

# Part I : Theory . Adrian Piñeda Sanchez A00839710

## Part I: Theory

### Problem 0: OLS as Orthogonal Projection and the Hat Matrix

Let  $\mathbf{X} \in \mathbb{R}^{n \times p}$  have full column rank and consider the OLS problem

$$\hat{\beta} = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\beta\|_2^2.$$

(a) **Normal equations.** Derive the normal equations  $\mathbf{X}^\top \mathbf{X} \hat{\beta} = \mathbf{X}^\top \mathbf{y}$  by differentiating the objective with respect to  $\beta$ .

$$\text{Let } f(\beta) = \|y - X\beta\|_2^2 = (y - X\beta)^\top (y - X\beta)$$

$$\text{Expand: } f(\beta) = y^\top y - 2\beta^\top X^\top y + \beta^\top X^\top X\beta$$

Differentiate and set to 0:

$$\nabla_{\beta} f(\beta) = -2X^\top y + 2X^\top X\beta = 0 \Rightarrow X^\top X\hat{\beta} = X^\top y$$

□

(b) **Orthogonality of residuals.** Let  $e = y - X\hat{\beta}$ . Prove that  $X^\top e = 0$ . Interpret this condition geometrically in words (1-3 sentences).

$$X^\top e = X^\top (y - X\hat{\beta}) = X^\top y - X^\top X\hat{\beta} = 0$$

Geometric meaning:  $\hat{y} = X\hat{\beta}$  is the orthogonal projection of  $y$  onto  $\text{Col}(X)$ , the residual  $e = y - \hat{y}$  lies in  $\text{Col}(X)^\perp$ , hence is the orthogonal to every column of  $X$ . □

(c) **Hat matrix.** Define the hat matrix  $\mathbf{H} := \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$  and  $\hat{y} := \mathbf{Hy}$ . Prove that  $\mathbf{H}$  is symmetric and idempotent, i.e.

$$\mathbf{H}^\top = \mathbf{H}, \quad \mathbf{H}^2 = \mathbf{H}.$$

Conclude that  $\hat{y}$  is the orthogonal projection of  $y$  onto  $\text{Col}(\mathbf{X})$ .

$$\mathbf{H} := \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top, \quad \hat{y} = \mathbf{Hy}$$

Symmetry:

$$H^T = (X(X^T X)^{-1} X^T)^T = X((X^T X)^{-1})^T X^T = X(X^T X)^{-1} X^T = H, \text{ since } X^T X$$

is symmetric hence  $(X^T X)^{-1}$  is symmetric

Dempotency:

$$H^2 = X(X^T X)^{-1} X^T X(X^T X)^{-1} X^T = X(X^T X)^{-1} (X^T X) (X^T X)^{-1} X^T = H$$

Therefore  $H$  is an orthogonal projector and  $\hat{y}$  is the orthogonal projection of  $y$  onto  $\text{Col}(X)$ .  $\square$

(d) One-line consequence. Show that if  $X$  includes an intercept column  $1$ , then the residuals satisfy

$$\sum_{i=1}^n e_i = 0.$$

(You should be able to do this in one line from part (b).)

If  $X$  contains an intercept column  $1$ , then from  $X^T e = 0$  we get

$$1^T e = 0 \Rightarrow \sum_{i=1}^n e_i = 0.$$

$\square$

### Problem 1: Ridge Regression via SVD (Matrix Form)

Consider the Ridge regression estimator:

$$\hat{\beta}_{\text{Ridge}}(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}.$$

Let  $\mathbf{X} = \mathbf{UDV}^\top$  be the SVD, where  $\mathbf{D} = \text{diag}(d_1, \dots, d_p)$ ,  $\mathbf{U} \in \mathbb{R}^{n \times p}$  has orthonormal columns, and  $\mathbf{V} \in \mathbb{R}^{p \times p}$  is orthogonal.

(a) Guided derivation (matrix form). Starting from the Ridge formula and using the SVD identities

$$\mathbf{X}^\top \mathbf{X} = \mathbf{V} \mathbf{D}^2 \mathbf{V}^\top, \quad \mathbf{X}^\top \mathbf{y} = \mathbf{V} \mathbf{D} \mathbf{U}^\top \mathbf{y},$$

show step by step that

$$\hat{\beta}_{\text{Ridge}}(\lambda) = \mathbf{V} (\mathbf{D}^2 + \lambda \mathbf{I})^{-1} \mathbf{D} \mathbf{U}^\top \mathbf{y}.$$

**Hint:** Use  $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$  and the fact that

$$(\mathbf{V} \mathbf{A} \mathbf{V}^\top)^{-1} = \mathbf{V} \mathbf{A}^{-1} \mathbf{V}^\top \quad \text{when } \mathbf{V} \text{ is orthogonal and } \mathbf{A} \text{ is invertible.}$$

$$\hat{\beta}_{\text{Ridge}}(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

Using the SVD identities  $\mathbf{X}^\top \mathbf{X} = \mathbf{V} \mathbf{D}^2 \mathbf{V}^\top$  and  $\mathbf{X}^\top \mathbf{y} = \mathbf{V} \mathbf{D} \mathbf{U}^\top \mathbf{y}$ ,

$$\hat{\beta} = (\mathbf{V} \mathbf{D}^2 \mathbf{V}^\top + \lambda \mathbf{I})^{-1} (\mathbf{V} \mathbf{D} \mathbf{U}^\top \mathbf{y})$$

Since  $\lambda \mathbf{I} = \mathbf{V} (\lambda \mathbf{I}) \mathbf{V}^\top$

$$\mathbf{V} \mathbf{D}^2 \mathbf{V}^\top + \lambda \mathbf{I} = \mathbf{V} (\mathbf{D}^2 + \lambda \mathbf{I}) \mathbf{V}^\top.$$

Using  $(\mathbf{V} \mathbf{A} \mathbf{V}^\top)^{-1} = \mathbf{V} \mathbf{A}^{-1} \mathbf{V}^\top$

$$\hat{\beta} = \mathbf{V} (\mathbf{D}^2 + \lambda \mathbf{I})^{-1} \mathbf{V}^\top \mathbf{V} \mathbf{D} \mathbf{U}^\top \mathbf{y} = \mathbf{V} (\mathbf{D}^2 + \lambda \mathbf{I})^{-1} \mathbf{D} \mathbf{U}^\top \mathbf{y},$$

because  $\mathbf{V}^\top \mathbf{V} = \mathbf{I}$



(b) Shrinkage intuition. Explain (in 4-8 lines) why Ridge shrinks directions associated with small singular values  $d_j$  the most. Connect your explanation to multicollinearity / ill-conditioning of  $\mathbf{X}^\top \mathbf{X}$ .

In the V-basis, ridge scales each singular direction  $v_i$  by:  $\frac{d_i}{d_i^2 + \lambda}$

when  $d_i$  is small, this factor becomes very small, so ridge heavily shrinks components along poorly-identified directions. Small  $d_i$  correspond to near-collinearity and/or ill-conditioning of  $\mathbf{X}^\top \mathbf{X}$ , where OLS would amplify noise and produce unstable (high-variance) coefficients. Adding  $\lambda \mathbf{I}$  improves conditioning and stabilizes the solution by shrinking those fragile directions the most.

**(Optional/Bonus) Expansion into singular components.** Starting from the matrix expression in part (a), expand  $\hat{\beta}_{\text{Ridge}}(\lambda)$  as a sum over singular directions (i.e., an expression involving  $d_j$ ,  $u_j^\top y$ , and  $v_j$ ).

$$\hat{\beta}_{\text{Ridge}}(\lambda) = \sum_{j=1}^p \frac{d_j}{d_j^2 + \lambda} (u_j^\top y) v_j.$$

$V(D^2 + \lambda I)^{-1} V^\top$  is a diagonal in the singular basis.

### Problem 2: Understanding LASSO Sparsity (Geometric View)

Recall that LASSO solves:

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1.$$

**(a) 1D warm-up (very short).** In the 1D case with a single parameter  $\beta$ , the objective has the form

$$L(\beta) = \underbrace{\frac{1}{2n} \sum_{i=1}^n (y_i - \beta x_i)^2}_{\text{a parabola in } \beta} + \underbrace{\lambda |\beta|}_{\text{a V-shape in } \beta}.$$

In 2-3 sentences (no calculus required), explain why increasing  $\lambda$  makes the minimizer more likely to be exactly  $\hat{\beta} = 0$ .

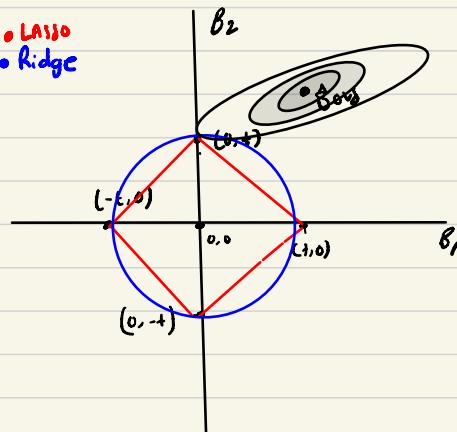
The squared-error term is a smooth parabola in  $\beta$ , while  $\lambda |\beta|$  is a V-shaped function with a sharp corner at  $\beta = 0$ . As  $\lambda$  increases, the V-shape becomes more influential and pulls the minimizer toward the corner. Because of the corner (non-differentiability) at zero, the optimum can land exactly at  $\hat{\beta} = 0$  over a range of  $\lambda$ .

**(b) Geometric interpretation in 2D.** Assume  $p = 2$  and consider the constrained form of LASSO:

$$\min_{\beta \in \mathbb{R}^2} \frac{1}{2n} \|y - X\beta\|_2^2 \quad \text{s.t.} \quad |\beta_1| + |\beta_2| \leq t.$$

- Sketch the feasible region  $|\beta_1| + |\beta_2| \leq t$  (a diamond in the  $(\beta_1, \beta_2)$  plane).
- Sketch a few elliptical contours of the RSS objective.
- Explain (4-8 lines) why the optimum often occurs at a *corner* of the diamond, implying that one coordinate is exactly zero.
- Briefly compare with Ridge, where the constraint is  $\beta_1^2 + \beta_2^2 \leq t$  (a circle). Why does Ridge typically *not* set coefficients exactly to zero?

- Lasso
- Ridge



The constrained optimum occurs where the smallest RSS ellipse first touches the feasible region. Because the L1-ball (diamond) has sharp corners aligned with the coordinate axes, the tangency often happens at corner. At a corner, one coordinate is exactly zero ( $B_j = 0$ ) which yields sparsity. In contrast, "ridge" uses the L2 ball  $B_1^2 + B_2^2 \leq t$ , which is smooth (a circle) so tangency typically happens off the axes and doesn't force coefficients to be exactly 0.