# The Mathematics of Machine Learning 2025

Lecture 3

**Tim van Erven**

University of Amsterdam

# Outline

## Linear Regression

1. Least Squares

## Model Selection for Regression

2. Intro
3. Best-subset Selection
4. Ridge Regression
5. Lasso

## Discussion

6. Getting Rid of the Intercept
7. Comparison: Ridge vs Lasso
8. Formula for Ridge / Least Squares
9. Ridge and Lasso as Shrinkage Methods
10. Equivalence of the Two Formulations

# Least Squares for Linear Regression

$$\begin{pmatrix} Y_1 \\ X_1 \end{pmatrix}, \begin{pmatrix} Y_2 \\ X_2 \end{pmatrix}, \ldots, \begin{pmatrix} Y_N \\ X_N \end{pmatrix}$$

**Linear Regression:**

- $X_i \in \mathbb{R}^p$, $Y_i \in \mathbb{R}$
- Model $\mathcal{F}$ can be represented as linear functions:

$$f_\beta(X_i) = \beta_0 + \sum_{j=1}^{p} X_{i,j}\beta_j \qquad \beta = (\beta_0, \beta_1, \ldots, \beta_p)$$

**Least Squares/ERM:** Choose $\hat{\beta}$ to minimize *residual sum of squares*

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (Y_i - f_\beta(X_i))^2 = \sum_{i=1}^{N} (Y_i - X_i^\intercal \beta)^2$$

# Interpretation: Least Squares as Projection

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{1,1} & \dots & X_{1,p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{N,1} & \dots & X_{N,p} \end{pmatrix} = \begin{pmatrix} 1 & X_1^\intercal \\ \vdots & \vdots \\ 1 & X_N^\intercal \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

**Least Squares:** Choose $\hat{\beta}$ to minimize

$$\text{RSS}(\beta) = \sum_{i=1}^{N} (Y_i - X_i^\intercal \beta)^2 = \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2$$

# Interpretation: Least Squares as Projection

$$\boldsymbol{X} = \begin{pmatrix} \tilde{X}_0 & \tilde{X}_1 & & \tilde{X}_p \\ 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{N,1} & \cdots & X_{N,p} \end{pmatrix} = \begin{pmatrix} 1 & X_1^\intercal \\ \vdots & \vdots \\ 1 & X_N^\intercal \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

**Least Squares:** Choose $\hat{\beta}$ to minimize

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (Y_i - X_i^\intercal \beta)^2 = \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2$$

$\boldsymbol{X}\hat{\beta}$ is the **projection** of $\boldsymbol{y}$ onto
the span of the per-feature vectors $\tilde{X}_0, \tilde{X}_1, \ldots, \tilde{X}_p$.

# Interpretation: Least Squares as Projection

$$\boldsymbol{X} = \begin{pmatrix} \tilde{X}_0 & \tilde{X}_1 & & \tilde{X}_p \\ 1 & X_{1,1} & \cdots & X_{1,p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{N,1} & \cdots & X_{N,p} \end{pmatrix} = \begin{pmatrix} 1 & X_1^\intercal \\ \vdots & \vdots \\ 1 & X_N^\intercal \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

**Least Squares:** Choose $\hat{\beta}$ to minimize

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (Y_i - X_i^\intercal \beta)^2 = \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2$$

$\boldsymbol{X}\hat{\beta}$ is the **projection** of $\boldsymbol{y}$ onto
the span of the per-feature vectors $\tilde{X}_0, \tilde{X}_1, \ldots, \tilde{X}_p$.

**Corollary:** The least-squares solution $\hat{\beta}$ is unique if and only if $\tilde{X}_0, \tilde{X}_1, \ldots, \tilde{X}_p$ are **linearly independent**.
(Otherwise there would exist infinitely many $\hat{\beta}'$ such that $\boldsymbol{X}\hat{\beta}' = \boldsymbol{X}\hat{\beta}$.)

# Outline

## Linear Regression

1. Least Squares

## Model Selection for Regression

2. Intro
3. Best-subset Selection
4. Ridge Regression
5. Lasso

## Discussion

6. Getting Rid of the Intercept
7. Comparison: Ridge vs Lasso
8. Formula for Ridge / Least Squares
9. Ridge and Lasso as Shrinkage Methods
10. Equivalence of the Two Formulations

# Model Selection for Regression: Intro

Suppose **many features** $p$. Maybe even $p \gg N$!

Often not satisfied with least squares, because we want:

1. **Better prediction accuracy:**
   - Least squares prone to overfitting when $p$ too large compared to $N$
2. **Better interpretability:**
   - If many features, which ones are most important?
     - E.g. If we have 20 000 genes, want to know which ones are (most) relevant for prediction

# Best-subset Selection

Want to use only $m$ **most important features.**

Select $m \leq p$ features as follows:

1. Run least squares for each of the $\binom{p}{m} \approx \left(\frac{p}{m}\right)^m$ subsets of $m$ features.
2. Choose the subset such that RSS($\hat{\beta}$) is smallest.

# Best-subset Selection

Want to use only $m$ **most important features.**

Select $m \leq p$ features as follows:

1. Run least squares for each of the $\binom{p}{m} \approx \left(\frac{p}{m}\right)^m$ subsets of $m$ features.
2. Choose the subset such that RSS($\hat{\beta}$) is smallest.

Different ways to choose $m$. E.g. use cross-validation.

# Best-subset Selection

Want to use only $m$ **most important features.**

Select $m \leq p$ features as follows:

1. Run least squares for each of the $\binom{p}{m} \approx \left(\frac{p}{m}\right)^m$ subsets of $m$ features.
2. Choose the subset such that RSS($\hat{\beta}$) is smallest.

Different ways to choose $m$. E.g. use cross-validation.

**Issues:**

- **Computation not feasible:** exponentially many subsets in $m$
- **Potential instability:** if features are (nearly) linearly dependent, then multiple subsets can be (nearly) equally good, so which one is chosen depends on minor noise $\rightarrow$ challenge for interpretability

# Best-subset Selection

Want to use only $m$ **most important features.**

Select $m \leq p$ features as follows:

1. Run least squares for each of the $\binom{p}{m} \approx \left(\frac{p}{m}\right)^m$ subsets of $m$ features.
2. Choose the subset such that RSS($\hat{\beta}$) is smallest.

Different ways to choose $m$. E.g. use cross-validation.

**Computable Approximations:**

- **Forward stepwise:** start with **no** features, greedily choose one feature to add until we have $m$ features.
- **Backward stepwise:** start with **all** features, greedily choose one feature to remove until we have $m$ features left.

Much less attractive because of greedy/myopic approach.

# Ridge Regression

**Notation change:** separate intercept $\beta_0 \in \mathbb{R}$ from rest of parameters $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$.

**Definition:** For hyperparameter $\lambda \geq 0$, minimize

$$\text{RSS}(\beta_0, \beta) + \lambda \|\beta\|_2^2 \quad = \quad \sum_{i=1}^{N} (Y_i - X_i^\intercal \beta - \beta_0)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

# Ridge Regression

**Notation change:** separate intercept $\beta_0 \in \mathbb{R}$ from rest of parameters $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$.

Penalize larger $\beta$ but not intercept $\beta_0$

**Definition:** For hyperparameter $\lambda \geq 0$, minimize

$$\mathrm{RSS}(\beta_0, \beta) + \lambda \|\beta\|_2^2 \quad = \quad \sum_{i=1}^{N} (Y_i - X_i^\intercal \beta - \beta_0)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

# Ridge Regression

**Notation change:** separate intercept $\beta_0 \in \mathbb{R}$ from rest of parameters $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$.

Penalize larger $\beta$ but not intercept $\beta_0$

**Definition:** For hyperparameter $\lambda \geq 0$, minimize

$$\mathrm{RSS}(\beta_0, \beta) + \lambda \|\beta\|_2^2 \quad = \quad \sum_{i=1}^{N} (Y_i - X_i^\intercal \beta - \beta_0)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

**Alternative formulation:** For hyperparameter $t \geq 0$,

$$\begin{aligned} &\text{Minimize} && \mathrm{RSS}(\beta_0, \beta) \\ &\text{Subject to} && \|\beta\|_2^2 \leq t \end{aligned}$$

- Will see: two formulations are equivalent!
- Choose $\lambda$ or $t$ by cross-validation, or based on theoretical analysis.

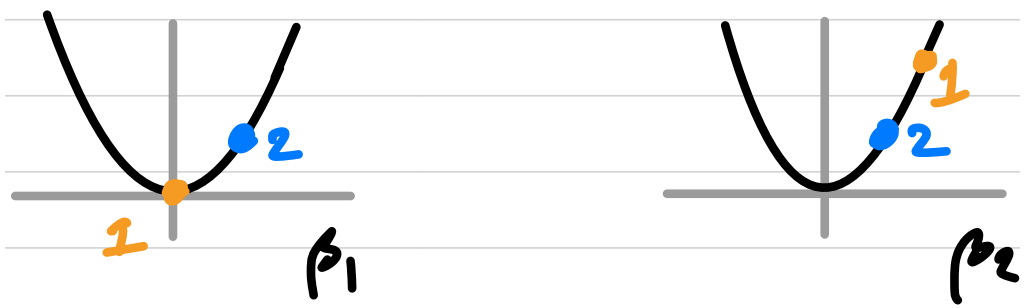
# Ridge Motivation

More stable / less variance than least squares for highly correlated features

**Example:**
Suppose two features are the same: $X_1 = X_2$.
Then RSS$(\beta_0, \beta)$ the same for:

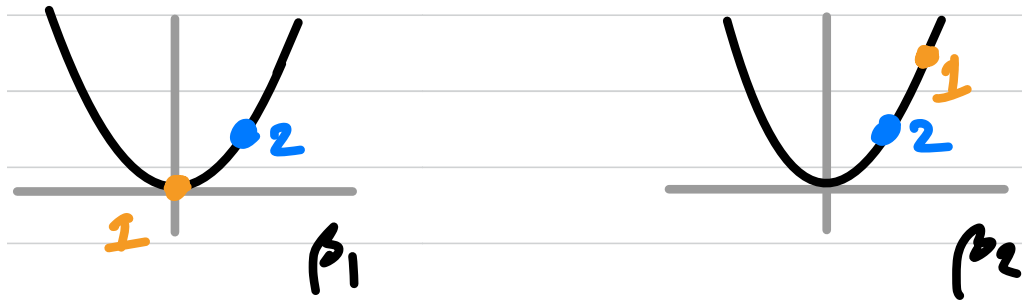
1. $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$
2. $\beta = (\beta_1 + a, \beta_2 - a, \ldots, \beta_p)$ for any $a$.

So least squares does not distinguish between:



# Ridge Motivation

More stable / less variance than least squares for highly correlated features

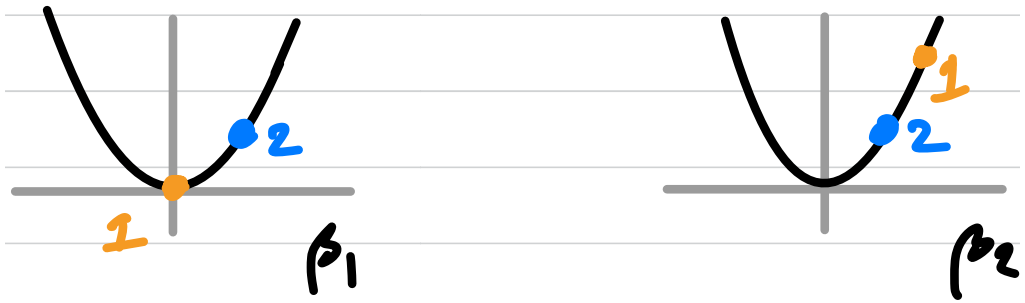



Ridge regression prefers equal weights (situation 2)[1].

[1] Because $0 + z^2 \geq \left(\frac{z}{2}\right)^2 + \left(\frac{z}{2}\right)^2$

# Ridge Motivation

More stable / less variance than least squares for highly correlated features




Ridge regression prefers equal weights (situation 2)[1].

- For highly correlated features it makes the weights approximately equal.
- $p \gg N$ allowed!
- Ridge solution is always unique for $\lambda > 0$, because

$$\text{RSS}(\beta_0, \beta) + \lambda \|\beta\|_2^2 \qquad \text{is strictly convex.}$$

[1] Because $0 + z^2 \geq \left(\frac{z}{2}\right)^2 + \left(\frac{z}{2}\right)^2$

# Ridge Remarks

**Warning:** Need to **normalize features**!

- If we make features $X_j$ 10 times as small, then least squares makes $\beta_j$ 10 times as large and its predictions do not change.
- This does **not** hold for ridge regression, so need to rescale features to standard range.

# Ridge Remarks

**Warning:** Need to **normalize features**!

- If we make features $X_j$ 10 times as small, then least squares makes $\beta_j$ 10 times as large and its predictions do not change.
- This does **not** hold for ridge regression, so need to rescale features to standard range.

**Computation:**
Closed-form formula (see later)

# Lasso

**Definition:** For hyperparameter $\lambda \geq 0$, minimize

$$\mathrm{RSS}(\beta_0, \beta) + \lambda \|\beta\|_1 \quad = \quad \sum_{i=1}^{N} (Y_i - X_i^{\mathsf{T}} \beta - \beta_0)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

**Notation:** $\|\beta\|_q = \left( \sum_{j=1}^{p} |\beta|^q \right)^{1/q}$ is $\ell_q$-norm of $\beta$ for $q \geq 1$.

# Lasso

**Definition:** For hyperparameter $\lambda \geq 0$, minimize

$$\mathrm{RSS}(\beta_0, \beta) + \lambda \|\beta\|_1 \quad = \quad \sum_{i=1}^{N} (Y_i - X_i^{\mathsf{T}} \beta - \beta_0)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

**Notation:** $\|\beta\|_q = \left( \sum_{j=1}^{p} |\beta|^q \right)^{1/q}$ is $\ell_q$-norm of $\beta$ for $q \geq 1$.

# Lasso

**Definition:** For hyperparameter $\lambda \geq 0$, minimize

$$\mathrm{RSS}(\beta_0, \beta) + \lambda \|\beta\|_1 \quad = \quad \sum_{i=1}^{N} (Y_i - X_i^\mathsf{T} \beta - \beta_0)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

**Notation:** $\|\beta\|_q = \left( \sum_{j=1}^{p} |\beta|^q \right)^{1/q}$ is $\ell_q$-norm of $\beta$ for $q \geq 1$.

**Alternative formulation:** For hyperparameter $t \geq 0$,

$$\begin{array}{ll} \text{Minimize} & \mathrm{RSS}(\beta_0, \beta) \\ \text{Subject to} & \|\beta\|_1 \leq t \end{array}$$

- Will see: two formulations are equivalent!
- Choose $\lambda$ or $t$ by cross-validation; to get a predetermined number of non-zero parameters; or based on theoretical analysis.

# Lasso Motivation

Sets **many parameters $\beta_j$ to** $0$ to
find only the "most important" features

- Good for prediction and interpretation!
- $p \gg N$ allowed
- NB Need to normalize features (for same reasons as for Ridge)

# Lasso Motivation

Sets **many parameters $\beta_j$ to 0** to
find only the "most important" features

- Good for prediction and interpretation!
- $p \gg N$ allowed
- NB Need to normalize features (for same reasons as for Ridge)

**Computation:**

- Exist multiple efficient algorithms
- The LARS algorithm computes solutions for all values of $t$ in one pass

# Outline

## Linear Regression

1. Least Squares

## Model Selection for Regression

2. Intro
3. Best-subset Selection
4. Ridge Regression
5. Lasso

## Discussion

6. Getting Rid of the Intercept
7. Comparison: Ridge vs Lasso
8. Formula for Ridge / Least Squares
9. Ridge and Lasso as Shrinkage Methods
10. Equivalence of the Two Formulations

# Getting Rid of the Intercept

**Centering features:**

$$\tilde{X}_i = X_i - \bar{X}, \qquad \text{where } \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

## Theorem

*For any method of the form*

$$\hat{\beta}_0, \hat{\beta} = \underset{\beta_0,\beta}{\arg\min}\, \mathrm{RSS}(\beta_0, \beta) + \lambda \mathrm{pen}(\beta),$$

*centering the features only changes $\hat{\beta}_0$, but not $\hat{\beta}$. Moreover,* **after** *centering, the intercept is always*

$$\hat{\beta}_0 = \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i.$$

Proof.

Homework. □

# Getting Rid of the Intercept

**Centering features:**

$$\tilde{X}_i = X_i - \bar{X}, \qquad \text{where } \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

## Theorem

*For any method of the form* ($\beta_0$ not penalized)

$$\hat{\beta}_0, \hat{\beta} = \underset{\beta_0,\beta}{\arg\min}\, \mathrm{RSS}(\beta_0, \beta) + \lambda \mathrm{pen}(\beta),$$

*centering the features only changes $\hat{\beta}_0$, but not $\hat{\beta}$. Moreover,* **after** *centering, the intercept is always*

$$\hat{\beta}_0 = \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i.$$

Effect: centering $Y_i$

## Proof.

Homework. □

# Getting Rid of the Intercept

## Theorem

*For any method of the form* $\beta_0$ not penalized

$$\hat{\beta}_0, \hat{\beta} = \underset{\beta_0, \beta}{\arg\min}\, \mathrm{RSS}(\beta_0, \beta) + \lambda \mathrm{pen}(\beta),$$

*centering the features only changes* $\hat{\beta}_0$*, but not* $\hat{\beta}$*. Moreover,* **after** *centering, the intercept is always*

$$\hat{\beta}_0 = \bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i.$$

Effect: centering $Y_i$

**Interpretation:** Ridge or Lasso with intercept on original data

$\Updownarrow$

Ridge or Lasso without intercept after centering both $X_i$ and $Y_i$.

# Getting Rid of the Intercept

## Theorem

*For any method of the form* $\beta_0$ not penalized

$$\hat{\beta}_0, \hat{\beta} = \underset{\beta_0, \beta}{\arg\min}\, \mathrm{RSS}(\beta_0, \beta) + \lambda \mathrm{pen}(\beta),$$

*centering the features only changes* $\hat{\beta}_0$*, but not* $\hat{\beta}$*. Moreover,* **after** *centering, the intercept is always*

$$\hat{\beta}_0 = \bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i.$$

Effect: centering $Y_i$

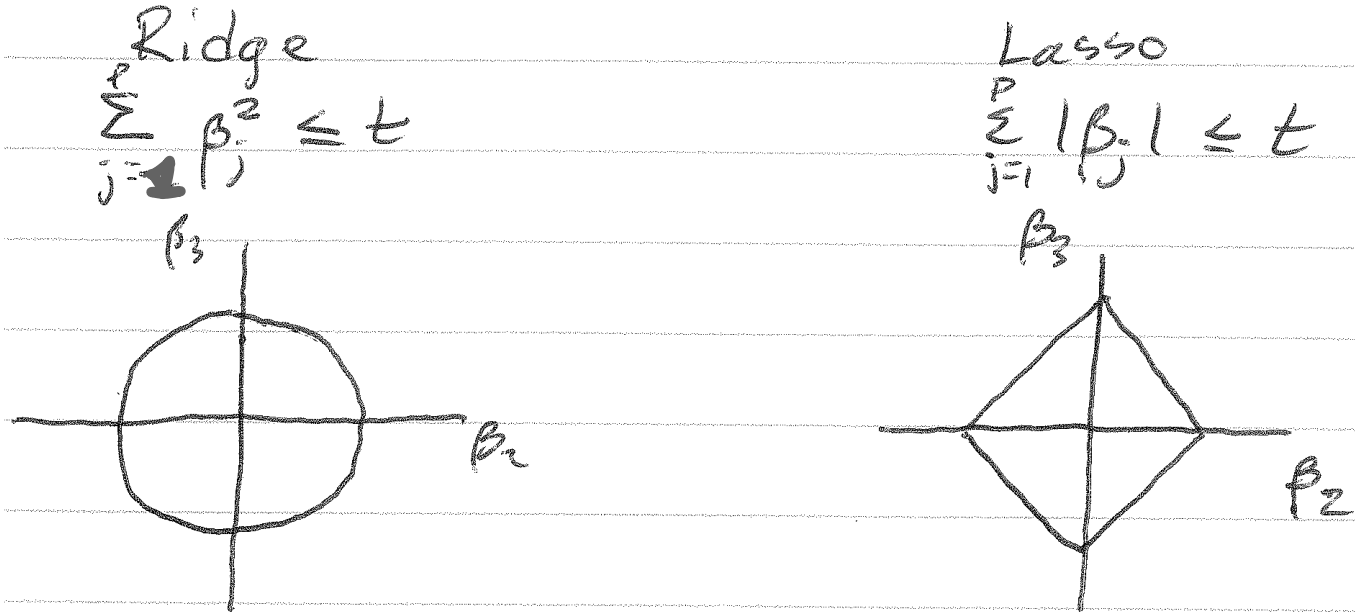
**Interpretation:** Ridge or Lasso with intercept on original data

$$\Updownarrow$$

Ridge or Lasso without intercept after centering both $X_i$ and $Y_i$.

**For rest of the lecture: Assume $X_i$ and $Y_i$ have been centered and there is no intercept.**

# Comparison: Ridge vs Lasso



# Comparison: Ridge vs Lasso

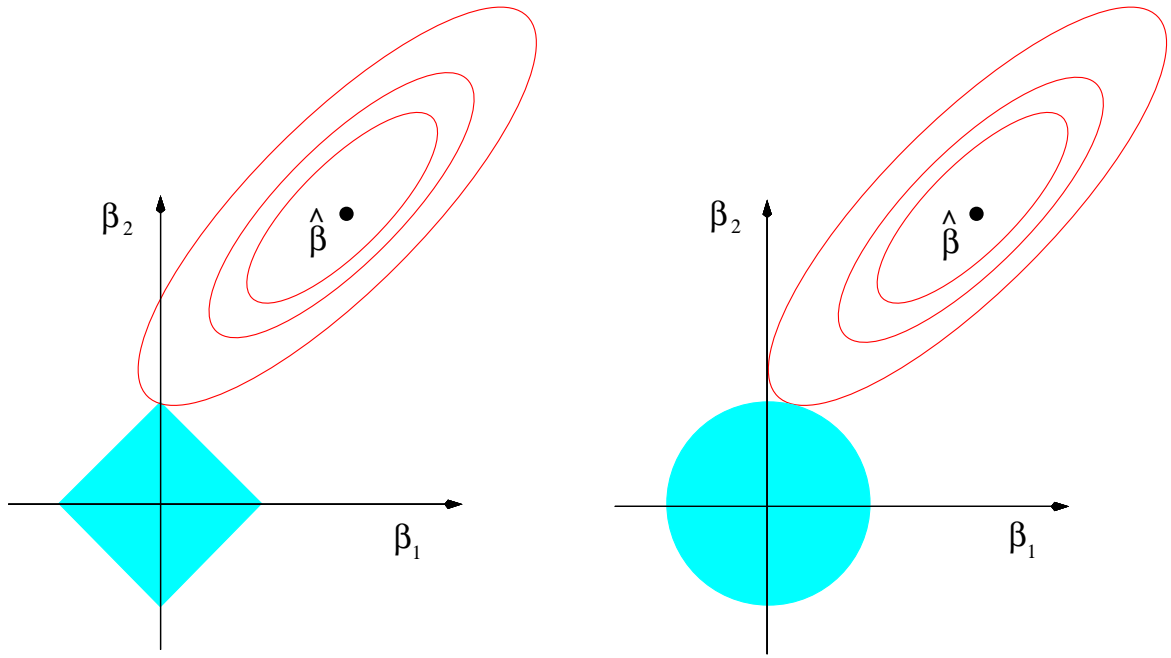
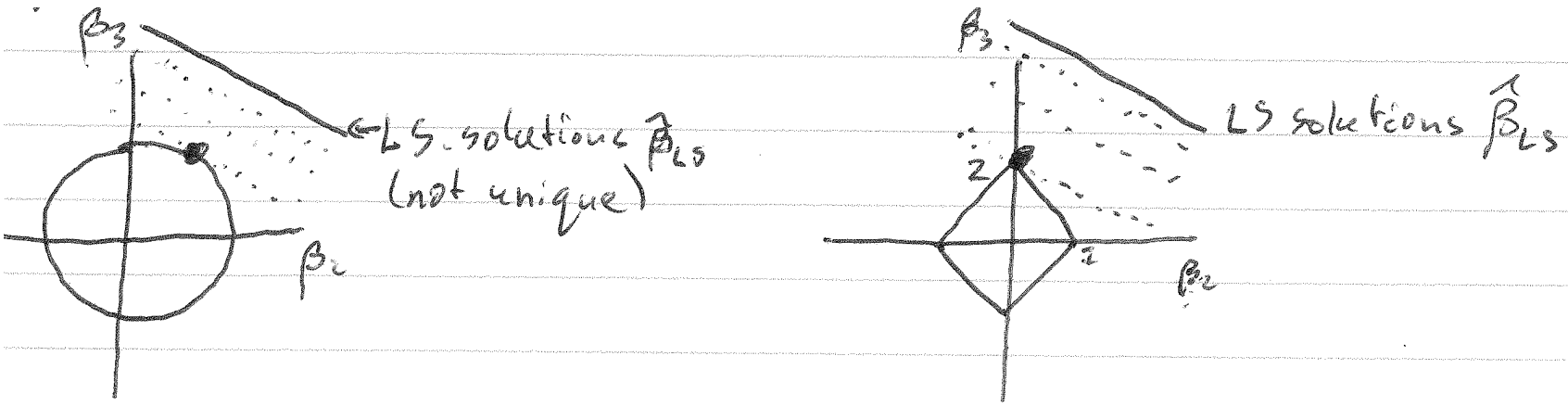



**FIGURE 3.11.** *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \leq t$ *and* $\beta_1^2 + \beta_2^2 \leq t^2$, *respectively, while the red ellipses are the contours of the least squares error function.*

# Comparison: Ridge vs Lasso

For **highly correlated features**:




**Ridge:**

- $\beta_1$ and $\beta_2$ get about equal weight

**Lasso:**

- Minor noise determines which of corners 1 and 2 is chosen.
- One variable gets all weight; the other 0.

# Formula for Ridge / Least Squares

$$\boldsymbol{X} = \begin{pmatrix} X_{1,1} & \dots & X_{1,p} \\ \vdots & & \vdots \\ X_{N,1} & \dots & X_{N,p} \end{pmatrix} = \begin{pmatrix} X_1^\intercal \\ \vdots \\ X_N^\intercal \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

The Ridge estimator $\hat{\beta}$ minimizes

$$\mathrm{RSS}(\beta) + \lambda \|\beta\|_2^2 = \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_2^2$$

## Lemma

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

# Formula for Ridge / Least Squares

$$\boldsymbol{X} = \begin{pmatrix} X_{1,1} & \dots & X_{1,p} \\ \vdots & & \vdots \\ X_{N,1} & \dots & X_{N,p} \end{pmatrix} = \begin{pmatrix} X_1^\intercal \\ \vdots \\ X_N^\intercal \end{pmatrix}, \qquad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$$

The Ridge estimator $\hat{\beta}$ minimizes

$$\mathrm{RSS}(\beta) + \lambda \|\beta\|_2^2 = \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda \|\beta\|_2^2$$

## Lemma

$$\hat{\beta}_{\text{Ridge}} = \left(\boldsymbol{X}^\intercal \boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}$$

- Always invertible for $\lambda > 0$, so Ridge solution unique
- $\lambda = 0$: least squares; not always invertible, because solution not always unique

# Formula for Ridge / Least Squares

The Ridge estimator $\hat{\beta}$ minimizes

$$\text{RSS}(\beta) + \lambda\|\beta\|_2^2 = \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$$

## Lemma

$$\hat{\beta}_{\text{Ridge}} = \left(\boldsymbol{X}^\intercal\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^\intercal\boldsymbol{y}$$

**Proof:** Objective convex, so can find minimum by setting gradient to 0:

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\beta}\left(\|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_2^2\right) \\ &= -2\boldsymbol{X}^\intercal\left(\boldsymbol{y} - \boldsymbol{X}\beta\right) + 2\lambda\beta \end{aligned}$$

$$2\boldsymbol{X}^\intercal\boldsymbol{y} = 2\boldsymbol{X}^\intercal\boldsymbol{X}\beta + 2\lambda\beta = 2(\boldsymbol{X}^\intercal\boldsymbol{X} + \lambda\boldsymbol{I})\beta$$

$$\beta = (\boldsymbol{X}^\intercal\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\intercal\boldsymbol{y}$$

# Ridge and Lasso as Shrinkage Methods

Compared to least squares, both Ridge and Lasso **"shrink"** parameters towards 0.

- We quantify this in (very) special case where per-feature vectors in $\boldsymbol{X}$ are orthonormal: $\boldsymbol{X}^\intercal \boldsymbol{X} = \boldsymbol{I}$

**Theorem**

*If $\boldsymbol{X}^\intercal \boldsymbol{X} = \boldsymbol{I}$, then, for $\hat{\beta}$ the least squares estimator,*

a) $\hat{\beta}_{\text{Ridge},j} = \dfrac{1}{1+\lambda}\hat{\beta}_j$

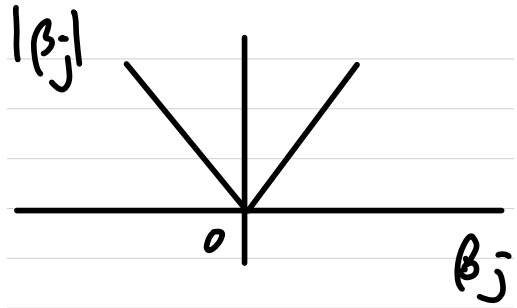
b) $\hat{\beta}_{\text{Lasso},j} = \text{sign}(\hat{\beta}_j)\left(|\hat{\beta}_j| - \dfrac{\lambda}{2}\right)_+ = \begin{cases} \hat{\beta}_j - \frac{\lambda}{2} & \text{if } \hat{\beta}_j > \frac{\lambda}{2} \\ \hat{\beta}_j + \frac{\lambda}{2} & \text{if } \hat{\beta}_j < -\frac{\lambda}{2} \\ 0 & \text{if } |\hat{\beta}_j| \leq \frac{\lambda}{2} \end{cases}$

# Proof for Ridge

If $\boldsymbol{X}^\intercal \boldsymbol{X} = \boldsymbol{I}$, then

$$\begin{aligned}
\hat{\beta}_{\text{Ridge}} &= \left(\boldsymbol{X}^\intercal \boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^\intercal \boldsymbol{y} \\
&= \left(\boldsymbol{I} + \lambda \boldsymbol{I}\right)^{-1} (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y} \\
&= \frac{1}{1+\lambda} \hat{\beta}
\end{aligned}$$

# Proof for Lasso




**Subgradients:**

- $|\beta_j|$ differentiable except at 0
- Would still like to 'differentiate' $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.
- For **convex** functions, can work with set of all possible **subgradients = possible slopes of tangent lines**.
- $g(\beta)$ is a subgradient of $\|\beta\|_1$ if

$$g(\beta)_j \in s(\beta_j) := \begin{cases} -1 & \text{if } \beta_j < 0 \\ +1 & \text{if } \beta_j > 0 \\ [-1, +1] & \text{if } \beta_j = 0 \end{cases}$$

# Proof for Lasso

- $g(\beta)$ is a subgradient of $\|\beta\|_1$ if

$$g(\beta)_j \in s(\beta_j) := \begin{cases} -1 & \text{if } \beta_j < 0 \\ +1 & \text{if } \beta_j > 0 \\ [-1, +1] & \text{if } \beta_j = 0 \end{cases}$$

If $\boldsymbol{X}^\intercal \boldsymbol{X} = \boldsymbol{I}$, then

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\beta}\left(\|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right) \\ &= -2\boldsymbol{X}^\intercal(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda g(\beta) \\ &= -2\boldsymbol{X}^\intercal \boldsymbol{y} + 2\boldsymbol{X}^\intercal \boldsymbol{X}\beta + \lambda g(\beta) \\ &= -2(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{y} + 2\beta + \lambda g(\beta) \\ &= -2\hat{\beta} + 2\beta + \lambda g(\beta) \end{aligned}$$

# Proof for Lasso

- $g(\beta)$ is a subgradient of $\|\beta\|_1$ if

$$g(\beta)_j \in s(\beta_j) := \begin{cases} -1 & \text{if } \beta_j < 0 \\ +1 & \text{if } \beta_j > 0 \\ [-1, +1] & \text{if } \beta_j = 0 \end{cases}$$

If $\boldsymbol{X}^\intercal \boldsymbol{X} = \boldsymbol{I}$, then

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\beta}\left(\|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|\right) \\ &= -2\boldsymbol{X}^\intercal(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda g(\beta) \\ &= -2\boldsymbol{X}^\intercal \boldsymbol{y} + 2\boldsymbol{X}^\intercal \boldsymbol{X}\beta + \lambda g(\beta) \\ &= -2(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{y} + 2\beta + \lambda g(\beta) \\ &= -2\hat{\beta} + 2\beta + \lambda g(\beta) \end{aligned}$$

$\beta$ minimum $\Leftrightarrow 2\beta - 2\hat{\beta} + \lambda g(\beta) = 0$ for one of the possible subgradients $g(\beta)$

$$\Leftrightarrow \beta_j \in \hat{\beta}_j - \frac{\lambda}{2} s(\beta_j) \quad \text{for all } j$$

# Proof for Lasso

- $g(\beta)$ is a subgradient of $\|\beta\|_1$ if

$$g(\beta)_j \in s(\beta_j) := \begin{cases} -1 & \text{if } \beta_j < 0 \\ +1 & \text{if } \beta_j > 0 \\ [-1, +1] & \text{if } \beta_j = 0 \end{cases}$$

$\beta$ minimum $\Leftrightarrow 2\beta - 2\hat{\beta} + \lambda g(\beta) = 0$ for one of the possible subgradients $g(\beta)$

$$\Leftrightarrow \beta_j \in \hat{\beta}_j - \frac{\lambda}{2} s(\beta_j) \quad \text{for all } j$$

i) $|\hat{\beta}_j| \leq \frac{\lambda}{2} \rightarrow \beta_j = 0$

ii) $\hat{\beta}_j > \frac{\lambda}{2} \rightarrow \beta_j = \hat{\beta}_j - \frac{\lambda}{2}$

iii) $\hat{\beta}_j < -\frac{\lambda}{2} \rightarrow \beta_j = \hat{\beta}_j + \frac{\lambda}{2}$

# Equivalence of the Two Formulations

$$\hat{\beta}_0, \hat{\beta} \in \underset{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p}{\arg\min}\ h(\beta_0, \beta) + \lambda\, \text{pen}(\beta) \tag{1}$$

$$\hat{\beta}_0, \hat{\beta} \in \underset{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p}{\arg\min}\ h(\beta_0, \beta) \text{ subject to } \text{pen}(\beta) \leq t \tag{2}$$

Ridge: $h(\beta_0, \beta) = \text{RSS}(\beta_0, \beta)$, $\text{pen}(\beta) = \|\beta\|_2^2$
Lasso: $h(\beta_0, \beta) = \text{RSS}(\beta_0, \beta)$, $\text{pen}(\beta) = \|\beta\|_1$

## Theorem

*Suppose*

- $h$ *and* pen *are both convex*
- $\text{pen}(\beta) \geq 0$, *with equality if* $\beta = 0$.

*Then*

a) *For any* $\lambda \in [0, \infty]$ *and* $\hat{\beta}_0, \hat{\beta}$ *satisfying* (1), *there exists* $t \in [0, \infty]$ *such that* $\hat{\beta}_0, \hat{\beta}$ *satisfy* (2).

b) *For any* $t \in [0, \infty]$ *and* $\hat{\beta}_0, \hat{\beta}$ *satisfying* (2), *there exists* $\lambda \in [0, \infty]$ *such that* $\hat{\beta}_0, \hat{\beta}$ *satisfy* (1).

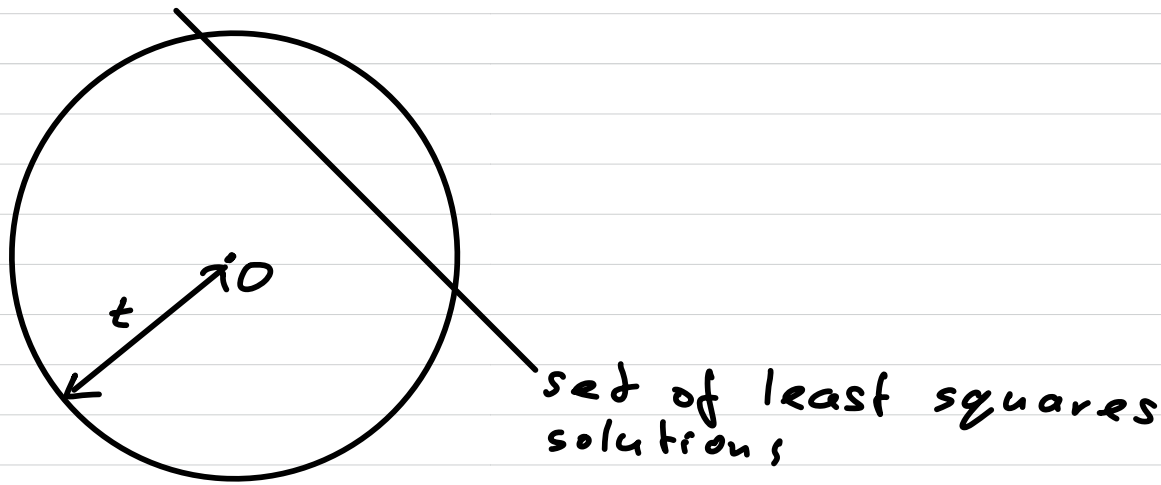
The remaining slides provide **additional discussion and proofs** for the equivalence between the two formulations

- It is optional to look at them
- They are not part of the official study materials for the exam

# Subtle Issue

Subtle issue: It is not true that for any $t$ there exists $\lambda$ such that (1) and (2) have the same solution sets.

Counterexample: Ridge with two equal features




Then primal (2) has subset of least squares solutions as solutions, but (1) either has all least squares solutions (for $\lambda=0$) or no least squares solutions ($\lambda>0$).

# Proof (using KKT conditions discussed in the April 7 Lecture)

Proof:

Primal: $\min_{\beta_0, \beta} \; h(\beta_0, \beta)$

$$\text{s.t.} \quad g_t(\beta) := \text{pen}(\beta) - t \leq 0$$

Lagrangian:

$$\mathcal{L}(\beta_0, \beta, \lambda) = h(\beta_0, \beta) + \lambda g_t(\beta)$$

For $0 < t < \infty$, strong duality holds by Slater's condition, because $g_t(0) < 0$!

# Proof (using KKT conditions discussed in the April 7 Lecture)

KKT conditions on optimal parameters $(\beta_0^*, \beta^*, \lambda^*)$:

- $g_t(\beta^*) \le 0$, $\lambda^* \ge 0$
- $\lambda^* g_t(\beta^*) = 0$
- $\beta_0^*, \beta^* \in \underset{\beta_0, \beta}{\operatorname{argmin}} \; h(\beta_0, \beta) + \lambda^* g_t(\beta)$

b) If $t = 0$, then $\lambda = \infty$ works. If $t = \infty$, then $\lambda = 0$ works. If $0 < t < \infty$, then by strong duality KKT conditions are necessary, so if $\beta_0^*, \beta^*$ are primal optimal, then there exists $\lambda^* \ge 0$ s.t.

$$\beta_0^*, \beta^* \in \underset{\beta_0, \beta}{\operatorname{argmin}} \; h(\beta_0, \beta) + \lambda^* g_t(\beta) = \underset{\beta_0, \beta}{\operatorname{argmin}} \; h(\beta_0, \beta) + \lambda^* \operatorname{pen}(\beta)$$

# Proof (using KKT conditions discussed in the April 7 Lecture)

a) Suppose $\beta_0^*, \beta^*$ satisfy (1).

If $\lambda^* = \infty$ or $\text{pen}(\beta^*) = 0$, then $t = 0$ works.

Otherwise, let $t = \text{pen}(\beta^*) > 0$.

Then
- $g_t(\beta^*) = 0$
- $\beta_0^*, \beta^* \in \underset{\beta_0, \beta}{\text{argmin}}\ h(\beta_0, \beta) + \lambda^* g_t(\beta^*)$

So KKT conditions satisfied and therefore

$\beta_0^*, \beta^*$ are primal optimal