Remember ERM

$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} l_s(h_w(x_i), y_i) + \lambda r(\tilde{w})$$

How should we set λ (regulates the bias/variance)?

Overfitting and Underfitting

Underfitting: The classifier learned on the training set is **NOT** expressive enough to even account for the data provided. Both training **AND** testing error are likely to be high, as the classifier does not account for relevant information present in the training set.

Overfitting: The classifier learned on the training set is too specific, and cannot be used to accurately infer anything about the unseen data. Although training error is likely to decrease over time, the test error will begin to increase as the classifier begins to make predictions on patterns which exist only in the training set and **NOT** in the broader distribution.

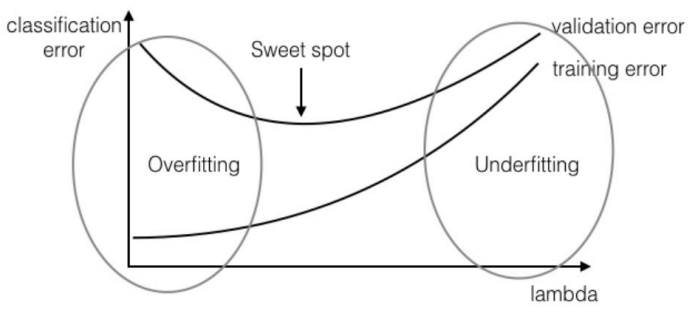

Figure 1: overfitting and underfitting

Identifying the Sweet Spot

① Divide data into training and validation sets. Train your data on the "training split" and evaluate it on the "validation split", for various values of λ

② Perform Telescopic search

(a) First find order of magnitude

(b) then, experiment in that order of magnitude

Early Stopping

Stop your optimization after $M > 0$ number of gradient steps, even if the optimization has not converged yet

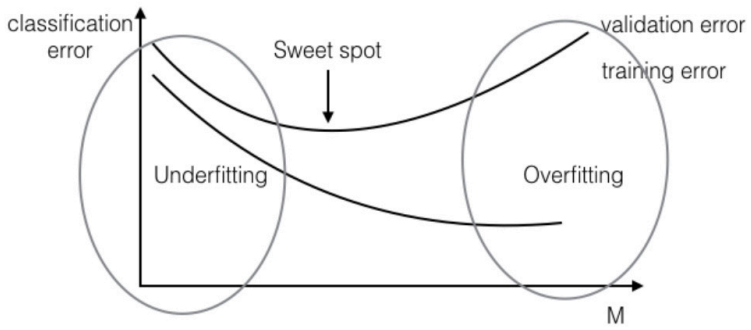

Figure 2: Early stopping