Computational Data Analysis (02582)

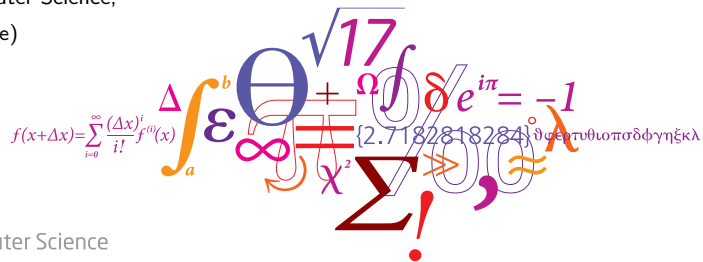
LEARNING FROM DATA (EPE, Bias, Variance and Complexity)

Sneha Das

Assistant Professor

Department of Applied Mathematics and Computer Science,

Technical University of Denmark (DTU Compute)



DTU Compute

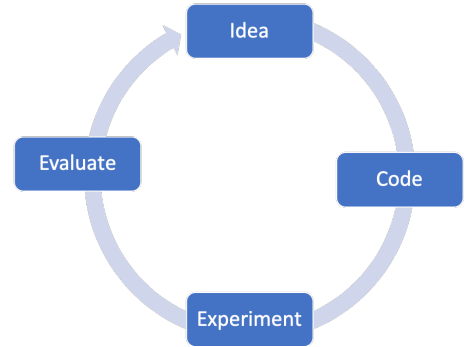
Department of Applied Mathematics and Computer Science

Outline I

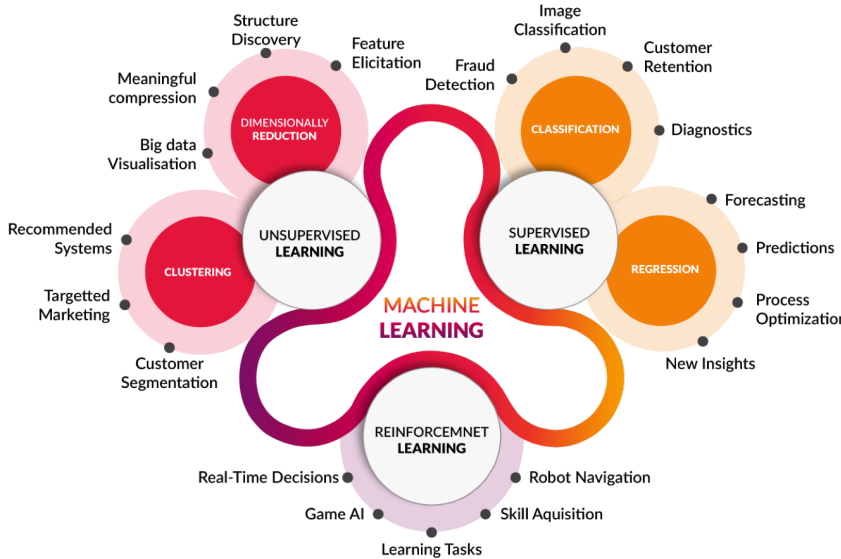
- Part I - What are we trying to learn?
  - Learning targets and finite data
    - Training vs test error
    - Instability under resampling
    - Expected prediction error (EPE)
  - Bias, variance, and data limitations
    - Bias-variance tradeoff
    - Controlling the bias-variance tradeoff

Learning targets and finite data

- **The Goal:** Learning from data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$.
- **The Split:** Learning $\leftarrow$ Training vs. Test.
- Why do we need the split?
- Finite data constraints on the unobservable target.



Landscape of terminology



Terminology

Methods are divided into categories depending on the nature of the response variable.
No response variable available - **unsupervised**

- Principal Component Analysis
- Cluster analysis

Response variable(s) available - **supervised**

- Categorical response - **classification**
  - Linear Discriminant Analysis
  - K Nearest Neighbors
  - Support Vector Machine
- Continuous response - **regression**
  - Ordinary Least Squares
  - Ridge regression
  - K-Nearest-Neighbors

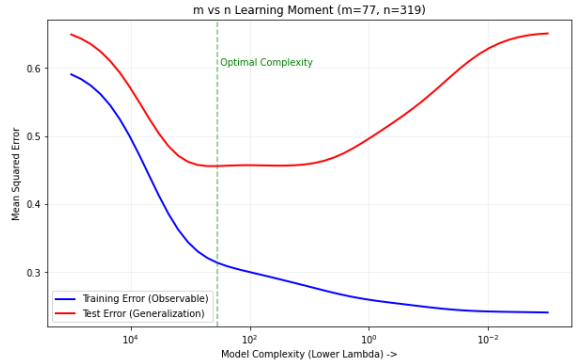
Inputs, predictors, features...

What we call our variables differs on the domain we come from!

Machine learning	**Statistics**	**Pattern recognition**
X inputs	X predictors	X features
Y outputs	Y responses	Y responses

Error types

- **Training Error:** Optimistic estimate.
- **Test Error:** Generalization performance.
- **Generalization Gap:** The difference between the two.



Terminology - a suggestion

The **data matrix** X. (Capital letters denote matrices)

- Rows correspond to **samples**
- Columns correspond to **variables**
- Size of X is $(n \times m)$, n samples and m variables.

Terminology The **response variable** y (small letters denote vectors)

- Usually a $(n \times 1)$ vector
- Also known as the **output**

The **model coefficients** β

- A $(m \times 1)$ vector

The **model error** vector e

- An $(n \times 1)$ vector

The **prediction error** vector ϵ

- An $(n \times 1)$ vector
- Also known as the **residual**

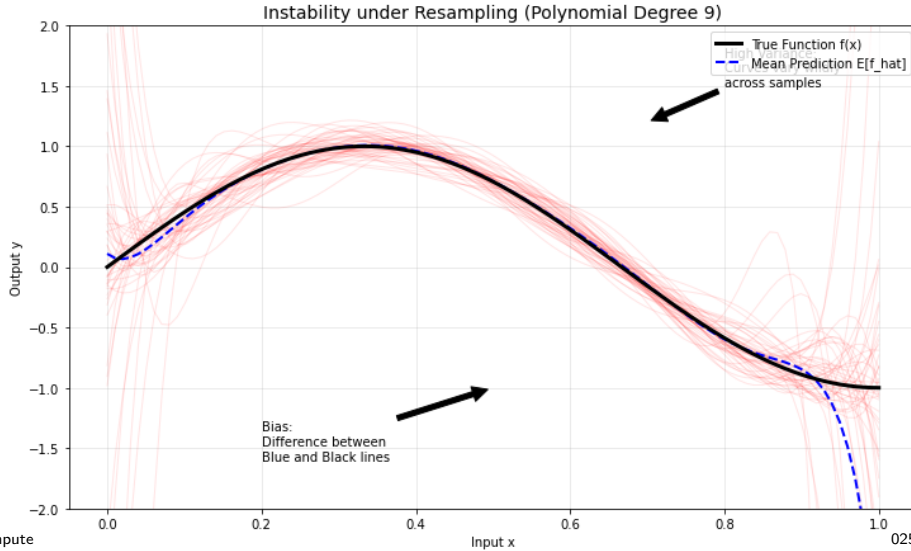
A **regression model** that is linear in β can be written $y = X\beta + e$

Instability under resampling

- **The Observation:** Small changes in $\mathcal{D}$ lead to large changes in $\hat{f}$.
- **The Problem:** If $\hat{f}$ is different in every 'parallel universe,' which one is the 'real' performance?
- **Training Error** is too specific to one dataset (optimistic).
- **Test Error** is a snapshot of one dataset.
- **The Phenomenon:** Same method, different result.

Instability under resampling

Multiple fitted curves showing variation across different samples.



Performance on unseen world (data)

How well does my learning algorithm perform across all possible data, on average?!
$rightarrow$ Expected prediction error

Expected prediction error

The **EPE** is our theoretical 'Gold Standard' target.

$$EPE(\hat{f}) = \mathbb{E}[(Y - \hat{f}(X))^2]$$

- **Integration:** It averages over the entire population distribution $P(X, Y)$.
- **The Dilemma:** We don't know $P(X, Y)$. We only have n samples.
- **The Goal:** We want a model $\hat{f}$ that minimizes this EPE, not just training error.

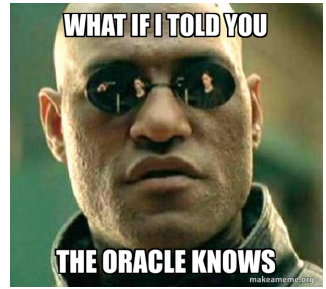
EPE Lower bound: Irreducible Noise

Why is $EPE > 0$ even for a perfect model?

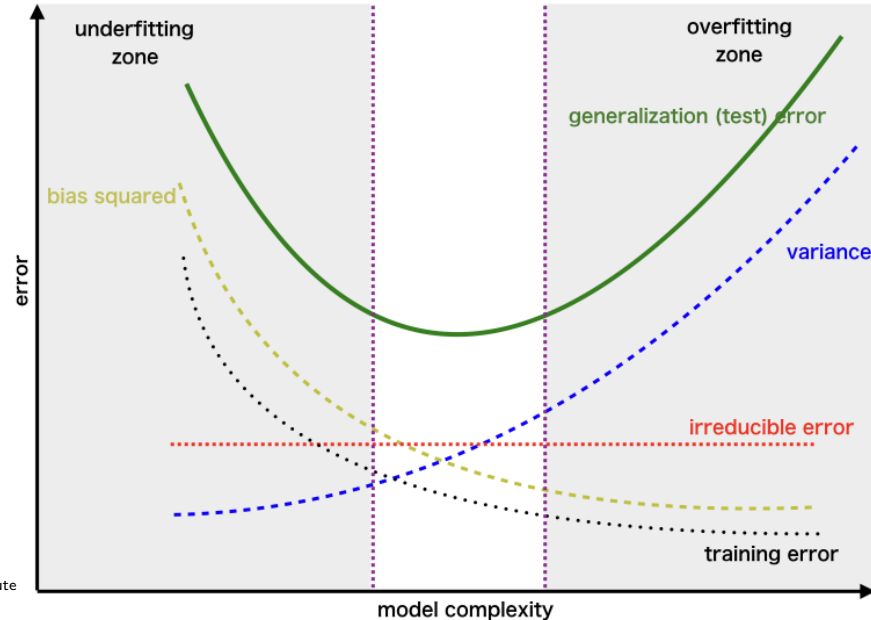
- **True Process:** $Y = f(X) + \epsilon$
- **Noise (ϵ):** Randomness we cannot model (measurement error, hidden variables).
- **Irreducible Error:** $\sigma^2 = \text{Var}(\epsilon)$.

$$\text{Total Error} = \text{Reducible Error} + \sigma^2$$

Even an 'Oracle' who knows $f(X)$ perfectly will have an error of σ^2.



EPE lower bound



Decomposing EPE

Three sources of errors contribute to the EPE,

$$\begin{aligned} EPE(x_0) &= E_{y,D|x_0}||y(x_0) - \hat{f}(x_0; D)||^2 \\ &= \sigma_e^2 + \text{bias}^2(\hat{f}(x_0; D)) + \text{variance}(\hat{f}(x_0; D)) \end{aligned}$$

where $\hat{f}(x_0)$ is the estimate of $f(x_0)$ with observed $y(x_0) = f(x_0) + e$, $E(e) = 0$, variance$(e) = \sigma_e^2$ and D is training data.

Decomposing EPE

Three sources of errors contribute to the EPE,

$$\begin{aligned} EPE(x_0) &= E_{y,D|x_0}||y(x_0) - \hat{f}(x_0; D)||^2 \\ &= \sigma_e^2 + \text{bias}^2(\hat{f}(x_0; D)) + \text{variance}(\hat{f}(x_0; D)) \end{aligned}$$

1. irreducible error $\sigma_e^2 = E_y \left(y(x_0) - f(x_0)\right)^2$

where $\hat{f}(x_0)$ is the estimate of $f(x_0)$ with observed $y(x_0) = f(x_0) + e$, $E(e) = 0$, variance$(e) = \sigma_e^2$ and D is training data.

Decomposing EPE

Three sources of errors contribute to the EPE,

$$\begin{aligned} EPE(x_0) &= E_{y,D|x_0}||y(x_0) - \hat{f}(x_0; D)||^2 \\ &= \sigma_e^2 + \text{bias}^2(\hat{f}(x_0; D)) + \text{variance}(\hat{f}(x_0; D)) \end{aligned}$$

1. irreducible error $\sigma_e^2 = E_y \left(y(x_0) - f(x_0)\right)^2$
2. $\text{bias}^2(\hat{f}(x_0; D)) = \Big(E_D(\hat{f}(x_0; D)) - f(x_0)\Big)^2$

where $\hat{f}(x_0)$ is the estimate of $f(x_0)$ with observed $y(x_0) = f(x_0) + e$, $E(e) = 0$, $\text{variance}(e) = \sigma_e^2$ and D is training data.

Decomposing EPE

Three sources of errors contribute to the EPE,

$$\begin{aligned} EPE(x_0) &= E_{y,D|x_0}||y(x_0) - \hat{f}(x_0; D)||^2 \\ &= \sigma_e^2 + \text{bias}^2(\hat{f}(x_0; D)) + \text{variance}(\hat{f}(x_0; D)) \end{aligned}$$

1. irreducible error $\sigma_e^2 = E_y \left(y(x_0) - f(x_0)\right)^2$
2. $\text{bias}^2(\hat{f}(x_0; D)) = \Big(E_D(\hat{f}(x_0; D)) - f(x_0)\Big)^2$
3. $\text{variance}(\hat{f}(x_0; D)) = E_D \left(\hat{f}(x_0; D) - E_D(\hat{f}(x_0; D))\right)^2$

where $\hat{f}(x_0)$ is the estimate of $f(x_0)$ with observed $y(x_0) = f(x_0) + e$, $E(e) = 0$, $\text{variance}(e) = \sigma_e^2$ and D is training data.

What is *Statistical Bias* ?

Q. What is it?

- The difference between an expected value and the true value.

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta \tag{1}$$

Q. Bias of what?

- Could be of β the model parameters.
- Could also be of the predictions $\hat{y}$.

Q. How to get an unbiased estimate?

- Repeat experiment, take average of β or $\hat{y}$.
- Will be equal to the true value or unbiased.

Bias: Systematic error, could be due to incorrect assumptions in the model.

What is *Variance*?

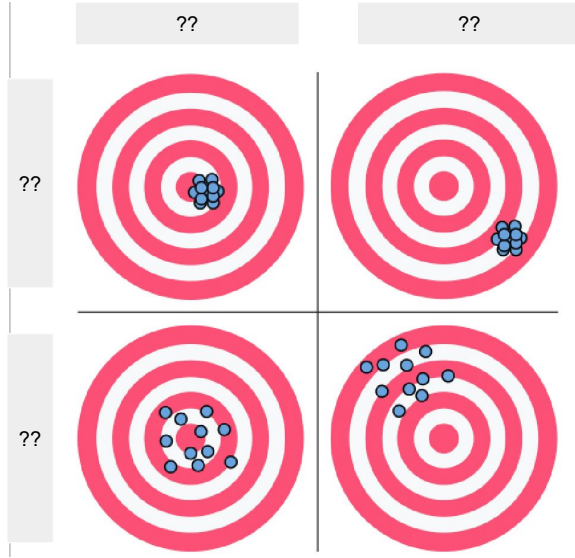
We might be right on average (unbiased) but we only do one experiment.
Variance $\rightarrow$ Quantifies: How far are we from the true value, if we do only one experiment?

$$\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2] \tag{2}$$

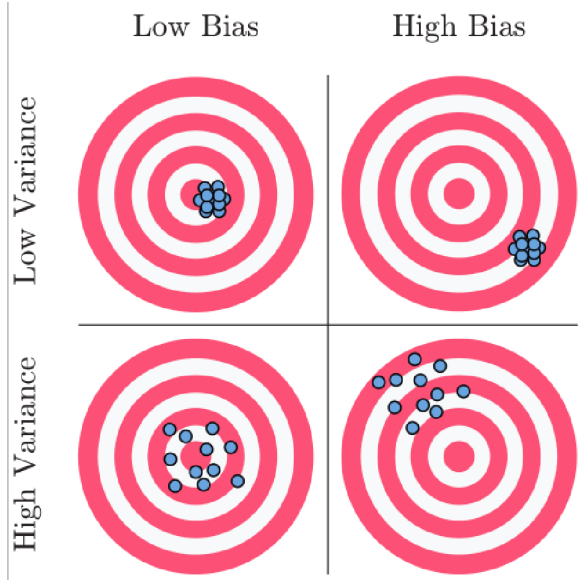
- **High variance,** we might end up far from the true value.
- **Low variance,** we get almost the same result every time, how far it is from the true value depends on the bias.

Variance: Error due to sensitivity to small fluctuations in the training set (eg: sampling differences; sensor noise in observational data).

Bias & variance



Bias & variance



Linear regression (Review)

Given a continuous response measurement, find the relation of this variable to a set of input variables. Model:

$$y = X\beta + e \tag{3}$$

Examples:

- Predict ice-cream sale, y, based on measurement of temperature, x_1, and sunshine hours, x_2, $X = [x_1, x_2]$. The relation is given by β.
- Predict house prices based on interest rate, unemployment rate, region, size and building year.
- Predict life expectancy given smoking habits, age, BMI etc.

Ordinary least suares (OLS)

(Error types, generalization gap, EPE, Noise and irreducible error, Bias-variance tradeoff wrt complexity) Find the β that minimizes the residual error (RSS - residual sum of squares), $y - X\beta$. Positive and negative errors are equally bad and large errors are worse than small errors. Hence, minimize:

$$||y - X\beta||_2^2 = \sum_{i=1}^{n}(y_i - X_i\beta)^2 \tag{4}$$

$$\hat{\beta}_{\text{OLS}} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2$$

Set gradient to zero:

$$\frac{\partial}{\partial\beta}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

$$-2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

$$\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\beta = 0$$

$$\hat{\beta}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- **Requirement:** $\mathbf{X}^T\mathbf{X}$ must be invertible (non-singular).
- **Instability:** If features are highly correlated, $\mathbf{X}^T\mathbf{X}$ is near-singular, causing high variance.

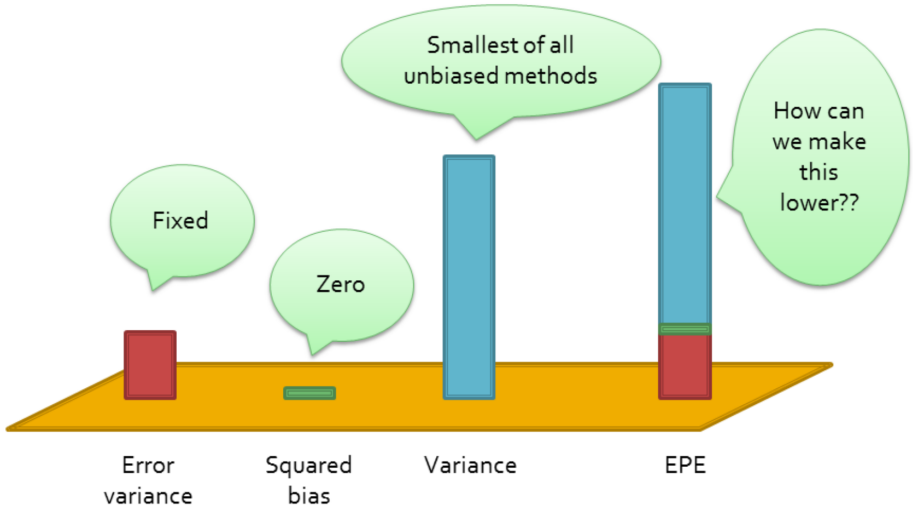
Properties of OLS

Ordinary least squares (OLS) is great!
OLS is the **best linear unbiased estimate** (BLUE)

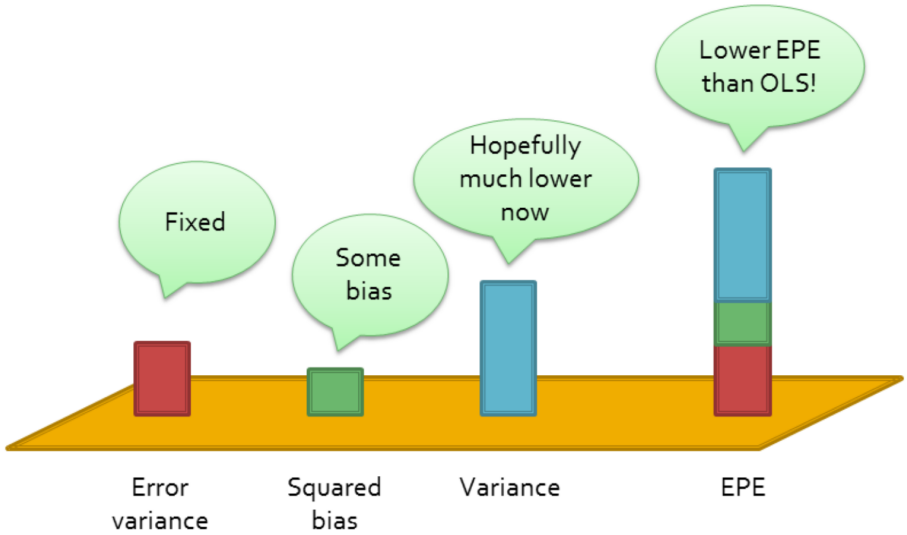
- Unbiased: $E(\beta_{OLS}) = \beta$
- Best unbiased: $Var(\beta_{OLS}) \leq Var(\beta_{linear})$

Back to EPE

OLS - Best Unbiases linear estimator



Back to EPE



introduce bias ... operationalizing tradeoff (overfitting & underfitting)

Controlling the bias-variance tradeoff - Model complexity

Model complexity refers to the **flexibility** or the **Degrees of Freedom** of the estimator $\hat{f}$.

- **For parametric models:** Often tied to the number of parameters m.
- **For non-parametric models (eg: knn):** Tied to the size of the neighborhood or the strength of a penalty.

The Complexity Trade-off:

- **Low Complexity:** High Bias (assumes too much), Low Variance (stable) $\rightarrow$ underfitting.
- **High Complexity:** Low Bias (fits (training) data well), High Variance (unstable) $\rightarrow$ overfitting.

The m vs. n Learning Moment

Complexity is not an absolute property of the model; it is **relative** to the amount of data n.

Stable Regime ($n \gg m$)

- Data constrains the model.
- Training and Test error are close.
- Overfitting is difficult.

Overfitting Regime ($n \approx m$)

- Model 'fits' noise.
- Large Generalization Gap.
- Instability explodes.

(*Exercise 4 Example:* 11 *features for* 1600 *wine samples (Stable) vs.* 77 *features for* 300 *wine samples (Overfitting).*)

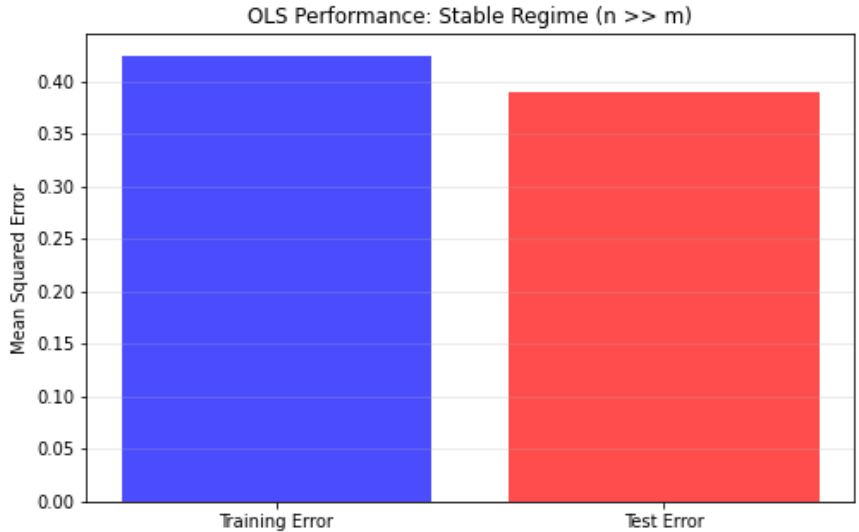
UCI-Wine dataset

df - DataFrame

Index	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	class
0	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
4	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5

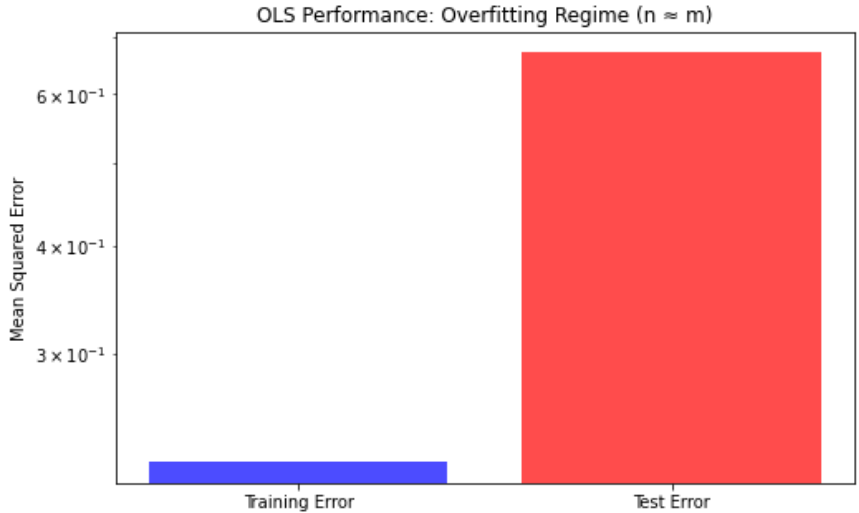
OLS

Generalization gap for $M << N$



OLS

Generalization gap for $M \sim N$



Example 2 - Ridge regeression

Ridge (Error types, generalization gap, EPE, Noise and irreducible error, Bias-variance tradeoff wrt complexity)

1. We wish to lower the variance of $\hat{y} = X\beta$.
2. Lowering the size of β will lower the variance of $\hat{y}$.
3. Lower the size of β by **shrinkage**,
   - $\hat{\beta}_{\text{ridge}} = \arg\min_\beta ||y - X\beta||^2 + \lambda||\beta||^2$
   - OLS criterion plus extra term.
   - λ controls the amount of shrinkage.

Ridge regression

Ridge regression has a closed-form solution!

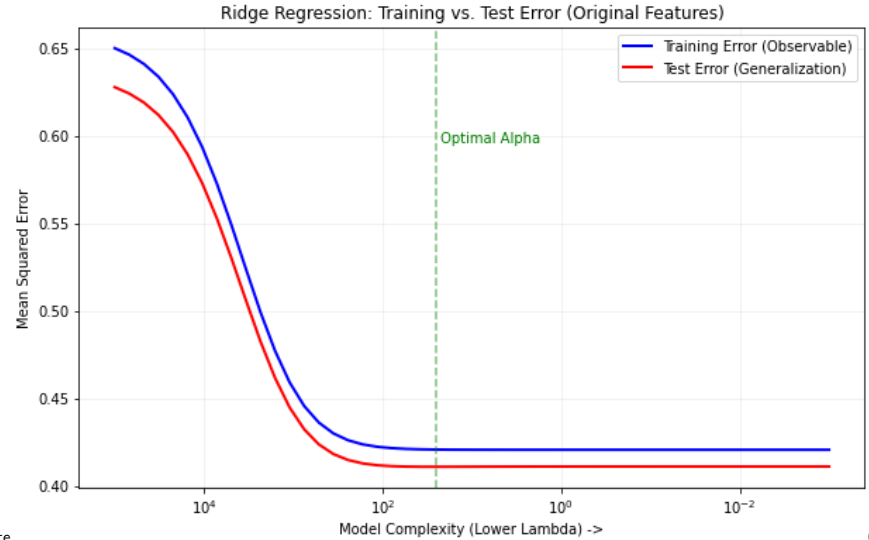
$$\beta_{ridge} = (X^T X + \lambda I)^{-1} X^T y \tag{5}$$

- (You will derive this expression for β_{ridge} in the exercises.)
- Intuitively, we are adding a small number to the diagonal of the matrix to invert.
  - Hence, the name 'ridge' regression.
  - Stabilizes the inverse numerically.
  - Ridge regression solutions are available even when $m > n$!

Important: Data are centered (mean of each variable is zero) and normalized (variables scaled such that standard deviation equals 1).
Centering data removes the mean, causes the shrinkage to shrink towards zero.
Normalizing data puts equal importance to all variables due to the penalty term $||\beta||^2$.

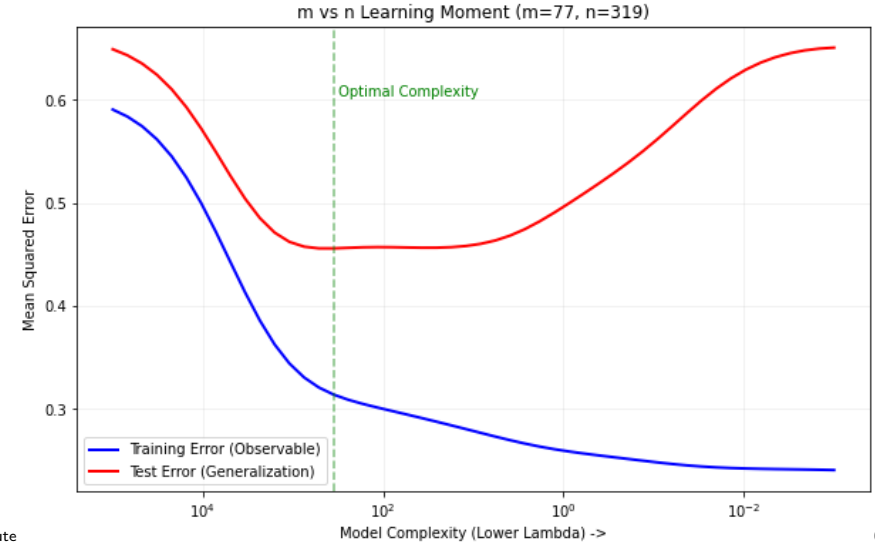
Ridge regression

Error with respect to complexity ($M << N$)



Ridge regression

Error with respect to complexity ($M \approx N$)



Controlling the Complexity 'Knob'

How do we move along the Bias-Variance curve?

Method	Complexity Knob	High Complexity
OLS	Fixed (m)	Large m
Ridge	Regularization λ	Small $\lambda \to 0$

- **Regularization:** Deliberately adds bias to reduce variance.

Model selection: Finding the Sweet Spot

How do we choose the optimal level of complexity?

- We cannot observe EPE directly (requires the Oracle).
- We use **Validation Sets** or **Cross-Validation** to estimate test performance.

Next Lecture:
Cross-validation, AIC/BIC, and Model Selection strategies.

Today's exercises

1.1 The OLS 'Wobble'

Focus: Instability.
How sampling noise and collinearity ($r = 0.98$) explode coefficient variance, proving that "Unbiased $\neq$ Reliable."

1.2 Ridge Tradeoff

Focus: The Knob (λ).
Using regularization to 'break' the model (add bias) to tame variance and find the EPE 'Sweet Spot'.

1.3 The M vs N Paradox

Focus: Wine Quality.
Stable: $n \gg m$ (Flat error).
Overfitting: $n \approx m$ (U-Shape).
Complexity is a property of the whole system.