Ridge Regression

Yingzhen Li

Department of Computing
Imperial College London

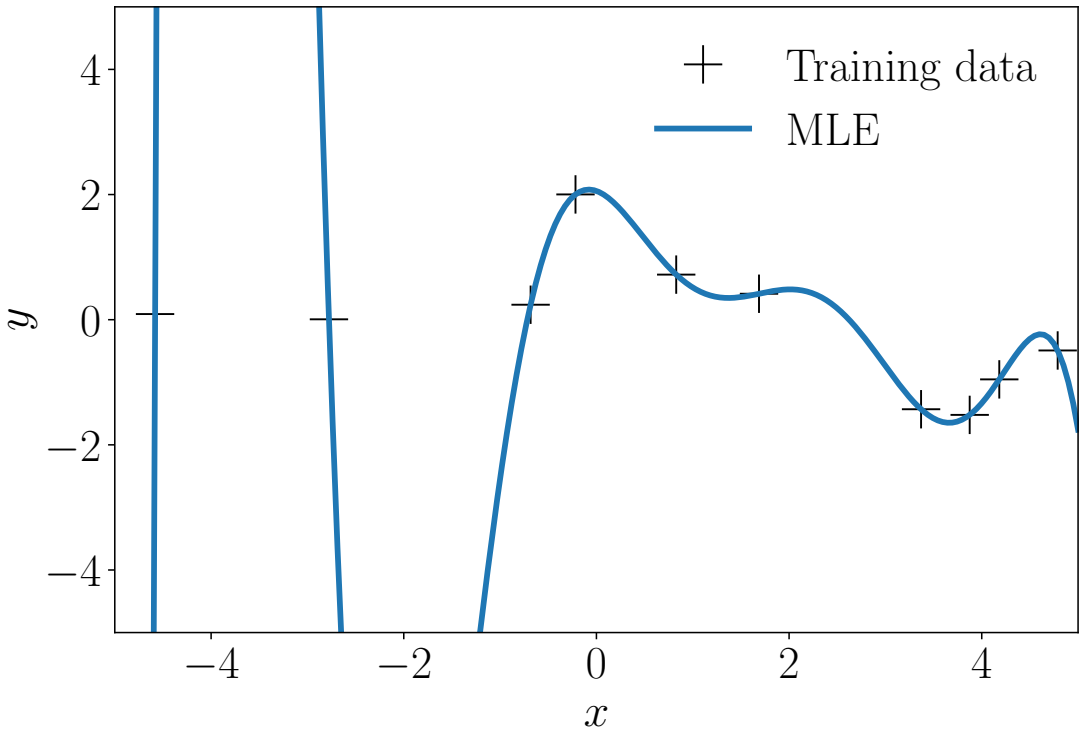
@liyzhen2
yingzhen.li@imperial.ac.uk

Nov 2, 2021

Reading for this week

Read MML book: Sections 4.5, 9.2, 10.1-10.3, 10.6
Do MML book exercises: Exercises 4.8 - 4.12
An extra exercise will be uploaded to course materials.

Overfitting



$$\phi(x) = [1 \; x \; x^2 \; x^3, \ldots]^\top \tag{1}$$

When the model is too flexible, risk of overfitting!

Overfitting

To help avoid overfitting:

- Choose model with the right complexity (using validation data)

Overfitting

To help avoid overfitting:

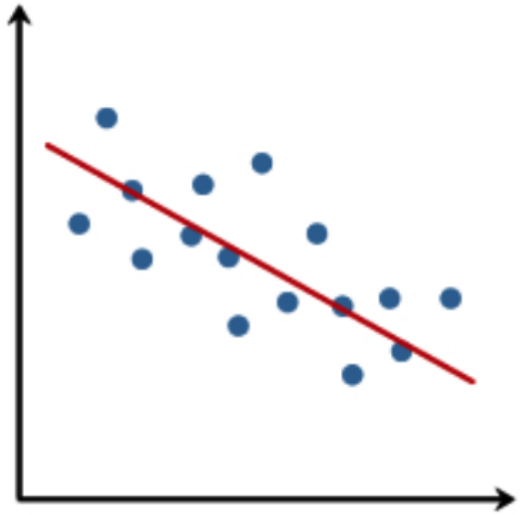
- Choose model with the right complexity (using validation data)
- **Regularise the model** (this lecture)

Recap: Linear regression

Fitting linear regression models:

- Dataset: $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$,
  $\mathbf{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times D}$,
  $\mathbf{y} = [y_1, ..., y_N]^\top \in \mathbb{R}^{N \times 1}$
- Goal: find $\boldsymbol{\theta} \in \mathbb{R}^{D \times 1}$ such that

$$\mathbf{y} \approx \mathbf{X}\boldsymbol{\theta}$$



Recap: Linear regression

A typical linear regression model:

- $\boldsymbol{x} \in \mathbb{R}^{D\times 1}$: input features; $y \in \mathbb{R}$: output value
- Model and loss:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^\top \boldsymbol{\theta}, \quad y = f(\boldsymbol{x}; \boldsymbol{\theta}) + \epsilon, \ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$L(\boldsymbol{\theta}) = \frac{1}{2\sigma^2} \sum_n (f(\boldsymbol{x}_n; \boldsymbol{\theta}) - y_n)^2$$

- Rewriting the loss in matrix form:

$$L(\boldsymbol{\theta}) = \frac{1}{2\sigma^2} ||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||_2^2$$

Ridge regression

A typical **ridge regression** model:

- $\boldsymbol{x} \in \mathbb{R}^{D\times 1}$: input features; $y \in \mathbb{R}$: output value
- Model and **loss**:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^\top \boldsymbol{\theta}, \quad y = f(\boldsymbol{x}; \boldsymbol{\theta}) + \epsilon, \ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$L(\boldsymbol{\theta}) = \frac{1}{2\sigma^2} \sum_n (f(\boldsymbol{x}_n; \boldsymbol{\theta}) - y_n)^2 + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2$$

- Rewriting the loss in matrix form:

$$L(\boldsymbol{\theta}) = \frac{1}{2\sigma^2} ||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2$$

Ridge regression

Optimal solution of $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_R^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2 := \arg\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$$

- Setting $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = 0$:

Ridge regression

Optimal solution of $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_R^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{2\sigma^2}||\mathbf{y}-\mathbf{X}\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2 := \arg\min_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})$$

- Setting $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = 0$:

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}-\mathbf{y}) + \lambda\boldsymbol{\theta} = 0$$

$$\Rightarrow (\lambda\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X})\boldsymbol{\theta}^* = \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{y}$$

$$\Rightarrow \boldsymbol{\theta}_R^* = (\sigma^2\lambda\mathbf{I} + \mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

Gradient descent for ridge regression

Gradient descent to find $\boldsymbol{\theta}_R^*$:

Assume constant step-sizes $\gamma_t = \gamma$:

1. Define **starting point** $\boldsymbol{\theta}_0$, set $t \leftarrow 0$
2. Set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma_t \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t)$, $t \leftarrow t+1$

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &= \boldsymbol{\theta}_t - \gamma_t \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}_t) \\ &= \boldsymbol{\theta}_t - \gamma \left(\frac{1}{\sigma^2} \mathbf{X}^\top (\mathbf{X}\boldsymbol{\theta}_t - \mathbf{y}) + \lambda \boldsymbol{\theta}_t \right) \\ &= ((1-\gamma\lambda)\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})\boldsymbol{\theta}_t + \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y}\end{aligned}$$

3. Repeat 1 until stopping criterion.

Gradient descent for ridge regression

Gradient descent to find $\boldsymbol{\theta}_R^*$:

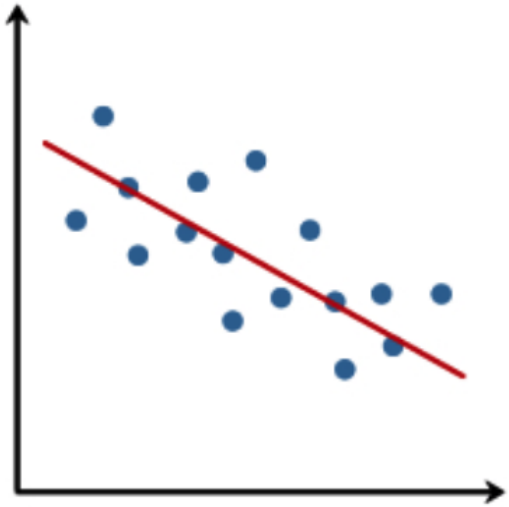
Assume constant step-sizes $\gamma_t = \gamma$:

- GD returns the following iterative updates:

$$\boldsymbol{\theta}_{t+1} = ((1-\gamma\lambda)\mathbf{I} - \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{X})\boldsymbol{\theta}_t + \frac{\gamma}{\sigma^2}\mathbf{X}^\top\mathbf{y}$$

- An exercise for you in this week:
  1. Show that if GD converges, it would converge to $\boldsymbol{\theta}_R^*$
  2. Derive the "safe threshold" for the constant step size γ

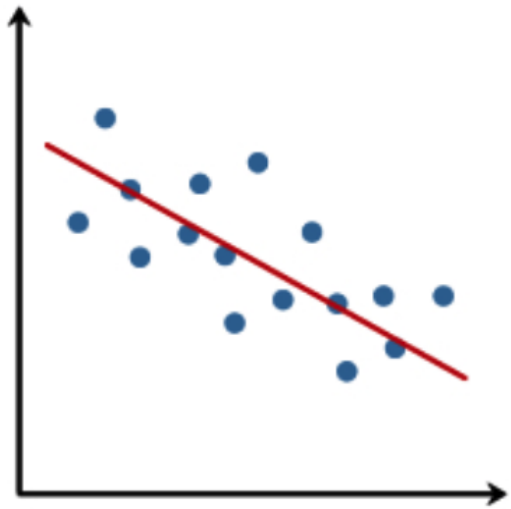
Regression with non-linear features



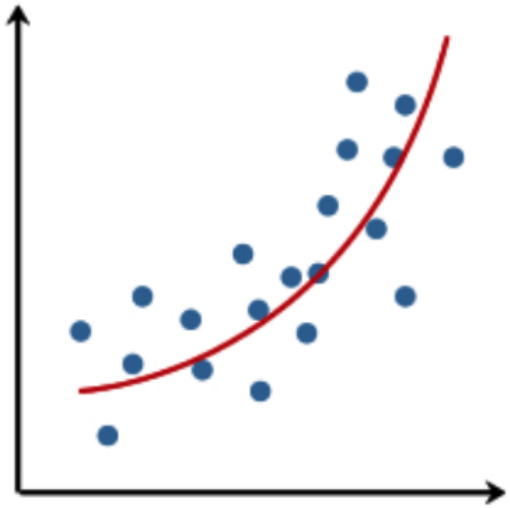
Linear regression

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^\top \boldsymbol{\theta}$$

Regression with non-linear features



$\Rightarrow$



Linear regression

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{x}^\top \boldsymbol{\theta}$$

Non-linear regression

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \phi(\boldsymbol{x})^\top \boldsymbol{\theta}$$

Regression with non-linear features

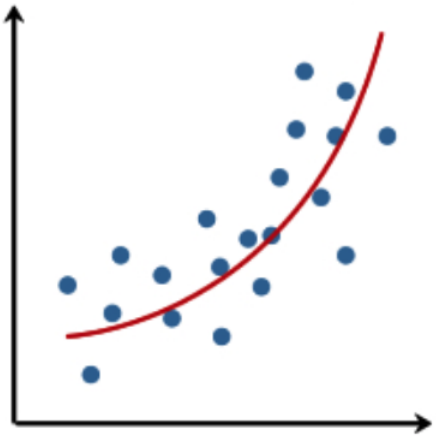
Extending to **non-linear regression**:

- Key idea: using a non-linear feature mapping: $\phi(\cdot) : \mathbb{R}^D \to \mathbb{R}^p$
- The non-linear regression model:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \phi(\boldsymbol{x})^\top \boldsymbol{\theta}$$

$$y = f(\boldsymbol{x}, \boldsymbol{\theta}) + \epsilon, \ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Recover linear regression when $\phi(\boldsymbol{x}) = \boldsymbol{x}$



$$\phi(x) = [1, x, x^2]$$

Regression with non-linear features

Fitting ridge regression:

$$L(\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\sum_n (f(\boldsymbol{x}_n, \boldsymbol{\theta}) - y_n)^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2$$

- **Write** $\Phi = [\phi(\boldsymbol{x}_1), ..., \phi(\boldsymbol{x}_N)]^\top \in \mathbb{R}^{N\times p}$:

$$\boldsymbol{\theta}_R^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \frac{1}{2\sigma^2}||\mathbf{y} - \Phi\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2}||\boldsymbol{\theta}||_2^2$$

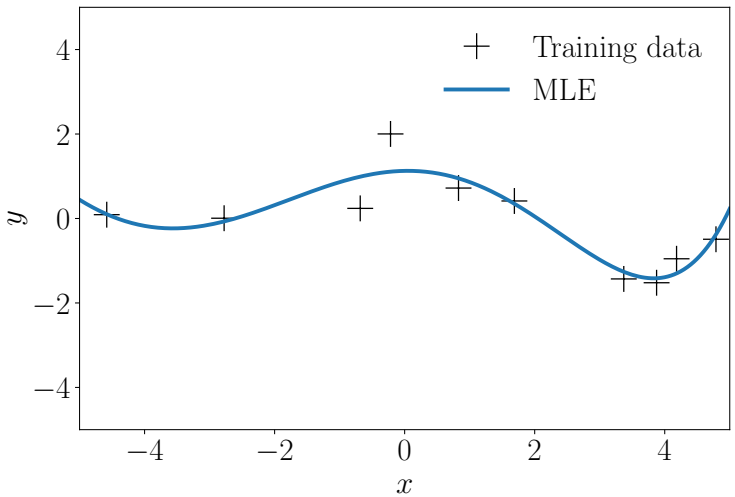
- Optimal solution for $\boldsymbol{\theta}$:

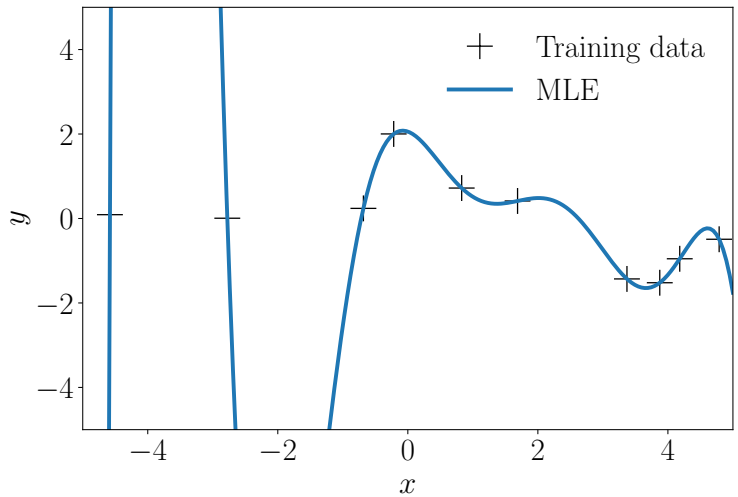
$$\boldsymbol{\theta}_R^* = (\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top\mathbf{y}$$

Intuition behind ridge regression

Regression with polynomial functions as an example:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{p} \theta_i \boldsymbol{x}^{i-1}$$





Several solutions fit the training data almost equally well.

$\Rightarrow$ How to choose a model?

Intuition behind ridge regression

Regression with polynomial functions as an example:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{p} \theta_i \boldsymbol{x}^{i-1}$$

The ℓ_2 regulariser used in ridge regression:

$$R(\boldsymbol{\theta}) = ||\boldsymbol{\theta}||_2^2 = \sum_{i=1}^{p} \boldsymbol{\theta}_i^2$$

- shrinks elements of $\boldsymbol{\theta}$ to zero

Intuition behind ridge regression

Regression with polynomial functions as an example:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{p} \theta_i \boldsymbol{x}^{i-1}$$

The ℓ_2 regulariser used in ridge regression:

$$R(\boldsymbol{\theta}) = ||\boldsymbol{\theta}||_2^2 = \sum_{i=1}^{p} \boldsymbol{\theta}_i^2$$

- shrinks elements of $\boldsymbol{\theta}$ to zero
- if $\theta_i = 0$, then feature $\boldsymbol{x}^{i-1}$ is not in use
  $\Rightarrow$ simpler model!
- Ridge regression balances between data fit and model simplicity

Bias-variance trade-off

Ridge regression can return estimator of $\boldsymbol{\theta}$ with **smaller variance**.

In such case the (expected) test error can be reduced.

Bias-variance trade-off

Estimating model parameters:

- Assume **no model mismatch**: $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ generated by:

$$y_n = f(\boldsymbol{x}_n; \boldsymbol{\theta}_0) + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

- The above process also writes as $\mathcal{D} \sim p_{data}^N$.
- Test data $(\boldsymbol{x}_{test}, y_{test})$ generated from **the same process**

Bias-variance trade-off

Estimating model parameters:

- Assume **no model mismatch**: $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ generated by:

$$y_n = f(\boldsymbol{x}_n; \boldsymbol{\theta}_0) + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

- The above process also writes as $\mathcal{D} \sim p_{data}^N$.
- Test data $(\boldsymbol{x}_{test}, y_{test})$ generated from **the same process**
- Goal: estimate $\boldsymbol{\theta}$ using data such that $\boldsymbol{\theta}^* \approx \boldsymbol{\theta}_0$
  - Both linear regression and ridge regression are ways to estimate $\boldsymbol{\theta}$
  - $\boldsymbol{\theta}^*$ depends on the dataset $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$
    (so we will also write the solution as $\boldsymbol{\theta}^*(\mathcal{D})$ when dataset is given)

Bias-variance trade-off

Notice that $\boldsymbol{\theta}^*$ depends on the dataset $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$
Consider expected ℓ_2 error for estimate:

$$error(\boldsymbol{\theta}^*) = \mathbb{E}_{\mathcal{D} \sim p_{data}^N}[||\boldsymbol{\theta}^*(\mathcal{D}) - \boldsymbol{\theta}_0||_2^2]$$

Bias-variance trade-off

Notice that $\boldsymbol{\theta}^*$ depends on the dataset $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$
Consider expected ℓ_2 error for estimate:

$$\begin{aligned} error(\boldsymbol{\theta}^*) &= \mathbb{E}_{\mathcal{D}\sim p_{data}^N}[||\boldsymbol{\theta}^*(\mathcal{D}) - \boldsymbol{\theta}_0||_2^2] \\ &= \underbrace{||\mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}^*(\mathcal{D})] - \boldsymbol{\theta}_0||_2^2}_{bias^2} + \underbrace{\mathrm{tr}[\mathbb{V}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}^*(\mathcal{D})]]}_{variance} \end{aligned}$$

Bias-variance trade-off

Notice that $\boldsymbol{\theta}^*$ depends on the dataset $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$
Consider expected ℓ_2 error for estimate:

$$\begin{aligned} error(\boldsymbol{\theta}^*) &= \mathbb{E}_{\mathcal{D}\sim p_{data}^N}[||\boldsymbol{\theta}^*(\mathcal{D}) - \boldsymbol{\theta}_0||_2^2] \\ &= \underbrace{||\mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}^*(\mathcal{D})] - \boldsymbol{\theta}_0||_2^2}_{bias^2} + \underbrace{\mathrm{tr}[\mathbb{V}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}^*(\mathcal{D})]]}_{variance} \end{aligned}$$

- Ideally: want unbiased estimator (bias=0) and small variance

Bias-variance trade-off

Notice that $\boldsymbol{\theta}^*$ depends on the dataset $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$
Consider expected ℓ_2 error for estimate:

$$\begin{aligned} error(\boldsymbol{\theta}^*) &= \mathbb{E}_{\mathcal{D}\sim p_{data}^N}[||\boldsymbol{\theta}^*(\mathcal{D}) - \boldsymbol{\theta}_0||_2^2] \\ &= \underbrace{||\mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}^*(\mathcal{D})] - \boldsymbol{\theta}_0||_2^2}_{bias^2} + \underbrace{\mathrm{tr}[\mathbb{V}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}^*(\mathcal{D})]]}_{variance} \end{aligned}$$

- Ideally: want unbiased estimator (bias=0) and small variance
- In practice: bias-variance trade-off

Bias-variance trade-off

Notice that $\boldsymbol{\theta}^*$ depends on the dataset $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$
Consider expected error in **matrix form**:

$$Error(\boldsymbol{\theta}^*) = \mathbb{E}_{\mathcal{D} \sim p_{data}^N}[(\boldsymbol{\theta}^*(\mathcal{D}) - \boldsymbol{\theta}_0)(\boldsymbol{\theta}^*(\mathcal{D}) - \boldsymbol{\theta}_0)^\top]$$

Bias-variance trade-off

Notice that $\boldsymbol{\theta}^*$ depends on the dataset $\mathcal{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$
Consider expected error in **matrix form**:

$$\begin{aligned} Error(\boldsymbol{\theta}^*) &= \mathbb{E}_{\mathcal{D}\sim p_{data}^N}[(\boldsymbol{\theta}^*(\mathcal{D}) - \boldsymbol{\theta}_0)(\boldsymbol{\theta}^*(\mathcal{D}) - \boldsymbol{\theta}_0)^\top] \\ &:= \mathbf{b}(\boldsymbol{\theta}^*)\mathbf{b}(\boldsymbol{\theta}^*)^\top + \mathbf{V}(\boldsymbol{\theta}^*) \end{aligned}$$

$$\begin{aligned} \text{bias:} \quad & \mathbf{b}(\boldsymbol{\theta}^*) = \mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}^*(\mathcal{D})] - \boldsymbol{\theta}_0 \\ \text{variance:} \quad & \mathbf{V}(\boldsymbol{\theta}^*) = \mathbb{V}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}^*(\mathcal{D})] \end{aligned}$$

- Ideally: want unbiased estimator (bias=$\mathbf{0}$) and small variance
- In practice: bias-variance trade-off

Bias-variance trade-off

How bias-variance trade-off is relevant to over-fitting:
Expected prediction error:

$$error_{pred}(\boldsymbol{\theta}^*) = \mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\mathbb{E}_{(\boldsymbol{x}_{test},y_{test})\sim p_{data}}[||y_{test} - f(\boldsymbol{x}_{test}, \boldsymbol{\theta}^*(\mathcal{D}))||_2^2]]$$

Bias-variance trade-off

How bias-variance trade-off is relevant to over-fitting:
Expected prediction error:

$$\begin{aligned} error_{pred}(\boldsymbol{\theta}^*) &= \mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\mathbb{E}_{(\boldsymbol{x}_{test},y_{test})\sim p_{data}}[||y_{test} - f(\boldsymbol{x}_{test}, \boldsymbol{\theta}^*(\mathcal{D}))||_2^2]] \\ &= \mathbb{E}_{\boldsymbol{x}_{test}}[\boldsymbol{\phi}(\boldsymbol{x}_{test})^\top Error(\boldsymbol{\theta}^*)\boldsymbol{\phi}(\boldsymbol{x}_{test})] + \sigma^2 \end{aligned}$$

Bias-variance trade-off

How bias-variance trade-off is relevant to over-fitting:
Expected prediction error:

$$\begin{aligned} error_{pred}(\boldsymbol{\theta}^*) &= \mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\mathbb{E}_{(\boldsymbol{x}_{test},y_{test})\sim p_{data}}[||y_{test} - f(\boldsymbol{x}_{test}, \boldsymbol{\theta}^*(\mathcal{D}))||_2^2]] \\ &= \mathbb{E}_{\boldsymbol{x}_{test}}[\phi(\boldsymbol{x}_{test})^\top Error(\boldsymbol{\theta}^*)\phi(\boldsymbol{x}_{test})] + \sigma^2 \end{aligned}$$

If we have two estimators $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$:

$$Error(\boldsymbol{\theta}_1) \preceq Error(\boldsymbol{\theta}_2) \quad \Rightarrow \quad error_{pred}(\boldsymbol{\theta}_1) \leqslant error_{pred}(\boldsymbol{\theta}_2)$$

Bias-variance trade-off

How bias-variance trade-off is relevant to over-fitting:
Expected prediction error:

$$\begin{aligned} error_{pred}(\boldsymbol{\theta}^*) &= \mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\mathbb{E}_{(\boldsymbol{x}_{test},y_{test})\sim p_{data}}[||y_{test} - f(\boldsymbol{x}_{test}, \boldsymbol{\theta}^*(\mathcal{D}))||_2^2]] \\ &= \mathbb{E}_{\boldsymbol{x}_{test}}[\boldsymbol{\phi}(\boldsymbol{x}_{test})^\top Error(\boldsymbol{\theta}^*)\boldsymbol{\phi}(\boldsymbol{x}_{test})] + \sigma^2 \end{aligned}$$

If we have two estimators $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$:

$$Error(\boldsymbol{\theta}_1) \preceq Error(\boldsymbol{\theta}_2) \quad \Rightarrow \quad error_{pred}(\boldsymbol{\theta}_1) \leqslant error_{pred}(\boldsymbol{\theta}_2)$$

- Smaller estimation error $\Rightarrow$ smaller prediction error
- Depends on bias-variance trade-off

Linear regression returns an unbiased estimator

Reminder for solving linear/ridge regression:

- Write $\Phi = [\phi(\boldsymbol{x}_1), ..., \phi(\boldsymbol{x}_N)]^\top \in \mathbb{R}^{N \times p}$:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta} \in \Theta} \frac{1}{2\sigma^2} ||\mathbf{y} - \Phi\boldsymbol{\theta}||_2^2 + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2$$

- Optimal solution for $\boldsymbol{\theta}$ in ridge regression:

$$\boldsymbol{\theta}_R^* = (\sigma^2 \lambda \mathbf{I} + \Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}$$

- Optimal solution for $\boldsymbol{\theta}$ in linear regression ($\lambda = 0$):

$$\boldsymbol{\theta}_L^* = (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}$$

Linear regression returns an unbiased estimator

Optimal solution for linear regression: $\boldsymbol{\theta}_L^* = (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}$

- Assuming no model error:

$$\mathbf{y} = \Phi \boldsymbol{\theta}_0 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} = [\epsilon_1, ..., \epsilon_N]^\top, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

Linear regression returns an unbiased estimator

Optimal solution for linear regression: $\boldsymbol{\theta}_L^* = (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}$

- Assuming no model error:

$$\mathbf{y} = \Phi \boldsymbol{\theta}_0 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} = [\epsilon_1, ..., \epsilon_N]^\top, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

- Leading to optimal solution as: $\boldsymbol{\theta}_L^* = (\Phi^\top \Phi)^{-1} \Phi^\top (\Phi \boldsymbol{\theta}_0 + \boldsymbol{\epsilon})$

Linear regression returns an unbiased estimator

Optimal solution for linear regression: $\boldsymbol{\theta}_L^* = (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}$

- Assuming no model error:

$$\mathbf{y} = \Phi \boldsymbol{\theta}_0 + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} = [\epsilon_1, ..., \epsilon_N]^\top, \quad \epsilon_n \sim \mathcal{N}(0, \sigma^2)$$

- Leading to optimal solution as: $\boldsymbol{\theta}_L^* = (\Phi^\top \Phi)^{-1} \Phi^\top (\Phi \boldsymbol{\theta}_0 + \boldsymbol{\epsilon})$
- **Unbiased estimator**:

$$\mathbb{E}_{\mathcal{D} \sim p_{data}^N} [\boldsymbol{\theta}_L^*(\mathcal{D})] = \mathbb{E}_{\mathcal{D} \sim p_{data}^N} [(\Phi^\top \Phi)^{-1} \Phi^\top (\Phi \boldsymbol{\theta}_0 + \boldsymbol{\epsilon})] = \boldsymbol{\theta}_0$$

Ridge regression returns a biased estimator

The ridge regression estimator: $\boldsymbol{\theta}_R^* = (\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top(\Phi\boldsymbol{\theta}_0 + \boldsymbol{\epsilon})$

- Compute the mean of $\boldsymbol{\theta}_R^*$ for $\mathcal{D} \sim p_{data}^N$:

Ridge regression returns a biased estimator

The ridge regression estimator: $\boldsymbol{\theta}_R^* = (\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top(\Phi\boldsymbol{\theta}_0 + \boldsymbol{\epsilon})$

- Compute the mean of $\boldsymbol{\theta}_R^*$ for $\mathcal{D} \sim p_{data}^N$:

$$\mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}_R^*(\mathcal{D})] = (\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top\Phi\boldsymbol{\theta}_0$$

$\Rightarrow$ Ridge regression returns a **biased estimator**

Ridge regression returns a biased estimator

The ridge regression estimator: $\boldsymbol{\theta}_R^* = (\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top(\Phi\boldsymbol{\theta}_0 + \boldsymbol{\epsilon})$

- Compute the mean of $\boldsymbol{\theta}_R^*$ for $\mathcal{D} \sim p_{data}^N$:

$$\mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}_R^*(\mathcal{D})] = (\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top\Phi\boldsymbol{\theta}_0$$

  $\Rightarrow$ Ridge regression returns a **biased estimator**

- Compute the variance of $\boldsymbol{\theta}_R^*$ for $\mathcal{D} \sim p_{data}^N$:

$$\mathbb{V}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}_R^*(\mathcal{D})] = \mathbb{V}_{\mathcal{D}\sim p_{data}^N}[(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top(\Phi\boldsymbol{\theta}_0 + \boldsymbol{\epsilon})]$$

Ridge regression returns a biased estimator

The ridge regression estimator: $\boldsymbol{\theta}_R^* = (\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top(\Phi\boldsymbol{\theta}_0 + \boldsymbol{\epsilon})$

- Compute the mean of $\boldsymbol{\theta}_R^*$ for $\mathcal{D} \sim p_{data}^N$:

$$\mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}_R^*(\mathcal{D})] = (\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top\Phi\boldsymbol{\theta}_0$$

  $\Rightarrow$ Ridge regression returns a **biased estimator**

- Compute the variance of $\boldsymbol{\theta}_R^*$ for $\mathcal{D} \sim p_{data}^N$:

$$\begin{aligned}\mathbb{V}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}_R^*(\mathcal{D})] &= \mathbb{V}_{\mathcal{D}\sim p_{data}^N}[(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top(\Phi\boldsymbol{\theta}_0 + \boldsymbol{\epsilon})] \\ &= \mathbb{V}_{\mathcal{D}\sim p_{data}^N}[(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top\boldsymbol{\epsilon}]\end{aligned}$$

Ridge regression returns a biased estimator

The ridge regression estimator: $\boldsymbol{\theta}_R^* = (\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top(\Phi\boldsymbol{\theta}_0 + \boldsymbol{\epsilon})$

- Compute the mean of $\boldsymbol{\theta}_R^*$ for $\mathcal{D} \sim p_{data}^N$:

$$\mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}_R^*(\mathcal{D})] = (\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top\Phi\boldsymbol{\theta}_0$$

  $\Rightarrow$ Ridge regression returns a **biased estimator**

- Compute the variance of $\boldsymbol{\theta}_R^*$ for $\mathcal{D} \sim p_{data}^N$:

$$\begin{aligned}\mathbb{V}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}_R^*(\mathcal{D})] &= \mathbb{V}_{\mathcal{D}\sim p_{data}^N}[(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top(\Phi\boldsymbol{\theta}_0 + \boldsymbol{\epsilon})] \\ &= \mathbb{V}_{\mathcal{D}\sim p_{data}^N}[(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top\boldsymbol{\epsilon}] \\ &= \sigma^2(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top\Phi(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\end{aligned}$$

Ridge regression returns a biased estimator

Bias of ridge regression estimator ($\lambda > 0$):

$$\begin{aligned}\mathbf{b}(\lambda) := \mathbb{E}_{\mathcal{D}\sim p_{data}^N}[\boldsymbol{\theta}_R^*(\mathcal{D})] - \boldsymbol{\theta}_0 &= (\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top\Phi\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0 \\ &= -\sigma^2\lambda(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\boldsymbol{\theta}_0\end{aligned}$$

Bias of linear regression estimator ($\lambda = 0$):

$$\mathbf{b}(0) = \mathbf{0}$$

Ridge regression returns a biased estimator

Bias of ridge regression estimator ($\lambda > 0$):

$$\mathbf{b}(\lambda) := \mathbb{E}_{\mathcal{D} \sim p_{data}^N}[\boldsymbol{\theta}_R^*(\mathcal{D})] - \boldsymbol{\theta}_0 = (\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top\Phi\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0$$
$$= -\sigma^2\lambda(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\boldsymbol{\theta}_0$$

Bias of linear regression estimator ($\lambda = 0$):

$$\mathbf{b}(0) = \mathbf{0}$$

Variance of ridge regression estimator ($\lambda > 0$):

$$\mathbf{V}(\lambda) := \sigma^2(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}\Phi^\top\Phi(\sigma^2\lambda\mathbf{I} + \Phi^\top\Phi)^{-1}$$

Variance of linear regression estimator ($\lambda = 0$):

$$\mathbf{V}(0) = \sigma^2(\Phi^\top\Phi)^{-1}$$

Ridge regression can perform better in prediction

Expected prediction error of ridge regression ($\lambda > 0$):

$$error_{pred}(\boldsymbol{\theta}_R^*) = \mathbb{E}_{\boldsymbol{x}_{test}}[\boldsymbol{\phi}(\boldsymbol{x}_{test})^\top Error(\boldsymbol{\theta}_R^*)\boldsymbol{\phi}(\boldsymbol{x}_{test})] + \sigma^2$$

$$Error(\boldsymbol{\theta}_R^*) = \mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top + \mathbf{V}(\lambda)$$

Ridge regression can perform better in prediction

Expected prediction error of ridge regression ($\lambda > 0$):

$$error_{pred}(\boldsymbol{\theta}_R^*) = \mathbb{E}_{\boldsymbol{x}_{test}}[\boldsymbol{\phi}(\boldsymbol{x}_{test})^\top Error(\boldsymbol{\theta}_R^*)\boldsymbol{\phi}(\boldsymbol{x}_{test})] + \sigma^2$$
$$Error(\boldsymbol{\theta}_R^*) = \mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top + \mathbf{V}(\lambda)$$

Expected prediction error of linear regression ($\lambda = 0$):

$$error_{pred}(\boldsymbol{\theta}_L^*) = \mathbb{E}_{\boldsymbol{x}_{test}}[\boldsymbol{\phi}(\boldsymbol{x}_{test})^\top Error(\boldsymbol{\theta}_L^*)\boldsymbol{\phi}(\boldsymbol{x}_{test})] + \sigma^2$$
$$Error(\boldsymbol{\theta}_L^*) = \mathbf{b}(0)\mathbf{b}(0)^\top + \mathbf{V}(0) = \mathbf{V}(0)$$

Ridge regression can perform better in prediction

Expected prediction error of ridge regression ($\lambda > 0$):

$$error_{pred}(\boldsymbol{\theta}_R^*) = \mathbb{E}_{\boldsymbol{x}_{test}}[\boldsymbol{\phi}(\boldsymbol{x}_{test})^\top Error(\boldsymbol{\theta}_R^*)\boldsymbol{\phi}(\boldsymbol{x}_{test})] + \sigma^2$$
$$Error(\boldsymbol{\theta}_R^*) = \mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top + \mathbf{V}(\lambda)$$

Expected prediction error of linear regression ($\lambda = 0$):

$$error_{pred}(\boldsymbol{\theta}_L^*) = \mathbb{E}_{\boldsymbol{x}_{test}}[\boldsymbol{\phi}(\boldsymbol{x}_{test})^\top Error(\boldsymbol{\theta}_L^*)\boldsymbol{\phi}(\boldsymbol{x}_{test})] + \sigma^2$$
$$Error(\boldsymbol{\theta}_L^*) = \mathbf{b}(0)\mathbf{b}(0)^\top + \mathbf{V}(0) = \mathbf{V}(0)$$

This means if there exists some $\lambda > 0$ such that:

$$\mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top + \mathbf{V}(\lambda) \preceq \mathbf{V}(0) \quad \Rightarrow \quad error_{pred}(\boldsymbol{\theta}_R^*) \leqslant error_{pred}(\boldsymbol{\theta}_L^*)$$

Ridge regression can perform better in prediction

Derivations available in the extra lecture notes:
We can indeed choose e.g. $0 \leqslant \lambda \leqslant \frac{2}{||\boldsymbol{\theta}_0||_2^2}$ such that:

$$\mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top + \mathbf{V}(\lambda) \preceq \mathbf{V}(0) \quad \Rightarrow \quad error_{pred}(\boldsymbol{\theta}_R^*) \leqslant error_{pred}(\boldsymbol{\theta}_L^*)$$

Ridge regression can perform better in prediction

Derivations available in the extra lecture notes:
We can indeed choose e.g. $0 \leqslant \lambda \leqslant \frac{2}{||\boldsymbol{\theta}_0||_2^2}$ such that:

$$\mathbf{b}(\lambda)\mathbf{b}(\lambda)^\top + \mathbf{V}(\lambda) \preceq \mathbf{V}(0) \quad \Rightarrow \quad error_{pred}(\boldsymbol{\theta}_R^*) \leqslant error_{pred}(\boldsymbol{\theta}_L^*)$$

For $\lambda > 0$, we can also show that

$$\mathbf{V}(\lambda) - \mathbf{V}(0) \preceq 0$$

$\Rightarrow$ The smaller prediction error of $\boldsymbol{\theta}_R^*$ comes from
having **smaller variance**!
$\Rightarrow$ λ needs to be chosen carefully so that bias is not too large

Bias-variance trade-off

Ridge regression can return estimator of $\boldsymbol{\theta}$ with **smaller variance**.

In such case the (expected) test error can be reduced.

Bias-variance trade-off

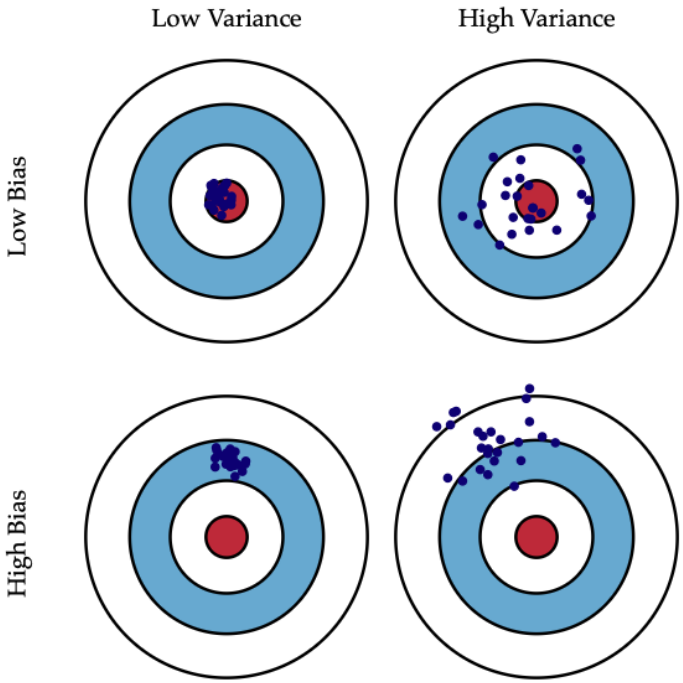
Ridge regression can return estimator of $\boldsymbol{\theta}$ with **smaller variance**.

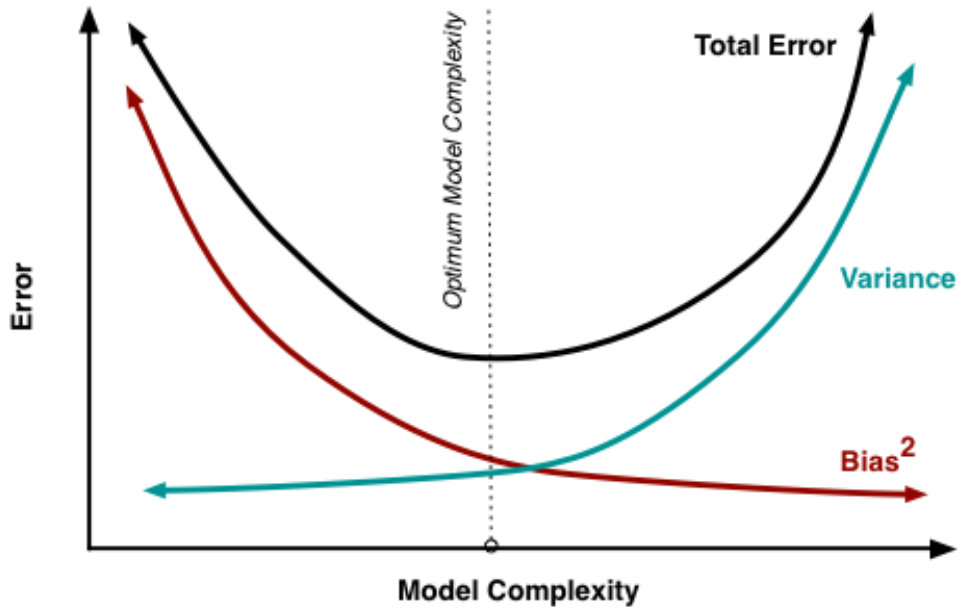
In such case the (expected) test error can be reduced.

- $\boldsymbol{\theta}_R^*$ is a biased estimator of $\boldsymbol{\theta}_0$ when $\lambda > 0$
- There exists λ such that $\mathbf{V}(\lambda) \preceq \mathbf{V}(0)$
- ... and it leads to $error_{pred}(\boldsymbol{\theta}_R^*) \leqslant error_{pred}(\boldsymbol{\theta}_L^*)$

Bias-variance trade-off

Visualising Bias-variance trade-off:





Figures from `http://scott.fortmann-roe.com/docs/BiasVariance.html`