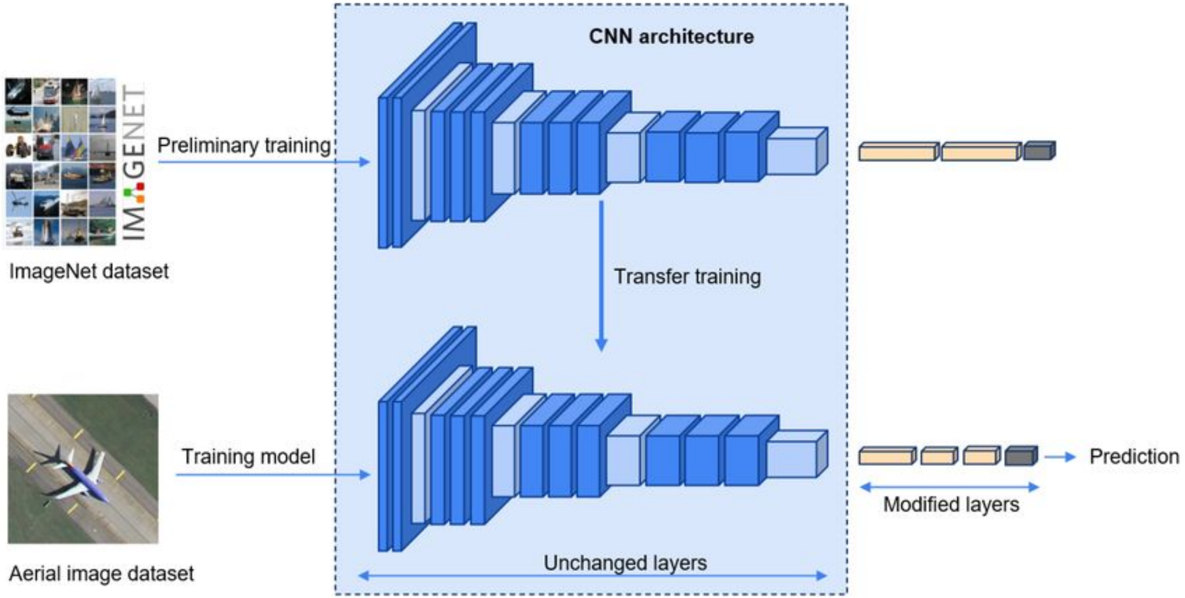
CSE4261: Neural Network and Deep Learning

Lecture: 21.05.2025

Sangeeta Biswas, Ph.D.
Associate Professor,
University of Rajshahi, Rajshahi-6205, Bangladesh

Pre-Trained Based Classifier



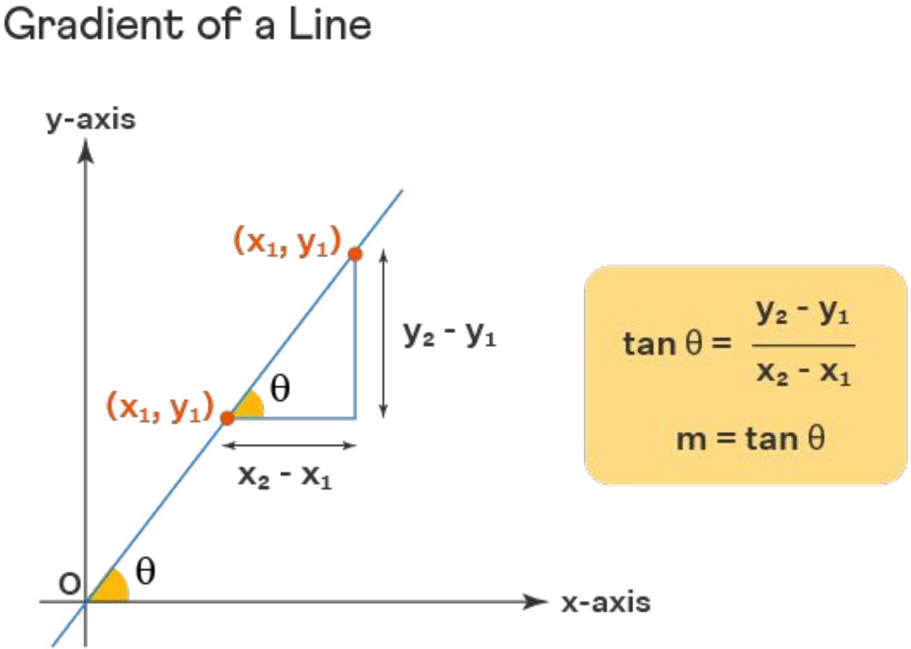
- Unchanged pre-trained model can be considered as a good feature-extractor

Data Generator

- In by-default situation, we need to load whole dataset into memory which cause OOM problem for a large dataset or a device with a smaller memory.
- We need to complete all preprocessing tasks and save preprocessed data in the hard disks.
    - It demands extra storage space specially when we need to figure out appropriate preprocessing steps.
- Data generator is used to:
    - generate customized batch.
    - Handle large dataset while avoiding OOM (Out of Memory) problem
    - Pre-process data on run-time
- Code help: https://stanford.edu/~shervine/blog/keras-how-to-generate-data-on-the-fly

Gradient

- Gradient (GD) means the change in the value of a quantity with change in a given variable
- the sign of a GD represents the direction of greatest change in a scalar function

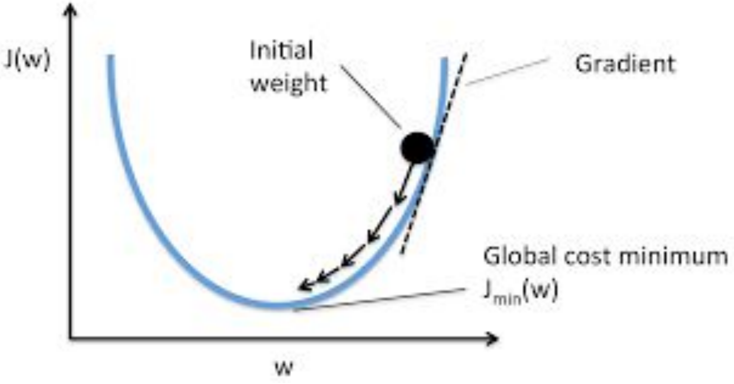


Gradient Descent Algorithm

$$\text{repeat until convergence } \{$$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J$$

$$\}$$

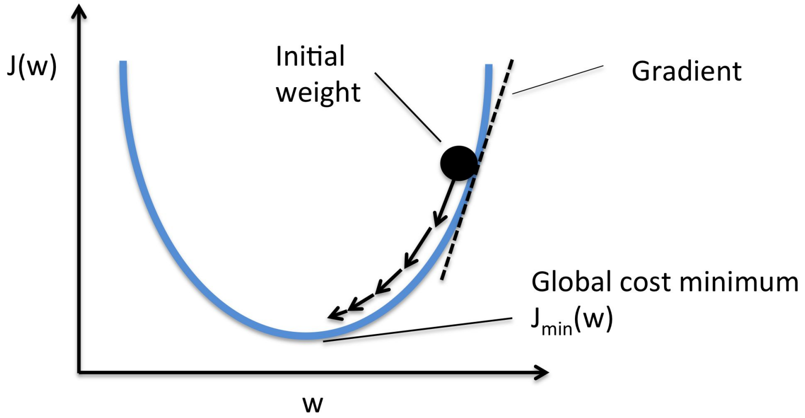


Gradient Descent Algorithm

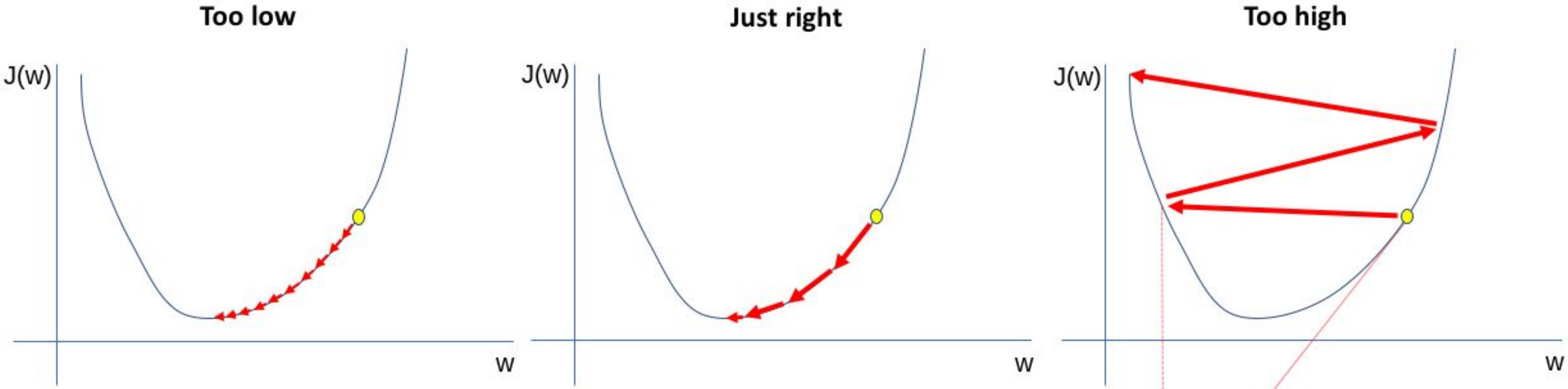
- Initialize parameters randomly
- Estimate outputs for inputs in a forward pass
- Estimate loss
- Estimate gradient of the loss function with respect to all parameters
- Move parameters in the opposite direction of gradient

$$w = w - \alpha \frac{d}{dw} J(w)$$

- Repeat these steps until loss is minimum



Effect of Learning Rate, α



A small learning rate requires many updates before reaching the minimum point

The optimal learning rate swiftly reaches the minimum point

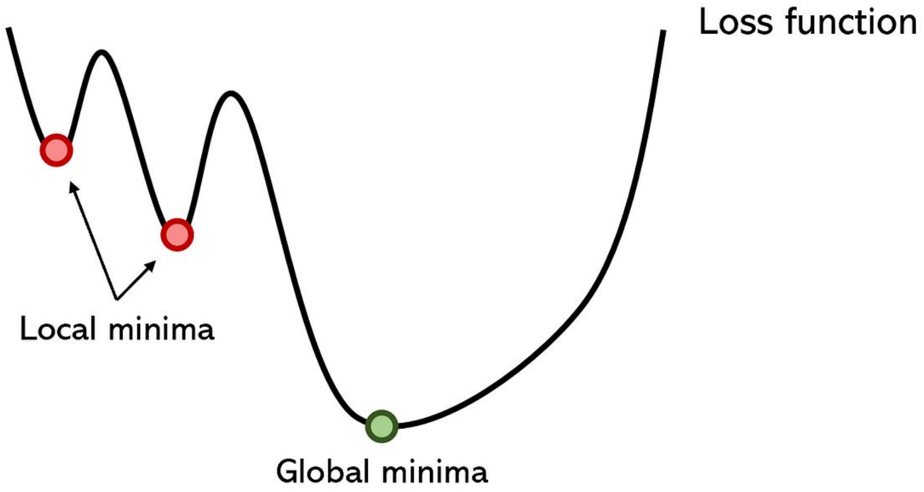
Too large of a learning rate causes drastic updates which lead to divergent behaviors

$$w = w - \alpha \frac{d}{dw} J(w)$$

Local Vs. Global Minima

- **Local Minimum**
  A point where a function's value for a variable parameter is smaller than nearby points, but possibly larger than at a distant point.

- **Global Minimum**
  A point where a function's value for a variable parameter is smaller than at all other feasible points

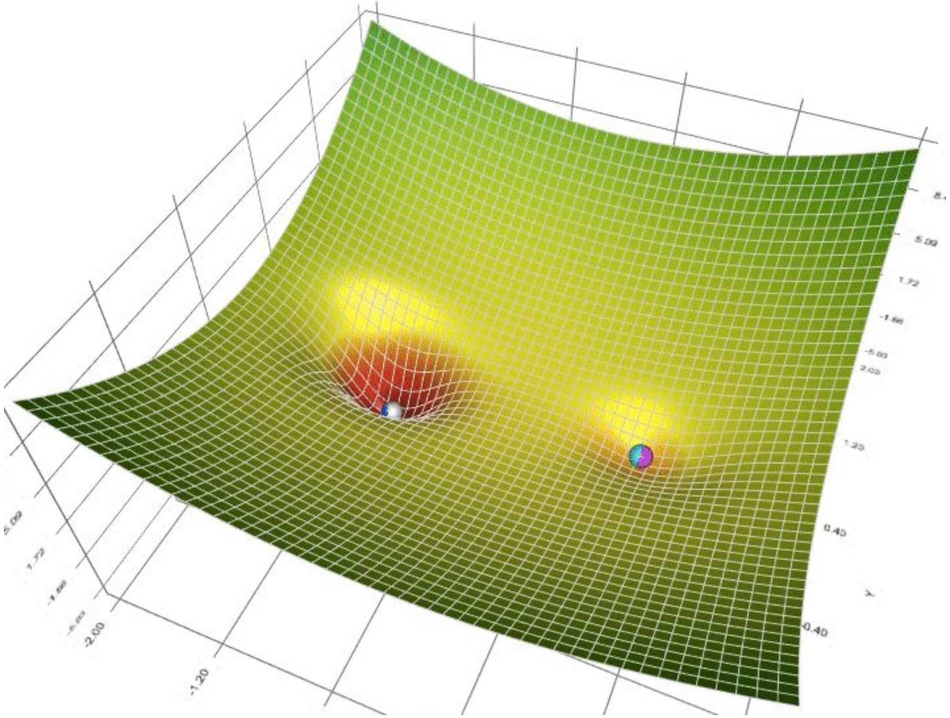


Our target is to reach the global minima, but unfortunately we end up at a local minima.

Gradient Descent Algorithm

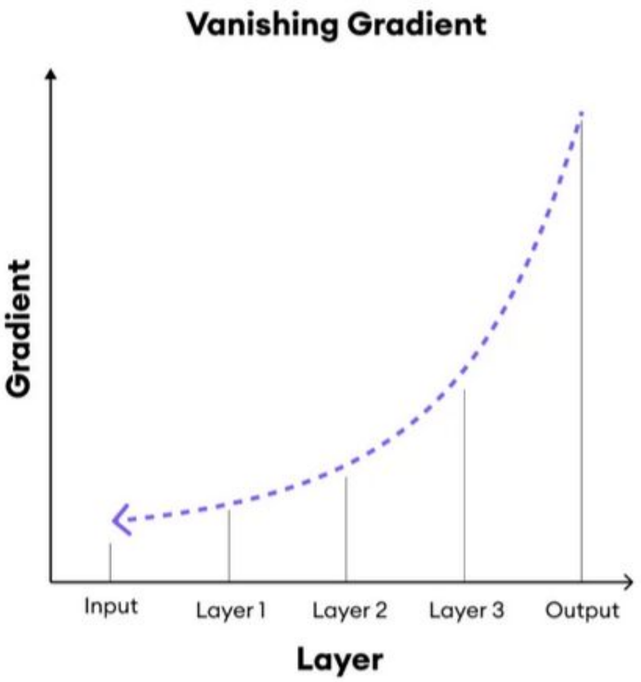
Different Optimization Algorithms:

- AdaGrad
- RMSProp
- Adam



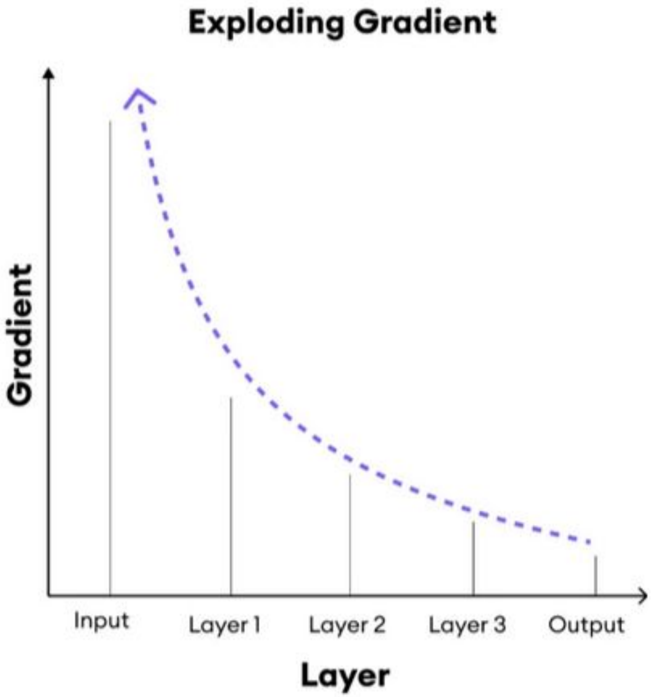
Vanishing Gradient

- gradients often get smaller and smaller and approach zero
- therefore, the weights of the initial or lower layers nearly unchanged
- as a result, the gradient descent never converges to the optimum
- model learns very slowly and perhaps the training stagnates at a very early stage just after a few iterations



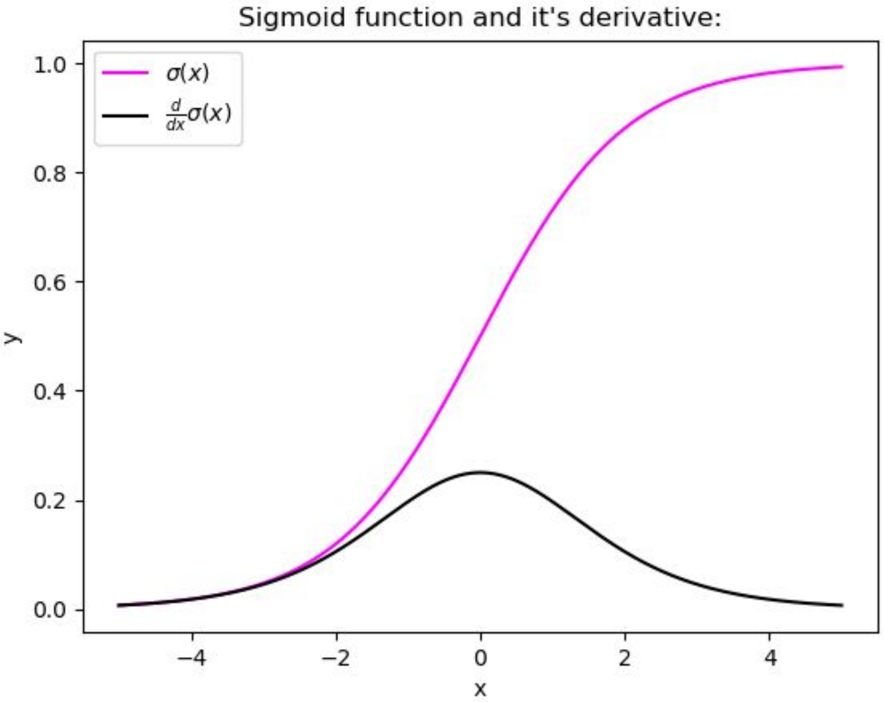
Exploding Gradient

- gradients keep on getting larger and larger
- it causes very large weight updates
- an exponential growth in the model parameters
- model weights may become NaN during training
- it causes the gradient descent to diverge
- model experiences avalanche learning



Ways to Handle GD Vanishing and Explosion

- Proper weight initialization
  - Glorot / He / He initialization
- Non-saturating Activation Functions
  - ReLU, ELU
- Batch normalization
  - Zero centering and normalizing inputs
- Gradient clipping to mitigate gradient explosion
  - clip all the partial derivatives of the loss with respect to each trainable parameter between –1.0 and 1.0

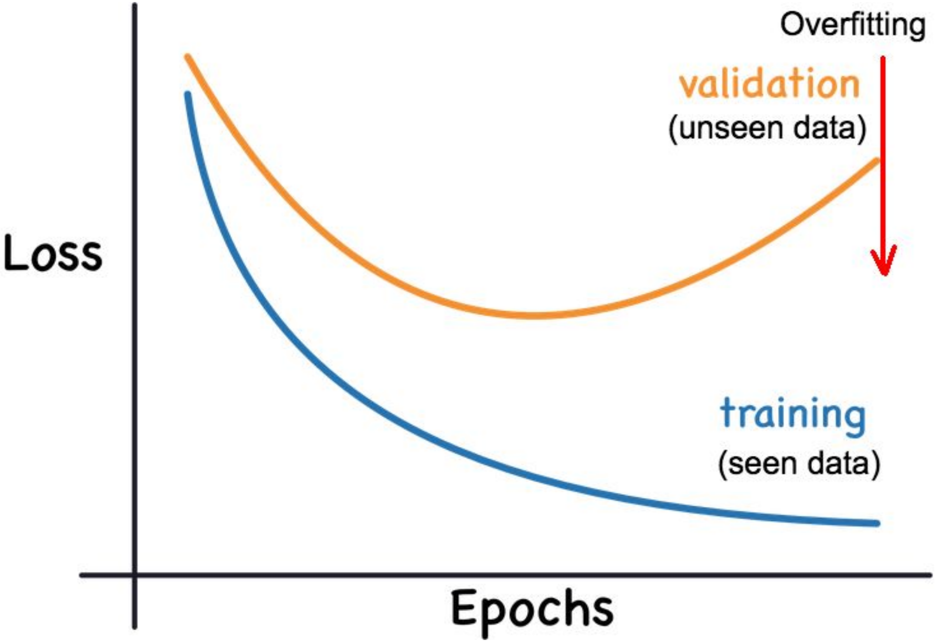


Overfitting

Overfitting occurs when:

- the model is so closely aligned to the training data that it does not know how to respond to new data.
- the model memorizes training data

Memorizing is not learning



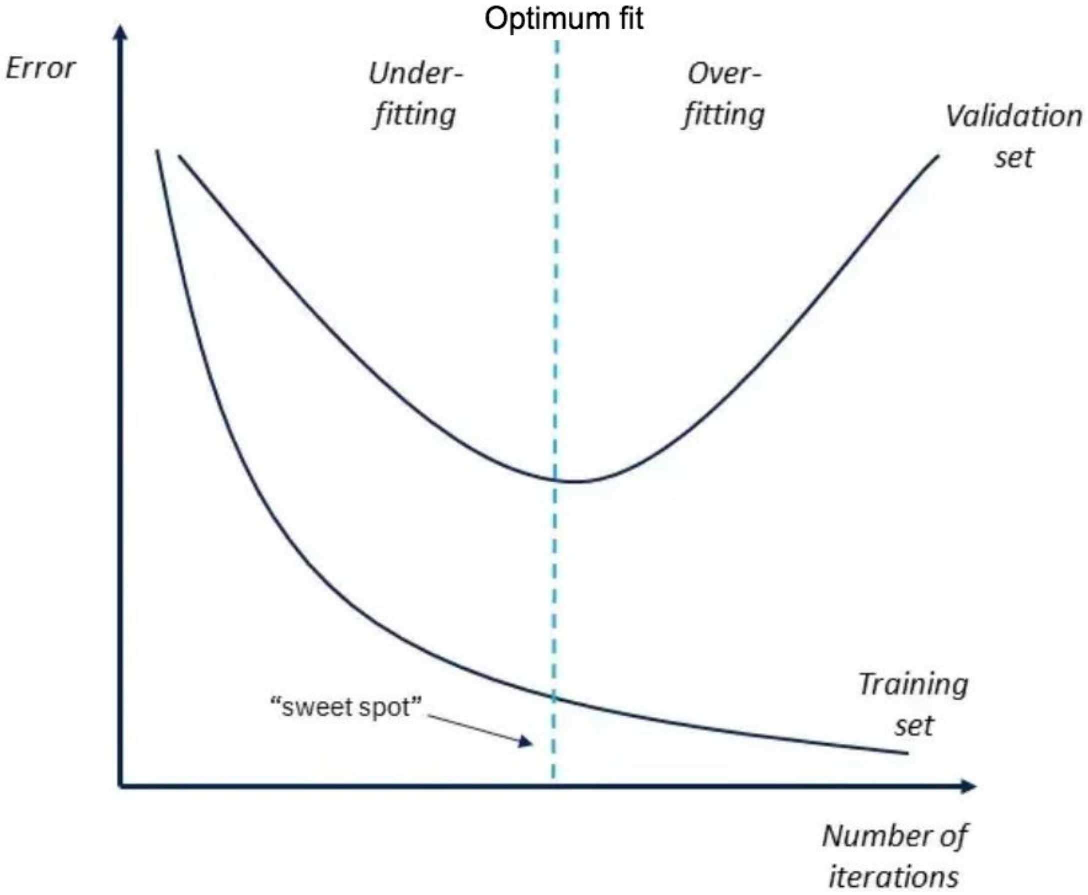
Reasons behind Overfitting

Overfitting can happen because:

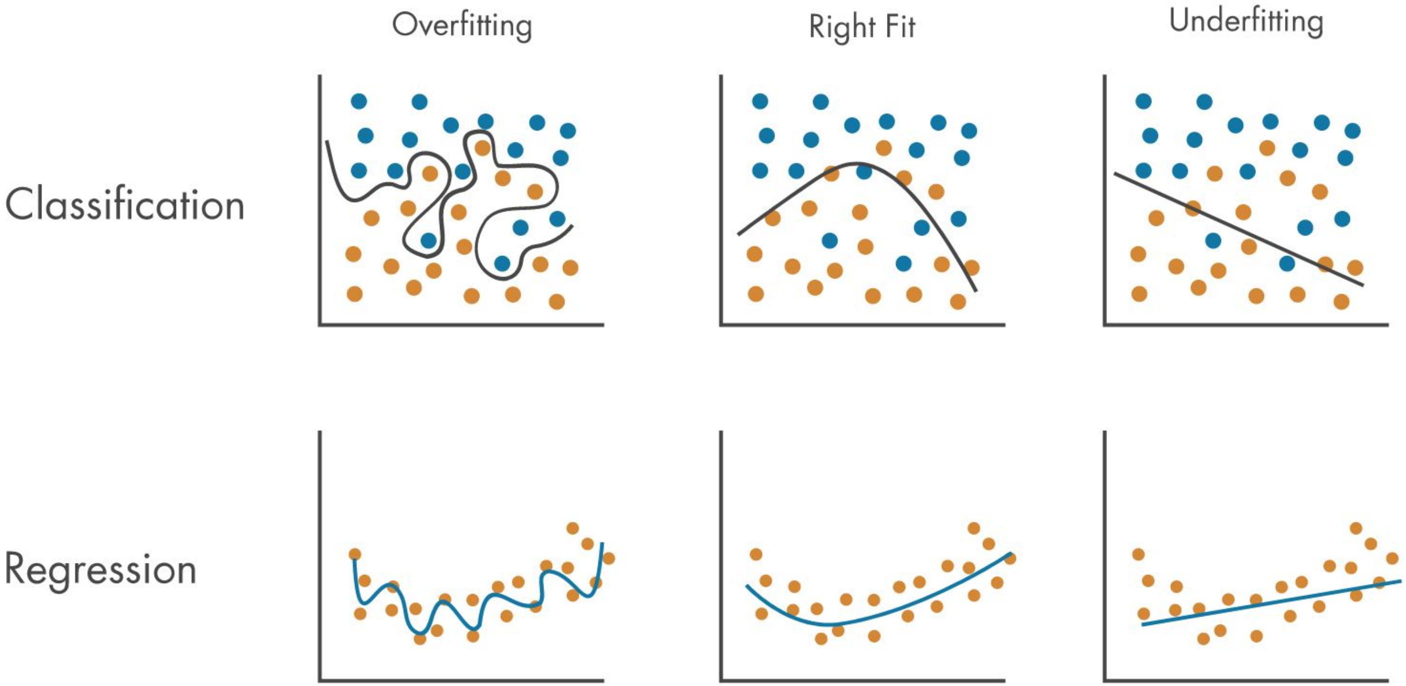
- Model is too complex; it memorizes very subtle patterns in the training data that don’t generalize well.
  - Model has too many parameters
- The training data size is too small for the model complexity and/or contains large amounts of irrelevant information.
- During training model sees a small dataset for too many times i.e., too many epochs were used during training.

Underfitting

- Underfitting is the opposite concept of overfitting
- When the prediction error on both the training and validation dataset is high, and the difference between them is very minimal, the model is said to have under fitted.

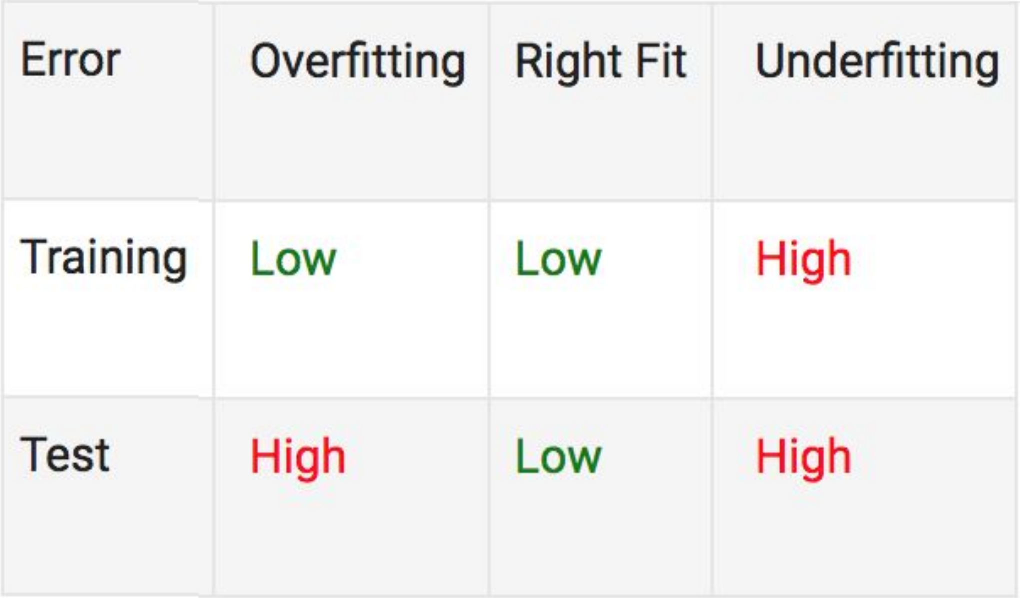


Overfitting - Underfitting in Classification and Regression



Overfitting vs Underfitting

When only looking at the computed error of a model for the training data, overfitting is harder to detect than underfitting.



Error	Overfitting	Right Fit	Underfitting
Training	Low	Low	High
Test	High	Low	High

More details: https://www.mathworks.com/discovery/overfitting.html