13-08-2024

## Data pre-processing

- Mean centering
- Standardization
  - Mean zero and variance one
- Normalization: (0, 1)
- Whitening: axis rotation (decorrelate)
  - Principal component analysis
    - Issue with covariance matrix for large data
    - Estimate correlation matrix

## Weights/Parameters initialization

- Random values: Gaussian distribution with 0 mean and small  $(10^{-2})$  / large  $(10^2)$  standard deviation
  - Problem: Vanishing and exploding gradient
  - $E[o^{l-1}] = E[o^l] \text{ and } Var[o^{l-1}] = Var[o^l]$
  - what about the number of inputs to a neuron?
    - 2 vs 100 inputs to a neuron?
  - Solution: sample from a normal Gaussian distribution with 0 mean and  $\sqrt{1/n_{in}}$  standard deviation
  - $ightharpoonup n_{in}$  number of inputs to a neuron
  - can you see any problem further?
  - what about the number of outputs at a particular layer?
  - \* Xavier initialization: sample from a normal Gaussian distribution with 0 mean and  $\sqrt{2/(n_{in}+n_{out})}$  standard deviation
  - $rac{1}{2}$  number of outputs to a neuron

## Gradient descent strategies

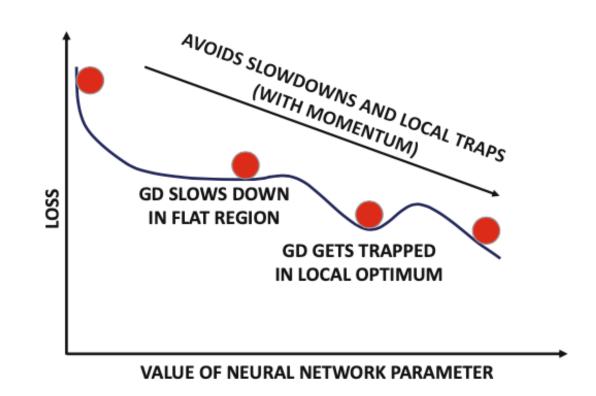
$$\bullet \ W = W - \eta \frac{\partial L}{\partial W}$$

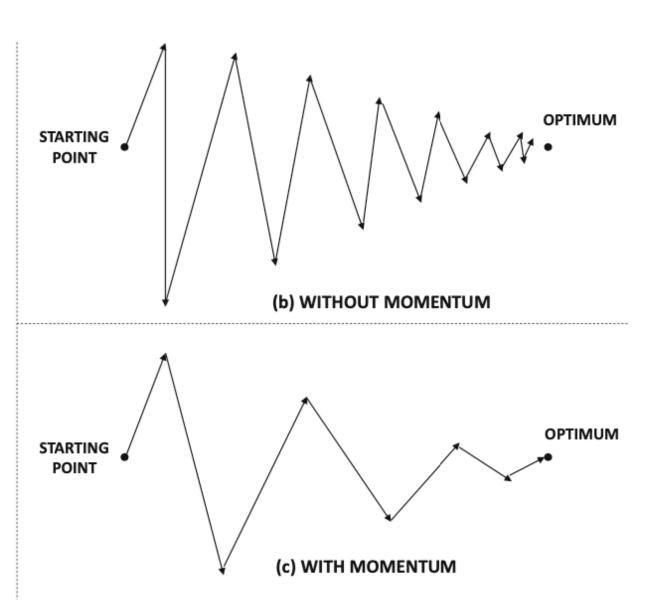
$${
m W}=W+V;\,V=-\eta {\partial L\over \partial W}$$
, velocity

- Problems ?
- Stuck at local optima
- Moment based update
  - Record past velocity and used that in current update

$$V = \beta V - \eta \frac{\partial L}{\partial W}, W = W + V, \beta \in (0,1)$$

$$- \text{ If } \beta = 0 ?$$





▶ Nesterov Momentum: 
$$V = \beta V - \eta \frac{\partial L(W + \beta V)}{\partial W}$$
,  $W = W + V$