Bias-Variance Tradeoff

CMPUT 296: Basics of Machine Learning

Textbook §9.2-9.3

Logistics

- Midterm and assignment 2 marking are in progress
- Assignment #3 will be available today

Recap: Regularization

- **Regularization:** minimize the training cost plus a complexity penalty
  - $c(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \text{cost}(f(\mathbf{x}_i; \mathbf{w}), y_i) + \lambda\, \text{penalty}(\mathbf{w})$
  - Only make a model more complex if it improves loss "enough"
  - The **hyperparameter** λ controls our notion of "enough"
- **L2 Regularization:** penalty is sum of squared weights: $\text{penalty}(\mathbf{w}) = \sum_{j=1}^{d} w_j^2$
  - L2 regularized linear regression corresponds to **MAP inference** with independent zero-mean **Gaussian priors** on each weight (except w_0)
- **L1 Regularization:** Penalty is sum of absolute values: $\text{penalty}(\mathbf{w}) = \sum_{j=1}^{d} |w_j|$
  - Corresponds to MAP inference with independent **Laplacian prior** on weights
  - Produces **sparse** solutions (many entries of $\mathbf{w}$ are set to **exactly 0**)

Outline

1. Recap & Logistics
2. Bias and Variance in Linear Regression / Parameter Estimation
3. Bias and Variance in General / Function Outputs

Bias, Variance, and Error

Suppose we are estimating a quantity μ using an **estimator** $\hat{X}$.

$$\text{Bias}(\hat{X}) = \mathbb{E}[\hat{X} - \mu]$$

$$\text{Var}(\hat{X}) = \mathbb{E}\left[(\hat{X} - \mathbb{E}[\hat{X}])^2\right]$$

Recall that an estimator's mean squared error **decomposes** into bias and variance:

$$MSE(\hat{X}) = \mathbb{E}\left[(\hat{X} - \mu)^2\right] = \text{Bias}^2(\hat{X}) + \text{Var}(\hat{X})$$

MLE for Linear Regression

Recall the **stochastic model** for linear regression with Gaussian errors:

$$Y = \omega^T \mathbf{X} + \epsilon \qquad \text{where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Now recall the **MLE formulation** of the linear regression problem:

$$\mathbf{w}_{\text{MLE}} = \arg \min_{\mathbf{w} \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

Question: What quantity is being estimated?

$$MSE(\hat{X}) = \mathbb{E}\left[(\hat{X} - \mu)^2\right] = \text{Bias}^2(\hat{X}) + \text{Var}(\hat{X})$$

$$MSE(\mathbf{w}_{\text{MLE}}) = \mathbb{E}\left[(\mathbf{w}_{\text{MLE}} - \omega)^2\right] = \text{Bias}^2(\mathbf{w}_{\text{MLE}}) + \text{Var}(\mathbf{w}_{\text{MLE}})$$

MLE for Linear Regression: Bias

What is the bias of the MLE estimator? Let's consider the 1D case:

Recall:

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T\right)\mathbf{w}_{\text{MLE}}(\mathscr{D}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i Y_i$$

where $\mathscr{D}$ is random

$$\implies w_{\text{MLE}}(\mathscr{D}) = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

for one-dimensional x

Bias of w_{MLE}

$$\mathbb{E}\left[w_{\text{MLE}}(\mathcal{D})\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right]$$

$$= \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i(\omega X_i + \epsilon_i)}{\sum_{i=1}^{n} X_i^2}\right]$$

$$= \mathbb{E}\left[\frac{\sum_{i=1}^{n} \omega X_i^2 + X_i \epsilon_i}{\sum_{i=1}^{n} X_i^2}\right]$$

$$= \mathbb{E}\left[\frac{\omega \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} X_i \epsilon_i}{\sum_{i=1}^{n} X_i^2}\right]$$

$$= \mathbb{E}\left[\frac{\omega \sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i^2}\right] + \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i \epsilon_i}{\sum_{i=1}^{n} X_i^2}\right]$$

$$= \mathbb{E}\left[\omega\right] + \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i \epsilon_i}{\sum_{i=1}^{n} X_i^2}\right]$$

$$= \mathbb{E}\left[\omega\right] + \sum_{i=1}^{n} \mathbb{E}\left[\frac{X_i \epsilon_i}{\sum_{i=1}^{n} X_i^2}\right]$$

$$= \mathbb{E}\left[\omega\right] + \sum_{i=1}^{n} \mathbb{E}\left[\epsilon_i \frac{X_i}{\sum_{i=1}^{n} X_i^2}\right]$$

$$= \mathbb{E}\left[\omega\right] + \sum_{i=1}^{n} \mathbb{E}[\epsilon_i]\, \mathbb{E}\left[\frac{X_i}{\sum_{i=1}^{n} X_i^2}\right]$$

$$= \mathbb{E}[\omega] \quad \blacksquare$$

$\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

MLE for Linear Regression: Variance

$$\mathrm{Var}\left[w_{\mathsf{MLE}}(\mathscr{D})\right] = \mathbb{E}\left[\left(w_{\mathsf{MLE}}(\mathscr{D}) - \omega\right)^2\right]$$

$$= \mathbb{E}\left[\left(w_{\mathsf{MLE}}(\mathscr{D})\right)^2 - 2\omega w_{\mathsf{MLE}}(\mathscr{D}) - \omega^2\right]$$

$$= \mathbb{E}\left[\left(w_{\mathsf{MLE}}(\mathscr{D})\right)^2\right] - \mathbb{E}\left[2\omega w_{\mathsf{MLE}}(\mathscr{D})\right] + \mathbb{E}\left[\omega^2\right]$$

$$= \mathbb{E}\left[\left(w_{\mathsf{MLE}}(\mathscr{D})\right)^2\right] - 2\omega\mathbb{E}\left[w_{\mathsf{MLE}}(\mathscr{D})\right] + \omega^2$$

$$= \mathbb{E}\left[\left(w_{\mathsf{MLE}}(\mathscr{D})\right)^2\right] - 2\omega\omega + \omega^2$$

$$\boxed{= \mathbb{E}\left[\left(w_{\mathsf{MLE}}(\mathscr{D})\right)^2\right] - \omega^2}$$

MLE for Linear Regression: Variance (2)

$$\mathbb{E}[w_{\text{MLE}}] = \mathbb{E}[\omega] + \sum_{i=1}^{n} \mathbb{E}\left[\epsilon_i \frac{X_i}{\sum_{i=1}^{n} X_i^2}\right]$$

$$\iff \quad w_{\text{MLE}} = \omega + \sum_{i=1}^{n} \epsilon_i \frac{X_i}{\sum_{i=1}^{n} X_i^2}$$

$$\begin{aligned}
\text{Var}\left[w_{\text{MLE}}(\mathcal{D})\right] &= \mathbb{E}\left[\left(w_{\text{MLE}}(\mathcal{D})\right)^2\right] - \omega^2 \\
&= \mathbb{E}\left[\left(\omega + \sum_{i=1}^{n} \epsilon_i \frac{X_i}{\sum_{i=1}^{n} X_i^2}\right)^2\right] - \omega^2 \\
&= \omega^2 + 2\omega\mathbb{E}\left[\sum_{i=1}^{n} \epsilon_i \frac{X_i}{\sum_{i=1}^{n} X_i^2}\right] + \mathbb{E}\left[\left(\sum_{i=1}^{n} \epsilon_i \frac{X_i}{\sum_{i=1}^{n} X_i^2}\right)^2\right] - \omega^2 \\
&= 2\omega \sum_{i=1}^{n} \underbrace{\mathbb{E}\left[\epsilon_i \frac{X_i}{\sum_{i=1}^{n} X_i^2}\right]}_{=0} + \mathbb{E}\left[\left(\sum_{i=1}^{n} \epsilon_i \frac{X_i}{\sum_{i=1}^{n} X_i^2}\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} \epsilon_i \frac{X_i}{\sum_{i=1}^{n} X_i^2}\right)^2\right] = \sigma^2\mathbb{E}\left[\frac{1}{\sum_{i=1}^{n} X_i^2}\right]
\end{aligned}$$

(exercise)

MLE for Linear Regression: Bias vs. Variance

$$MSE(\mathbf{w}_{\text{MLE}}) = \mathbb{E}\left[(\mathbf{w}_{\text{MLE}} - \omega)^2\right] = \text{Bias}^2(\mathbf{w}_{\text{MLE}}) + \text{Var}(\mathbf{w}_{\text{MLE}})$$

$$\text{Bias}(w_{\text{MLE}}) = 0$$

$$\text{Var}(w_{\text{MLE}}) = \sigma^2 \mathbb{E}\left[\frac{1}{\sum_{i=1}^{n} X_i^2}\right]$$

- w_{MLE} is **unbiased**
- But the **variance** can be **very large**
  - Especially when n is small

MAP for Linear Regression

Recall that the MAP formulation of the linear regression problem with a Gaussian prior on the weights is equivalent to L2-regularized linear regression:

$$\mathbf{w}_{\text{MAP}} = \arg \min_{\mathbf{w} \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \sum_{i=1}^{n} w_i^2$$

Again restricting to the 1D case, we can solve the MAP regression problem analytically:

$$w_{\text{MAP}}(\mathscr{D}) = \frac{\sum_{i=1}^{n} X_i Y_i}{\lambda + \sum_{i=1}^{n} X_i^2}$$

MAP for Linear Regression: Bias

$$\mathbb{E}\left[w_{\text{MAP}}(\mathscr{D})\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i Y_i}{\lambda + \sum_{i=1}^{n} X_i^2}\right]$$

$$= \mathbb{E}\left[\frac{\omega \sum_{i=1}^{n} X_i^2}{\lambda + \sum_{i=1}^{n} X_i^2}\right] + \underbrace{\mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i \epsilon_i}{\lambda + \sum_{i=1}^{n} X_i^2}\right]}_{= 0}$$

$$= \omega \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i^2}{\lambda + \sum_{i=1}^{n} X_i^2}\right]$$

$$\neq \omega$$

MAP for Linear Regression: Bias vs. Variance

$$\mathbb{E}\left[w_{\text{MAP}}(\mathscr{D})\right] = \omega\mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i^2}{\lambda + \sum_{i=1}^{n} X_i^2}\right] \neq \omega$$

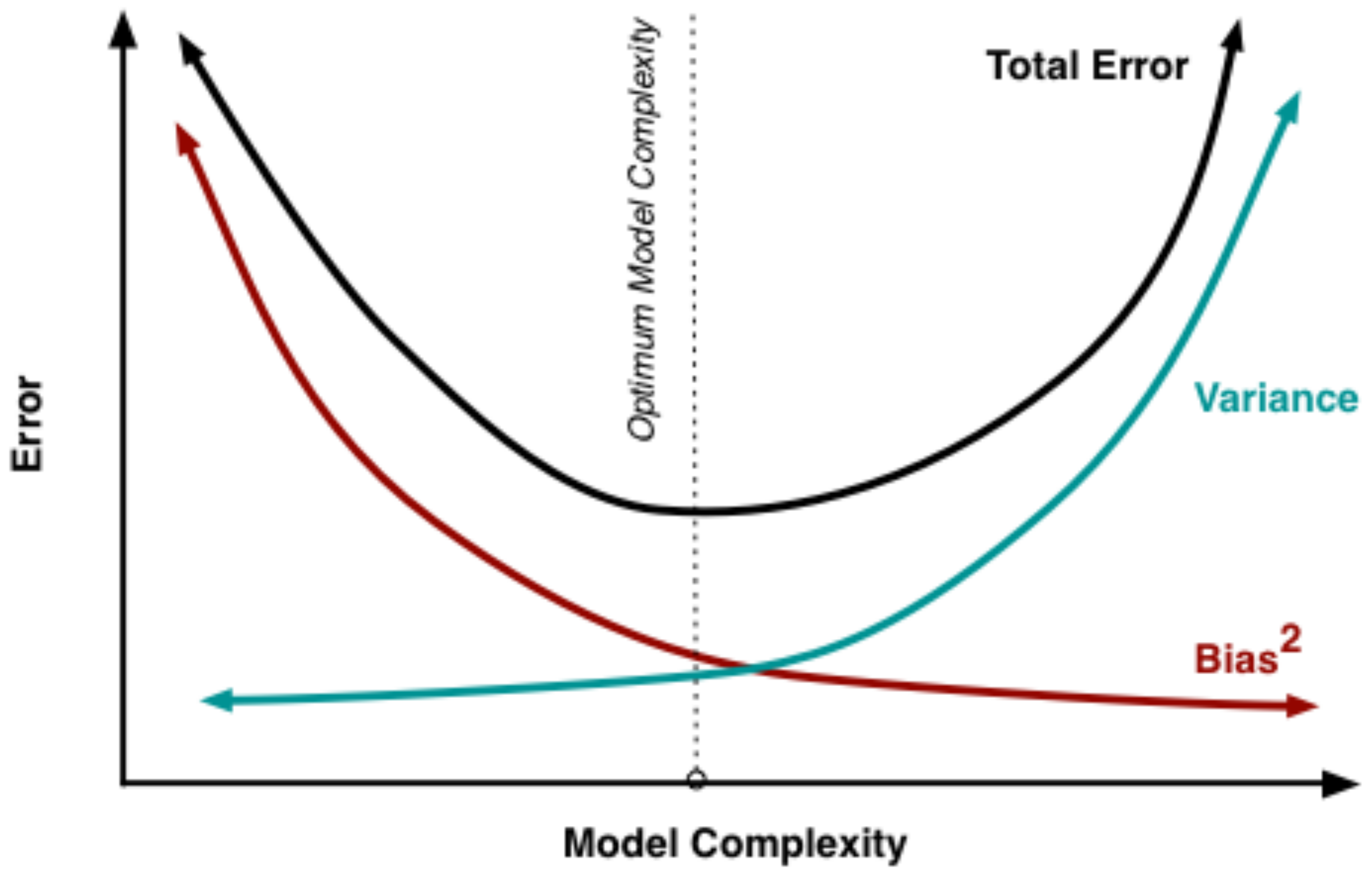
$$\text{Var}\left(w_{\text{MAP}}\right) = \sigma^2\mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i^2}{\left(\lambda + \sum_{i=1}^{n} X_i^2\right)^2}\right]$$

$$\mathbb{E}\left[w_{\text{MLE}}\right] = \omega$$

$$\text{Var}(w_{\text{MLE}}) = \sigma^2\mathbb{E}\left[\frac{1}{\sum_{i=1}^{n} X_i^2}\right]$$

- w_{MAP} is biased downwards (**why?**)
- But $\text{Var}\left(w_{\text{MAP}}\right) < \frac{1}{\lambda}$ even when $\sum_{i=1}^{n} X_i^2$ is very small (**why?**)

Choosing λ



- There exists an **optimal** λ for which total generalization error is **minimized**
- **Question:** Can we find that λ by directly optimizing generalization error?

(Image: http://scott.fortmann-roe.com/docs/BiasVariance.html)

Hypothesis Class Might Not Contain the "Real" Function

- The preceding treatment of linear regression assumes that there is a **true parameter** ω
- Suppose that instead the true model is **quadratic**:

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \epsilon$$

  but we are nevertheless performing **linear regression**

- **Question:** How can we apply the bias/variance argument?

Outputs vs Parameters

- We can perform a very similar analysis by comparing predictor **outputs** instead of predictor **parameters**

- Recall that error for a predictor $f(X)$ decomposes into reducible and irreducible error:

$$MSE(f(X)) = \mathbb{E}\left[\left(f(X) - f^*(X)\right)^2\right] + \mathbb{E}\left[\left(f^*(X) - Y\right)^2\right]$$

Reducible error   Irreducible error

- We can treat the predictor itself as a random variable $f_{\mathcal{D}}$ and reason about the expected value of the reducible error

Bias vs. Variance for Outputs

$$\mathbb{E}\left[\left(f_{\mathscr{D}}(X) - f^*(X)\right)^2\right] = \left(\mathbb{E}\left[f_{\mathscr{D}}(X)\right] - f^*(X)\right)^2 + \mathrm{Var}\left[f_{\mathscr{D}}(X)\right]$$

- We can decompose the **reducible error** into bias and variance of the **outputs**
  - (Very similar to our derivation for parameters)
- Note that $f^*(X)$ is the **optimal predictor**; it **need not** be part of our **hypothesis class**
- $f_{\mathscr{D}}(X)$ is the **predictor** that will be chosen from our **hypothesis class** based on the dataset $\mathscr{D}$ (so when we treat $\mathscr{D}$ as a random variable, $f_{\mathscr{D}}$ is also random)
- Regularization changes how we choose $f_{\mathscr{D}}$ from a given hypothesis class
- Choosing a different hypothesis class **can change** both the bias and variance of $f_{\mathscr{D}}$

Hypothesis Class Selection

$$\mathbb{E}\left[\left(f_{\mathscr{D}}(X) - f^*(X)\right)^2\right] = \left(\mathbb{E}\left[f_{\mathscr{D}}(X)\right] - f^*(X)\right)^2 + \mathrm{Var}\left[f_{\mathscr{D}}(X)\right]$$

- When the hypothesis class does not contain the true model, the hypothesis class itself **introduces bias** (**why?**)
    - Larger hypothesis classes will have smaller bias, but may also have higher variance (**why?**)

Prior Knowledge

- **Balancing between bias and variance** is a core problem in machine learning
- We accomplish this by encoding **prior knowledge** in various ways:
  - Choice of hypothesis class
  - Choice of regularization
  - Prior distributions over parameters
- Some prior knowledge is **domain specific**: e.g., prior distribution over parameters based on data that we've already seen; knowledge of physical processes that suggests a given family of functions
- Some prior knowledge is **not**: e.g., preferring small sets of features or small weights

Summary

- Expected generalization error can be **decomposed** into **bias** and **variance**
  - Using a biased estimator can be better than an unbiased one if it **sufficiently reduces variance**
- Worked example: **linear regression**
  - **MLE estimator** is **unbiased** but can have **high variance**
  - **MAP estimator** is **biased** but has a **controllable maximum variance**
- This same principle applies to the choice of **hypothesis class**
  - Bigger hypothesis class can be less biased, but higher variance
- In all cases, exploiting **prior knowledge** is the key to controlling bias vs. variance