

# Lasso Regression

---

Nipun Batra

IIT Gandhinagar

September 17, 2025

# Outline

1. Introduction and Motivation
2. Mathematical Formulation
3. Why Lasso Gives Sparsity
  - 3.1 Geometric Interpretation
  - 3.2 Gradient Descent Interpretation
4. Geometric Interpretation
5. Regularization Effects
6. Feature Selection Properties
7. Subgradient Methods
8. Coordinate Descent Algorithm
9. Worked Example
10. Visual Coordinate Descent
11. Failure of Coordinate Descent
12. Mathematical Derivation
13. Lasso vs Ridge Comparison
14. Summary and Applications

# Introduction and Motivation

# What is Lasso Regression?

## Definition: LASSO

Least **A**bsolute **S**hrinkage and **S**election **O**perator

## Key Points: Key Properties

- Uses L1 penalty (absolute values) instead of L2 penalty
- Leads to **sparse solutions** (many coefficients become exactly zero)
- Performs automatic feature selection
- Popular for high-dimensional problems

# Mathematical Formulation

# Problem: Why Not Just Use Ridge?

## Important: Limitation of Ridge Regression

Ridge regression shrinks coefficients but **never makes them exactly zero**

## Example: High-Dimensional Problem

- 1000 features, only 50 are truly relevant
- Ridge gives tiny but non-zero coefficients for irrelevant features
- Model is not interpretable
- Need automatic feature selection!

# Lasso Objective Function

## Definition: Constrained Form

$$\boldsymbol{\theta}_{\text{opt}} = \arg \min_{\boldsymbol{\theta}} \|(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\|_2^2 \text{ subject to } \|\boldsymbol{\theta}\|_1 \leq s$$

## Theorem: Penalized Form (Using Lagrangian Duality)

Constrained form is equivalent to:

$$\boldsymbol{\theta}_{\text{opt}} = \arg \min_{\boldsymbol{\theta}} \underbrace{\|(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1}_{\text{Lasso Objective}}$$

## L1 Norm (Manhattan Distance)

$$\|\boldsymbol{\theta}\|_1 = |\theta_1| + |\theta_2| + \cdots + |\theta_d| = \sum_{j=1}^d |\theta_j|$$

# The Challenge: Non-Differentiability

## Important: Problem

The L1 norm  $\|\boldsymbol{\theta}\|_1 = \sum_j |\theta_j|$  is **not differentiable** at  $\theta_j = 0$

Cannot Use Standard Calculus

$$\frac{\partial}{\partial \boldsymbol{\theta}} [\|(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1] = 0$$

This fails because  $\frac{\partial |\theta_j|}{\partial \theta_j}$  is undefined at  $\theta_j = 0$

## Key Points: Solution Approaches

- **Coordinate Descent:** Optimize one coefficient at a time
- **Subgradient Methods:** Generalize derivatives to non-smooth functions



# Why Lasso Gives Sparsity

# Sparsity: The Key Question

## Important: Central Question

Why does Lasso produce sparse solutions while Ridge doesn't?

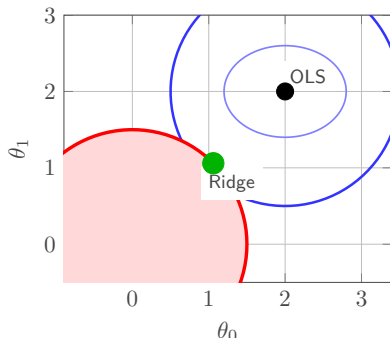
## Key Points: Two Perspectives

- **Geometric:** Shape of constraint regions
- **Algorithmic:** Behavior of optimization algorithms

## Example: Preview

We'll see why  $L_p$  norms with  $p < 2$  promote sparsity

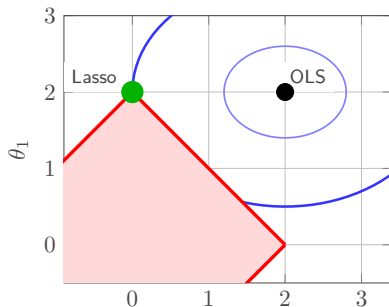
# L2 Norm: Ridge Constraint



## Key Points: L2 Properties

- **Shape:** Perfect circle
- **Constraint:**  
 $\theta_0^2 + \theta_1^2 \leq c$
- **Boundary:**  
Smooth everywhere
- **Intersection:**  
Rarely on axes
- **Result:** No sparsity

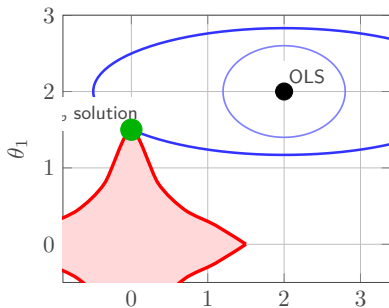
# L1 Norm: Lasso Constraint



## Key Points: L1 Properties

- **Shape:** Diamond/rhombus
- **Constraint:**  
 $|\theta_0| + |\theta_1| \leq c$
- **Corners:** Sharp at axes
- **Intersection:**  
High probability on axes
- **Result:**  
Automatic sparsity!

## $L_p$ Norm: Even More Sparsity ( $p < 1$ )



### Key Points: $L_p$ Properties ( $p < 1$ )

- **Shape:** Highly concave
- **Constraint:**  
 $(|\theta_0|^p + |\theta_1|^p)^{1/p} \leq c$
- **Corners:**  
Ultra-sharp at axes
- **Sparsity:**  
Extremely high
- **Problem:**  
Non-convex!

## Sparsity Progression: $L_2 \rightarrow L_1 \rightarrow L_p$

### Theorem: Key Insight

As  $p$  decreases from 2 to 1 to  $p < 1$ :

- Constraint regions become more **pointed** at axes
- Probability of intersection at axes **increases**
- Sparsity **increases**
- Optimization difficulty **increases**

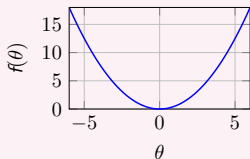
### Example: Why $p = 1$ is Special

- Still promotes sparsity (sharp corners)
- Remains convex (unlike  $p < 1$ ) and Computationally tractable
- Perfect balance of sparsity and solvability

## L2 vs L1: Gradient Behavior

### Key Points: L2 Penalty:

$$f(\theta) = \frac{1}{2}\theta^2$$

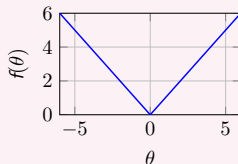


Gradient:  $\frac{df}{d\theta} = \theta$

Shrinks proportionally to current value

### Key Points: L1 Penalty:

$$f(\theta) = |\theta|$$



Subgradient:  $\text{sign}(\theta) = \pm 1$

Constant push toward zero

## L2 vs L1: Gradient Behavior

**Example: Example: Start at  $\theta = 5$**

**L2:**  $5 \rightarrow 2.5 \rightarrow 1.25 \rightarrow 0.625 \rightarrow \dots$  (never exactly zero)

**L1:**  $5 \rightarrow 4.5 \rightarrow 4.0 \rightarrow 3.5 \rightarrow \dots \rightarrow 0$  (reaches zero in finite steps)

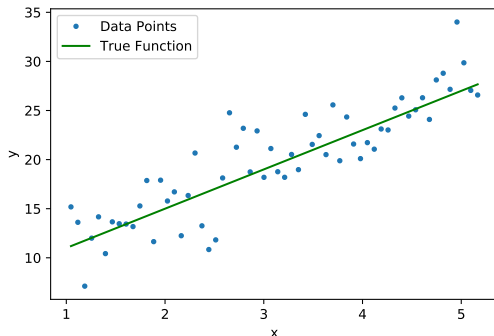


# Geometric Interpretation

# Sample Dataset for Demonstration

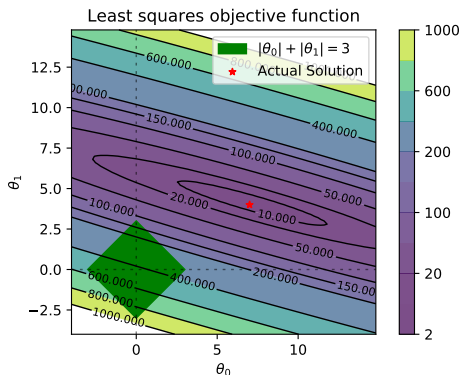
## Example: True Function

We'll demonstrate Lasso on a simple linear relationship:  $y = 4x + 7$



Sample data from  $y = 4x + 7$  with noise

# Geometric Interpretation: L1 vs L2 Constraints



L1 vs L2 constraint regions

## Key Points: Key Insight

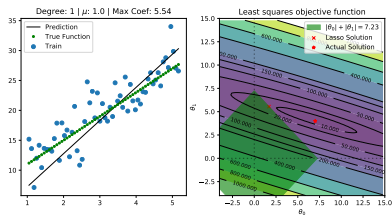
Diamond corners  $\Rightarrow$  exact zeros! Circle  $\Rightarrow$  no sparsity.

# Regularization Effects

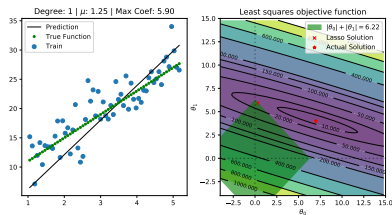
# Effect of $\lambda$ on Solution Path

## Important: Regularization Parameter

$\lambda$  controls fit vs sparsity trade-off

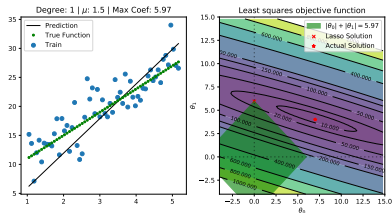


$\lambda = 1.0$  - Moderate

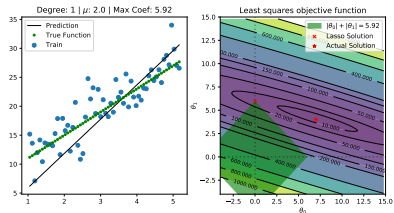


$\lambda = 1.25$  - Higher

# Increasing Regularization Strength



$\lambda = 1.5$  - Strong

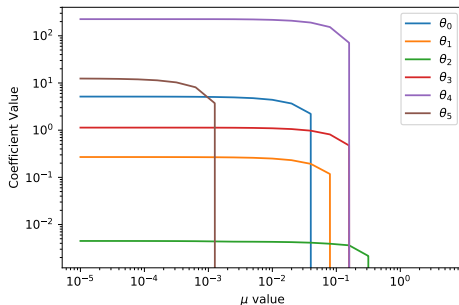


$\lambda = 2.0$  - Very strong

## Key Points: Observation

As  $\lambda$  increases  $\rightarrow$  more coefficients become exactly zero (automatic feature selection)

# Lasso Regularization Path



Coefficient values vs  $\lambda$

## Key Points: Key Observations

- Coefficients shrink to zero as  $\lambda$  increases
- Natural feature selection ordering

# Feature Selection Properties



# Lasso for Automatic Feature Selection

## Definition: Automatic Feature Selection

Lasso performs regression and feature selection simultaneously by setting irrelevant coefficients to exactly zero

## Key Points: Key Advantages

- **Sparsity:** Many coefficients  $\rightarrow$  exactly zero
- **Interpretability:** Understand which features matter
- **Efficiency:** Fewer parameters, faster prediction

# Subgradient Methods

# What is a Subgradient?

A subgradient generalizes the concept of gradient to convex but non-differentiable functions

## Example: Classic Example

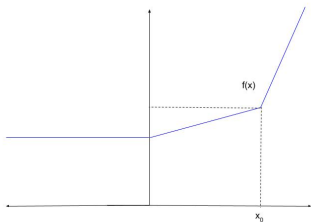
For  $f(x) = |x|$ :

- $f'(x) = 1$  when  $x > 0$
- $f'(x) = -1$  when  $x < 0$
- $f'(0)$  is undefined, but subgradient  $\in [-1, 1]$

## Important: Why Important for Lasso?

The L1 penalty  $|\theta_j|$  is non-differentiable at  $\theta_j = 0$

# Subgradient: Visual Intuition



Non-differentiable function at  $x_0$

## Important: Task

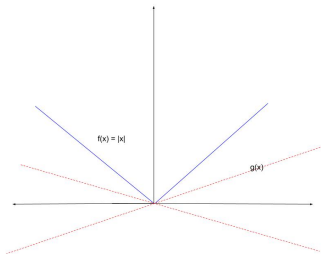
Find the "derivative" of  $f(x)$  at the non-differentiable point  $x = x_0$

## Construction

Find differentiable  $g(x)$  such that:

- $g(x_0) = f(x_0)$
- $g(x) \leq f(x)$  for all  $x$

# Subgradient of $|x|$ at $x = 0$



Supporting lines with slopes in  $[-1, 1]$

## Subgradient Set

For  $f(x) = |x|$  at  $x = 0$ :

$$\partial f(0) = [-1, 1]$$

## Key Points: Key Insight

Multiple supporting lines  
 $\Rightarrow$  set of valid subgradients

## Important: Lasso Connection

This subgradient concept is exactly what we need for the L1 penalty term!

# Coordinate Descent Algorithm

# Introduction to Coordinate Descent

## Definition: Coordinate Descent

Optimization method: minimize one coordinate at a time

## Key Points: Key Idea

- Hard: optimize all coordinates together
- Easy: optimize one coordinate at a time
- Perfect for non-differentiable Lasso!

## Algorithm Overview

$$\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) \text{ becomes } \min_{\theta_j} f(\theta_1, \dots, \theta_{j-1}, \theta_j, \theta_{j+1}, \dots, \theta_d)$$

# Coordinate Descent Properties

## Key Points: Advantages

- **No step-size:** Exact 1D minimization
- **Convergence:** Guaranteed for convex Lasso
- **Efficient:** Closed-form updates

## Selection Strategies

Cyclic, Random, or Greedy coordinate selection

## Important: Process

Cycle through coordinates, optimizing one at a time until convergence



# Worked Example

# Coordinate Descent Example Setup

Learn  $y = \theta_0 + \theta_1 x$  using coordinate descent on the dataset below

x	y
1	1
2	2
3	3

## Setup

- Initial parameters:  $(\theta_0, \theta_1) = (2, 3)$
- $$\text{MSE} = \frac{14 + 3\theta_0^2 + 14\theta_1^2 - 12\theta_0 - 28\theta_1 + 12\theta_0\theta_1}{3}$$
- Using standard least squares (no regularization for simplicity)

# Coordinate Descent Iterations

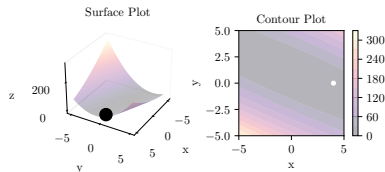
## Iteration 1:

INIT:  $\theta_0 = 2$  and  $\theta_1 = 3$

Fix  $\theta_1 = 3$ , optimize  $\theta_0$ :

$$\frac{\partial \text{MSE}}{\partial \theta_0} = 6\theta_0 + 24 = 0$$

$$\theta_0 = -4$$



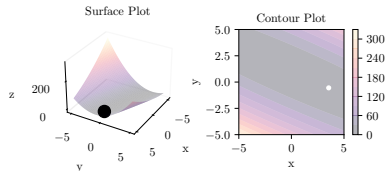
Starting point

## Iteration 2:

INIT:  $\theta_0 = -4$  and  $\theta_1 = 3$

Fix  $\theta_0 = -4$ , optimize  $\theta_1$ :

$$\theta_1 = 2.7$$



After 2 iterations

# Visual Coordinate Descent

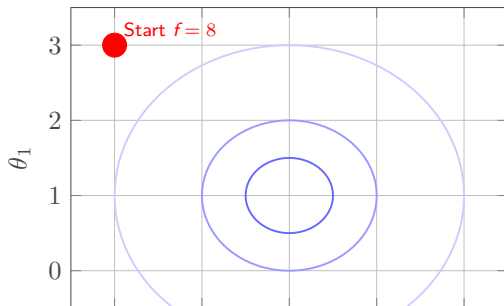
# Coordinate Descent: Setup

## Example: Problem

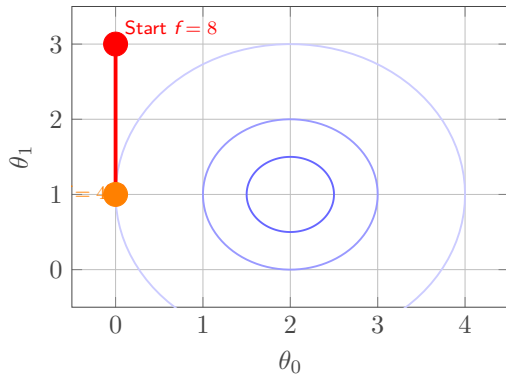
Minimize

$$f(\theta_0, \theta_1) = (\theta_0 - 2)^2 + (\theta_1 - 1)^2$$

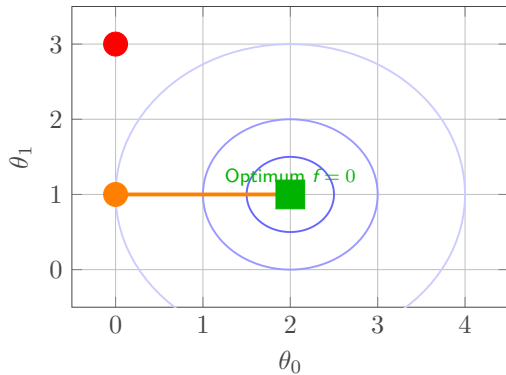
starting from  $(0, 3)$



# Coordinate Descent: Step 1



## Coordinate Descent: Step 2



# Failure of Coordinate Descent



# Mathematical Derivation

# Lasso Coordinate Descent: Setup

## Lasso Objective

$$\text{Minimize } \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=0}^d |\theta_j|$$

## Key Points: Key Definitions

- $\rho_j = \sum_{i=1}^n x_{ij}(y_i - \hat{y}_i^{(-j)})$  (partial residual correlation)
- $z_j = \sum_{i=1}^n x_{ij}^2$  (feature norm squared)
- $\hat{y}_i^{(-j)}$  = prediction without  $j$ -th feature

# Lasso Coordinate Descent: Setup

## Coordinate Update Rule

Fix all  $\theta_k$  for  $k \neq j$ , minimize w.r.t.  $\theta_j$ :

$$\min_{\theta_j} \sum_{i=1}^n (y_i - \hat{y}_i^{(-j)} - \theta_j x_{ij})^2 + \lambda |\theta_j|$$

# Subgradient Analysis

Subgradient of Lasso Objective w.r.t.  $\theta_j$

$$\frac{\partial}{\partial \theta_j}(\text{Lasso}) = -2\rho_j + 2\theta_j z_j + \lambda \frac{\partial}{\partial \theta_j} |\theta_j|$$

**Theorem: Subgradient of  $|\theta_j|$**

$$\frac{\partial}{\partial \theta_j} |\theta_j| = \begin{cases} +1 & \text{if } \theta_j > 0 \\ [-1, +1] & \text{if } \theta_j = 0 \\ -1 & \text{if } \theta_j < 0 \end{cases}$$

# Soft-Thresholding Solution

## Theorem: Complete Lasso Update Rule

$$\theta_j = \begin{cases} \frac{\rho_j + \lambda/2}{z_j} & \text{if } \rho_j < -\lambda/2 \\ 0 & \text{if } |\rho_j| \leq \lambda/2 \\ \frac{\rho_j - \lambda/2}{z_j} & \text{if } \rho_j > \lambda/2 \end{cases}$$

## Important: Sparsity Mechanism

If correlation  $|\rho_j| \leq \lambda/2$  is weak, set  $\theta_j = 0$ !

## Key Points: Soft-Thresholding Properties

- **Shrinkage:** Coefficients pulled toward zero
- **Selection:** Small coefficients  $\rightarrow$  exactly zero

# Lasso vs Ridge Comparison

# Lasso vs Ridge: Key Differences

Property	Ridge (L2)	Lasso (L1)
Penalty	$\sum \theta_j^2$	$\sum  \theta_j $
Sparsity	Never exactly zero	Can be exactly zero
Feature Selection	No	Yes
Differentiable	Yes	No (at $\theta_j = 0$ )
Solution Method	Closed form	Coordinate descent
Constraint Shape	Circle	Diamond
Best for	Multicollinearity	Feature selection

## Key Points: When to Use Each

**Lasso:** High-dimensional data, need interpretable model, expect few relevant features

**Ridge:** All features somewhat relevant, multicollinearity issues, want stable solution

# Summary and Applications



# Lasso Regression: Summary

## Theorem: Three-Part Understanding

**Visual:** L1 diamond constraint  $\rightarrow$  sparsity at sharp corners

**Algorithmic:** Coordinate descent + soft-thresholding  $\rightarrow$  exact zeros

**Mathematical:** Subgradients handle non-differentiability elegantly

## Key Points: Key Advantages

- Regression + feature selection simultaneously
- Sparse, interpretable models
- Handles high-dimensional data well

# Lasso Regression: Summary

## Key Points: Limitations

- Arbitrary selection among correlated features
- May underperform when all features are relevant

# Applications and Extensions

## Example: Real-World Applications

- **Genomics:** 20,000+ genes  $\rightarrow$  identify disease markers
- **Text Mining:** 100k+ words  $\rightarrow$  sentiment analysis features
- **Signal Processing:** Sparse signal reconstruction
- **Finance:** Risk factor selection from hundreds of indicators
- **Marketing:** Customer segmentation with key attributes

## Key Points: Extensions

- **Elastic Net:** Combines  $L1 + L2$  penalties
- **Group Lasso:** Selects groups of related features
- **Fused Lasso:** Enforces smoothness in ordered features