Parameter Initialization

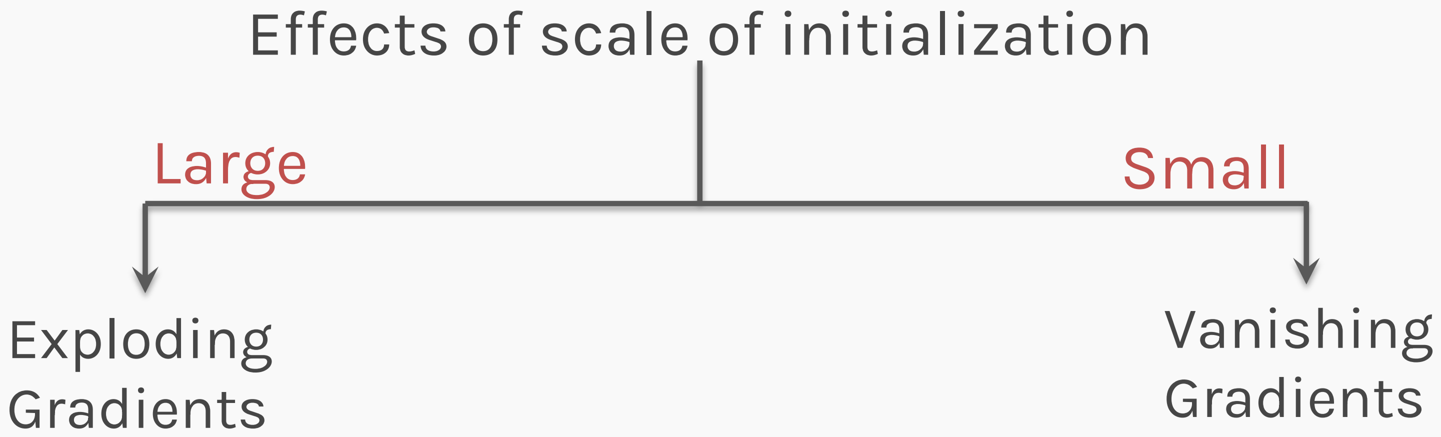
Pavlos Protopapas

Parameter Initialization

Aim:

Break symmetry between units to ensure each unit computes a different function

For this, initialize all weights (not biases) randomly – Gaussian or Uniform



Effects of scale of initialization

Large

Small

Exploding Gradients

Vanishing Gradients

Xavier Initialization

- Heuristics for all outputs have unit variance

- For a fully-connected layer with m inputs:

$$\boxed{W_{ij} \approx N\left(0, \frac{1}{m}\right)}$$

- For ReLU units, it is recommended to have:

$$\boxed{W_{ij} \approx N\left(0, \frac{2}{m}\right)}$$

Normalized Initialization - Kaiming He initialization

- For a fully-connected layer with m inputs and n outputs :

$$\boxed{W_{ij} \approx U\left(-\sqrt{\frac{6}{m+n}}, \sqrt{\frac{6}{m+n}}\right)}$$

- Heuristic trades off between initializing all layers with the same activation and variable variance.

The variance of a W_{ij} is different for different m's and n's

- Sparse variant when m is large
  - Initialize k non-zero weights in each unit

Bias Initialization

Output unit bias

- Marginal statistics of the output in the training set

Hidden unit bias

- Avoid saturation at initialization

Ex: In ReLU, initialize bias to 0.001 instead of 0

Bias Initialization

Output unit bias

- Marginal statistics of the output in the training set

Hidden unit bias

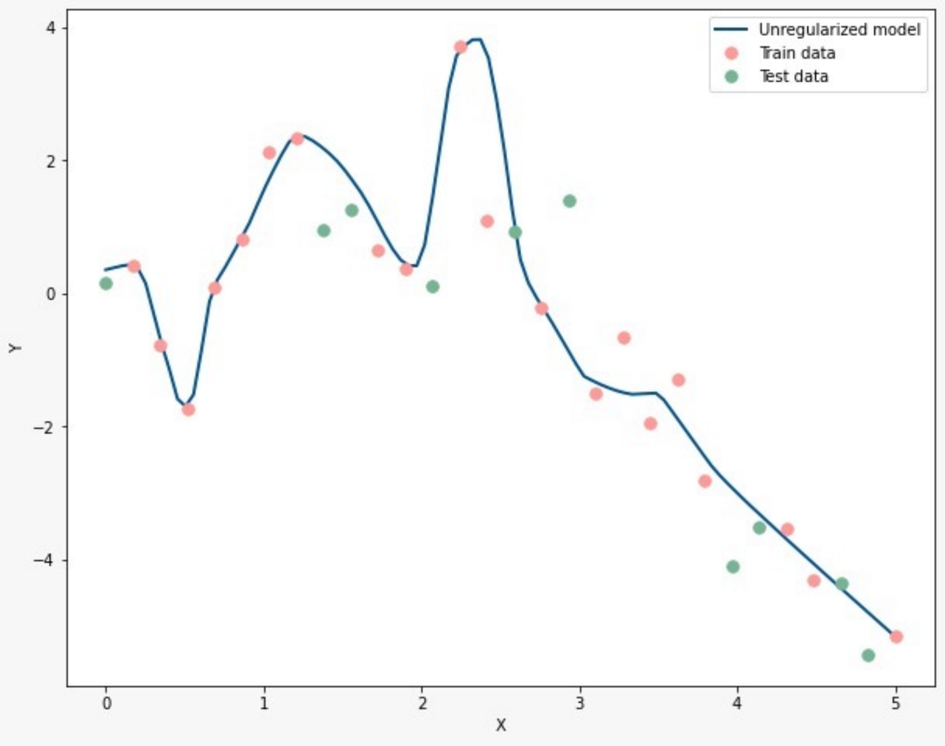
- Avoid saturation at initialization

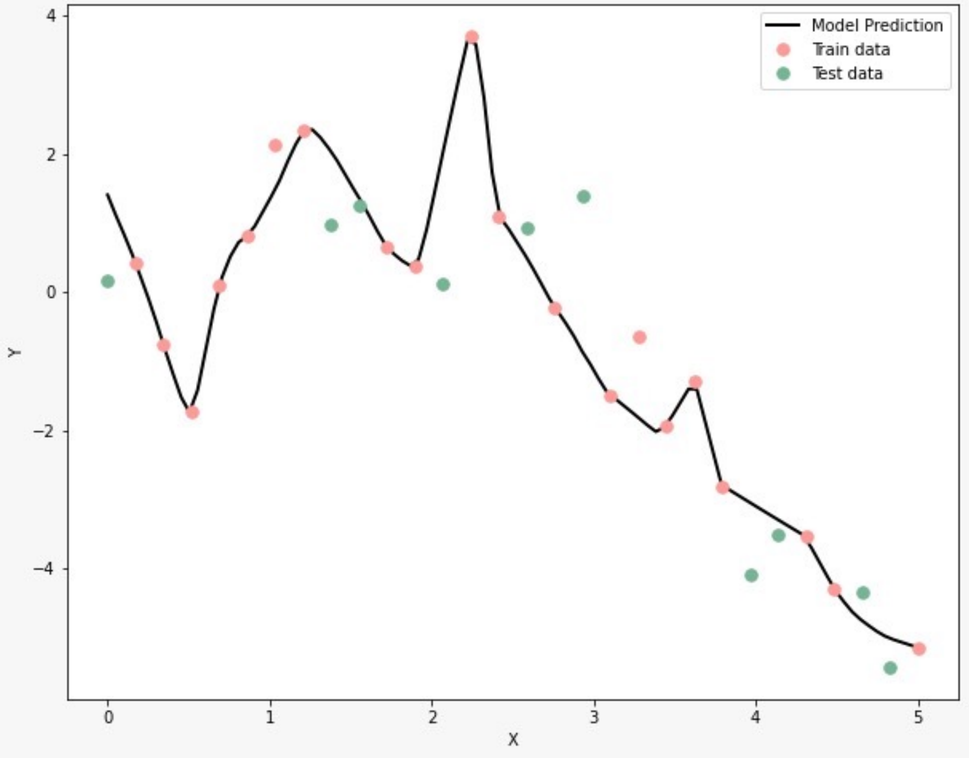
Ex: In ReLU, initialize bias to 0.001 instead of 0

This ensures that all ReLU units fire in the beginning and therefore obtain and propagate some gradient

Synthetic data generated using $y = x\sin x + \epsilon,\ \epsilon \sim \mathrm{N}(0,1)$

Data fitted with a FCNN with 3 hidden layers with 100 nodes per layer, using tanh activation



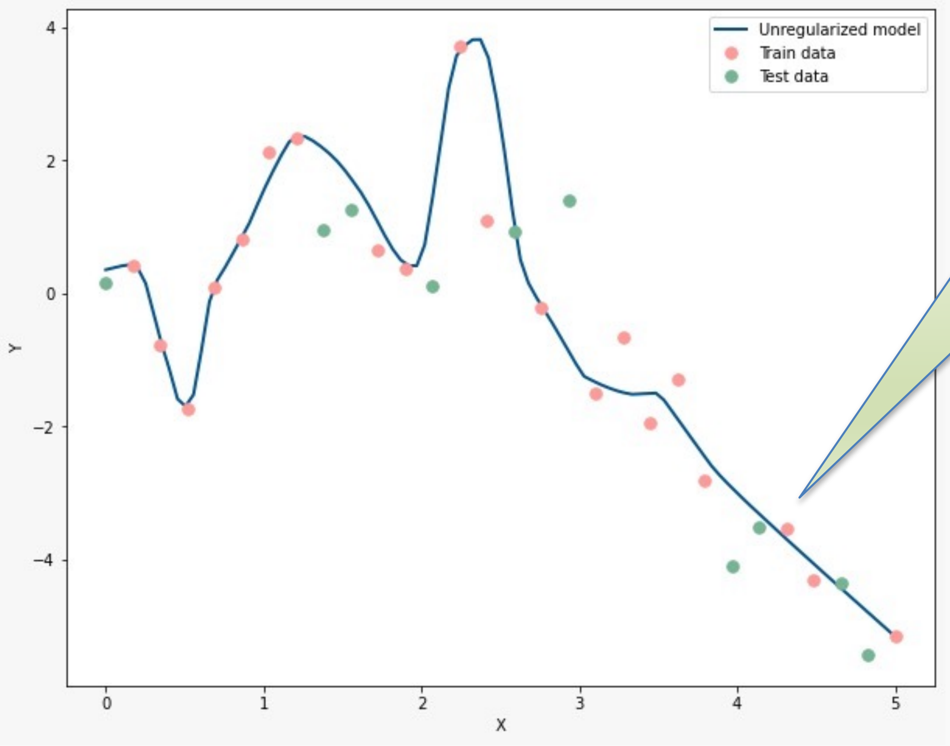


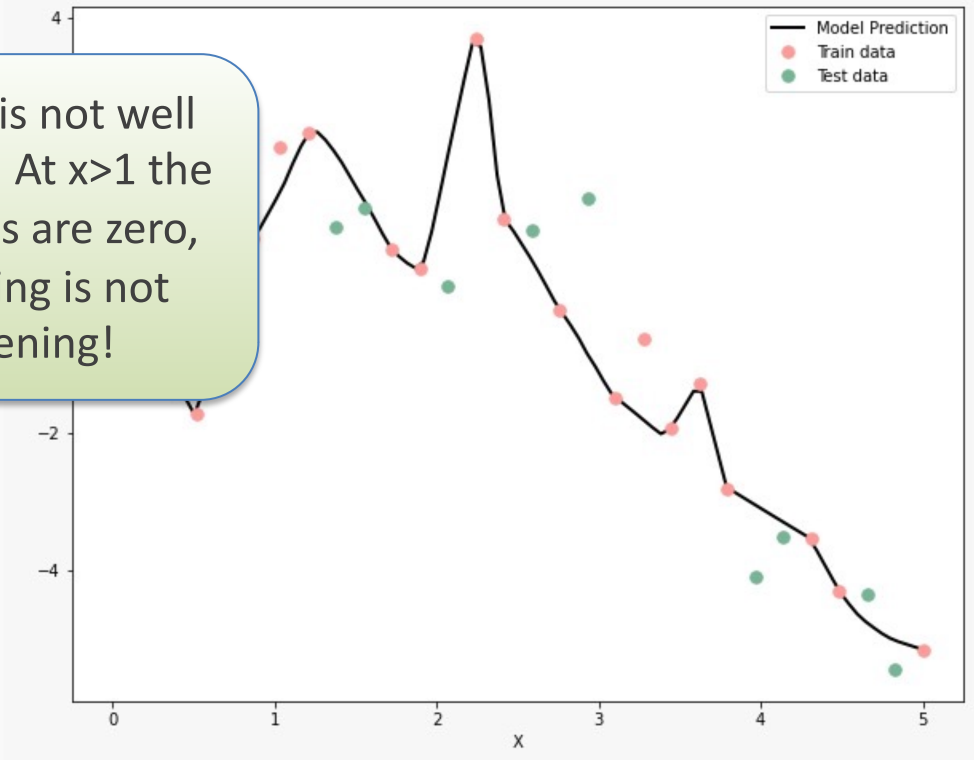
Parameter initialization with Normalized initialization:
$W \sim U[-1,1]$

Parameter initialization with Normalized initialization:
$W \sim U[-5,5]$

Synthetic data generated using $y = x \sin x + \epsilon,\ \epsilon \sim \mathrm{N}(0,1)$

Data fitted with a FCNN with 3 hidden layers with 100 nodes per layer, using tanh activation





This area is not well well fitted. At x>1 the derivatives are zero, so learning is not happening!

Parameter initialization with Normalized initialization: $W \sim U[-1,1]$

Parameter initialization with Normalized initialization: $W \sim U[-5,5]$