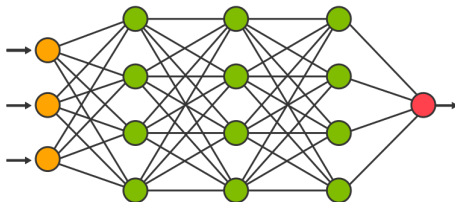


Introduction to Neural Networks

ML Instruction Team, Fall 2022

CE Department
Sharif University of Technology



Problem: OverFitting in a Neural Network

- Why does overfitting happen in a neural network?
 - ▷ There are **Too many free parameters**.

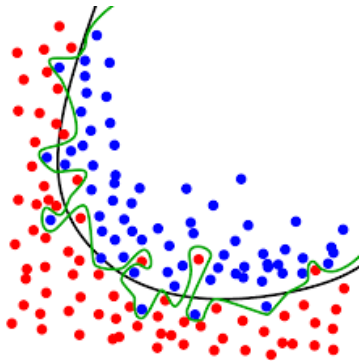


Figure: OverFitting in a neural network, [Source](#)

Solution 1: L1/L2 Regularization

- It is like a linear regression regularizer.
- Sum the regularizer term for every **layer weight**!

$$L = \frac{1}{N} \sum_{i=1}^N L(\phi(x_i), y_i) + \lambda \sum_{i,j,k} R(W_{j,k}^{(i)})$$

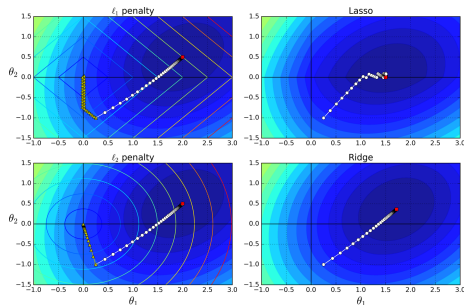


Figure: Convergence diagram for different losses,

Source

L1/L2 Regularization

L1/L2 regularizer functions (review)

$$L1 : R(w) = |w|$$

$$L2 : R(w) = w^2$$

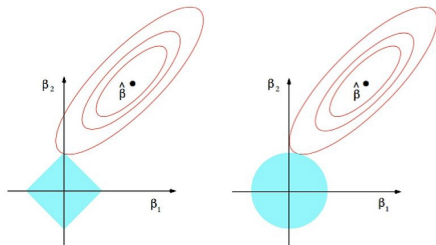


Figure: L1/L2 regularizers' solution diagram, [Source](#)

You can also combine the two different regularizers (Elastic Net).

$$R(w) = \beta w^2 + |w|$$

Solution 2: Early Stopping

- Stop the training procedure when the validation error is **minimum**.

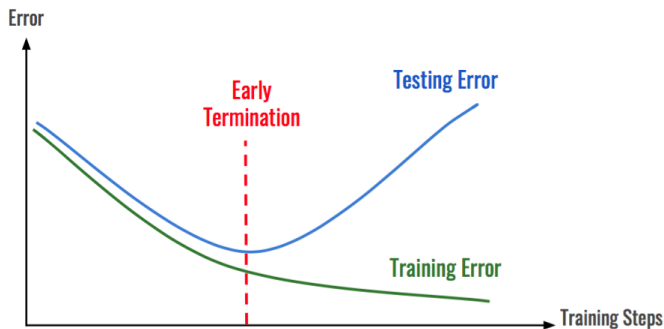


Figure: Early Stopping diagram, [Source](#)

Dropout: Training Time

- In each forward pass, **randomly** set some neurons to zero.
- The probability of dropping out for each neuron, which is called **dropout rate**, is a hyperparameter.
 - ▷ 0.5 is a common dropout rate.
- The probability of not dropping out is also called the **keep probability**.

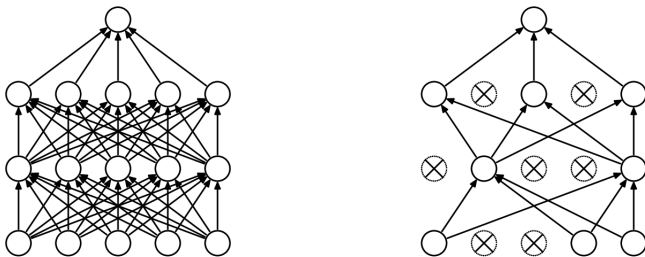


Figure: Behavior of dropout at training time, Source

Dropout: Why can this possibly be a good idea?

- Dropout-trained neurons are unable to **co-adapt** with their surrounding neurons.
- They also can't depend too heavily on a small number of input neurons.
- They become less responsive to even little input changes.
- The result is a stronger network that **generalizes** better.



Figure: Discrimination of neurons at training time. [1]

Dropout: Why can this possibly be a good idea?

- Dropout trains a **large ensemble of models** that share parameters.
- Every possible dropout state for neurons of a network, which is called a **mask**, is one model.
- A fully connected network with 4096 neurons has $2^{4096} \sim 10^{1233}$ possible masks! There are only 10^{82} atoms in the universe!

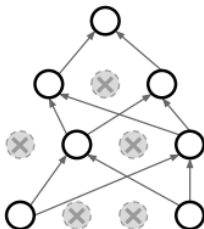


Figure: Behavior of dropout at training time. [1]

Dropout: Test Time

- Dropout makes our output random at training time.

$$y = f_W(x, \underbrace{z}_{\text{random mask}})$$

- We want to **average out** the randomness at test time,

$$y = f(x) = E_z[f(x, z)] = \int p(z) f(x, z) dz$$

- But this integral seems complicated.
- Let's approximate the integral for a superficial layer where dropout rate is 0.5.

Dropout: Test Time

$$\begin{aligned}
 E_{train}[a] &= \frac{1}{4}(w_1x + w_2y) + \frac{1}{4}(w_1x + 0y) \\
 &\quad + \frac{1}{4}(0x + w_2y) + \frac{1}{4}(0x + 0y) \\
 &= \frac{1}{2}(w_1x + w_2y)
 \end{aligned}$$

$$E_{test}[a] = w_1x + w_2y$$

$$\Rightarrow E_{test}[a] = \underbrace{0.5}_{\text{keep probability}} E_{train}[a]$$

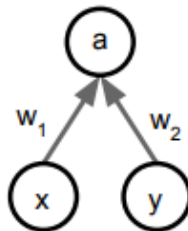


Figure: Simple neural network. [1]

Solution: Batch Norm Layer

- It is used to **normalize** the data.

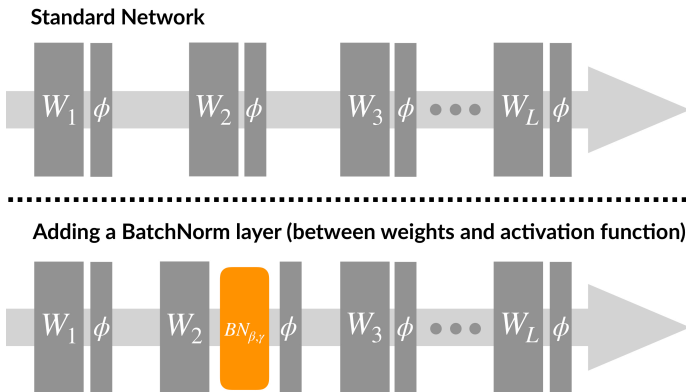


Figure: The suggested place to put a BatchNorm layer, [Source](#)

Batch Norm: Training Time

- First, it zero-centers and normalizes the batch.

$$\mu_B := \frac{1}{N_B} \sum x_B^{(i)}$$
$$\sigma_B^2 := \frac{1}{N_B} \sum (x_B^{(i)} - \mu_B)^2$$
$$\hat{x}_B^{(i)} = \frac{x_B^{(i)} - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$$

- Then, scales and shifts the batch with two learnable parameters γ, β .

$$y_B^{(i)} = \gamma \hat{x}_B^{(i)} + \beta$$

Batch Norm: Test Time

- To zero-center and normalize the input, we need the average and variance of the whole data.
- Those parameters can be acquired during the training.
- Therefore, we need two more trainable parameters.

$$\mu_D := \frac{1}{N} \sum x^{(i)}$$
$$\sigma_D^2 := \frac{1}{N} \sum (x^{(i)} - \mu_D)^2$$

Batch Norm: Test Time

- Moving average //todo

Batch Norm: Performance

- Normalizing the data improves the convergence speed by a considerable amount.

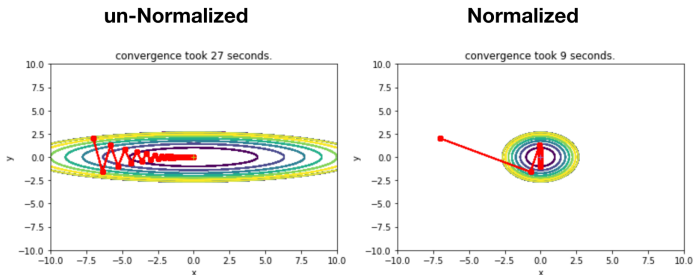


Figure: BatchNorm performance. Convergence speed is increased by 200%, [Source](#)

Batch Norm

■ Pros

- ▶ Vanishing/Exploding gradient problem is reduced by a considerable amount.
- ▶ You can use even saturating activation functions.
- ▶ The network is much less sensitive to the initial weight.
- ▶ We're able to use larger learning rates, which speeds up the training.
- ▶ It also acts as a regularizer.
 - There is no need for other regularizer techniques.

Batch Norm

Cons

- It increases model parameters and prediction latency.

- Solution: after the training procedure, we can mix the BatchNorm layer with its previous layer to hold the prediction latency.

$$\begin{aligned}
 x'^{(i)} &= Wx^{(i)} + b \\
 y^{(i)} &= \frac{x'^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta
 \end{aligned}
 \quad \Rightarrow \quad
 \begin{aligned}
 y^{(i)} &= W'x^{(i)} + b' \\
 W' &:= \frac{1}{\sqrt{\sigma^2 + \epsilon}} W \\
 b' &:= \beta + \frac{b - \mu}{\sqrt{\sigma^2 + \epsilon}}
 \end{aligned}$$

- Where μ and σ are the mean and variance for the test time.

Gradient Clipping

- What will happen in case of a large gradient value?
- The gradient descent will take us **far away** from our local position.

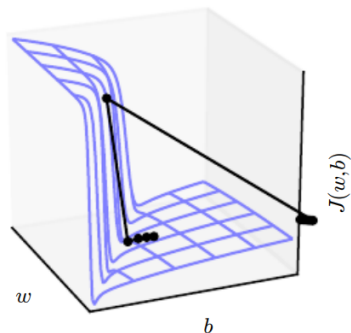


Figure: The problem of large gradient value [2].

Gradient Clipping

- Solve this problem simply by clipping gradient.
- Two approaches to do so:
 - ▷ Clipping by value
 - ▷ Clipping by norm

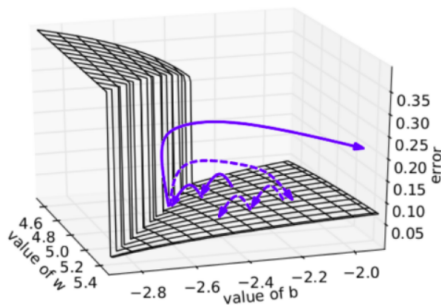


Figure: The effect of gradient clipping. Instead of solid line following dotted line will lead us to minimum, [Source](#)

Gradient Clipping by value

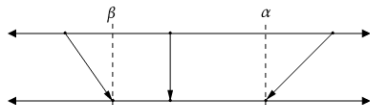
- Set a max (α) and min (β) threshold value.
- For each index of gradient g_i if it is lower or greater than your threshold clip it:

if $g_i > \alpha$:

$$g_i \leftarrow \alpha$$

else if $g_i < \beta$:

$$g_i \leftarrow \beta$$



Gradient Clipping by value

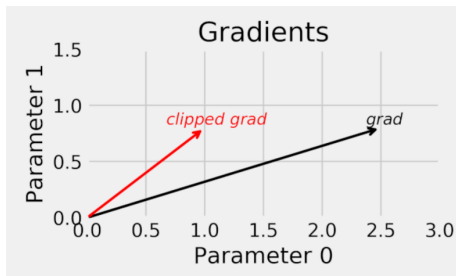


Figure: The effect of clipping by value. [Source](#)

- Clipping by value **will change gradient direction**.
- To preserve direction use clipping by norm.

Gradient Clipping by norm

- Clip the norm $\|g\|$ of the gradient g before updating parameters:

if $\|g\| > v$:

$$g \leftarrow \frac{g}{\|g\|} v$$

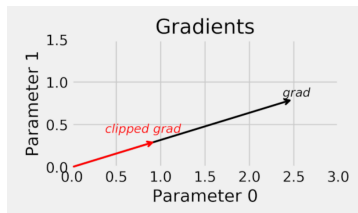


Figure: The effect of clipping by norm. [source](#)

v is the threshold for clipping which is a hyperparameter.

- Gradient clipping saves the direction of gradient and controls its norm.

Gradient Clipping

- Clipping by **value** will **change the direction** of the gradient, so it will send us to a bad neighborhood.
- Clipping by **norm** will **preserve the direction** and just control the value.
- So it is better to use clipping by **norm**.

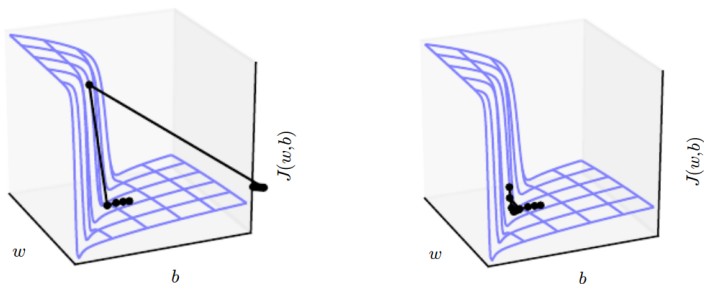


Figure: The "cliffs" landscape (left) without gradient clipping and (right) with gradient clipping [2].

Weight Initialization

- Is initialization really necessary?
- What are the impacts of initialization?
- A bad initialization may increase convergence time or even make optimization diverge.
- How to initialize?
 - ▷ Zero initialization
 - ▷ Random initialization

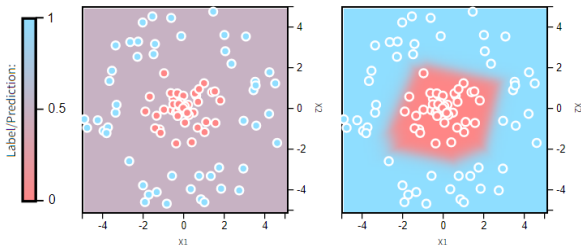


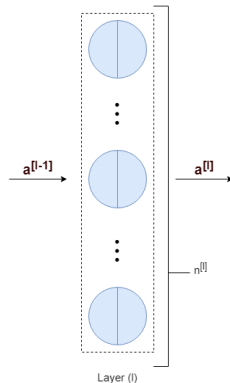
Figure: The output of a three layer network after about 600 epoch. (left) using a bad initialization method and (right) using an appropriate initialization [3].

Weight Initialization

Let's review some notations before we continue:

$$\begin{cases} n^{[l]} := \text{layer } l \text{ neurons number,} \\ W^{[l]} := \text{layer } l \text{ weights,} \\ b^{[l]} := \text{layer } l \text{ biases,} \\ a^{[l]} := \text{layer } l \text{ outputs} \end{cases}$$

$$\begin{cases} \text{fan}_{\text{in}}^{[l]} = n^{[l-1]} & (\text{layer } l \text{ number of inputs}), \\ \text{fan}_{\text{out}}^{[l]} = n^{[l]} & (\text{layer } l \text{ number of outputs}), \\ \text{fan}_{\text{avg}}^{[l]} = \frac{n^{[l-1]} + n^{[l]}}{2} \end{cases}$$



Weight Initialization: Zero Initialization

Zero Initialization method:

$$\begin{cases} W^{[l]} = \mathbf{0}, \\ b^{[l]} = \mathbf{0} \end{cases}$$

- Simple but perform very poorly. (why?)
- Zero initialization will lead each neuron to learn the same feature
- This problem is known as network **failing to break symmetry**
- In fact any constant initialization suffers from this problem.

Weight Initialization: Zero Initialization

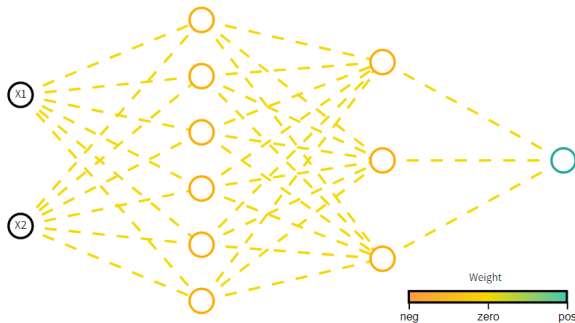


Figure: As we can see network has failed to break symmetry. There has been no improvement in weights after about 600 epochs of training [3].

- We need to break symmetry. How? using randomness.

Weight Initialization: Random Initialization

- To use randomness in our initialization we can use uniform or normal distribution:

General Uniform Initialization:

$$\begin{cases} W^{[l]} \sim U(-r, +r), \\ b^{[l]} = 0 \end{cases}$$

General Normal Initialization:

$$\begin{cases} W^{[l]} \sim \mathcal{N}(\mu = 0, \sigma^2), \\ b^{[l]} = 0 \end{cases}$$

- But this is really crucial to choose r or σ properly.

Weight Initialization: Random Initialization

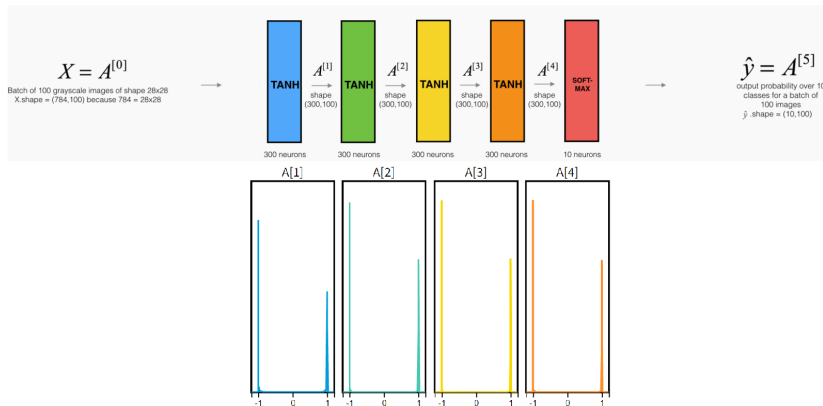


Figure: Uniform initialization problem. On the top, you can see the model architecture, and on the bottom, you can see the density of each layer's output. Model has trained on MNIST dataset for 4 epoch. Weights are initialized randomly from $U(\frac{-1}{\sqrt{n^{[l-1]}}}, \frac{1}{\sqrt{n^{[l-1]}}})$ [3].

Weight Initialization: Random Initialization

- How to choose r or σ ?
- We need to follow these rules:
 - ▷ keep the mean of the activations zero.
 - ▷ keep the variance of the activations same across every layer.

Xavier Initialization:

- For Uniform distribution use:

$$r = \sqrt{\frac{3}{\text{fan}_{\text{avg}}}}$$

- For Normal distribution use:

$$\sigma^2 = \frac{1}{\text{fan}_{\text{avg}}}$$

(You can read about why this method works at [3].)

Weight Initialization: Xavier Initialization

- Xavier initialization works well on **Tanh**, **Logestic** or **Sigmoid** activation function.

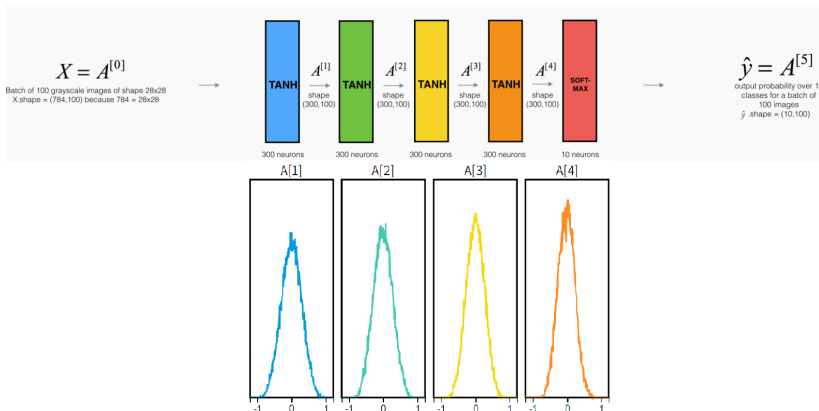


Figure: Vanishing gradient is no longer problem using Xavier initialization. Model has trained on MNIST dataset for 4 epoch. [3].

Weight Initialization: He Initialization

- Different method has proposed for different activation functions.

He Initialization:

- For Normal distribution:

$$\sigma^2 = \frac{2}{n^{[l]}}$$

- For Uniform distribution:

$$r = \sqrt{3\sigma^2}$$

- He initialization works well on **ReLU and its variants**.

Various GD types

- So far you got familiar with gradient-based optimization.
- If $\mathbf{g} = \nabla_{\boldsymbol{\theta}} \mathcal{J}$, then we will update parameters with this simple rule:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \mathbf{g}$$

- But there is one question here, how to compute \mathbf{g} ?
- Based on how we calculate \mathbf{g} we will have different types of gradient descent:
 - ▷ Batch Gradient Descent
 - ▷ Stochastic Gradient Descent
 - ▷ Mini-Batch Gradient Descent

Various GD types

Recap:

Training cost function (\mathcal{J}) over a dataset usually is the average of loss function (\mathcal{L}) on entire training set, so for a dataset $\mathcal{D} = \{d_i\}_{i=1}^n$ we have:

$$\mathcal{J}(\mathcal{D}) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(d_i; \theta)$$

For example:

$$H(p, q) = -\frac{1}{m} \sum_{i=1}^m \sum_{j=1}^k y_j^{(i)} \log(p(y_j^{(i)}))$$

$$\text{MSE}(y, \hat{y}) = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2$$

$$\text{MAE}(y, \hat{y}) = \frac{1}{m} \sum_{i=1}^m |(y_i - \hat{y}_i)|$$

Various GD types: Batch Gradient Descent

- In this type we use **entire training set** to calculate gradient.

Batch Gradient:

$$\mathbf{g} = \frac{1}{n} \sum_{i=1}^n \nabla_{\boldsymbol{\theta}} \mathcal{L}(d_i, \boldsymbol{\theta})$$

- Using this method with very large training set:
 - ▷ Your data can be too large to process in your memory.
 - ▷ It requires a lot of processing to compute gradient for all samples.
- Using exact gradient may lead us to local minima.
- Moving noisy may help us get out of this local minimas.

Various GD types: Batch Gradient Descent

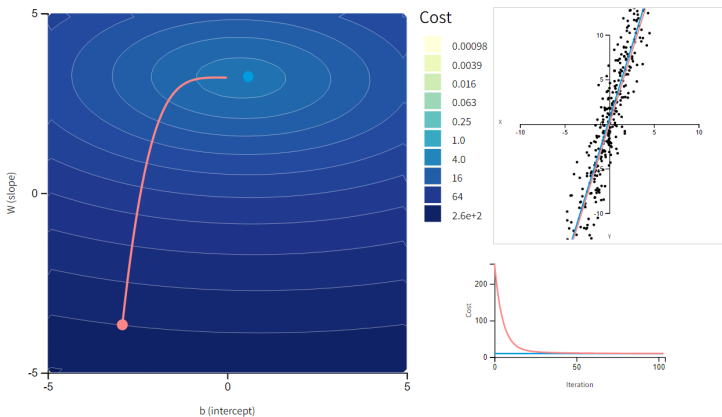


Figure: Optimization of parameters using BGD. Movement is very smooth [4].

Various GD types: Stochastic Gradient Descent

- Instead of calculating exact gradient, we can estimate it using our data.
- This is exactly what SGD does, it estimates gradient using **only single data point**.

Stochastic Gradient:

$$\hat{g} = \nabla_{\theta} \mathcal{L}(d_i, \theta)$$

- As we use an approximation of gradient, instead of gently decreasing, the cost function will bounce up and down and decrease only on average.
- This method is really computationally efficient cause we only need to calculate gradient for one point per iteration.

Various GD types: Stochastic Gradient Descent

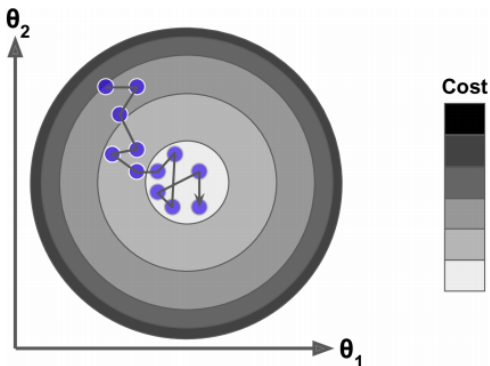


Figure: Optimization of parameters using SGD. As we expect, the movement is not that smooth [4].

Various GD types: Mini-Batch Gradient Descent

- In this method we still use estimation idea But use **a batch of data** instead of one point.

Mini-Batch Gradient:

$$\hat{\mathbf{g}} = \frac{1}{|\mathcal{B}|} \sum_{d \in \mathcal{B}} \nabla_{\boldsymbol{\theta}} \mathcal{L}(d, \boldsymbol{\theta}), \quad \mathcal{B} \subset \mathcal{D}$$

- This is a better estimation than SGD.
- With this way we can get a performance boost from hardware optimization, especially when using GPUs.
- Batch size ($|\mathcal{B}|$) is a hyperparameter you need to tune.

Various GD types: Mini-Batch Gradient Descent

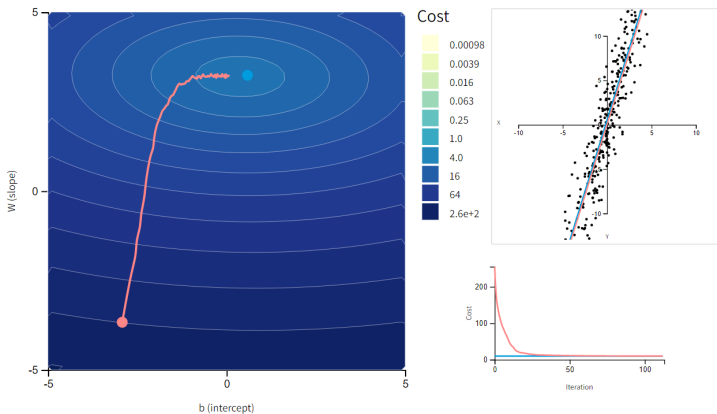


Figure: Optimization of parameters using MBGD. The movement is much smoother than SGD and behave like BGD [4].

Various GD types

- Now that we know what a batch is, we can define epoch and iteration:
 - ▷ One **Epoch** is when an entire dataset is passed forward and backward through the network only once.
 - ▷ One **Iteration** is when a batch is passed forward and backward through the network.

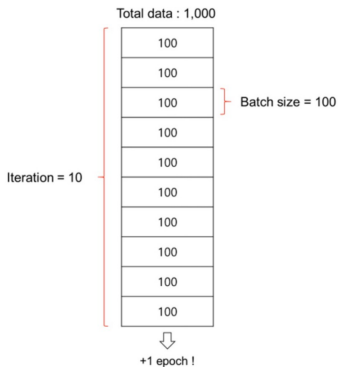


Figure: Epoch vs Iteration, [source](#).

Various GD types

- So we got familiar with different types of GD.
 - ✓ Batch Gradient Descent (BGD)
 - ✓ Stochastic Gradient Descent (SGD)
 - ✓ Mini-Batch Gradient Descent (MBGD)
- The most recommended one is MBGD, because it is computational efficient.
- Choosing the right batch size is important to ensure convergence of the cost function and parameter values, and to the generalization of your model.

Thank You!

Any Question?

References



F.-F. L. . J. J. . S. Yeung, “Training neural networks,” 2018.

http://cs231n.stanford.edu/slides/2018/cs231n_2018_lecture07.pdf.



I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*.

MIT Press, 2016.

<http://www.deeplearningbook.org>.



K. Katanforoosh and D. Kunin, “Initializing neural networks,” 2019.

<https://www.deeplearning.ai/ai-notes/initialization/>.



K. Katanforoosh and D. Kunin, “Parameter optimization in neural networks,” 2019.

<https://www.deeplearning.ai/ai-notes/optimization/>.