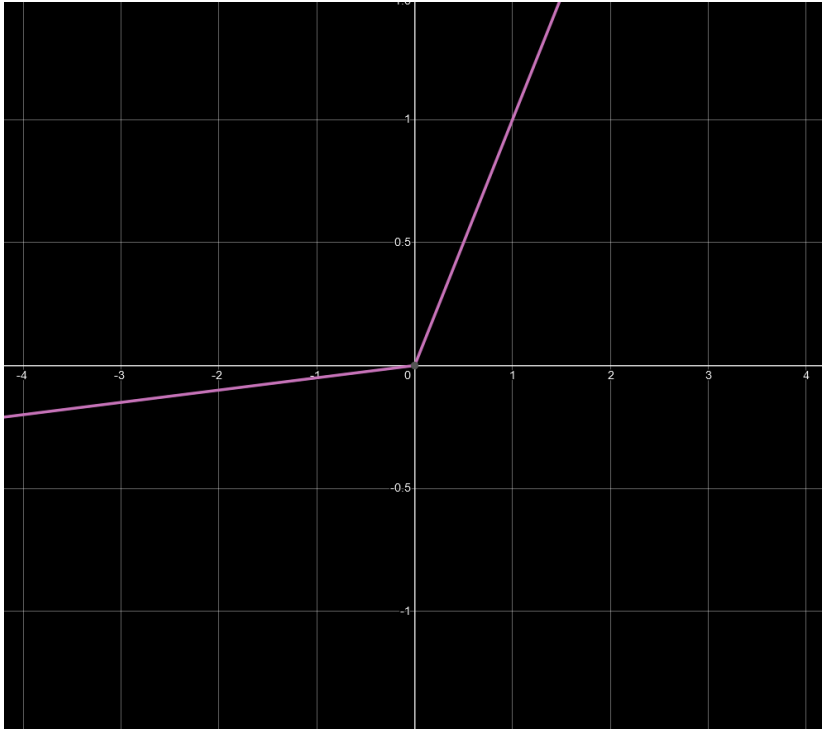
$$h^{\ell} = tanh(h^{\ell-1})$$

- **Min:** 0, **Max:** Infinity
- Output always **positive**
- **No saturation** on positive end!
- **Gradients**
  - $\mathbf{0}$ **if** $\mathbf{x} \leq \mathbf{0}$ (dead ReLU)
  - Constant otherwise (does not vanish)
- **Cheap to compute (max)**



$$h^{\ell} = max(0, h^{\ell-1})$$

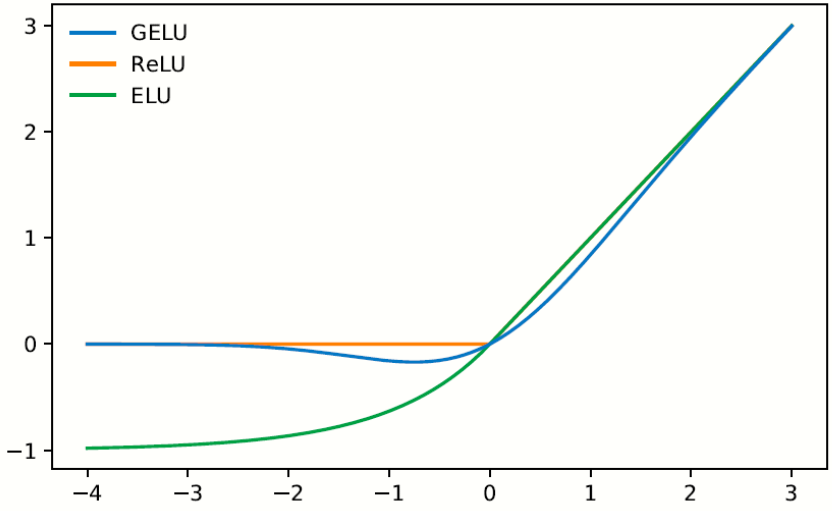
- **Min:** -Infinity, **Max:** Infinity
- **Learnable parameter!**
- **No saturation**
- **Gradients**
  - No dead neuron
- **Still cheap to compute**



$$h^{\ell} = max(\alpha h^{\ell-1}, h^{\ell-1})$$

Leaky ReLU

- **Activation functions is still area of research!**
  - Though many don’t catch on
- **In Transformer architectures, other activations such as GeLU is common**



From "Gaussian Error Linear Units (GELUs)”, Hendrycks & Gimpel

Variations: ELU, GeLU, etc.

Selecting a Non-Linearity

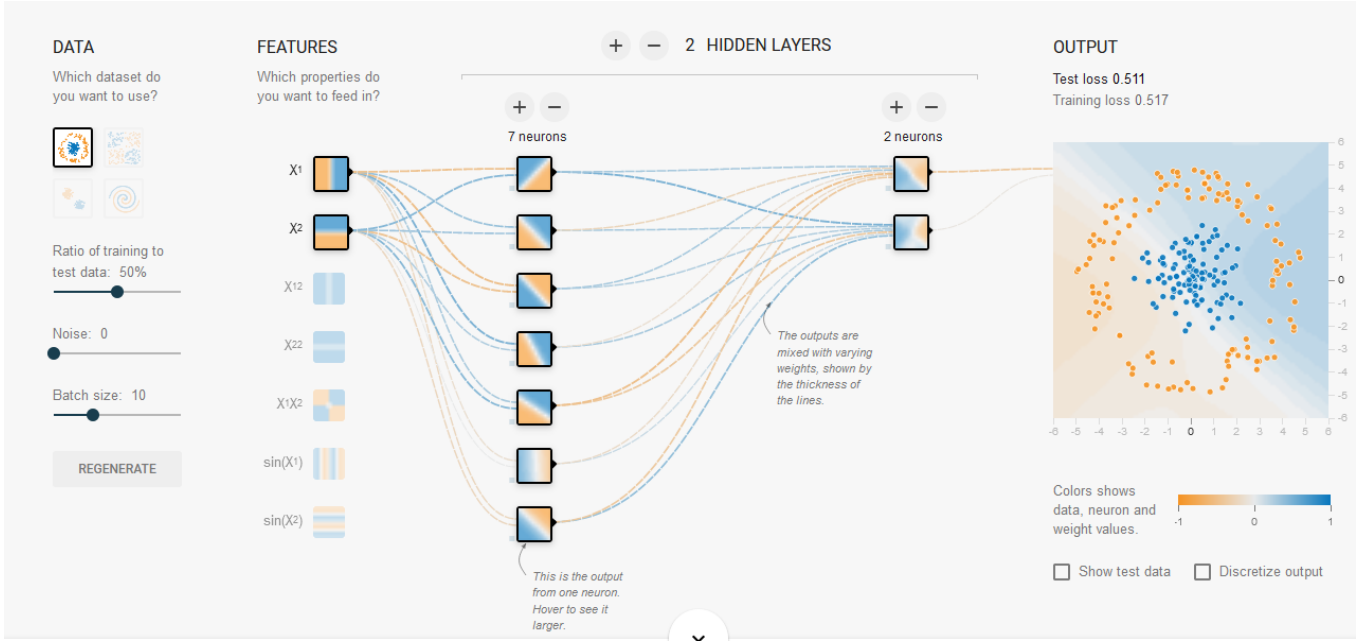
Which **non-linearity** should you select?

- Unfortunately, **no one activation function is best** for all applications
- **ReLU** is most common starting point
  - Sometimes leaky ReLU can make a big difference
- **Sigmoid** is typically avoided unless clamping to values from [0,1] is needed

Georgia Tech

Demo

- http://playground.tensorflow.org



Initialization

Initializing the Parameters

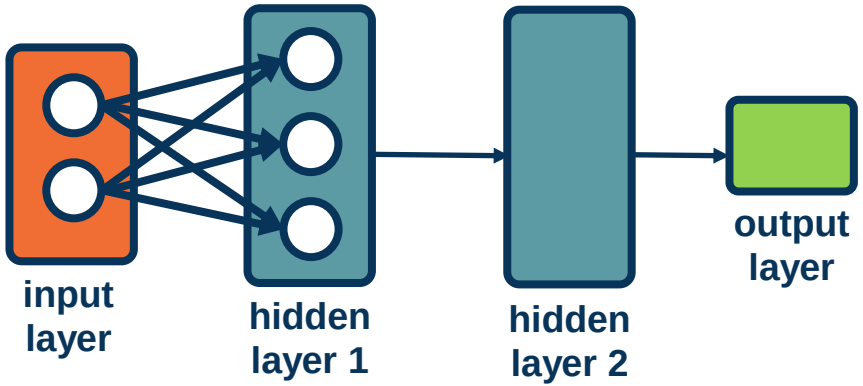
The parameters of our model must be **initialized to something**

- Initialization is **extremely important**!
  - Determines how **statistics of outputs** (given inputs) behave
  - Determines how well **gradients flow** in the beginning of training (important)
  - Could **limit use of full capacity** of the model if done improperly
- Initialization that is **close to a good (local) minima** will converge faster and to a better solution

Initializing values to a constant value leads to **a degenerate solution**!

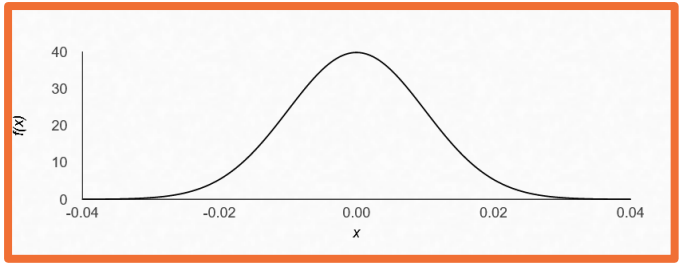
- What happens to the **weight updates?**
- Each node has the same input from previous layers so gradients **will be the same**
- As a results**, all weights will be updated** to the same exact values

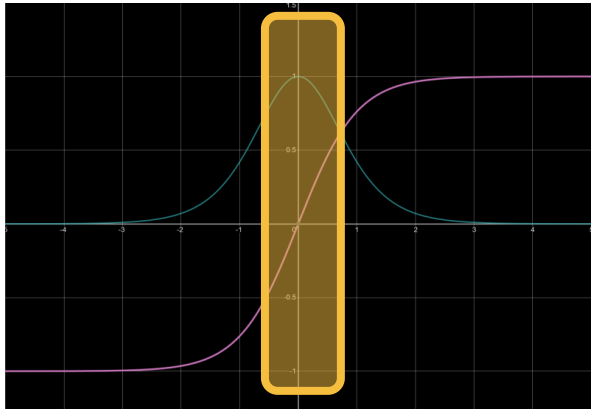
$$w_i = c \quad \forall i$$



Common approach is **small normally distributed random numbers**

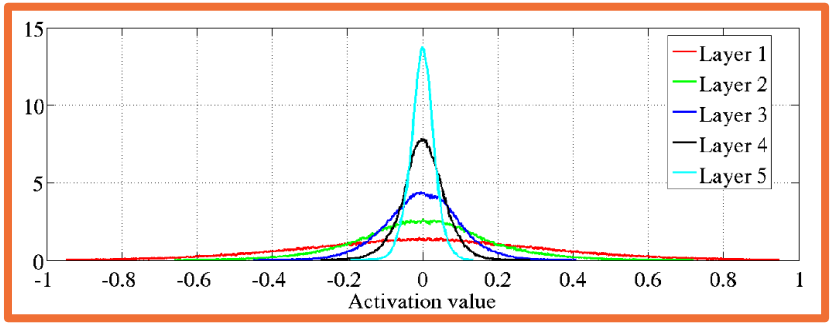
- E.g. $N(\mu, \sigma)$ $where$ $\mu = 0, \sigma = 0.01$
- **Small weights** are preferred since no feature/input has prior importance
- Keeps the model within the **linear region of most activation functions**





Deeper networks (with many layers) are more sensitive to initialization

- With a deep network, **activations (outputs of nodes) get smaller**
  - Standard deviation reduces significantly
- **Leads to small updates –** smaller values multiplied by upstream gradients



Distribution of activation values of a network with tanh non-linearities, for increasingly deep layers

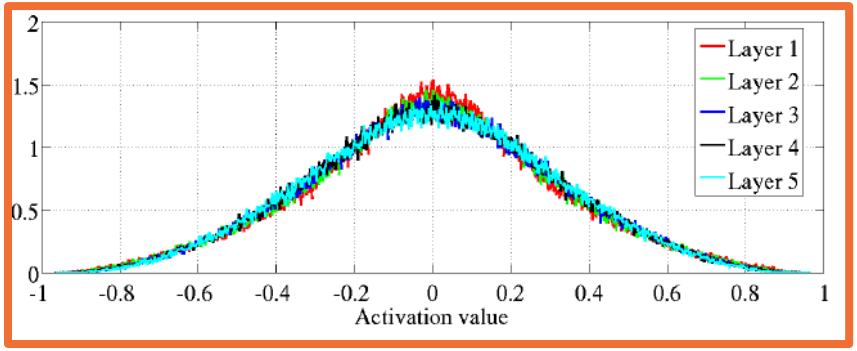
From "Understanding the difficulty of training deep feedforward neural networks." AISTATS, 2010.

Ideally, we'd like to maintain the variance at the output to be similar to that of input!

- This condition leads to a **simple initialization rule**, sampling from uniform distribution:

$$\text{Uniform}\left(-\frac{\sqrt{6}}{n_j+n_{j+1}}, +\frac{\sqrt{6}}{n_j+n_{j+1}}\right)$$

  - Where n_j is **fan-in** (number of input nodes) and n_{j+1} is **fan-out** (number of output nodes)



Distribution of activation values of a network with tanh non-linearities, for increasingly deep layers

From "Understanding the difficulty of training deep feedforward neural networks." AISTATS, **2010.**

In practice, **simpler versions** perform empirically well:

$$N(0,1) * \sqrt{\frac{1}{n_j}}$$

- This analysis holds for **tanh or similar activations**.
- Similar analysis for **ReLU activations** leads to:

$$N(0,1) * \sqrt{\frac{1}{n_j/2}}$$

"Delving Deep into Rectifiers:Surpassing Human-Level Performance on ImageNet Classification", ICCV, 2015.

Summary

Key takeaway: **Initialization matters!**

- Determines the **activation** (output) statistics, and therefore **gradient statistics**
- If gradients are **small**, no learning will occur and no improvement is possible!
- Important to reason about **output/gradient statistics** and analyze them for new layers and architectures

Georgia Tech

Normalization, Preprocessing, and Augmentation

Importance of Data

In deep learning, **data drives learning** of features and classifier

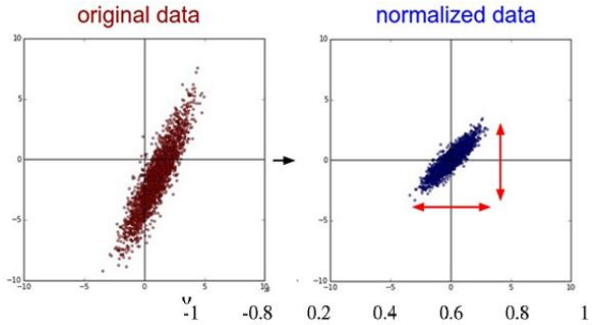
- Its **characteristics** are therefore extremely important
- Always **understand your data**!
- **Relationship** between output statistics, layers such as non-linearities, and gradients is important

Georgia Tech

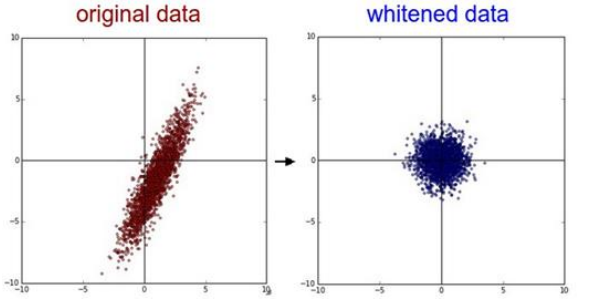
Just like initialization, **normalization** can **improve gradient flow and learning**

Typically **normalization methods** apply:

- Subtract mean, divide by standard deviation **(most common)**
  - This can be done **per** dimension
- Whitening, e.g. through Principle Component Analysis (PCA) **(not common)**



Data after subtracting mean, dividing by standard deviation



Data after whitening

Figure from slides by Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

- We can try to come up with a *layer* that can normalize the data across the neural network
- **Given:** A mini-batch of data $[\boldsymbol{B} \times \boldsymbol{D}]$ where $\boldsymbol{B}$ is batch size
- Compute mean and variance **for each dimension $\boldsymbol{d}$**

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

From: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, Sergey Ioffe, Christian Szegedy

Normalize data

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$$

Note: This part does not involve new parameters

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

From: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, Sergey Ioffe, Christian Szegedy

- We can give the model flexibility through **learnable parameters γ (scale) and β (shift)**
- Network can learn to **not normalize** if necessary!
- This layer is called a **Batch Normalization (BN) layer**

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

From: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, Sergey Ioffe, Christian Szegedy

Learnable Scaling and Offset

Some Complexities of BN

During inference, stored mean/variances calculated on training set are used

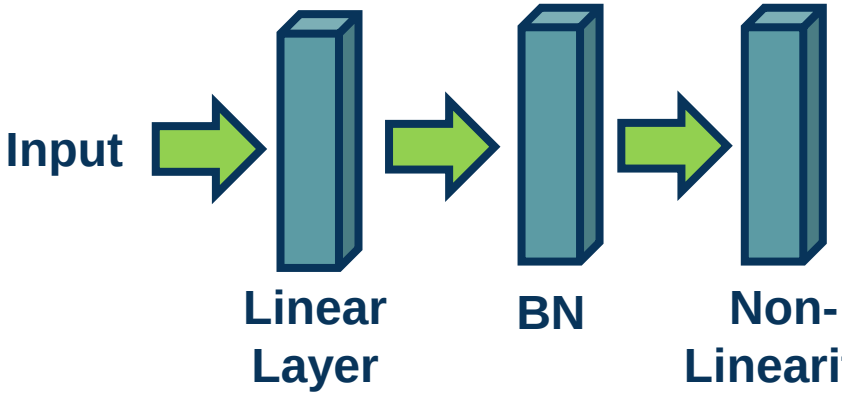
Sufficient batch sizes must be used to get stable per-batch estimates during training

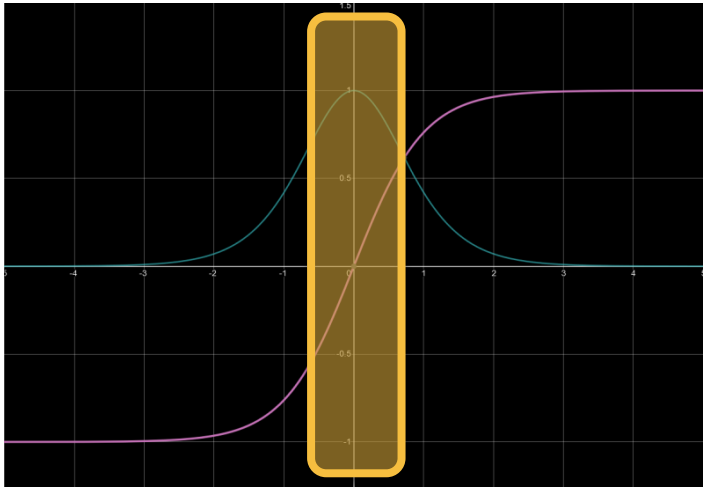
- This is especially an issue when **using multi-GPU or multi-machine training**
- **Use torch.nn.SyncBatchNorm** to estimate batch statistics in these settings

Georgia Tech

Normalization especially important before **non-linearities**!

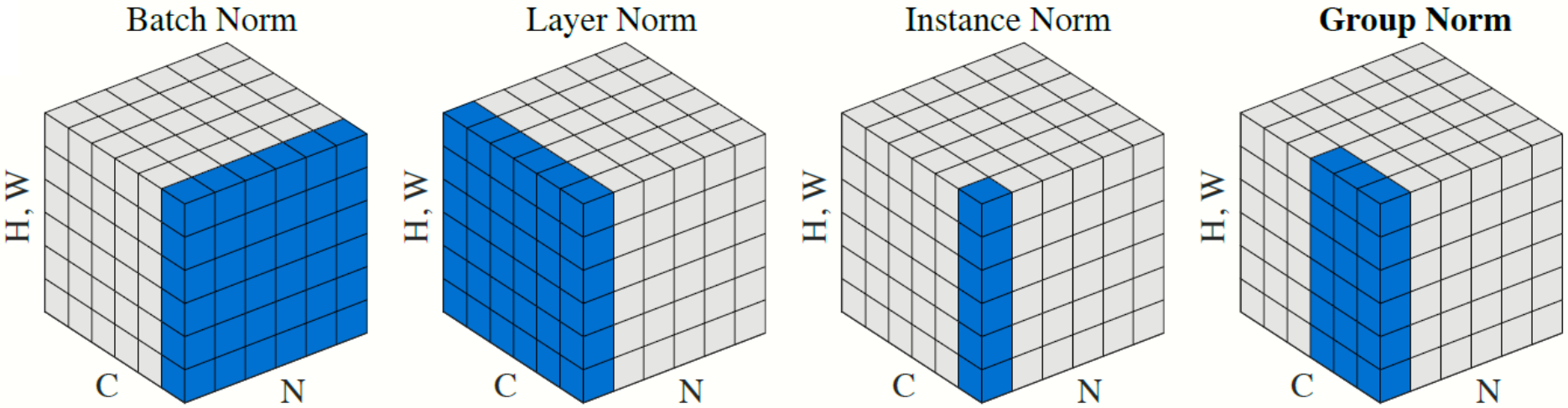
- Very low/high values (un-normalized/imbalanced data) cause **saturation**





Where to Apply BN

Batch normalization unstable for small batch sizes



From: Group Normalization, Wu et al.

Generalization of BN

There are **many variations of batch normalization**

- See Convolutional Neural Network lectures for an example

Resource:

- **Blog - Normalization**

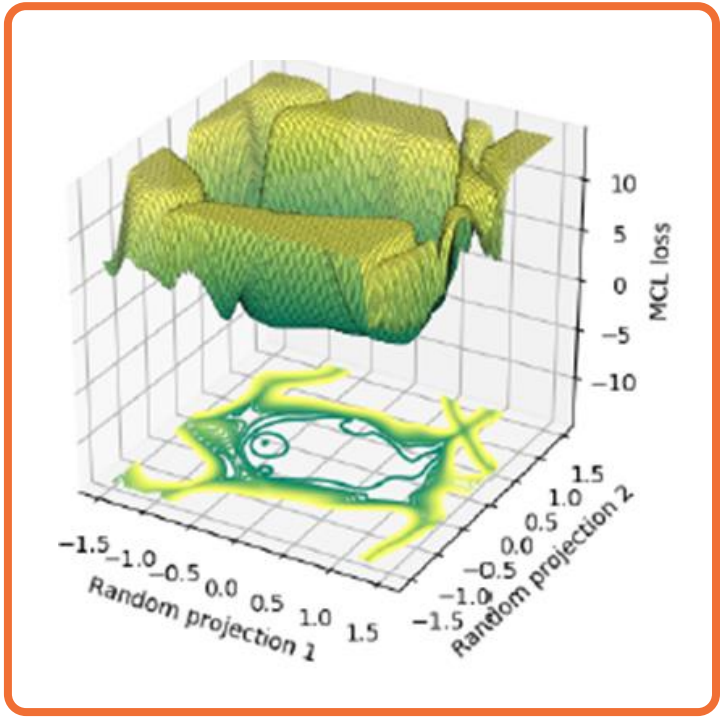
Optimizers

Deep learning involves **complex, compositional, non-linear functions**

The **loss landscape** is extremely **non-convex** as a result

There is **little direct theory** and a **lot of intuition/rules of thumbs** instead

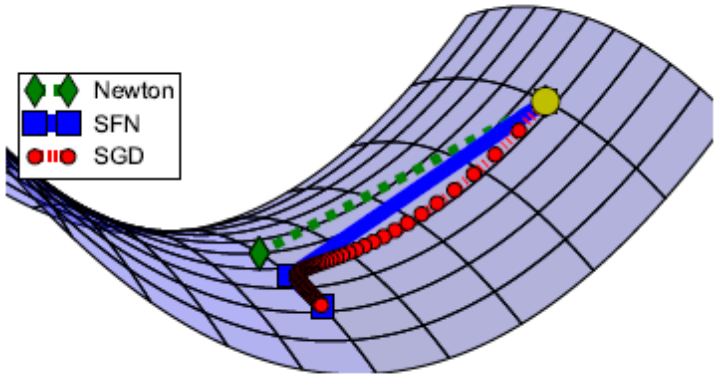
- Some insight can be gained via theory for simpler cases (e.g. convex settings)



It used to be thought that **existence of local minima is the main issue** in optimization

There are other **more impactful issues**:

- Noisy gradient estimates
- Saddle points
- Ill-conditioned loss surface

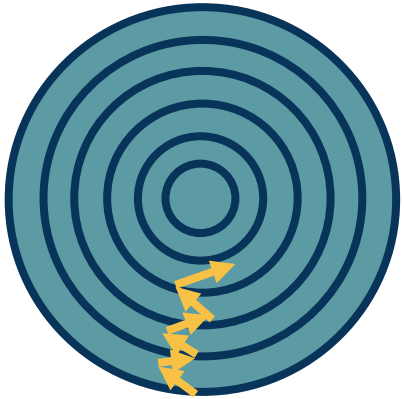


Saddle Point

From: Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, Dauphi et al., 2014.

- We use a **subset of the data at each iteration** to calculate the loss (& gradients)
- This is an **unbiased** estimator but can have high variance
- This results in **noisy steps** in gradient descent

$$L = \frac{1}{M}\sum L\left(f(x_i, W), y_i\right)$$



Noisy Gradients

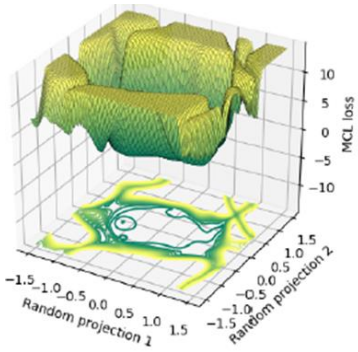
Georgia Tech

Several **loss surface geometries** are difficult for optimization

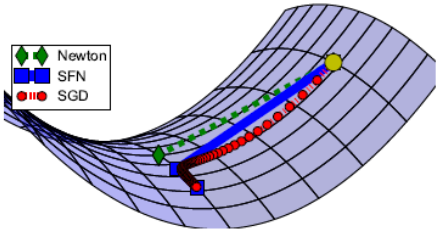
Several types of minima: Local minima, plateaus, saddle points

Saddle points are those where the gradient of orthogonal directions are zero

- But they **disagree** (it's min for one, max for another)



Plateau



Saddle Point

- Gradient descent takes a step in the **steepest direction** (negative gradient)

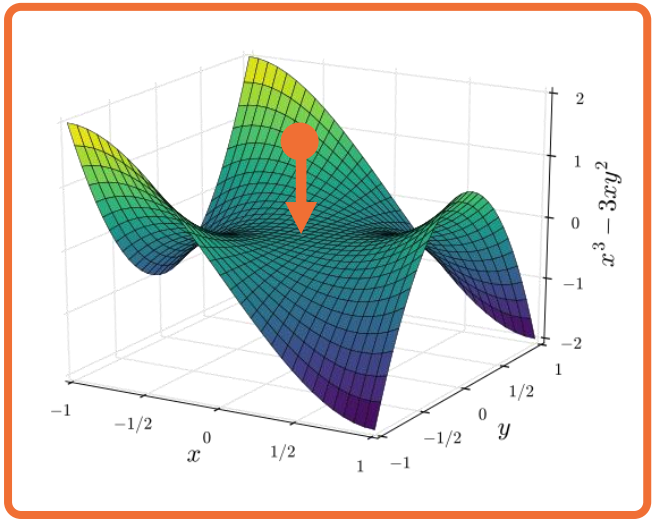
$$w_i = w_{i-1} - \alpha \frac{\partial L}{\partial w_i}$$

- **Intuitive idea:** Imagine a ball rolling down loss surface, and use **momentum** to pass flat surfaces

$$v_i = \beta v_{i-1} + \frac{\partial L}{\partial w_{i-1}}$$ **Update Velocity (starts as 0, $\beta = 0.99$)**

$$w_i = w_{i-1} - \alpha v_i$$ **Update Weights**

- Generalizes SGD ($\beta = 0$)



Adding Momentum

Georgia Tech

- Velocity term is an **exponential moving average** of the gradient

$$v_i = \beta v_{i-1} + \frac{\partial L}{\partial w_{i-1}}$$

$$v_i = \beta(\beta\, v_{i-2} + \frac{\partial L}{\partial w_{i-2}}) + \frac{\partial L}{\partial w_{i-1}}$$

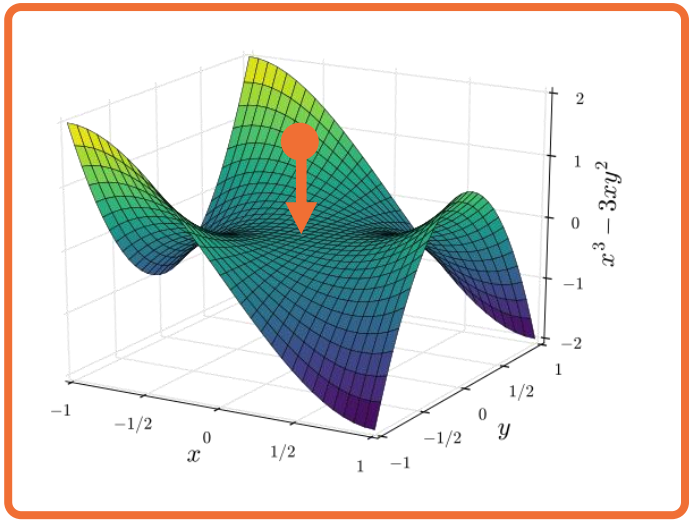
$$= \beta^2 v_{i-2} + \beta \frac{\partial L}{\partial w_{i-2}} + \frac{\partial L}{\partial w_{i-1}}$$

- There is a **general class of accelerated gradient methods**, with some theoretical analysis (under assumptions)

Equivalent formulation:

$$v_i = \beta v_{i-1} - \alpha \frac{\partial L}{\partial w_{i-1}}$$ **Update Velocity (starts as 0)**

$$w_i = w_{i-1} + v_i$$ **Update Weights**



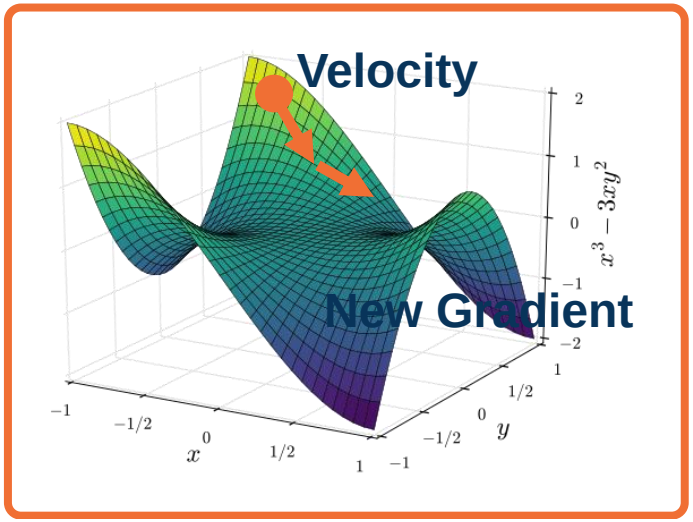
Key idea: Rather than combining velocity with current gradient, go along velocity **first** and then calculate gradient at new point

- We know velocity is probably a **reasonable direction**

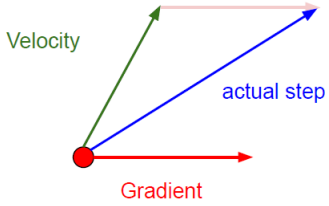
$$\hat{\boldsymbol{w}}_{\boldsymbol{i-1}} = \boldsymbol{w}_{\boldsymbol{i-1}} + \boldsymbol{\beta v}_{\boldsymbol{i-1}}$$

$$\boldsymbol{v}_{\boldsymbol{i}} = \boldsymbol{\beta v}_{\boldsymbol{i-1}} + \frac{\boldsymbol{\partial L}}{\boldsymbol{\partial \hat{w}}_{\boldsymbol{i-1}}}$$

$$\boldsymbol{w}_{\boldsymbol{i}} = \boldsymbol{w}_{\boldsymbol{i-1}} - \boldsymbol{\alpha}\, \boldsymbol{v}_{\boldsymbol{i}}$$



Momentum update:



Nesterov Momentum

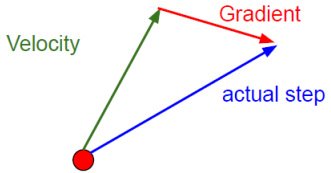


Figure Credit: Fei-Fei Li, Justin Johnson, Serena Yeung, CS 231n

Momentum

Note there are **several equivalent formulations** across deep learning frameworks!

Resource:
https://medium.com/the-artificial-impostor/sgd-implementation-in-pytorch-4115bcb9f02c

Georgia Tech

- **Activation Functions**: Use ReLU, GeLU, etc.
- **Initialization:** Important for initial activation and gradient statistics
- **Normalization:** Use dynamic normalization with learnable parts
- **Optimization:** Use momentum (helps w/ local minima, etc.)
  - **Next:** More sophisticated gradient history/statistics in update rule