5.4 Estimators, Bias and Variance

Table of Contents

Point Estimation

- Attempt to provide the single best prediction of some quantity of interest
- Quantity of interest can be:
  - A single parameter
  - A vector parameters
    - Weights in linear regression
  - A whole function

Point Estimator/Statistics

- To distinguish estimates of parameters from their true value, represented as $\hat{\theta}$
- Let $\{x^{(1)}, x^{(2)},..x^{(m)}\}$ b m i.i.d. data points
  - Then a point estimator or statistic is any function of the data $\hat{\theta} = g(x^{(1)},\ldots x^{(m)})$
    - Thus a statistic is any function of the data
    - It need not be close to the true theta
- A good estimator is a function whose output is close to the true underlying theta that generated the data

Function Estimation

- Here we predict a variable y given input x
  - We assume f(x) is the relationship between x and y
    - We may assume $y = f(x) + \varepsilon$
  - We are interested in approximating f with a model f hat
    - Function estimation is same as estimating a parameter theta
      - Where f hat is a point estimator in function space

1. Bias of an Estimator

- The bias of an estimator is defined as: $\mathrm{bias}(\hat{\boldsymbol{\theta}}_m) = \mathbb{E}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\theta}$
- The estimator is unbiased if $\mathrm{bias}(\hat{\theta}) = 0$
  - Implies $\mathbb{E}(\hat{\boldsymbol{\theta}}_m) = \boldsymbol{\theta}$.
- The estimator is asymptotically unbiased if $\lim_{m\to\infty} \mathrm{bias}(\hat{\boldsymbol{\theta}}_m) = \mathbf{0}$
  - Implies $\lim_{m\to\infty} \mathbb{E}(\hat{\boldsymbol{\theta}}_m) = \boldsymbol{\theta}$

Examples of Estimator Bias

- We look at common estimators of the following parameters to determine whether there is bias:
  - Bernoulli distribution - Θ
  - Gaussian Distribution - μ
  - Gaussian Distribution - σ^2

σ^2

Estimator of Bernoulli Mean

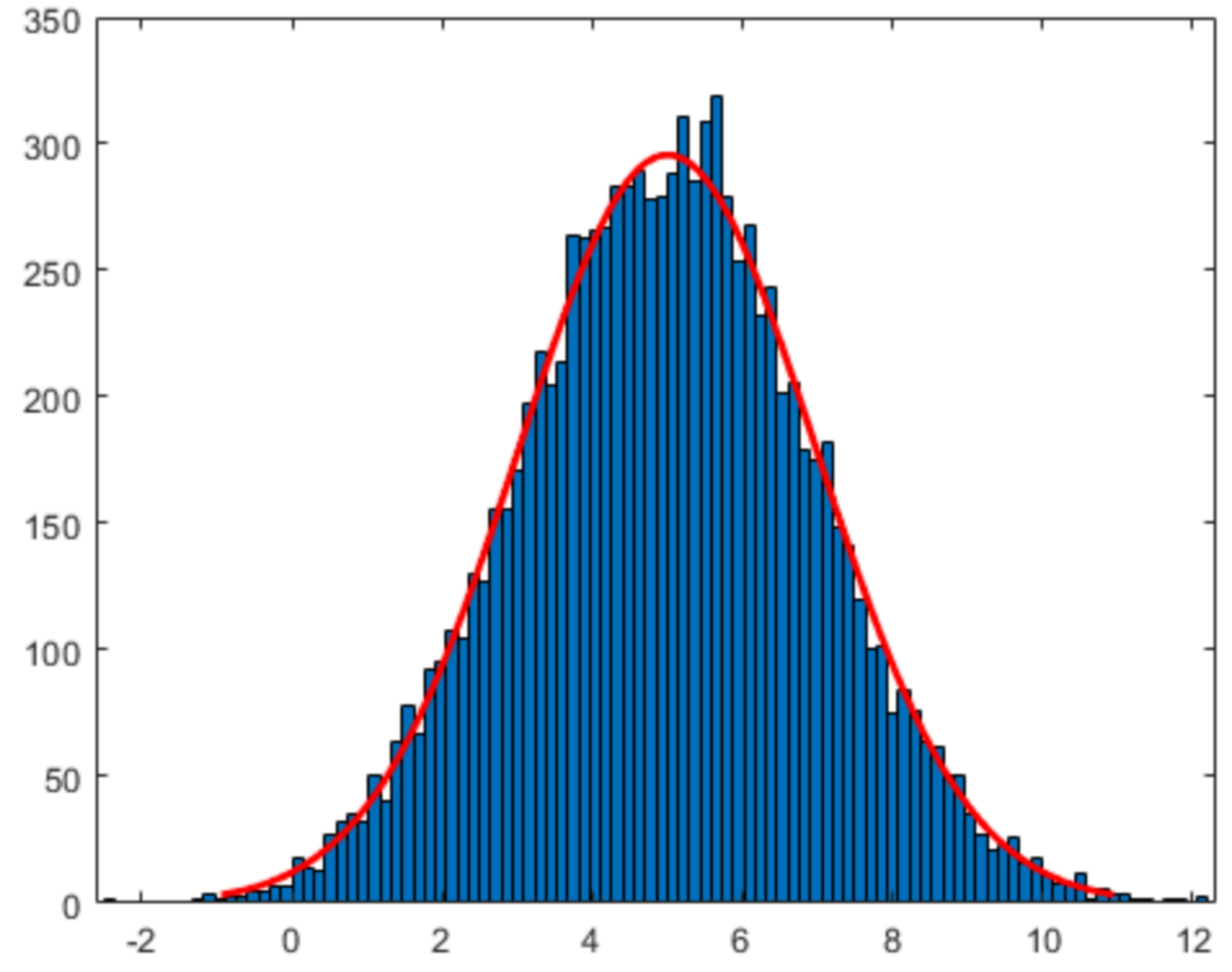
- Any events with 1 trial and 2 possible outcomes follows the Bernoulli distribution. ex. coin flip
- Bernoulli distribution for binary variable $x \in \{0,1\}$ with mean theta has the form $P(x;\theta)=\theta^x(1-\theta)^{1-x}$
- Estimator for theta given samples $\{x^{(1)},..,x^{(m)}\}$ is $\hat{\theta}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$
  - bias($\hat{\theta}_m$)=0 —> estimator is unbiased

$$
\begin{aligned}
\text{bias}(\hat{\theta}_m) &= \mathbb{E}[\hat{\theta}_m] - \theta \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] - \theta \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[x^{(i)}\right] - \theta \\
&= \frac{1}{m}\sum_{i=1}^{m} \sum_{x^{(i)}=0}^{1} \left(x^{(i)}\theta^{x^{(i)}}(1-\theta)^{(1-x^{(i)})}\right) - \theta \\
&= \frac{1}{m}\sum_{i=1}^{m} (\theta) - \theta \\
&= \theta - \theta = 0
\end{aligned}
$$

Estimator of Gaussian Mean

- Samples $\{x^{(1)},\ldots,x^{(m)}\}$ are i.i.d. according to $p(x^{(i)})=N(x^{(i)};\mu,\sigma^2)$
  - Sample mean is an estimator of the mean parameter $\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$
  - To determine bias of the sample mean:

$$\begin{aligned}
\text{bias}(\hat{\mu}_m) &= \mathbb{E}[\hat{\mu}_m] - \mu \\
&= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] - \mu \\
&= \left(\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[x^{(i)}\right]\right) - \mu \\
&= \left(\frac{1}{m}\sum_{i=1}^{m} \mu\right) - \mu \\
&= \mu - \mu = 0
\end{aligned}$$



  - Thus the sample mean is an unbiased estimator of the Gaussian mean

Estimator for Gaussian Variance

- Sample Variance: $\hat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2$
- Interested in computing $\text{bias}(\hat{\sigma}_m^2) = \mathbb{E}[\hat{\sigma}_m^2] - \sigma^2$
- We begin by evaluating —> $\mathbb{E}[\hat{\sigma}_m^2] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2\right] = \frac{m-1}{m}\sigma^2$
- Thus the bias of $\hat{\sigma}_m^2$ is $-\sigma^2/m$
- Thus the sample variance is a biased estimator
- The unbiased sample variance estimator: $\tilde{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2$

Variance and Standard Error

- Another property of an estimator:
  - How much we expect the estimator to vary as a function of the data sample
- Just as we computed the exception of the estimator to determine its bias, we can compute its variance
- The variance of an estimator is simply $\mathrm{Var}(\hat{\theta})$ where the random variable is the training set
- The square root of the variance is called the standard error, denoted $\mathrm{SE}(\hat{\theta})$
-

Importance of Standard Error

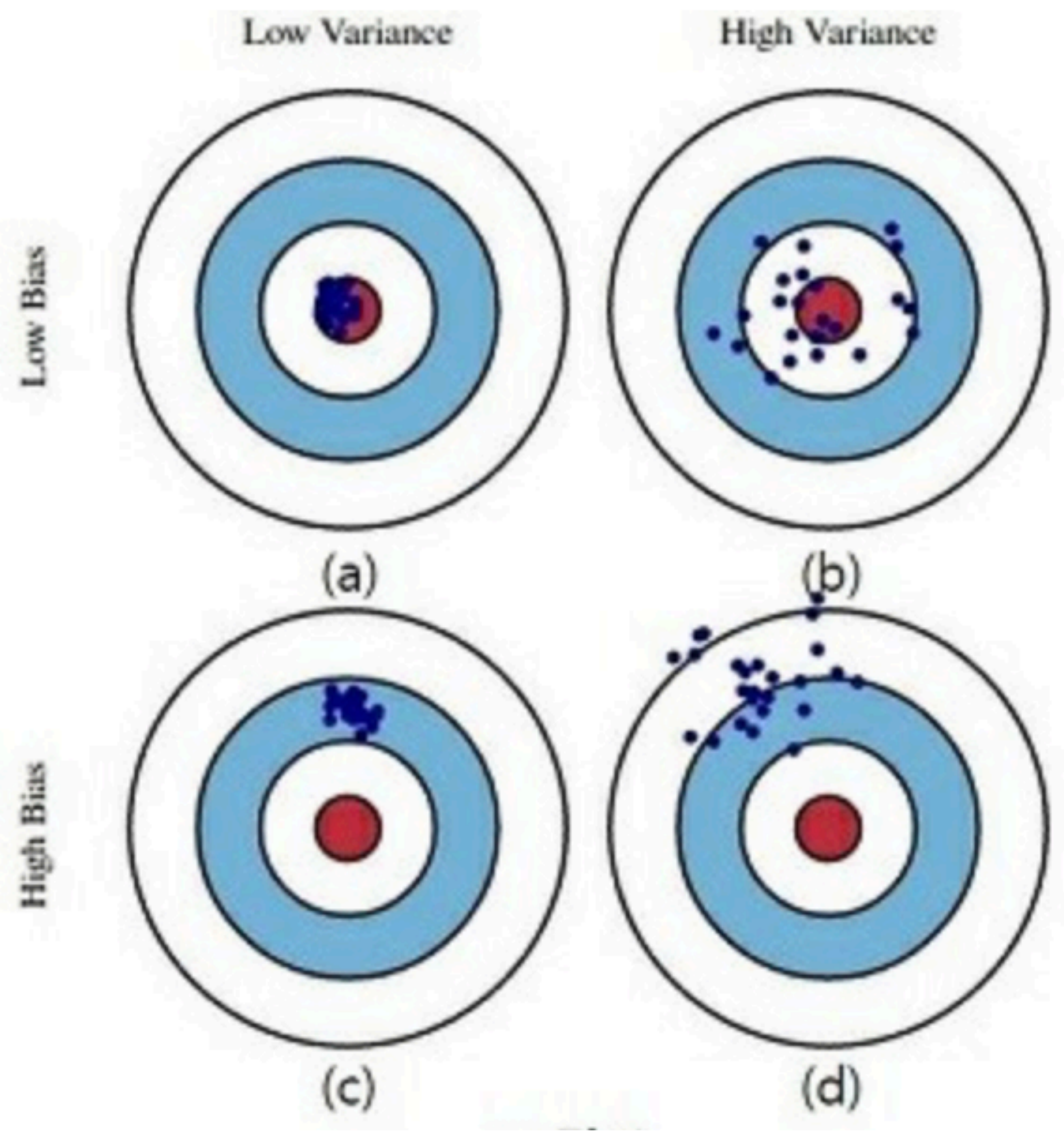
- It measures how we would expect the estimate to vary we obtain different samples from the same distribution
- The standard error of the mean is given by $$\mathrm{SE}(\hat{\mu}_m) = \sqrt{\mathrm{Var}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right]} = \frac{\sigma}{\sqrt{m}},$$
  - Where $–\sigma^2$ is the true variance of the samples x^(i)
  - Standard error often estimated using estimate of σ
    - Although not unbiased, approximate is reasonable

Standard Error in Machine Learning

- Often estimate generalization error - computing error on the test set
  - # of samples - accuracy
  - Mean normally distributed (central limit theorem), can compute probability that true expectation falls in any chosen interval
  - 95% condifence interval centered on mean is $(\hat{\mu}_m - 1.96\mathrm{SE}(\hat{\mu}_m), \hat{\mu}_m + 1.96\mathrm{SE}(\hat{\mu}_m))$
  - ML alg A > ML alg B if upper bound of A < lower bound of B

Trade-off Bias vs. Variance

- Bias measures: expected deviation from the true value of the function or parameter
- Variance measures: expected deviation that any particular sampling of the data is likely to cause

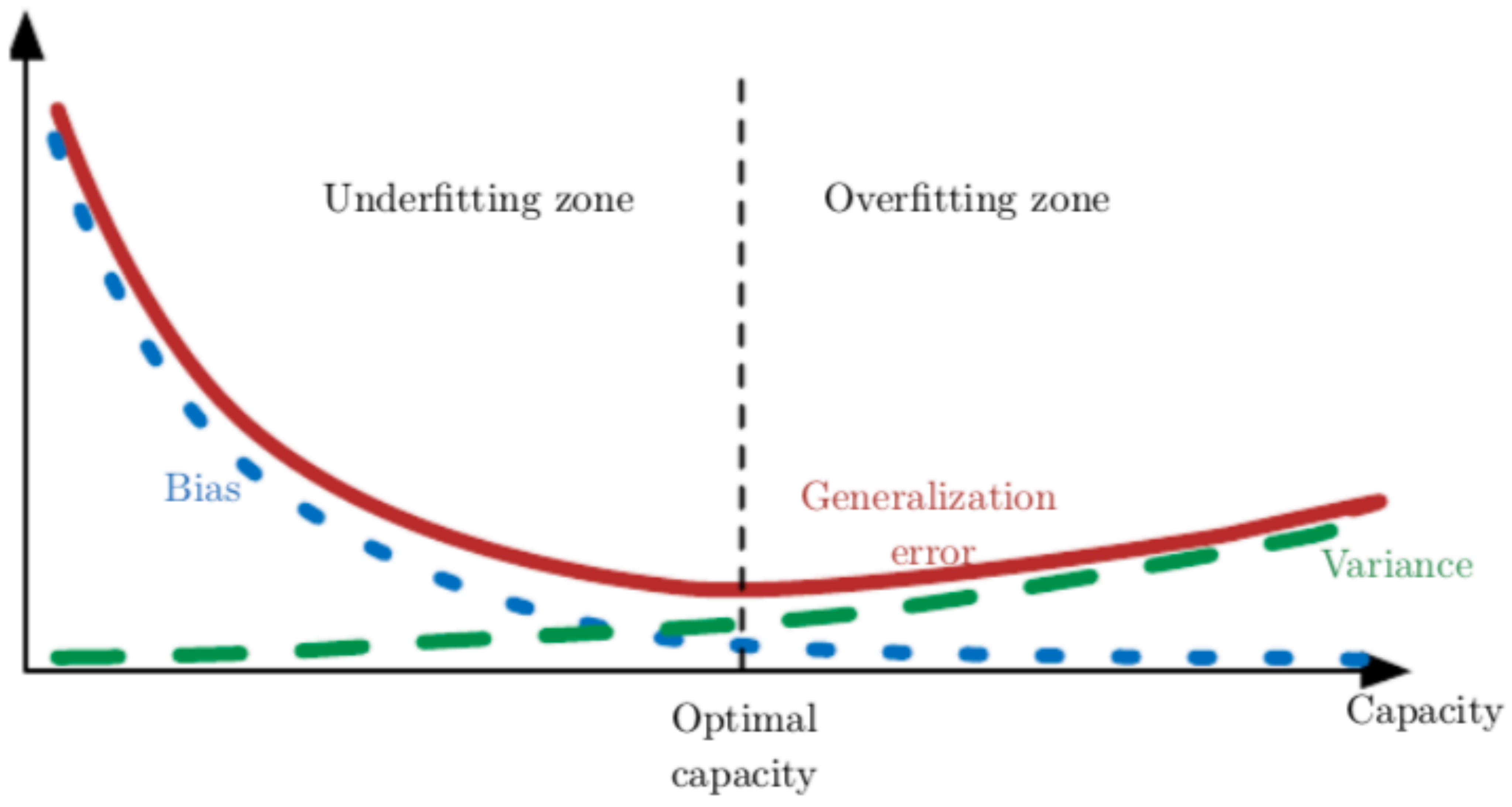


Mean Squared Error

- Mean Squared Error of an estimate is

$$\begin{aligned} \mathrm{MSE} &= \mathbb{E}[(\hat{\theta}_m - \theta)^2] \\ &= \mathrm{Bias}(\hat{\theta}_m)^2 + \mathrm{Var}(\hat{\theta}_m) \end{aligned}$$

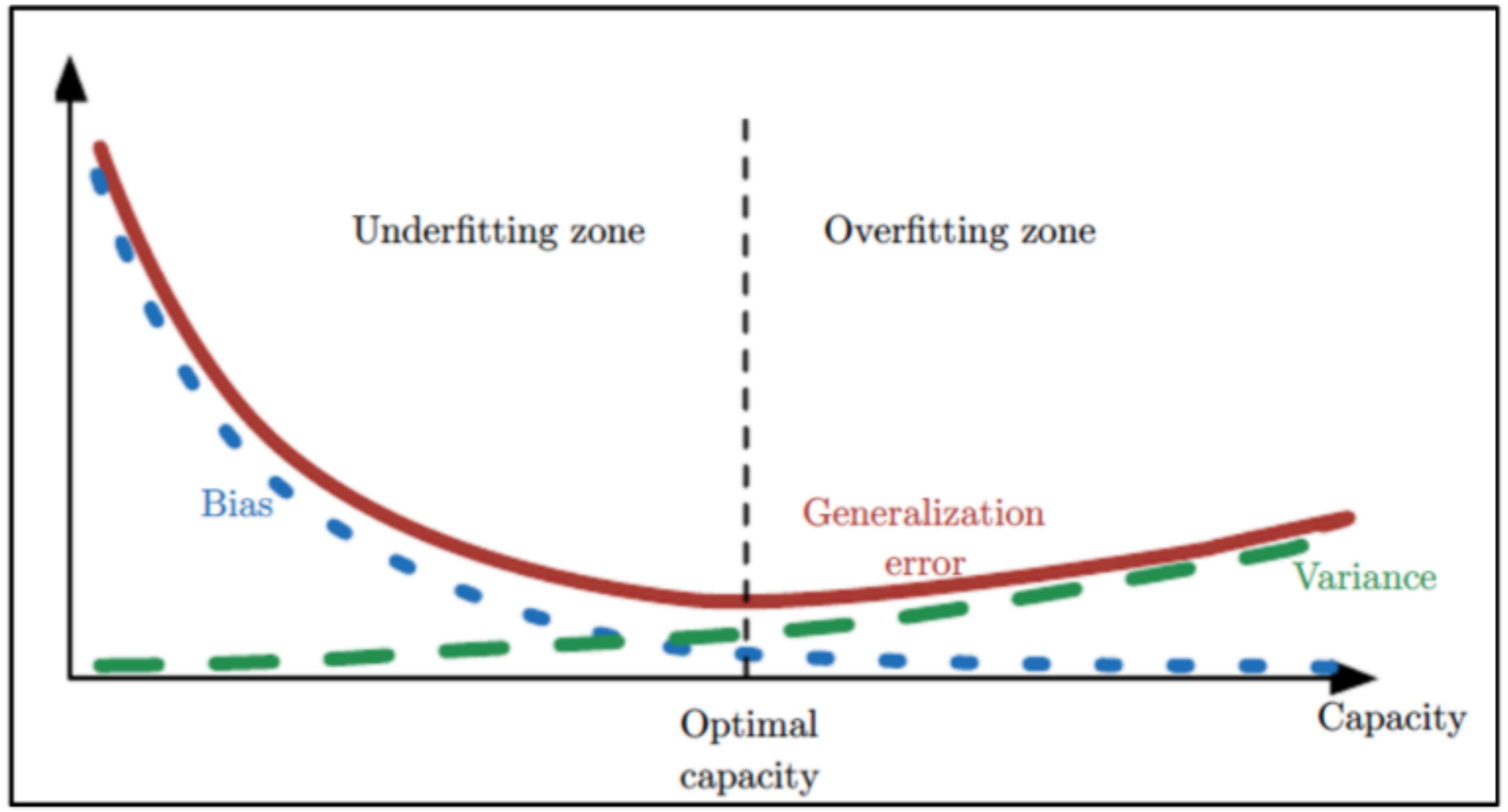
- Minimizing the MSE keeps both bias and variance in check



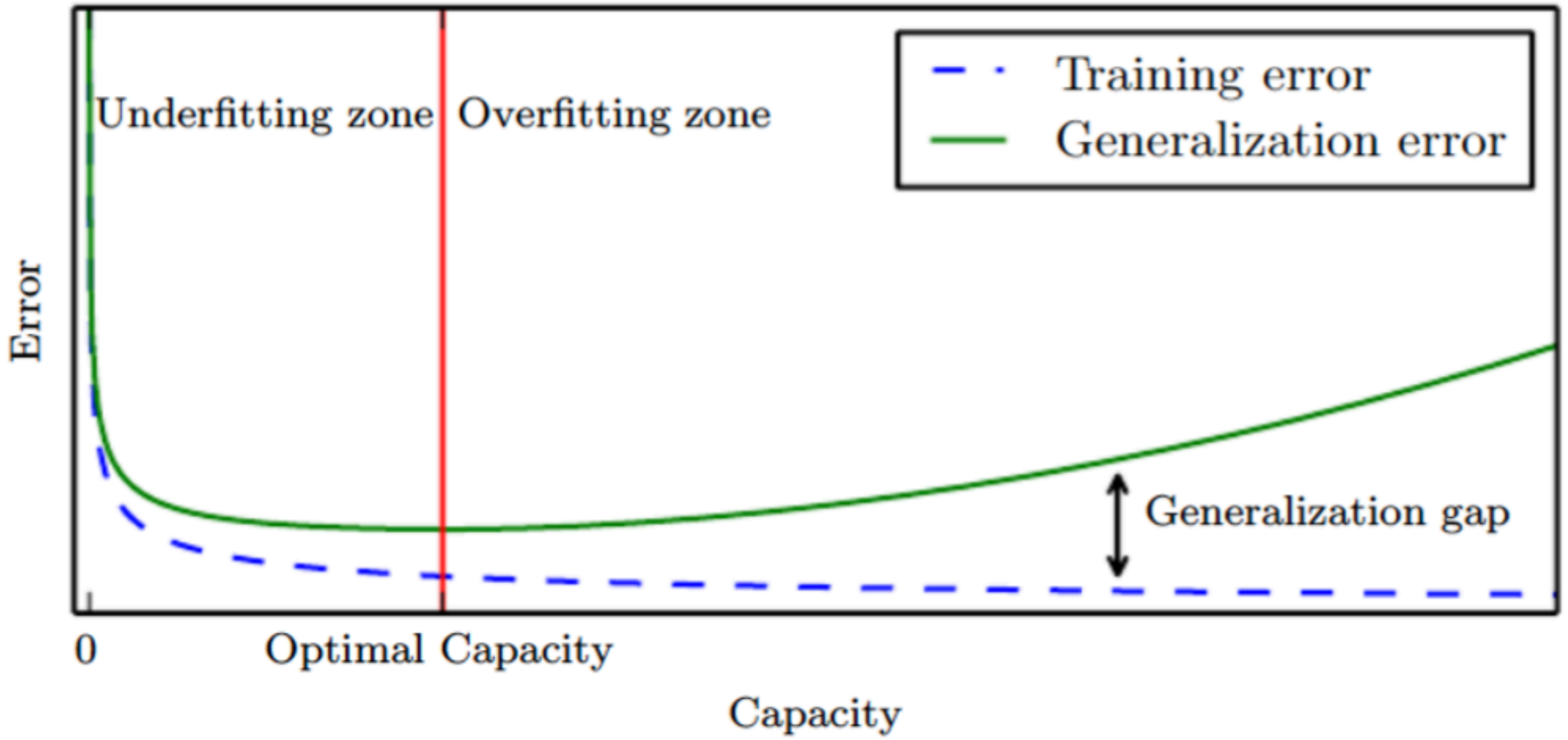
As capacity increases, bias tends to decrease, variance tends to increase

Underfit/Overfit - Bias-Variance

- Relationship of bias-variance to capacity is similar to undercutting and overfitting relationship to capacity



Bias-Variance to capacity



Model complexity to capacity

Consistency

- Behavior of the estimator as training set grows
- # of data points m in the training set grows, converge to the true value of the parameters: $\text{plim}_{m \to \infty} \hat{\theta}_m = \theta$
- Symbol plim: convergence in probability

Weak & Strong Consistency

- $\text{plim}_{m\to\infty}\hat{\theta}_m = \theta$ Means that
  - For any $\varepsilon > 0$, $P(|\hat{\theta} - \theta| > \varepsilon) \to 0$ as $m \to \infty$
  - Weak consistency
  - Convergence of $\hat{\theta}$ to θ
- Strong consistency: almost sure convergence of a sequence of random variables to a value x occurs when $p(\lim_{m\to\infty} \boldsymbol{x}^{(m)} = \boldsymbol{x}) = 1$
- Consistency ensures that the bias induced by the estimator decreases with m