Bias-Variance Tradeoff

Yingzhen Li

Department of Computing
Imperial College London

@liyzhen2
yingzhen.li@imperial.ac.uk

Nov 18, 2022

Regression with non-linear features

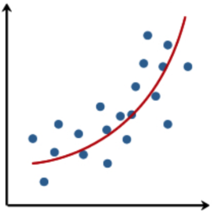
For **non-linear regression**:

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

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

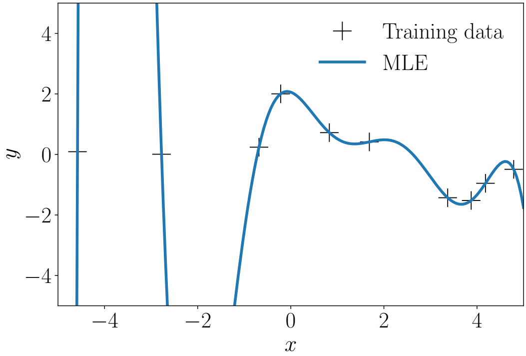
$$y = f(\boldsymbol{x}, \boldsymbol{\theta}) + \epsilon, \ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

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



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

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)
- **Regularise the model** (this lecture)
  - There's a bias-variance tradeoff here!

Regression with non-linear features

Fitting regression model with a **regulariser**:

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

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

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

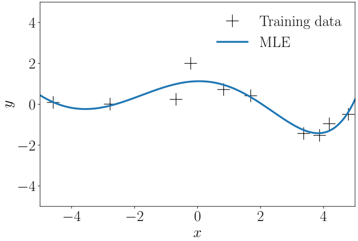
- Optimal solution for $\boldsymbol{\theta}$:

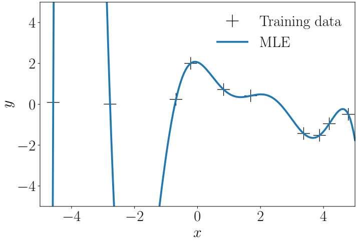
$$\boldsymbol{\theta}_R^* = (\sigma^2 \lambda \mathbf{I} + \Phi^\top \Phi)^{-1} \Phi^\top \mathbf{y}$$

Intuition behind the regulariser

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.
⇒ How to choose a model?

Intuition behind the regulariser

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 the regulariser

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

Intuition behind the regulariser

Potential questions on using regularisers:

- Do we obtain the ground truth parameters?
- Why regularised models can sometimes better fit the data (in terms of test error)?

To answer these: study Bias-variance tradeoff

Bias-variance tradeoff

The general concept of Bias-variance tradeoff:

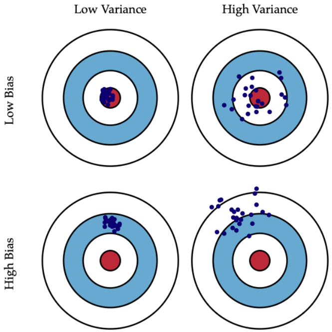
- Suppose there is an unknown quantity x_0 that we like to estimate;
- Assume we have a **stochastic estimator** X for x_0;
- Calculating the expected ℓ_2 error:

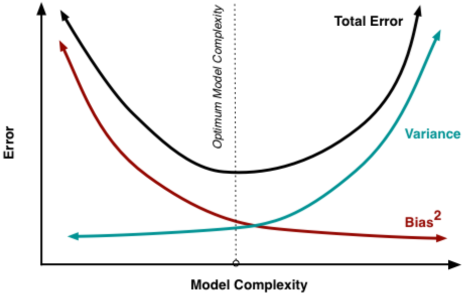
$$\mathbb{E}[||X - x_0||_2^2] = \underbrace{||\mathbb{E}[X] - x_0||_2^2}_{bias^2} + \underbrace{\mathrm{tr}[\mathbb{V}[X]]}_{variance}$$

  - **Unbiased** estimator: $bias = 0 \quad \Rightarrow \quad \mathbb{E}[X] = x_0$
  - **Low variance** estimator: variance is small

Bias-variance tradeoff

Visualising Bias-variance trade-off:





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

Bias-variance tradeoff in regression

Fact for Ridge regression (linear regression + ℓ_2 regulariser):
Ridge regression returns estimator of $\boldsymbol{\theta}$ which

- is **biased** (when $\lambda > 0$, unbiased only when $\lambda = 0$)
- has **smaller variance** than the MLE solution

With good choices of $\lambda > 0$, **the (expected) test error can be reduced**.

Bias-variance tradeoff in regression

How bias-variance tradeoff is relevant to overfitting:
Assuming **no model error**: ground truth parameter $\boldsymbol{\theta}_0$,

$$y = \phi(\boldsymbol{x})^\top \boldsymbol{\theta}_0 + \epsilon,\ \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Expected prediction error for $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\mathcal{D})$ over $\mathcal{D} \sim \pi^N$:

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

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

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

Bias-variance tradeoff in regression

How bias-variance tradeoff is relevant to overfitting:
Expected prediction error for $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\mathcal{D})$ over $\mathcal{D} \sim \pi^N$:

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

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

Bias-variance tradeoff in regression

How bias-variance tradeoff is relevant to overfitting:
Expected prediction error for $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\mathcal{D})$ over $\mathcal{D} \sim \pi^N$:

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

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

If we have two estimators $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ based on $\mathcal{D} \sim \pi^N$:

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

- 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:

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

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

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

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 $\mathcal{D} \sim \pi^N$:

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

$\Rightarrow$ 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 $\mathcal{D} \sim \pi^N$:

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

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

- Compute the covariance matrix of $\boldsymbol{\theta}_R^*$ for $\mathcal{D} \sim \pi^N$:

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

Ridge regression returns a biased estimator

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

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

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

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

Ridge regression returns a biased estimator

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

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

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

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

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

$$\mathbf{V}(\lambda) := \mathbb{E}_{\mathbf{X}_{\text{train}}}[\sigma^2(\sigma^2\lambda\mathbf{I} + \boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\Phi}(\sigma^2\lambda\mathbf{I} + \boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}]$$

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

$$\mathbf{V}(0) = \mathbb{E}_{\mathbf{X}_{\text{train}}}[\sigma^2(\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}]$$

Ridge regression can perform better in prediction

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

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

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

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

Ridge regression can perform better in prediction

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

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

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

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

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 exercises in the exercise sheet:

- For $\lambda > 0$, we can show reduced variance:

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

- We can choose e.g. $0 \leqslant \lambda \leqslant \frac{2}{||\boldsymbol{\theta}_0||_2^2}$ which leads to:

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

$\Rightarrow$ The smaller prediction error of $\boldsymbol{\theta}_R^*$ comes from having **smaller variance** in parameter estimate!

$\Rightarrow$ λ needs to be chosen carefully so that **the bias is not too large**

Bias-variance tradeoff in regression: Summary

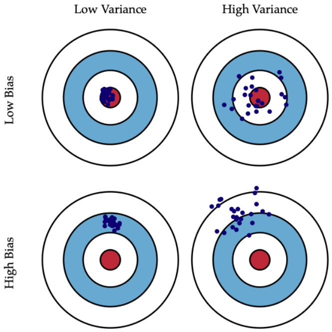
Ridge regression can return estimator of $\boldsymbol{\theta}$ with **smaller variance**.

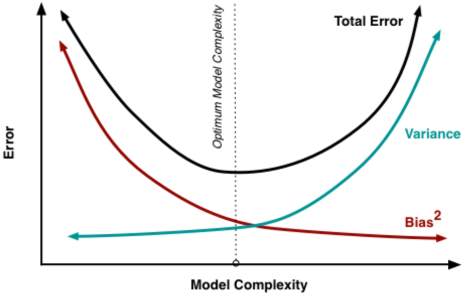
In such case the (expected) test error can be reduced.

- $\boldsymbol{\theta}_R^*$ is a biased estimator of $\boldsymbol{\theta}_0$ when $\lambda > 0$
- There exists λ such that
  - Variance is smaller: $\mathbf{V}(\lambda) \preceq \mathbf{V}(0)$
  - Bias is not too large
- ... and it leads to $error_{pred}(\boldsymbol{\theta}_R^*) \leqslant error_{pred}(\boldsymbol{\theta}_L^*)$

Bias-variance tradeoff

Visualising Bias-variance trade-off:





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