1 Problem Formulation

Input Space

$$\mathcal{X} \subseteq \mathbb{R}^p$$

Output Space

$$\mathcal{Y} \subseteq \mathbb{R}$$

Data Distribution

$$(x_i, y_i) \sim P(X, Y)$$

Learning Objective (Expected Risk Minimization)

Minimize expected squared loss:

$$R(\beta) = \mathbb{E}[(Y - X^T\beta)^2]$$

Empirical version:

$$\hat{R}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$

2 Model Specification

Hypothesis Function

$$h_\beta(x) = x^T\beta$$

Parameter Space

$$\beta \in \mathbb{R}^p$$

Structural Assumptions

- Linearity in parameters
- Homoscedastic errors
- Independence
- No perfect multicollinearity

3 Loss Function

Squared Loss

$$L(y, \hat{y}) = (y - \hat{y})^2$$

Why Squared Loss?

- Corresponds to Gaussian noise assumption
- Leads to closed-form solution
- Convex and differentiable

Convexity Proof

Squared loss Hessian:

$$\nabla^2 = 2X^T X$$

Since $X^T X$ is positive semi-definite → squared loss is convex.

4 Objective Function

Empirical Risk

$$J(\beta) = \|y - X\beta\|^2$$

Regularized (Ridge)

$$J(\beta) = \|y - X\beta\|^2 + \lambda\|\beta\|^2$$

---

5 Optimization Method

Closed Form (OLS)

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Gradient

$$\nabla J = -2X^T(y - X\beta)$$

Complexity

- Computing $X^T X$: $O(np^2)$
- Inversion: $O(p^3)$

6 Statistical Interpretation

Gaussian Noise Model

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$$

MLE

OLS = Maximum Likelihood under Gaussian noise.

7 Regularization & Generalization

Bias–Variance Tradeoff

- OLS: Low bias, high variance
- Ridge: Increased bias, reduced variance

Multicollinearity Effect

If predictors highly correlated → $X^T X$ nearly singular → inverse unstable → variance increas

Variance:

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Small eigenvalues → large variance.

8 Theoretical Properties

Unbiasedness Proof

$$\hat{\beta} = (X^T X)^{-1} X^T (X\beta + \epsilon)$$
$$= \beta + (X^T X)^{-1} X^T \epsilon$$

Taking expectation:

$$\mathbb{E}[\hat{\beta}] = \beta$$

Hence unbiased.

Covariance

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Gauss–Markov Theorem

OLS is BLUE (Best Linear Unbiased Estimator) if:

- Linearity
- Zero mean errors
- Constant variance
- No autocorrelation

Ridge Eigenvalue Shrinkage

Let:

$$X^T X = Q \Lambda Q^T$$

Ridge estimator:

$$(X^T X + \lambda I)^{-1} = Q(\Lambda + \lambda I)^{-1} Q^T$$

Each eigenvalue becomes:

$$\frac{1}{\lambda_i + \lambda}$$

→ Shrinks small eigenvalues.

Ridge vs Lasso Geometry

- Ridge → L2 ball (circular constraint)
- Lasso → L1 diamond (corners promote sparsity)

MAP Derivation

Assume prior:

$$\beta \sim N(0, \tau^2 I)$$

MAP gives Ridge.

Condition Number

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}$$

Large κ → unstable inverse.

When $p >> n$

- $X^T X$ singular
- Infinite solutions
- Need regularization

---

9 Computational Complexity

Stage	Complexity
Training	$O(np^2 + p^3)$
Inference	$O(p)$
Memory	$O(np)$

---

10 Limitations

- Sensitive to outliers
- Assumes linearity
- Fails with strong multicollinearity
- Poor when noise not Gaussian