Erasmus School of Economics

# Machine Learning

FEM31002

## Regularization

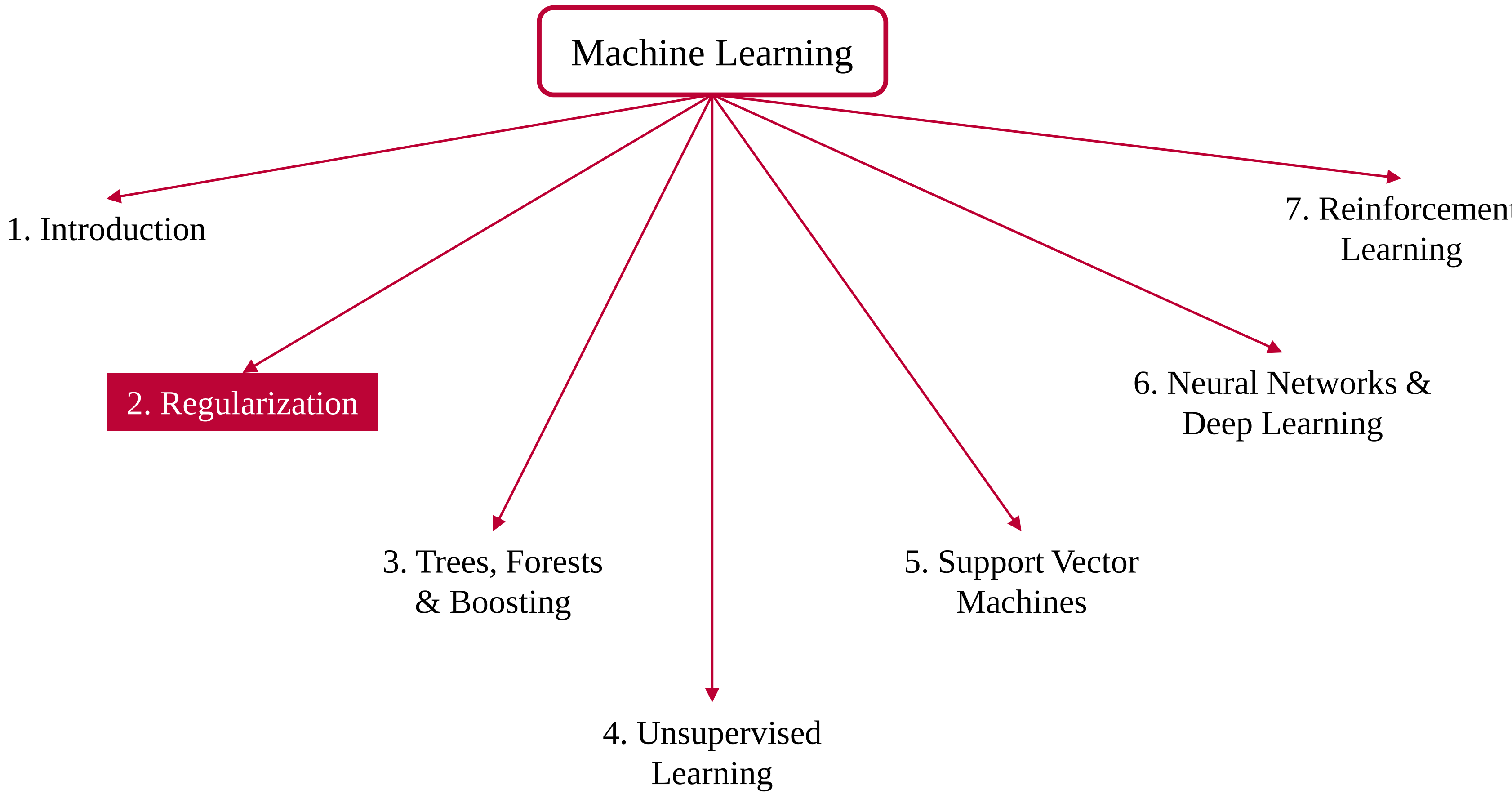
Part 1

**Ilker Birbil**

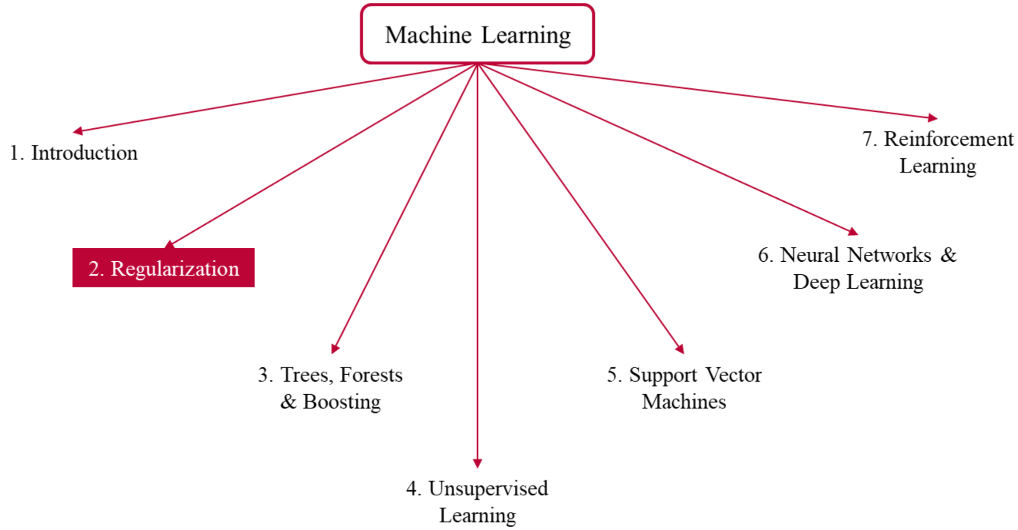
birbil@ese.eur.nl

ERASMUS UNIVERSITEIT ROTTERDAM

# Outline


Machine Learning
1. Introduction
2. Regularization
3. Trees, Forests & Boosting
4. Unsupervised Learning
5. Support Vector Machines
6. Neural Networks & Deep Learning
7. Reinforcement Learning

# Outline




- Shrinkage: Ridge Regression and Lasso
- Elastic Net
- Least Angle Regression
- Integer Programming Models

# Recall: Linear Regression

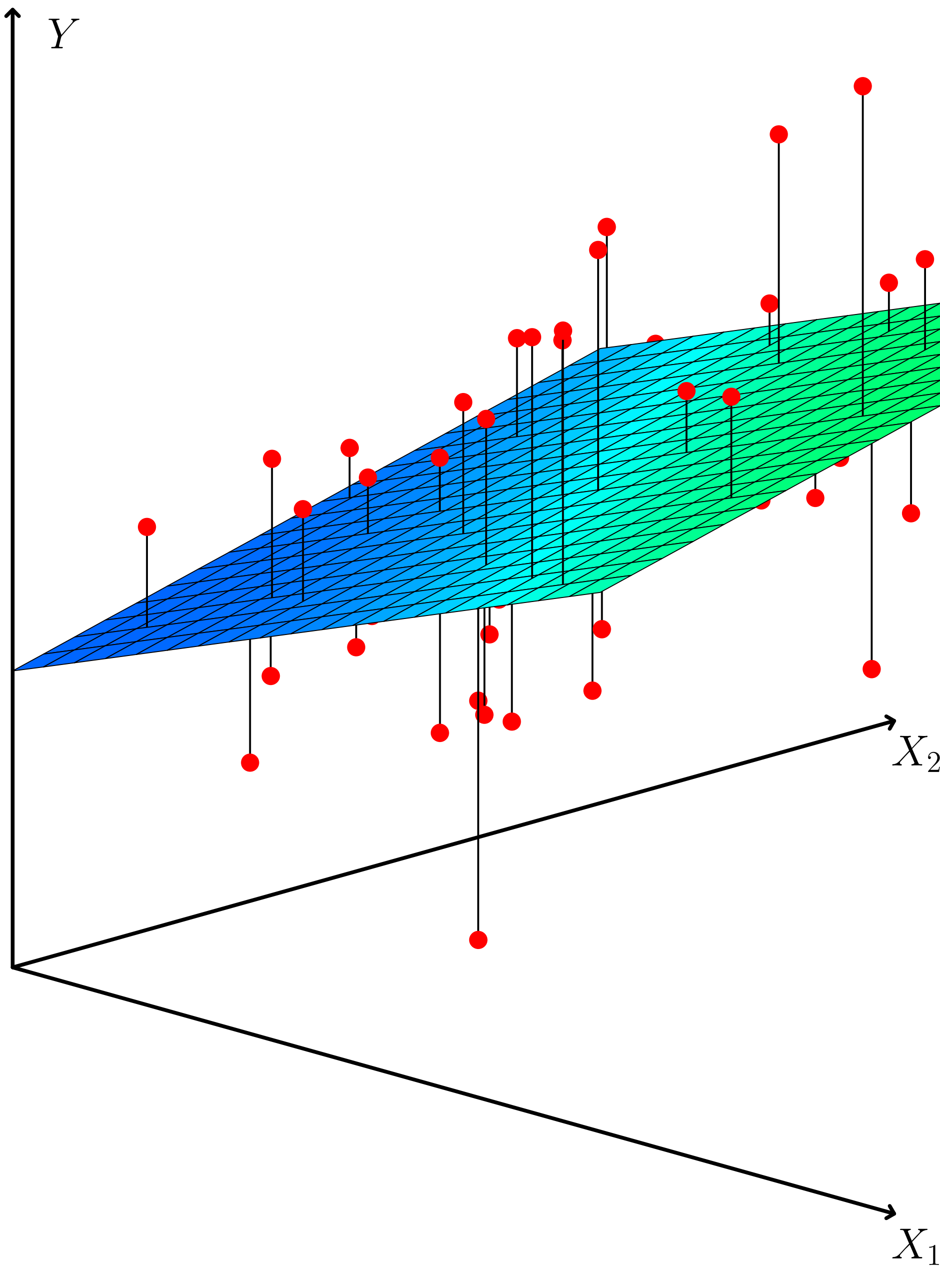
$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

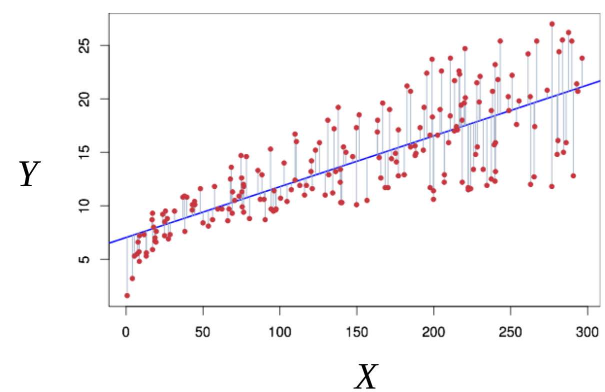
training data

$$\{(x_i, y_i) : 1, \ldots, n\}$$

$$y_i \approx \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}}_{\hat{y}_i}, \quad i = 1, \ldots, n$$

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$


Y
$X_2$
$X_1$


Y
X
5 10 15 20 25
0 50 100 150 200 250 300

## Least Squares Method (LSM)

$$\min_{\beta_0, \beta_1, \ldots, \beta_p} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

↓

Convex Optimization

↓

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$$

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^\intercal \\ 1 & x_2^\intercal \\ \vdots & \vdots \\ 1 & x_n^\intercal \end{bmatrix}_{n\times(p+1)}$$

$$\mathbf{y}^\intercal = (y_1, \ldots, y_n)$$

$$\boldsymbol{\beta}^\intercal = (\beta_0, \beta_1, \ldots, \beta_p)$$

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

(assuming full rank)

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y}$$

$$\hat{y}_0 = (1\ x_0^\intercal)\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$$

Features (Variables)

$$X_1, \cancel{X_2}, X_3, \ldots, \cancel{X_{p-1}}, X_p$$

$$Y \approx \beta_0 + \beta_1 X_1 + \underbrace{\cancel{\beta_2}}_{0} X_2 + \cdots + \underbrace{\cancel{\beta_{p-1}}}_{0} X_{p-1} + \beta_p X_p$$

**Prediction Accuracy**

**Model Interpretability**

$n >> p \longrightarrow$ LSM (unique solution) $\longrightarrow$ Low prediction variance

$p > n \longrightarrow$ LSM (multiple solutions) $\longrightarrow$ Prediction variance $\uparrow \infty$

# Shrinkage Methods

$$Y \approx \beta_0 + \beta_1 X_1 + \underset{0}{\cancel{\beta_2}} X_2 + \cdots + \underset{0}{\cancel{\beta_{p-1}}} X_{p-1} + \beta_p X_p$$

**Ridge Regression**

$$\min_{\beta_0, \beta_1, \ldots, \beta_p} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

λ : tuning (hyper)parameter

**Lasso**

$$\min_{\beta_0, \beta_1, \ldots, \beta_p} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

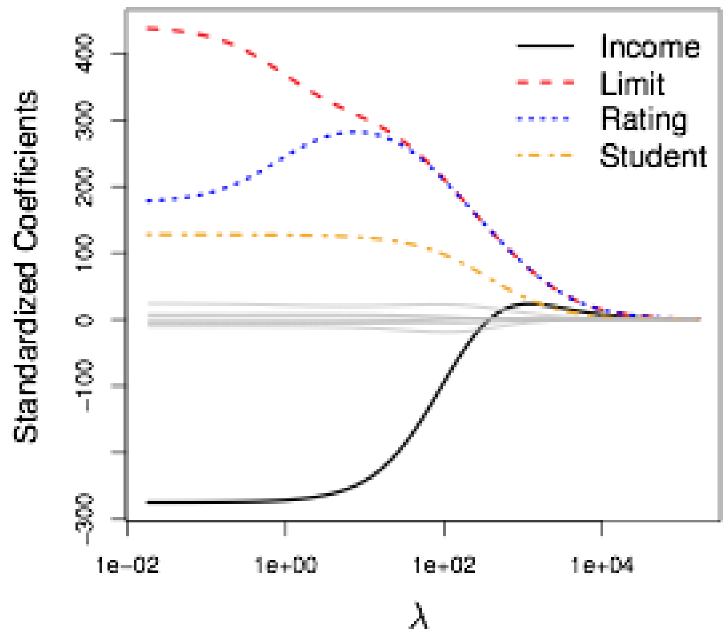
# Ridge Regression

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

$$\lambda \uparrow \infty \quad \longrightarrow \quad \beta_j \downarrow 0, \quad j = 1, \cdots, p$$

**Standardization**

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$$



$$\hat{\beta}_{\mathrm{R}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\intercal \boldsymbol{\beta}$$

$$\hat{\beta}_{\mathrm{LS}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y}$$

$$\boxed{\hat{\beta}_{\mathrm{R}} = (\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\intercal \mathbf{y}}$$

$\beta_0$ dropped (working with centered input)

$$\mathbf{X} = \begin{bmatrix} x_1^\intercal \\ x_2^\intercal \\ \vdots \\ x_n^\intercal \end{bmatrix}_{n \times p}$$

SVD approach

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\intercal$$

$$\mathbf{U} = [\mathbf{u}_1 \ \ldots \ \mathbf{u}_p]$$

$$\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_p)$$

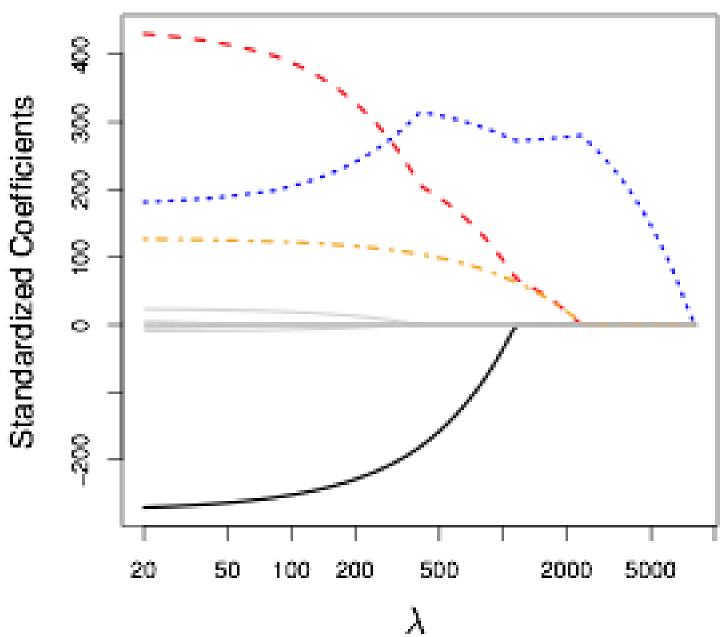
$$\begin{aligned} \mathbf{X}\hat{\beta}_{\mathrm{LS}} &= \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y} \\ &= \mathbf{U}\mathbf{U}^\intercal \mathbf{y} \\ &= \sum_{j=1}^{p} \mathbf{u}_j \mathbf{u}_j^\intercal \mathbf{y} \end{aligned}$$

$$\begin{aligned} \mathbf{X}\hat{\beta}_{\mathrm{R}} &= \mathbf{X}(\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\intercal \mathbf{y} \\ &= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda \mathbf{I})^{-1} \mathbf{D}\mathbf{U}^\intercal \mathbf{y} \\ &= \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda} \mathbf{u}_j^\intercal \mathbf{y} \end{aligned}$$

# Lasso

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

$$\lambda \uparrow \infty \quad \longrightarrow \quad \beta_j \downarrow 0, \quad j = 1, \cdots, p$$



$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{y}$$

$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = (\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^\intercal \mathbf{y}$$

$$\hat{\boldsymbol{\beta}}_L = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1$$

$\beta_0$ dropped (working with centered input)

$$\mathbf{X} = \begin{bmatrix} x_1^\intercal \\ x_2^\intercal \\ \vdots \\ x_n^\intercal \end{bmatrix}_{n \times p}$$

↓

Convex Optimization

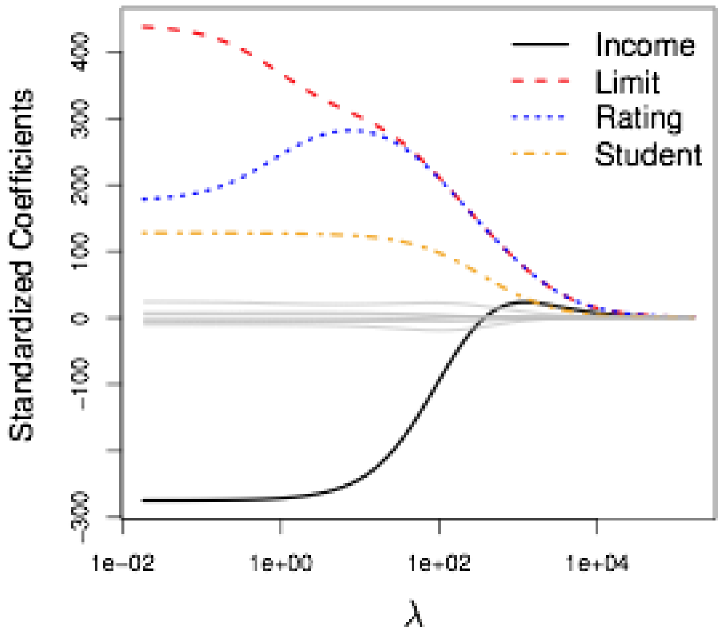
↓

No analytical solution

Very fast solution algorithms

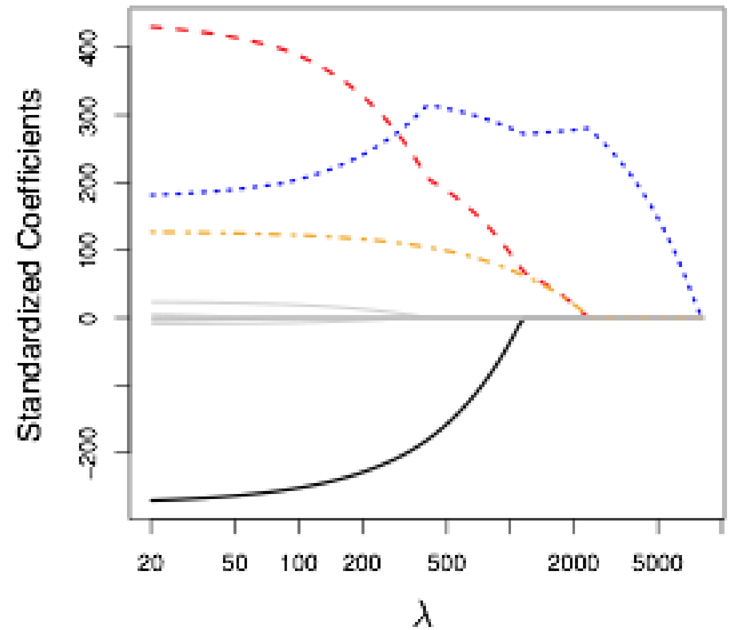
## Ridge Regression

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$




## Lasso

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$




## Example

$$\begin{bmatrix}\beta_1\\ \beta_2\\ \beta_3\end{bmatrix}=\begin{bmatrix}0.01\\ -0.01\\ 0.01\end{bmatrix}$$

$$\sum_{j=1}^{3}\beta_j^2 = 3(0.01)^2 = 0.0003$$

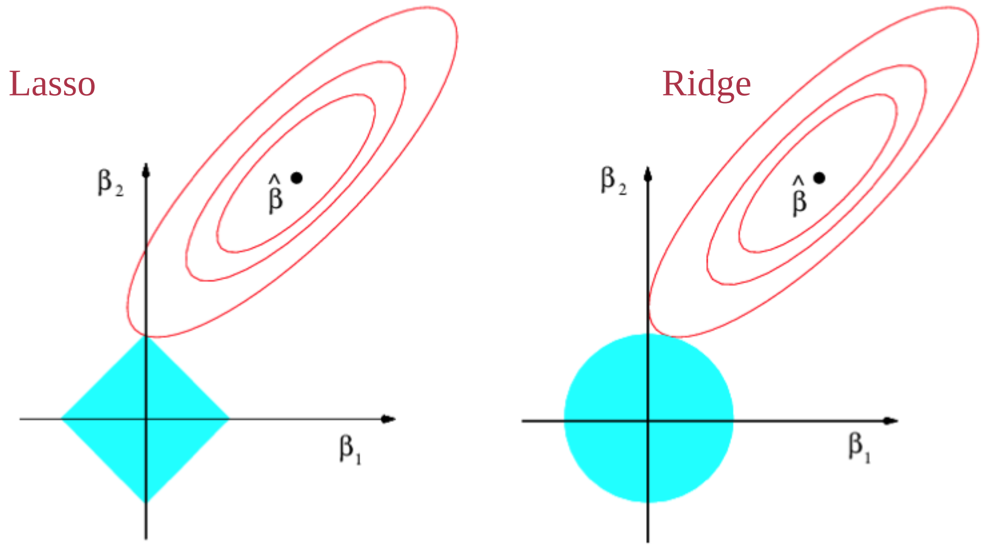
$$\sum_{j=1}^{3}|\beta_j| = 3(0.01) = 0.03$$

## Ridge

$$\min_{\boldsymbol{\beta}} \{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \ : \ \|\boldsymbol{\beta}\|_2^2 \leq \Delta\}$$

## Lasso

$$\min_{\boldsymbol{\beta}} \{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \ : \ \|\boldsymbol{\beta}\|_1 \leq \Delta\}$$

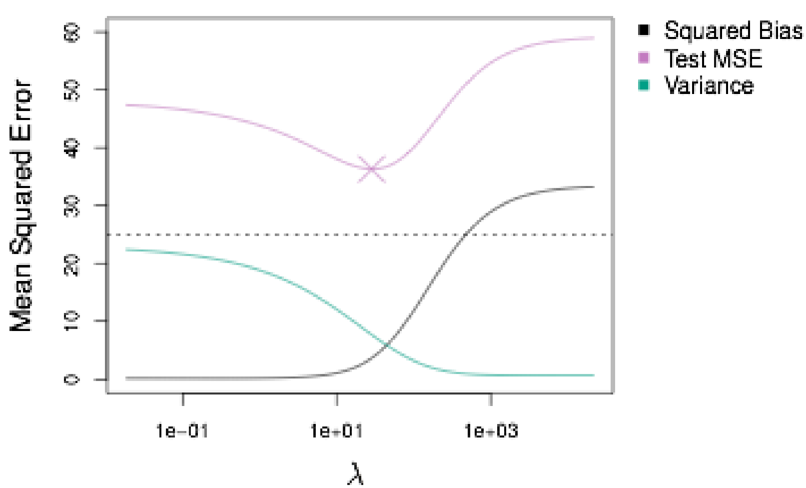



Least Squares Solution: $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$

# Ridge Regression – Lasso

- Sparse solution: Lasso
- Many parameters that are close to zero: Ridge regression
- Interpretability: Lasso
- Selecting the tuning parameter (λ): Grid Search **&** Cross-Validation

λ ↑ Model Complexity ↓ Variance ↓ Bias ↑



# Probabilistic Point of View

$y_i \sim N(\hat{y}_i, \sigma^2), \; i = 1, \ldots, n$ and i.i.d

$$\boxed{\mathbb{P}(\mathbf{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} P(y_i|\boldsymbol{\beta})}$$

$$\mathbb{P}(\boldsymbol{\beta}|\mathbf{y}) = \frac{\mathbb{P}(\mathbf{y}|\boldsymbol{\beta})\mathbb{P}(\boldsymbol{\beta})}{\mathbb{P}(\mathbf{y})}$$

$$\mathbb{P}(\boldsymbol{\beta}|\mathbf{y}) \propto \mathbb{P}(\mathbf{y}|\boldsymbol{\beta})\mathbb{P}(\boldsymbol{\beta})$$

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{MAP}} &= \arg\max_{\boldsymbol{\beta}} \; \mathbb{P}(\mathbf{y}|\boldsymbol{\beta})\mathbb{P}(\boldsymbol{\beta}) \\ &= \arg\max_{\boldsymbol{\beta}} \; \log \mathbb{P}(\mathbf{y}|\boldsymbol{\beta}) + \log \mathbb{P}(\boldsymbol{\beta}) \\ &= \arg\max_{\boldsymbol{\beta}} \; \sum_{i=1}^{n} \log \mathbb{P}(y_i|\boldsymbol{\beta}) + \log \mathbb{P}(\boldsymbol{\beta})\end{aligned}$$

constant prior, $\mathbb{P}(\boldsymbol{\beta})$

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\bullet} &= \arg\max_{\boldsymbol{\beta}} \; \sum_{i=1}^{n} \log \mathbb{P}(y_i|\boldsymbol{\beta}) + \log \mathbb{P}(\boldsymbol{\beta}) \\
&= \arg\max_{\boldsymbol{\beta}} \; \sum_{i=1}^{n} \log \mathbb{P}(y_i|\boldsymbol{\beta}) \\
&= \arg\max_{\boldsymbol{\beta}} \; \sum_{i=1}^{n} \log \left( \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2}{2\sigma^2}} \right) \\
&= \arg\max_{\boldsymbol{\beta}} \; -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 \\
&= \arg\min_{\boldsymbol{\beta}} \; \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2
\end{aligned}
$$

$$
\implies \hat{\boldsymbol{\beta}}_{\bullet} = \hat{\boldsymbol{\beta}}_{\mathrm{LS}}
$$

normally distributed prior, $\beta_j \sim N(0, \psi^2)$, $j = 1, \ldots, p$ and i.i.d

$$\mathbb{P}(\boldsymbol{\beta}) = \prod_{j=1}^{p} P(\beta_j)$$

$\beta_0$ dropped (working with centered input)

$$\mathbf{X} = \begin{bmatrix} x_1^\intercal \\ x_2^\intercal \\ \vdots \\ x_n^\intercal \end{bmatrix}_{n \times p}$$

$$\begin{aligned}
\hat{\beta}_\bullet &= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log \mathbb{P}(y_i|\boldsymbol{\beta}) + \log \mathbb{P}(\boldsymbol{\beta}) \\
&= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log \left( \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2}{2\sigma^2}} \right) + \sum_{j=1}^{p} \log \left( \frac{1}{\psi\sqrt{2\pi}} e^{-\frac{\beta_j^2}{2\psi^2}} \right) \\
&= \arg\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \frac{\sigma^2}{\psi^2} \sum_{j=1}^{p} \beta_j^2 \right) \\
&= \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{\sigma^2}{\psi^2} \boldsymbol{\beta}^\intercal \boldsymbol{\beta}
\end{aligned}$$

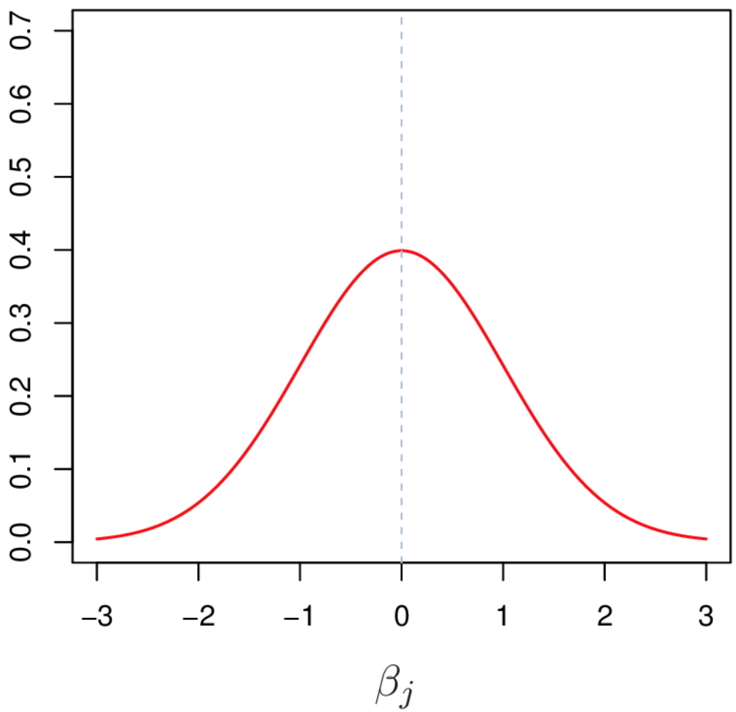
$$\implies \hat{\boldsymbol{\beta}}_\bullet = \hat{\boldsymbol{\beta}}_{\mathrm{R}} \text{ with } \lambda = \frac{\sigma^2}{\psi^2}$$

Laplace distributed prior, $\beta_j \sim \text{Laplace}(0, \phi)$, $j = 1, \ldots, p$ and i.i.d

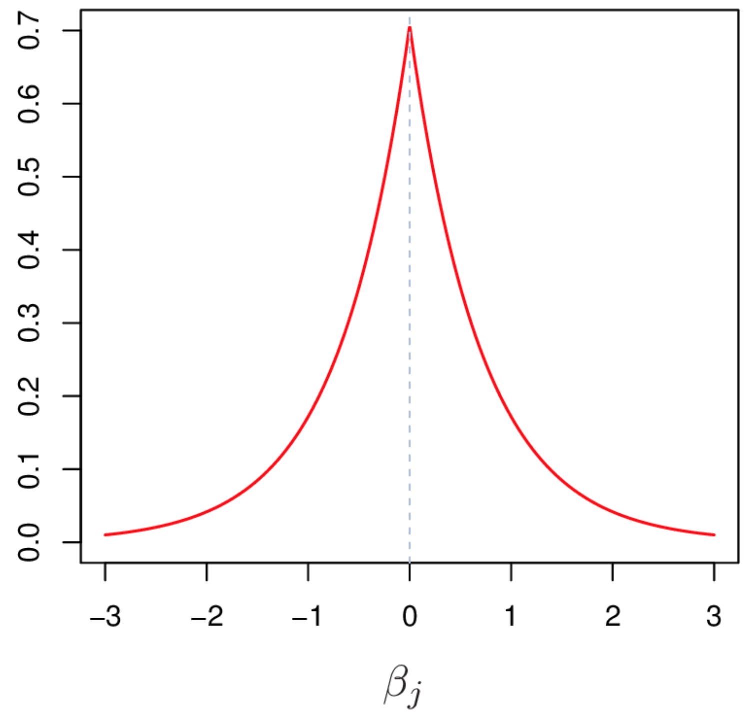
$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\bullet} &= \arg\max_{\boldsymbol{\beta}} \ \sum_{i=1}^{n} \log \mathbb{P}(y_i | \boldsymbol{\beta}) + \log \mathbb{P}(\boldsymbol{\beta}) \\
&= \arg\max_{\boldsymbol{\beta}} \ \sum_{i=1}^{n} \log \left( \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2}{2\sigma^2}} \right) + \sum_{j=1}^{p} \log \left( \frac{1}{2\phi} e^{-\frac{|\beta_j|}{2\phi}} \right) \\
&= \arg\min_{\boldsymbol{\beta}} \ \frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \frac{\sigma^2}{\phi} \sum_{j=1}^{p} |\beta_j| \right) \\
&= \arg\min_{\boldsymbol{\beta}} \ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\intercal}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{\sigma^2}{\phi} \|\boldsymbol{\beta}\|_1
\end{aligned}
$$

$$
\implies \hat{\boldsymbol{\beta}}_{\bullet} = \hat{\boldsymbol{\beta}}_{\mathrm{L}} \text{ with } \lambda = \frac{\sigma^2}{\phi}
$$

$\beta_j \sim N(0, \psi^2)$




$\beta_j \sim \text{Laplace}(0, \phi)$




$$\hat{\boldsymbol{\beta}}_{\mathrm{R}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\intercal \boldsymbol{\beta}$$

$$\hat{\boldsymbol{\beta}}_{L} = \arg\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1$$