Selected Topics in Mathematics of Learning

High-Dimensional Statistics

Lecturer: Marius Yamakou

Winter Semester 2024/25
Department of Data Science, FAU

November 12, 2024

Part III

Sparse Linear Models

Objectives:

- Recall the foundations of regression analysis, with a particular focus on the challenges and solutions in high-dimensional settings.

Part III

Sparse Linear Models

Objectives:

- Recall the foundations of regression analysis, with a particular focus on the challenges and solutions in high-dimensional settings.
- Cover linear regression basics, explores the issues that arise in large dimensions (such as overfitting and multicollinearity), and introduces regularization techniques like Ridge and Lasso regression to address these challenges.

Part III

Sparse Linear Models

Objectives:

- Recall the foundations of regression analysis, with a particular focus on the challenges and solutions in high-dimensional settings.
- Cover linear regression basics, explores the issues that arise in large dimensions (such as overfitting and multicollinearity), and introduces regularization techniques like Ridge and Lasso regression to address these challenges.
- Through comparisons, you will gain insights into the trade-offs and practical considerations in selecting appropriate regression methods for high-dimensional data.

Part III

Sparse Linear Models

Objectives:

- Recall the foundations of regression analysis, with a particular focus on the challenges and solutions in high-dimensional settings.
- Cover linear regression basics, explores the issues that arise in large dimensions (such as overfitting and multicollinearity), and introduces regularization techniques like Ridge and Lasso regression to address these challenges.
- Through comparisons, you will gain insights into the trade-offs and practical considerations in selecting appropriate regression methods for high-dimensional data.
- Understand different error metrics and prediction errors from a theoretical and practical perspective.
- How to choose the regularization parameter via cross-validation technique.

Outline

1. Linear regression setup
2. Recall what goes wrong in large dimensions
3. Ridge regression
4. Lasso regression
5. Comparison
6. Sparse linear models: A theoretical perspective
7. Sparse linear models: A practical perspective

1. Linear Regression: Setup in low dimension

- **Data Representation:**
  - $y \in \mathbb{R}^n$: Response vector (observed values)
  - $X \in \mathbb{R}^{n \times p}$: Design matrix (features/predictors)
  - $\beta \in \mathbb{R}^p$: Coefficient vector (parameters to estimate)

1. Linear Regression: Setup in low dimension

- **Data Representation:**
  - $y \in \mathbb{R}^n$: Response vector (observed values)
  - $X \in \mathbb{R}^{n \times p}$: Design matrix (features/predictors)
  - $\beta \in \mathbb{R}^p$: Coefficient vector (parameters to estimate)
- **Model Equation:**

$$y = X\beta + \epsilon$$

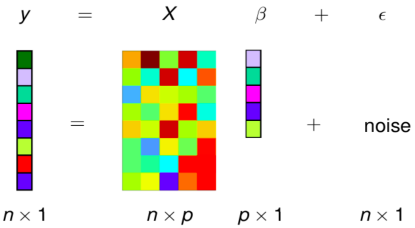
where $\epsilon \in \mathbb{R}^n$ represents the error or noise, assumed to follow a certain distribution [often $\epsilon \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$].

1. Linear Regression: Setup in low dimension

- **Data Representation:**
  - $y \in \mathbb{R}^n$: Response vector (observed values)
  - $X \in \mathbb{R}^{n \times p}$: Design matrix (features/predictors)
  - $\beta \in \mathbb{R}^p$: Coefficient vector (parameters to estimate)
- **Model Equation:**

$$y = X\beta + \epsilon$$

where $\epsilon \in \mathbb{R}^n$ represents the error or noise, assumed to follow a certain distribution [often $\epsilon \sim \mathcal{N}(0, \sigma^2 I_{n \times n})$].



- **Goal:** Estimate the β that minimize the sum of squared residuals: $\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2$.

  In other words, find the β to minimize the difference between the predicted and observed values of y.

2. Recall what goes wrong in large dimensions

Data setup: (y, X), where: $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$

Model assumption: $y = X\beta + \epsilon$ where: $\beta \in \mathbb{R}^p$ and $\epsilon \in \mathbb{R}^n$

Challenges in high dimensions: When $p > n$ we could have the following problems:

2. Recall what goes wrong in large dimensions

Data setup: (y, X), where: $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n\times p}$

Model assumption: $y = X\beta + \epsilon$ where: $\beta \in \mathbb{R}^p$ and $\epsilon \in \mathbb{R}^n$

Challenges in high dimensions: When $p > n$ we could have the following problems:

- The system becomes *underdetermined*, with infinitely many solutions. E.g., $y = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix}$, and assume $\epsilon = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

  The model is then: $y = X\beta$, which gives

$$\begin{cases} \beta_1 + \beta_2 + \beta_3 = 2, \\ \beta_1 + 2\beta_2 + 3\beta_3 = 3 \end{cases} \implies \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 1 + \beta_3 \\ 1 - 2\beta_3 \\ \beta_3 \end{pmatrix},$$

2. Recall what goes wrong in large dimensions

Data setup: (y, X), where: $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$

Model assumption: $y = X\beta + \epsilon$ where: $\beta \in \mathbb{R}^p$ and $\epsilon \in \mathbb{R}^n$

Challenges in high dimensions: When $p > n$ we could have the following problems:

- The system becomes *underdetermined*, with infinitely many solutions. E.g., $y = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, X = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix}$, and assume $\epsilon = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

  The model is then: $y = X\beta$, which gives

  $$\begin{cases} \beta_1 + \beta_2 + \beta_3 = 2, \\ \beta_1 + 2\beta_2 + 3\beta_3 = 3 \end{cases} \implies \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 1 + \beta_3 \\ 1 - 2\beta_3 \\ \beta_3 \end{pmatrix},$$

  which represents two equations (n) in three unknowns (p), making the system *underdetermined* and β_3 can take any value, leading to infinitely many solutions.

2. Recall what goes wrong in large dimensions

- Since there are infinitely many solutions for β, standard linear regression methods [e.g., ordinary least squares (OLS)] may fail to identify a unique solution for β because there are not enough observations (data points) to determine the effect of each feature/predictor on y.

2. Recall what goes wrong in large dimensions

- Since there are infinitely many solutions for β, standard linear regression methods [e.g., ordinary least squares (OLS)] may fail to identify a unique solution for β because there are not enough observations (data points) to determine the effect of each feature/predictor on y.
- Overfitting risk increases, impacting the model generalization.

  Why This Matters:

  - Without a unique solution for β, your model may fit perfectly to this data but fail to generalize to new unseen data.
  - In real applications, this can lead to misleading predictions and overfitting, as the model might rely on arbitrary combinations of features rather than meaningful patterns.

2. Recall what goes wrong in large dimensions

- From the model, $y = X\beta + \epsilon$, we have that $\epsilon = y - X\beta$, so:

  $\epsilon^\top \epsilon = (y - X\beta)^\top (y - X\beta) = y^\top y - y^\top X\beta - \beta^\top X^\top y + \beta^\top X^\top X\beta.$

2. Recall what goes wrong in large dimensions

- From the model, $y = X\beta + \epsilon$, we have that $\epsilon = y - X\beta$, so:

  $$\epsilon^\top \epsilon = (y - X\beta)^\top (y - X\beta) = y^\top y - y^\top X\beta - \beta^\top X^\top y + \beta^\top X^\top X\beta.$$

To minimize, take the derivative with respect to β and equate to zero:

$$\frac{d}{d\beta}\,\epsilon^\top \epsilon = 0 - y^\top X - X^\top y + 2X^\top X\beta = 0.$$

2. Recall what goes wrong in large dimensions

- From the model, $y = X\beta + \epsilon$, we have that $\epsilon = y - X\beta$, so:

$$\epsilon^\top \epsilon = (y - X\beta)^\top (y - X\beta) = y^\top y - y^\top X\beta - \beta^\top X^\top y + \beta^\top X^\top X\beta.$$

To minimize, take the derivative with respect to β and equate to zero:

$$\frac{d}{d\beta}\,\epsilon^\top \epsilon = 0 - y^\top X - X^\top y + 2X^\top X\beta = 0.$$

We use the facts that: (i) $y^\top X = (X^\top y)^\top$ and (ii) the derivative of $\beta^\top A\beta$ with respect to β is $2A\beta$ when A is symmetric, to have:

$$\frac{d}{d\beta}\,\epsilon^\top \epsilon = -2X^\top y + 2X^\top X\beta = 0 \quad \Longrightarrow \boxed{\beta = (X^\top X)^{-1} X^\top y}$$

2. Recall what goes wrong in large dimensions

- From the model, $y = X\beta + \epsilon$, we have that $\epsilon = y - X\beta$, so:

$$\epsilon^\top \epsilon = (y - X\beta)^\top (y - X\beta) = y^\top y - y^\top X\beta - \beta^\top X^\top y + \beta^\top X^\top X\beta.$$

To minimize, take the derivative with respect to β and equate to zero:

$$\frac{d}{d\beta}\,\epsilon^\top \epsilon = 0 - y^\top X - X^\top y + 2X^\top X\beta = 0.$$

We use the facts that: (i) $y^\top X = (X^\top y)^\top$ and (ii) the derivative of $\beta^\top A\beta$ with respect to β is $2A\beta$ when A is symmetric, to have:

$$\frac{d}{d\beta}\,\epsilon^\top \epsilon = -2X^\top y + 2X^\top X\beta = 0 \quad \Longrightarrow \boxed{\beta = (X^\top X)^{-1} X^\top y}$$

Problem: $X^\top X \in \mathbb{R}^{p\times p}$ known as **Gram matrix** is invertable only if $p < n$.

2. Recall what goes wrong in large dimensions

- From the model, $y = X\beta + \epsilon$, we have that $\epsilon = y - X\beta$, so:

$$\epsilon^\top \epsilon = (y - X\beta)^\top (y - X\beta) = y^\top y - y^\top X\beta - \beta^\top X^\top y + \beta^\top X^\top X\beta.$$

To minimize, take the derivative with respect to β and equate to zero:

$$\frac{d}{d\beta}\,\epsilon^\top \epsilon = 0 - y^\top X - X^\top y + 2X^\top X\beta = 0.$$

We use the facts that: (i) $y^\top X = (X^\top y)^\top$ and (ii) the derivative of $\beta^\top A\beta$ with respect to β is $2A\beta$ when A is symmetric, to have:

$$\frac{d}{d\beta}\,\epsilon^\top \epsilon = -2X^\top y + 2X^\top X\beta = 0 \quad \Longrightarrow \boxed{\beta = (X^\top X)^{-1} X^\top y}$$

Problem: $X^\top X \in \mathbb{R}^{p\times p}$ known as **Gram matrix** is invertable only if $p < n$. When $p \geq n$, the matrix $X^\top X$ is typically rank-deficient (i.e., not full rank with a rank of at most n) and thus not invertible, making it impossible to compute the OLS solution β directly.

2. Recall what goes wrong in large dimensions

- From the model, $y = X\beta + \epsilon$, we have that $\epsilon = y - X\beta$, so:

$$\epsilon^\top \epsilon = (y - X\beta)^\top (y - X\beta) = y^\top y - y^\top X\beta - \beta^\top X^\top y + \beta^\top X^\top X\beta.$$

To minimize, take the derivative with respect to β and equate to zero:

$$\frac{d}{d\beta}\,\epsilon^\top \epsilon = 0 - y^\top X - X^\top y + 2X^\top X\beta = 0.$$

We use the facts that: (i) $y^\top X = (X^\top y)^\top$ and (ii) the derivative of $\beta^\top A\beta$ with respect to β is $2A\beta$ when A is symmetric, to have:

$$\frac{d}{d\beta}\,\epsilon^\top \epsilon = -2X^\top y + 2X^\top X\beta = 0 \quad \Longrightarrow \boxed{\beta = (X^\top X)^{-1} X^\top y}$$

Problem: $X^\top X \in \mathbb{R}^{p\times p}$ known as **Gram matrix** is invertable only if $p < n$. When $p \geq n$, the matrix $X^\top X$ is typically rank-deficient (i.e., not full rank with a rank of at most n) and thus not invertible, making it impossible to compute the OLS solution β directly. In such cases, regularization methods, e.g., Ridge or Lasso regression are often needed to select one solution out of the many possible ones, which helps improve model stability and interpretability.

3. Ridge Regression

Ridge Regression is a regularization technique for addressing multicollinearity and overfitting in linear regression by adding a penalty term to the ordinary least squares (OLS) objective, shrinking the size of the coefficients to improve model generalization.

3. Ridge Regression

Ridge Regression is a regularization technique for addressing multicollinearity and overfitting in linear regression by adding a penalty term to the ordinary least squares (OLS) objective, shrinking the size of the coefficients to improve model generalization.

Ridge Estimator: For any $\lambda > 0$, we aim to minimize the objective function:

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 .$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and $\beta \in \mathbb{R}^p$.

3. Ridge Regression

Ridge Regression is a regularization technique for addressing multicollinearity and overfitting in linear regression by adding a penalty term to the ordinary least squares (OLS) objective, shrinking the size of the coefficients to improve model generalization.

Ridge Estimator: For any $\lambda > 0$, we aim to minimize the objective function:

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and $\beta \in \mathbb{R}^p$.
Solution: For any $\lambda > 0$, the solution to this minimization problem is:

$$\widehat{\beta} \overset{?}{=} (X^\top X + \lambda I)^{-1} X^\top y.$$

3. Ridge Regression

Ridge Regression is a regularization technique for addressing multicollinearity and overfitting in linear regression by adding a penalty term to the ordinary least squares (OLS) objective, shrinking the size of the coefficients to improve model generalization.

Ridge Estimator: For any $\lambda > 0$, we aim to minimize the objective function:

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and $\beta \in \mathbb{R}^p$.

Solution: For any $\lambda > 0$, the solution to this minimization problem is:

$$\widehat{\beta} \overset{?}{=} (X^\top X + \lambda I)^{-1} X^\top y.$$

where $I \in \mathbb{R}^{p \times p}$ is the identity matrix. Here, the term $(X^\top X + \lambda I)$ ensures that the matrix is invertible, even when X is not full-rank or suffers from multicollinearity.

3. Ridge Regression

Properties:

- **When $\lambda = 0$:** The ridge estimator reduces to the ordinary linear regression estimator.
- **Small λ (close to 0):** Ridge regression behaves similarly to ordinary least squares, with minimal shrinkage.
- **Large λ:** Coefficients are shrunk more aggressively, resulting in a simpler model with potentially better generalization but potentially higher bias.
- The value of λ is typically chosen using **cross-validation** to balance the trade-off between bias and variance.

3. Ridge Regression

Properties:

- **When $\lambda = 0$:** The ridge estimator reduces to the ordinary linear regression estimator.
- **Small λ (close to 0):** Ridge regression behaves similarly to ordinary least squares, with minimal shrinkage.
- **Large λ:** Coefficients are shrunk more aggressively, resulting in a simpler model with potentially better generalization but potentially higher bias.
- The value of λ is typically chosen using **cross-validation** to balance the trade-off between bias and variance.

Bias-Variance Trade-off:

- $\mathrm{Bias}(\widehat{\beta}) = \mathbb{E}[\widehat{\beta}] - \beta = -\lambda(X^\top X + \lambda I)^{-1}\beta$

3. Ridge Regression

Properties:

- **When $\lambda = 0$:** The ridge estimator reduces to the ordinary linear regression estimator.
- **Small λ (close to 0):** Ridge regression behaves similarly to ordinary least squares, with minimal shrinkage.
- **Large λ:** Coefficients are shrunk more aggressively, resulting in a simpler model with potentially better generalization but potentially higher bias.
- The value of λ is typically chosen using **cross-validation** to balance the trade-off between bias and variance.

Bias-Variance Trade-off:

- $\mathrm{Bias}(\widehat{\beta}) = \mathbb{E}[\widehat{\beta}] - \beta = -\lambda(X^\top X + \lambda I)^{-1}\beta$
- $\mathrm{Var}(\widehat{\beta}) = \mathbb{E}[(\widehat{\beta} - \mathbb{E}(\widehat{\beta}))^2] = (X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1}\sigma^2$

3. Ridge Regression

Properties:

- **When** $\lambda = 0$**:** The ridge estimator reduces to the ordinary linear regression estimator.
- **Small** λ **(close to 0):** Ridge regression behaves similarly to ordinary least squares, with minimal shrinkage.
- **Large** λ**:** Coefficients are shrunk more aggressively, resulting in a simpler model with potentially better generalization but potentially higher bias.
- The value of λ is typically chosen using **cross-validation** to balance the trade-off between bias and variance.

Bias-Variance Trade-off:

- $\mathrm{Bias}(\widehat{\beta}) = \mathbb{E}[\widehat{\beta}] - \beta = -\lambda(X^\top X + \lambda I)^{-1}\beta$
- $\mathrm{Var}(\widehat{\beta}) = \mathbb{E}[(\widehat{\beta} - \mathbb{E}(\widehat{\beta}))^2] = (X^\top X + \lambda I)^{-1} X^\top X (X^\top X + \lambda I)^{-1}\sigma^2$
- $\mathrm{MSE}(\widehat{\beta}) = \mathrm{Bias}^2(\widehat{\beta}) + \mathrm{Var}(\widehat{\beta}) = \mathrm{tr}\left[(X^\top X + \lambda I)^{-2}(\lambda^2 \beta\beta^\top + \sigma^2 X^\top X)\right]$

where σ^2 is the variance of the noise ϵ.

3. Ridge Regression

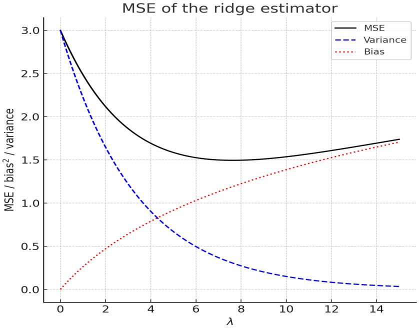
- $\mathrm{Bias}(\widehat{\beta}) \to \infty$ as $\lambda \to \infty$, since the regularization forces the coefficients closer to zero, possibly underfitting the data (Recall that $\lambda \to 0$ overfits the data as in OLS)

3. Ridge Regression

- $\text{Bias}(\widehat{\beta}) \to \infty$ as $\lambda \to \infty$, since the regularization forces the coefficients closer to zero, possibly underfitting the data (Recall that $\lambda \to 0$ overfits the data as in OLS)
- $\text{Var}(\widehat{\beta}) \to 0$ as $\lambda \to \infty$, making the model less sensitive to small fluctuations in data.

3. Ridge Regression

- $\mathrm{Bias}(\widehat{\beta}) \to \infty$ as $\lambda \to \infty$, since the regularization forces the coefficients closer to zero, possibly underfitting the data (Recall that $\lambda \to 0$ overfits the data as in OLS)
- $\mathrm{Var}(\widehat{\beta}) \to 0$ as $\lambda \to \infty$, making the model less sensitive to small fluctuations in data.
- The ideal λ minimizes the mean squared error $\mathrm{MSE}(\widehat{\beta})$, which depends on both the $\mathrm{Bias}(\widehat{\beta})$ and the $\mathrm{Var}(\widehat{\beta})$.



Large values of λ reduce overfitting by shrinking the coefficients, which reduces variance. However, this shrinkage can also introduce bias because the model may underfit the data if λ is too large.

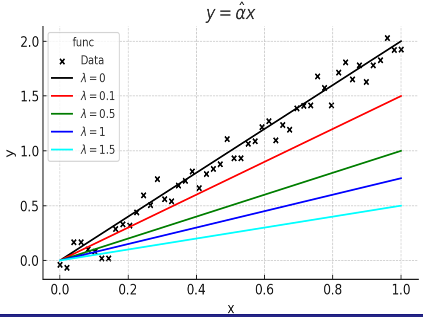
3. Ridge Regression: Shrinkage Effect

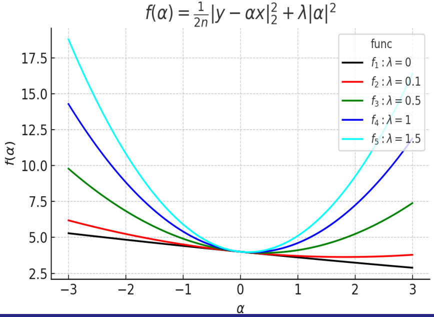
Observations: Ridge shrinks components of its estimate toward zero, but **never** set these components to be zero exactly (unless $\lambda = \infty$, in which case all components are zero and model becomes useless). Thus, ridge regression does not perform **variable selection** — the process of identifying and keeping only the most important features while discarding the others (by setting their coefficients to zero).

3. Ridge Regression: Shrinkage Effect

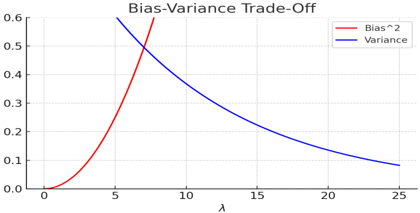
Observations: Ridge shrinks components of its estimate toward zero, but **never** set these components to be zero exactly (unless $\lambda = \infty$, in which case all components are zero and model becomes useless). Thus, ridge regression does not perform **variable selection** — the process of identifying and keeping only the most important features while discarding the others (by setting their coefficients to zero).

Example: One-dimensional case: $\widehat{\alpha} = \underset{\alpha}{\operatorname{argmin}} \frac{1}{2n}\|\mathrm{y} - \alpha \mathrm{x}\|_2^2 + \lambda|\alpha|^2$


$y = \hat{\alpha}x$


$f(\alpha) = \frac{1}{2n}|y - \alpha x|_2^2 + \lambda|\alpha|^2$

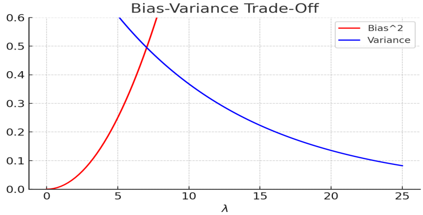
3. Ridge Regression: Example: Bias-Variance Trade-off and Test Error



Bias-Variance Trade-off

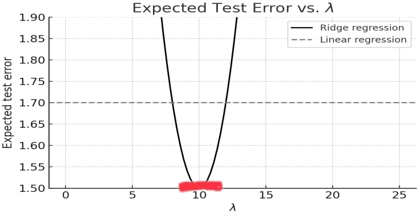
- Red curve (Bias^2): Bias increases with λ, as regularization shrinks coefficients, potentially leading to underfitting.
- Blue curve (Variance): Variance decreases as λ increases, making the model more stable but less flexible.
- Demonstrates the **bias-variance trade-off**: a balance between model complexity and stability.

3. Ridge Regression: Example: Bias-Variance Trade-off and Test Error



Bias-Variance Trade-off

- <span style="color:red">Red curve (Bias^2)</span>: Bias increases with λ, as regularization shrinks coefficients, potentially leading to underfitting.
- <span style="color:blue">Blue curve (Variance)</span>: Variance decreases as λ increases, making the model more stable but less flexible.
- Demonstrates the **bias-variance trade-off**: a balance between model complexity and stability.



Expected Test Error vs. λ

- U-shaped curve shows optimal λ minimizing test error by balancing bias and variance.
- **Dashed line (Linear Regression)**: Test error without regularization. Ridge Regression reduces error for λ in the highlighted range.
- Beyond optimal λ, error rises due to underfitting.

3.1 Sparsity

Understanding Sparsity in High-Dimensional Data

In many applications with high-dimensional data, where the number of features (p) is greater than the number of observations (n). Sparsity assumes that only a small subset of these features is relevant, allowing us to simplify the model by focusing on the most important predictors.

3.1 Sparsity

Understanding Sparsity in High-Dimensional Data

In many applications with high-dimensional data, where the number of features (p) is greater than the number of observations (n). Sparsity assumes that only a small subset of these features is relevant, allowing us to simplify the model by focusing on the most important predictors.

- In many applications, $p > n$, but not all features are important.
- Important prior knowledge: many extracted features in X are irrelevant to the outcome.
- Equivalently, this means many coefficients in β_0 are exactly zero.

3.1 Sparsity

Understanding Sparsity in High-Dimensional Data

In many applications with high-dimensional data, where the number of features (p) is greater than the number of observations (n). Sparsity assumes that only a small subset of these features is relevant, allowing us to simplify the model by focusing on the most important predictors.

- In many applications, $p > n$, but not all features are important.
- Important prior knowledge: many extracted features in X are irrelevant to the outcome.
- Equivalently, this means many coefficients in β_0 are exactly zero.
- For example, if y represents the size of a tumor, it might be reasonable to assume that y can be modeled as a linear combination of genetic information in X, which contains many genetic markers. However, most components of X (genes) will likely have zero or minimal impact, meaning only a few genes are truly relevant for predicting y.

3.1 Sparsity

Defining Sparsity

- **Sparsity**: The assumption that the "true" (unknown) coefficient vector β_0 has many entries that are exactly zero. Only a small number, s, of the p entries are non-zero, representing the important features.

3.1 Sparsity

Defining Sparsity

- **Sparsity**: The assumption that the "true" (unknown) coefficient vector β_0 has many entries that are exactly zero. Only a small number, s, of the p entries are non-zero, representing the important features.
- **Sparsity Assumption**: We define $S = \mathrm{supp}(\beta_0) = \{i \in \{1, \ldots, p\} \mid \beta_{0,i} \neq 0\}$, where $|S| = s$. This means that only s components are non-zero and are responsible for the outcome.

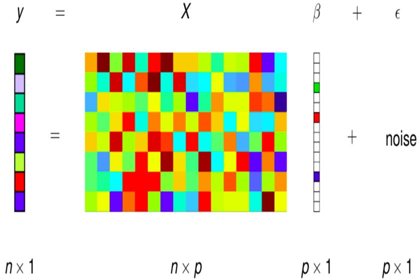
3.1 Sparsity

Defining Sparsity

- **Sparsity**: The assumption that the "true" (unknown) coefficient vector β_0 has many entries that are exactly zero. Only a small number, s, of the p entries are non-zero, representing the important features.

- **Sparsity Assumption**: We define $S = \text{supp}(\beta_0) = \{i \in \{1, \ldots, p\} \mid \beta_{0,i} \neq 0\}$, where $|S| = s$. This means that only s components are non-zero and are responsible for the outcome.

- Sparsity helps to reduce the effective dimension of the problem by focusing only on relevant features, making the model simpler and more interpretable.



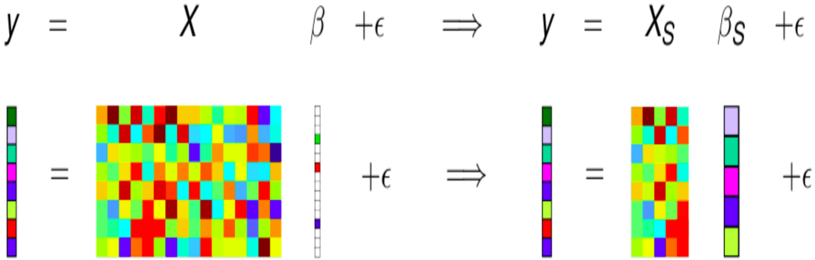
3.1 Sparsity

What does this dimension reduction look like?
Sparsity assumption: $|S| = s$ with $s \ll n$ and
$S = \mathrm{supp}(\beta_0) = \{i \in \{1, \ldots, p\} \mid \beta_{0,i} \neq 0\}$. The sparsity assumption allows us to reduce the effective dimension of the problem.

3.1 Sparsity

What does this dimension reduction look like?

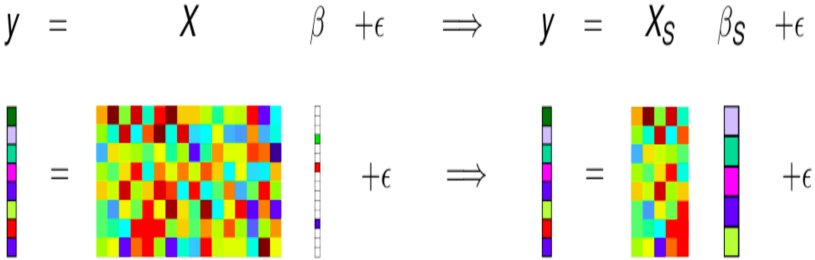
Sparsity assumption: $|S| = s$ with $s \ll n$ and $S = \mathrm{supp}(\beta_0) = \{i \in \{1, \dots, p\} \mid \beta_{0,i} \neq 0\}$. The sparsity assumption allows us to reduce the effective dimension of the problem.


$y = X \beta + \epsilon \implies y = X_S \beta_S + \epsilon$
$= \quad + \epsilon \quad \implies \quad = \quad + \epsilon$

3.1 Sparsity

What does this dimension reduction look like?

Sparsity assumption: $|S| = s$ with $s \ll n$ and $S = \mathrm{supp}(\beta_0) = \{i \in \{1, \ldots, p\} \mid \beta_{0,i} \neq 0\}$. The sparsity assumption allows us to reduce the effective dimension of the problem.



$y = X \beta + \epsilon \implies y = X_S \beta_S + \epsilon$
$= \quad + \epsilon \implies = \quad + \epsilon$

Hope: $X_s^\top X_s$ has full rank so that linear model can be applied.

Problem: We do not know the support: $S = \mathrm{supp}(\beta_0)$.

3.2 Sparsity Models

Motivation: How can we achieve sparsity in linear models? Sparsity is desirable in many applications to reduce model complexity and improve interpretability by selecting only the most important features.

3.2 Sparsity Models

Motivation: How can we achieve sparsity in linear models? Sparsity is desirable in many applications to reduce model complexity and improve interpretability by selecting only the most important features.

Subset Selection (Using the ℓ_0 "Norm")

$$\underset{\beta \in \mathbb{R}^q}{\text{argmin}} \|y - X\beta\|_2^2 \quad \text{subject to } \|\beta\|_0 \leq s \quad \text{where} \quad \|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}$$

3.2 Sparsity Models

Motivation: How can we achieve sparsity in linear models? Sparsity is desirable in many applications to reduce model complexity and improve interpretability by selecting only the most important features.

Subset Selection (Using the ℓ_0 "Norm")

$$\underset{\beta \in \mathbb{R}^q}{\text{argmin}} \|y - X\beta\|_2^2 \quad \text{subject to } \|\beta\|_0 \leq s \quad \text{where} \quad \|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}$$

Definition: β is **s-sparse** if it has at most s non-zero elements, i.e., $\|\beta\|_0 = s$.

- **Challenge:** $\|\beta\|_0$ (the ℓ_0 "norm" or "count norm") is not convex, making the optimization problem difficult to solve, since the process becomes computationally expensive and infeasible for large datasets.

3.2 Sparsity Models

Motivation: How can we achieve sparsity in linear models? Sparsity is desirable in many applications to reduce model complexity and improve interpretability by selecting only the most important features.

Subset Selection (Using the ℓ_0 "Norm")

$$\underset{\beta \in \mathbb{R}^q}{\text{argmin}} \|y - X\beta\|_2^2 \quad \text{subject to } \|\beta\|_0 \leq s \quad \text{where} \quad \|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}$$

Definition: β is **s-sparse** if it has at most s non-zero elements, i.e., $\|\beta\|_0 = s$.

- **Challenge:** $\|\beta\|_0$ (the ℓ_0 "norm" or "count norm") is not convex, making the optimization problem difficult to solve, since the process becomes computationally expensive and infeasible for large datasets.
- **Solution: Convex Relaxation**: Instead of a binary constraint on β (where each component is either 0 or non-zero i.e., $\beta_j \in \{0, 1\}$), we use a **convex relaxation** (such as the ℓ_1 Norm—sum of absolute values) that allows continuous values between 0 and 1, i.e., $\beta_j \in [0, 1]$.

3.2 Sparsity Models

Motivation: How can we achieve sparsity in linear models? Sparsity is desirable in many applications to reduce model complexity and improve interpretability by selecting only the most important features.

Subset Selection (Using the ℓ_0 "Norm")

$$\underset{\beta \in \mathbb{R}^q}{\text{argmin}} \|y - X\beta\|_2^2 \quad \text{subject to } \|\beta\|_0 \leq s \quad \text{where} \quad \|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}$$

Definition: β is **s-sparse** if it has at most s non-zero elements, i.e., $\|\beta\|_0 = s$.

- **Challenge:** $\|\beta\|_0$ (the ℓ_0 "norm" or "count norm") is not convex, making the optimization problem difficult to solve, since the process becomes computationally expensive and infeasible for large datasets.
- **Solution: Convex Relaxation**: Instead of a binary constraint on β (where each component is either 0 or non-zero i.e., $\beta_j \in \{0, 1\}$), we use a **convex relaxation** (such as the ℓ_1 Norm—sum of absolute values) that allows continuous values between 0 and 1, i.e., $\beta_j \in [0, 1]$.
- The Least Absolute Shrinkage and Selection Operator (LASSO) Estimator (using the ℓ_1 Norm) will be a natural relaxation of the problem.

4. LASSO Regression

Overview: The **LASSO (Least Absolute Shrinkage and Selection Operator)** is a regression technique that performs both regularization and feature selection, making it useful for creating sparse models.

4. LASSO Regression

Overview: The **LASSO (Least Absolute Shrinkage and Selection Operator)** is a regression technique that performs both regularization and feature selection, making it useful for creating sparse models.

LASSO Estimator: For any $\lambda > 0$, we define the LASSO estimator as:

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_1, \quad \text{where} \quad y \in \mathbb{R}^n \quad X \in \mathbb{R}^{n \times p}.$$

4. LASSO Regression

Overview: The **LASSO (Least Absolute Shrinkage and Selection Operator)** is a regression technique that performs both regularization and feature selection, making it useful for creating sparse models.

LASSO Estimator: For any $\lambda > 0$, we define the LASSO estimator as:

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_1, \quad \text{where} \quad y \in \mathbb{R}^n \quad X \in \mathbb{R}^{n \times p}.$$

- The ℓ_1 penalty term, $\lambda_n \|\beta\|_1$ (which is convex because it is the sum of absolute values, and the absolute value function is convex) encourages sparsity by shrinking some coefficients β_j exactly to zero, effectively selecting a subset of features.

4. LASSO Regression

Overview: The **LASSO (Least Absolute Shrinkage and Selection Operator)** is a regression technique that performs both regularization and feature selection, making it useful for creating sparse models.

LASSO Estimator: For any $\lambda > 0$, we define the LASSO estimator as:

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_1, \quad \text{where} \quad y \in \mathbb{R}^n \quad X \in \mathbb{R}^{n \times p}.$$

- The ℓ_1 penalty term, $\lambda_n \|\beta\|_1$ (which is convex because it is the sum of absolute values, and the absolute value function is convex) encourages sparsity by shrinking some coefficients β_j exactly to zero, effectively selecting a subset of features.
- This property of the ℓ_1 norm makes optimization problems involving it (such as LASSO) much easier to solve, as convex problems have well-behaved optimization properties, including a unique global minimum.

4. LASSO Regression

Overview: The **LASSO (Least Absolute Shrinkage and Selection Operator)** is a regression technique that performs both regularization and feature selection, making it useful for creating sparse models.

LASSO Estimator: For any $\lambda > 0$, we define the LASSO estimator as:

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \frac{1}{n} \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_1, \quad \text{where} \quad y \in \mathbb{R}^n \quad X \in \mathbb{R}^{n \times p}.$$

- The ℓ_1 penalty term, $\lambda_n \|\beta\|_1$ (which is convex because it is the sum of absolute values, and the absolute value function is convex) encourages sparsity by shrinking some coefficients β_j exactly to zero, effectively selecting a subset of features.
- This property of the ℓ_1 norm makes optimization problems involving it (such as LASSO) much easier to solve, as convex problems have well-behaved optimization properties, including a unique global minimum.
- λ_n is scaled according to the sample size n and controls the strength of regularization: higher λ_n results in more sparsity while lower λ_n reduces regularization and bias but increases variance.

4. LASSO Regression

- The factor $\frac{1}{n}$ normalizes the residual sum of squares by the number of observations, yielding an average squared error rather than the total squared error. This normalization is useful for comparing across datasets or sample sizes.

4. LASSO Regression

- The factor $\frac{1}{n}$ normalizes the residual sum of squares by the number of observations, yielding an average squared error rather than the total squared error. This normalization is useful for comparing across datasets or sample sizes.

Solution Properties:

- In general, **LASSO does not have an explicit solution**.
- However, if the design matrix X is orthogonal, i.e., $X^\top X = I_p$, we can derive an explicit solution:

$$\widehat{\beta}_j = \max\{\widehat{\beta}_j^{OLS} - \frac{\lambda}{2}, 0\} \quad \text{if } \widehat{\beta}_j^{OLS} > 0,$$

$$\widehat{\beta}_j = \min\{\widehat{\beta}_j^{OLS} + \frac{\lambda}{2}, 0\} \quad \text{if } \widehat{\beta}_j^{OLS} < 0$$

$$\widehat{\beta}_j = 0 \quad \text{if } |\widehat{\beta}_j^{\text{OLS}}| \leq \frac{\lambda}{2}$$

4. LASSO Regression

- The factor $\frac{1}{n}$ normalizes the residual sum of squares by the number of observations, yielding an average squared error rather than the total squared error. This normalization is useful for comparing across datasets or sample sizes.

Solution Properties:

- In general, **LASSO does not have an explicit solution**.
- However, if the design matrix X is orthogonal, i.e., $X^\top X = I_p$, we can derive an explicit solution:

$$\widehat{\beta}_j = \max\{\widehat{\beta}_j^{OLS} - \frac{\lambda}{2}, 0\} \quad \text{if } \widehat{\beta}_j^{OLS} > 0,$$

$$\widehat{\beta}_j = \min\{\widehat{\beta}_j^{OLS} + \frac{\lambda}{2}, 0\} \quad \text{if } \widehat{\beta}_j^{OLS} < 0$$

$$\widehat{\beta}_j = 0 \quad \text{if } |\widehat{\beta}_j^{\text{OLS}}| \leq \frac{\lambda}{2}$$

- These expressions show that LASSO applies a "soft thresholding" to the coefficients from ordinary least squares (OLS), i.e., not only shrinks coefficients but also sets some to zero based on λ, effectively selecting a subset of features in the model.

4.1 Illustrating the Bias-Variance Trade-off in LASSO

Implications of the Bias-Variance Trade-off in LASSO

- **Increasing λ (more shrinkage):** LASSO shrinks coefficients more, which **increases bias** by simplifying the model. This can lead to underfitting.

4.1 Illustrating the Bias-Variance Trade-off in LASSO

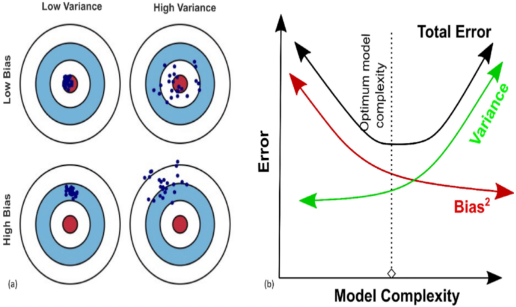
Implications of the Bias-Variance Trade-off in LASSO

- **Increasing λ (more shrinkage):** LASSO shrinks coefficients more, which **increases bias** by simplifying the model. This can lead to underfitting.
- **Decreasing λ (less shrinkage):** LASSO retains more complexity in the model, which **increases variance** by making the model more sensitive to fluctuations in the data. This can lead to overfitting.

4.1 Illustrating the Bias-Variance Trade-off in LASSO

Implications of the Bias-Variance Trade-off in LASSO

- **Increasing λ (more shrinkage):** LASSO shrinks coefficients more, which **increases bias** by simplifying the model. This can lead to underfitting.
- **Decreasing λ (less shrinkage):** LASSO retains more complexity in the model, which **increases variance** by making the model more sensitive to fluctuations in the data. This can lead to overfitting.



Key Insight: The ideal λ balances bias and variance to minimize the mean squared error (MSE) and achieve better generalization on new data.

5. Comparison of penalties: Ridge vs LASSO

Definition of the ℓ_1-Norm and its Effect on LASSO:

The ℓ_1-norm of a vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is defined as: $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

For a simple case with only two variables (β_1 and β_2), the ℓ_1-norm constraint $\|\beta\|_1 \leq t$ is equivalent to $|\beta_1| + |\beta_2| \leq t$, which defines a diamond-shaped region in 2D space.

5. Comparison of penalties: Ridge vs LASSO

Definition of the ℓ_1-Norm and its Effect on LASSO:

The ℓ_1-norm of a vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is defined as: $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

For a simple case with only two variables (β_1 and β_2), the ℓ_1-norm constraint $\|\beta\|_1 \leq t$ is equivalent to $|\beta_1| + |\beta_2| \leq t$, which defines a diamond-shaped region in 2D space.

Why the Diamond Shape Has Sharp Corners:

- In 2D, the constraint $|\beta_1| + |\beta_2| \leq t$ forms a square rotated by 45 degrees (a diamond), with the four corners aligned with the axes.

5. Comparison of penalties: Ridge vs LASSO

Definition of the ℓ_1-Norm and its Effect on LASSO:

The ℓ_1-norm of a vector $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ is defined as: $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

For a simple case with only two variables (β_1 and β_2), the ℓ_1-norm constraint $\|\beta\|_1 \leq t$ is equivalent to $|\beta_1| + |\beta_2| \leq t$, which defines a diamond-shaped region in 2D space.

Why the Diamond Shape Has Sharp Corners:

- In 2D, the constraint $|\beta_1| + |\beta_2| \leq t$ forms a square rotated by 45 degrees (a diamond), with the four corners aligned with the axes.
- The sharp corners arise due to the absolute values in the ℓ_1-norm. Each corner corresponds to cases where one of the coefficients is zero (e.g., $\beta_1 = 0$ or $\beta_2 = 0$), making it more likely for solutions to lie along these axes.

5. Comparison of penalties: Ridge vs LASSO

Definition of the ℓ_1-Norm and its Effect on LASSO:

The ℓ_1-norm of a vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is defined as: $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

For a simple case with only two variables (β_1 and β_2), the ℓ_1-norm constraint $\|\beta\|_1 \leq t$ is equivalent to $|\beta_1| + |\beta_2| \leq t$, which defines a diamond-shaped region in 2D space.

Why the Diamond Shape Has Sharp Corners:

- In 2D, the constraint $|\beta_1| + |\beta_2| \leq t$ forms a square rotated by 45 degrees (a diamond), with the four corners aligned with the axes.
- The sharp corners arise due to the absolute values in the ℓ_1-norm. Each corner corresponds to cases where one of the coefficients is zero (e.g., $\beta_1 = 0$ or $\beta_2 = 0$), making it more likely for solutions to lie along these axes.

Effect of Sharp Corners on LASSO Solutions:

- The sharp corners make it easier for the solution path to intersect the boundary of the ℓ_1-ball exactly at points where one or more coefficients are zero.

5. Comparison of penalties: Ridge vs LASSO

Definition of the ℓ_1-Norm and its Effect on LASSO:

The ℓ_1-norm of a vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is defined as: $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

For a simple case with only two variables (β_1 and β_2), the ℓ_1-norm constraint $\|\beta\|_1 \leq t$ is equivalent to $|\beta_1| + |\beta_2| \leq t$, which defines a diamond-shaped region in 2D space.

Why the Diamond Shape Has Sharp Corners:

- In 2D, the constraint $|\beta_1| + |\beta_2| \leq t$ forms a square rotated by 45 degrees (a diamond), with the four corners aligned with the axes.
- The sharp corners arise due to the absolute values in the ℓ_1-norm. Each corner corresponds to cases where one of the coefficients is zero (e.g., $\beta_1 = 0$ or $\beta_2 = 0$), making it more likely for solutions to lie along these axes.

Effect of Sharp Corners on LASSO Solutions:

- The sharp corners make it easier for the solution path to intersect the boundary of the ℓ_1-ball exactly at points where one or more coefficients are zero.
- In optimization, LASSO minimizes the objective function $\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$ subject to $\|\beta\|_1 \leq t$.

5. Comparison of penalties: Ridge vs LASSO

Definition of the ℓ_1-Norm and its Effect on LASSO:

The ℓ_1-norm of a vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is defined as: $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

For a simple case with only two variables (β_1 and β_2), the ℓ_1-norm constraint $\|\beta\|_1 \leq t$ is equivalent to $|\beta_1| + |\beta_2| \leq t$, which defines a diamond-shaped region in 2D space.

Why the Diamond Shape Has Sharp Corners:

- In 2D, the constraint $|\beta_1| + |\beta_2| \leq t$ forms a square rotated by 45 degrees (a diamond), with the four corners aligned with the axes.
- The sharp corners arise due to the absolute values in the ℓ_1-norm. Each corner corresponds to cases where one of the coefficients is zero (e.g., $\beta_1 = 0$ or $\beta_2 = 0$), making it more likely for solutions to lie along these axes.

Effect of Sharp Corners on LASSO Solutions:

- The sharp corners make it easier for the solution path to intersect the boundary of the ℓ_1-ball exactly at points where one or more coefficients are zero.
- In optimization, LASSO minimizes the objective function $\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$ subject to $\|\beta\|_1 \leq t$.
- When the solution lies on the boundary of the diamond (especially at a corner), some coefficients are driven to zero, resulting in **sparsity**: LASSO sets some coefficients to zero, effectively selecting a subset of features.

5. Comparison of penalties: Ridge vs LASSO

Definition of the ℓ_1-Norm and its Effect on LASSO:

The ℓ_1-norm of a vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is defined as: $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

For a simple case with only two variables (β_1 and β_2), the ℓ_1-norm constraint $\|\beta\|_1 \leq t$ is equivalent to $|\beta_1| + |\beta_2| \leq t$, which defines a diamond-shaped region in 2D space.

Why the Diamond Shape Has Sharp Corners:

- In 2D, the constraint $|\beta_1| + |\beta_2| \leq t$ forms a square rotated by 45 degrees (a diamond), with the four corners aligned with the axes.
- The sharp corners arise due to the absolute values in the ℓ_1-norm. Each corner corresponds to cases where one of the coefficients is zero (e.g., $\beta_1 = 0$ or $\beta_2 = 0$), making it more likely for solutions to lie along these axes.

Effect of Sharp Corners on LASSO Solutions:

- The sharp corners make it easier for the solution path to intersect the boundary of the ℓ_1-ball exactly at points where one or more coefficients are zero.
- In optimization, LASSO minimizes the objective function $\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$ subject to $\|\beta\|_1 \leq t$.
- When the solution lies on the boundary of the diamond (especially at a corner), some coefficients are driven to zero, resulting in **sparsity**: LASSO sets some coefficients to zero, effectively selecting a subset of features.

Comparison with the ℓ_2-Norm (Ridge Regression):

- The ℓ_2-norm, defined as $\|\beta\|_2 = \sqrt{\beta_1^2 + \beta_2^2}$, produces a circular constraint region in 2D (a sphere in higher dimensions).

5. Comparison of penalties: Ridge vs LASSO

Definition of the ℓ_1-Norm and its Effect on LASSO:

The ℓ_1-norm of a vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is defined as: $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

For a simple case with only two variables (β_1 and β_2), the ℓ_1-norm constraint $\|\beta\|_1 \leq t$ is equivalent to $|\beta_1| + |\beta_2| \leq t$, which defines a diamond-shaped region in 2D space.

Why the Diamond Shape Has Sharp Corners:

- In 2D, the constraint $|\beta_1| + |\beta_2| \leq t$ forms a square rotated by 45 degrees (a diamond), with the four corners aligned with the axes.
- The sharp corners arise due to the absolute values in the ℓ_1-norm. Each corner corresponds to cases where one of the coefficients is zero (e.g., $\beta_1 = 0$ or $\beta_2 = 0$), making it more likely for solutions to lie along these axes.

Effect of Sharp Corners on LASSO Solutions:

- The sharp corners make it easier for the solution path to intersect the boundary of the ℓ_1-ball exactly at points where one or more coefficients are zero.
- In optimization, LASSO minimizes the objective function $\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$ subject to $\|\beta\|_1 \leq t$.
- When the solution lies on the boundary of the diamond (especially at a corner), some coefficients are driven to zero, resulting in **sparsity**: LASSO sets some coefficients to zero, effectively selecting a subset of features.

Comparison with the ℓ_2-Norm (Ridge Regression):

- The ℓ_2-norm, defined as $\|\beta\|_2 = \sqrt{\beta_1^2 + \beta_2^2}$, produces a circular constraint region in 2D (a sphere in higher dimensions).
- This circular boundary does not have sharp corners, so it does not favor solutions where coefficients are exactly zero.

5. Comparison of penalties: Ridge vs LASSO

Definition of the ℓ_1-Norm and its Effect on LASSO:

The ℓ_1-norm of a vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is defined as: $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$.

For a simple case with only two variables (β_1 and β_2), the ℓ_1-norm constraint $\|\beta\|_1 \leq t$ is equivalent to $|\beta_1| + |\beta_2| \leq t$, which defines a diamond-shaped region in 2D space.

Why the Diamond Shape Has Sharp Corners:

- In 2D, the constraint $|\beta_1| + |\beta_2| \leq t$ forms a square rotated by 45 degrees (a diamond), with the four corners aligned with the axes.
- The sharp corners arise due to the absolute values in the ℓ_1-norm. Each corner corresponds to cases where one of the coefficients is zero (e.g., $\beta_1 = 0$ or $\beta_2 = 0$), making it more likely for solutions to lie along these axes.

Effect of Sharp Corners on LASSO Solutions:

- The sharp corners make it easier for the solution path to intersect the boundary of the ℓ_1-ball exactly at points where one or more coefficients are zero.
- In optimization, LASSO minimizes the objective function $\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$ subject to $\|\beta\|_1 \leq t$.
- When the solution lies on the boundary of the diamond (especially at a corner), some coefficients are driven to zero, resulting in **sparsity**: LASSO sets some coefficients to zero, effectively selecting a subset of features.

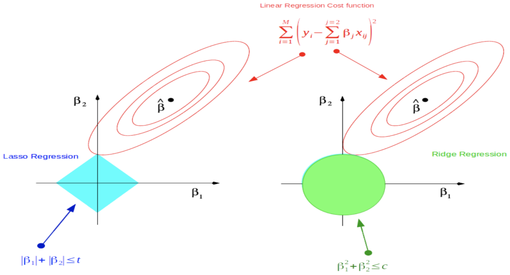
Comparison with the ℓ_2-Norm (Ridge Regression):

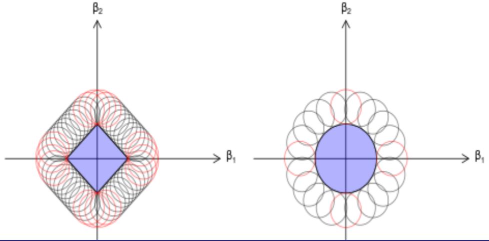
- The ℓ_2-norm, defined as $\|\beta\|_2 = \sqrt{\beta_1^2 + \beta_2^2}$, produces a circular constraint region in 2D (a sphere in higher dimensions).
- This circular boundary does not have sharp corners, so it does not favor solutions where coefficients are exactly zero.
- As a result, Ridge Regression shrinks coefficients but does not drive them to zero, retaining all features.

5. Comparison: Ridge vs LASSO

The major difference between ridge and lasso:

- the sharp, non-differentiable corners of the l_1-ball produce parsimonious models (models that prioritize simplicity and interpretability, aiming to capture the essential structure of the data without unnecessary complexity) for sufficiently large values of λ
- the lasso lacks an analytic solution, making both computation and theoretical results more difficult.


Linear Regression Cost function
$\sum_{i=1}^{M}\left(y_i - \sum_{j=1}^{j=2}\beta_j x_{ij}\right)^2$
β_2
$\hat{\beta}$
Lasso Regression
β_1
$|\beta_1| + |\beta_2| \leq t$
β_2
$\hat{\beta}$
Ridge Regression
β_1
$\beta_1^2 + \beta_2^2 \leq c$



5. Comparison: Variable selection in Ridge vs LASSO regressions

Definition of Variable Selection:

- Variable selection is the process of identifying the most relevant predictors or features in a model from a larger set while setting the coefficients of less relevant features to zero.
- The goal is to simplify the model, which can improve predictive accuracy and make the model easier to interpret.

5. Comparison: Variable selection in Ridge vs LASSO regressions

Definition of Variable Selection:

- Variable selection is the process of identifying the most relevant predictors or features in a model from a larger set while setting the coefficients of less relevant features to zero.
- The goal is to simplify the model, which can improve predictive accuracy and make the model easier to interpret.

LASSO vs. Ridge for Variable Selection:

- **LASSO regression** performs variable selection by encouraging some coefficients to be exactly zero, effectively removing those variables from the model. This makes it suitable for models where interpretability and identifying key features are essential.
- **Ridge regression** shrinks coefficients towards zero but does not set any to exactly zero. It reduces the influence of less important predictors but retains all variables in the model. This can be beneficial for prediction but less helpful for interpretation.

5. Comparison: Variable selection in Ridge vs LASSO regressions

Definition of Variable Selection:

- Variable selection is the process of identifying the most relevant predictors or features in a model from a larger set while setting the coefficients of less relevant features to zero.
- The goal is to simplify the model, which can improve predictive accuracy and make the model easier to interpret.

LASSO vs. Ridge for Variable Selection:

- **LASSO regression** performs variable selection by encouraging some coefficients to be exactly zero, effectively removing those variables from the model. This makes it suitable for models where interpretability and identifying key features are essential.
- **Ridge regression** shrinks coefficients towards zero but does not set any to exactly zero. It reduces the influence of less important predictors but retains all variables in the model. This can be beneficial for prediction but less helpful for interpretation.

Predictive Accuracy and Model Interpretation:

- Variable selection can enhance **predictive accuracy** by reducing overfitting, as unnecessary variables are removed from the model.
- It also improves **model interpretation** by simplifying the model, highlighting the most influential predictors.

5. Comparison: Variable selection in Ridge vs LASSO regressions

Definition of Variable Selection:

- Variable selection is the process of identifying the most relevant predictors or features in a model from a larger set while setting the coefficients of less relevant features to zero.
- The goal is to simplify the model, which can improve predictive accuracy and make the model easier to interpret.

LASSO vs. Ridge for Variable Selection:

- **LASSO regression** performs variable selection by encouraging some coefficients to be exactly zero, effectively removing those variables from the model. This makes it suitable for models where interpretability and identifying key features are essential.
- **Ridge regression** shrinks coefficients towards zero but does not set any to exactly zero. It reduces the influence of less important predictors but retains all variables in the model. This can be beneficial for prediction but less helpful for interpretation.

Predictive Accuracy and Model Interpretation:

- Variable selection can enhance **predictive accuracy** by reducing overfitting, as unnecessary variables are removed from the model.
- It also improves **model interpretation** by simplifying the model, highlighting the most influential predictors.

Performance of Ridge Regression When True Coefficients Are Zero:

- **Prediction:** Ridge Regression can still perform well in practice, as it shrinks all coefficients towards zero, mimicking the effect of setting small coefficients to zero.
- **Interpretation:** Ridge is less helpful for interpretation because it does not indicate which variables are irrelevant. LASSO is often preferred for interpretation because it creates a sparse model by setting some coefficients to zero.

Summary of penalization methods: Three canonical choices of penalization norms

$$\|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}, \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|, \quad \|\beta\|_2 = \left(\sum_{j=1}^{p} |\beta_j|^2\right)^{\frac{1}{2}}$$

- ℓ_0 norm: Counts the number of non-zero coefficients (sparsity).
- ℓ_1 norm: Sum of absolute values, promoting sparsity by shrinking some coefficients to zero.
- ℓ_2 norm: Sum of squared values, promoting smaller coefficients but not necessarily zero, reducing model sensitivity.

Summary of penalization methods: Three canonical choices of penalization norms

$$\|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}, \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|, \quad \|\beta\|_2 = \left(\sum_{j=1}^{p} |\beta_j|^2\right)^{\frac{1}{2}}$$

- ℓ_0 norm: Counts the number of non-zero coefficients (sparsity).
- ℓ_1 norm: Sum of absolute values, promoting sparsity by shrinking some coefficients to zero.
- ℓ_2 norm: Sum of squared values, promoting smaller coefficients but not necessarily zero, reducing model sensitivity.

Penalized Regression Forms:

- **Best Subset Selection (using ℓ_0 norm):** $\text{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda\|\beta\|_0$ Selects the smallest subset of predictors by directly counting non-zero coefficients, but is computationally intensive.

Summary of penalization methods: Three canonical choices of penalization norms

$$\|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}, \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|, \quad \|\beta\|_2 = \left(\sum_{j=1}^{p} |\beta_j|^2\right)^{\frac{1}{2}}$$

- ℓ_0 norm: Counts the number of non-zero coefficients (sparsity).
- ℓ_1 norm: Sum of absolute values, promoting sparsity by shrinking some coefficients to zero.
- ℓ_2 norm: Sum of squared values, promoting smaller coefficients but not necessarily zero, reducing model sensitivity.

Penalized Regression Forms:

- **Best Subset Selection (using ℓ_0 norm):** $\text{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda\|\beta\|_0$ Selects the smallest subset of predictors by directly counting non-zero coefficients, but is computationally intensive.
- **Lasso Regression (using ℓ_1 norm):** $\text{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1$ Shrinks some coefficients exactly to zero, providing both regularization and feature selection.

Summary of penalization methods: Three canonical choices of penalization norms

$$\|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}, \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|, \quad \|\beta\|_2 = \left(\sum_{j=1}^{p} |\beta_j|^2\right)^{\frac{1}{2}}$$

- ℓ_0 norm: Counts the number of non-zero coefficients (sparsity).
- ℓ_1 norm: Sum of absolute values, promoting sparsity by shrinking some coefficients to zero.
- ℓ_2 norm: Sum of squared values, promoting smaller coefficients but not necessarily zero, reducing model sensitivity.

Penalized Regression Forms:

- **Best Subset Selection (using ℓ_0 norm):** $\operatorname{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_0$
  Selects the smallest subset of predictors by directly counting non-zero coefficients, but is computationally intensive.
- **Lasso Regression (using ℓ_1 norm):** $\operatorname{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$
  Shrinks some coefficients exactly to zero, providing both regularization and feature selection.
- **Ridge Regression (using ℓ_2 norm):** $\operatorname{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$
  Shrinks coefficients but retains all features, reducing model complexity without feature elimination.

Summary of penalization methods: Three canonical choices of penalization norms

$$\|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}, \quad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|, \quad \|\beta\|_2 = \left(\sum_{j=1}^{p} |\beta_j|^2 \right)^{\frac{1}{2}}$$

- ℓ_0 norm: Counts the number of non-zero coefficients (sparsity).
- ℓ_1 norm: Sum of absolute values, promoting sparsity by shrinking some coefficients to zero.
- ℓ_2 norm: Sum of squared values, promoting smaller coefficients but not necessarily zero, reducing model sensitivity.

Penalized Regression Forms:

- **Best Subset Selection (using ℓ_0 norm):** $\text{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_0$ Selects the smallest subset of predictors by directly counting non-zero coefficients, but is computationally intensive.
- **Lasso Regression (using ℓ_1 norm):** $\text{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$ Shrinks some coefficients exactly to zero, providing both regularization and feature selection.
- **Ridge Regression (using ℓ_2 norm):** $\text{argmin}_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$ Shrinks coefficients but retains all features, reducing model complexity without feature elimination.

Sparsity Assumption: Assume β_0 is s-sparse, i.e., $\|\beta_0\|_0 = s$, meaning it has at most s non-zero elements.

Summary

What are possible penalizations and their advantages and disadvantages?

- Ridge:
  - penalizes with l_2-norm $\|\beta\|_2 = \left(\sum_{j=1}^{p} |\beta_j|^2\right)^{\frac{1}{2}}$
  - explicit representation
  - shrinks towards small values
  - does not do model selection
- Best subset selection:
  - penalizes with l_0-norm $\|\beta\|_0 = \sum_{j=1}^{p} 1_{\{\beta_j \neq 0\}}$,
  - not convex
- LASSO:
  - penalizes with l_1-norm $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$
  - shrinks exactly to zero
  - convex

What is the bias-variance trade-off?

- The bias increases as λ (amount of shrinkage) increases.
- The variance decreases as λ (amount of shrinkage) increases.