Part I : Theory. Adrian Pinedo 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

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

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

Let $f(\beta) = \|y - X\beta\|_2^2 = (y - X\beta)^T (y - X\beta)$

Expand: $f(\beta) = y^T y - 2\beta^T X^T y + \beta^T X^T X \beta$

Differentiate and set to 0:

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

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

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

Geometric meaning: $\hat{y} = X\hat{\beta}$ is the orthogonal projection of y onto $Col(X)$ the residual $e = y - \hat{y}$ lies in $Col(X)^\perp$, 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 $\hat{\mathbf{y}} := \mathbf{H}\mathbf{y}$. Prove that $\mathbf{H}$ is symmetric and idempotent, i.e.

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

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

$$H := X(X^T X)^{-1} X^T, \qquad \hat{y} = Hy$$

Symmetry:

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

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

Idempotent:

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

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

(d) One-line consequence. Show that if $\mathbf{X}$ includes an intercept column $\mathbf{1}$, 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 $X^Te = 0$ we get $1^Te = 0 \Rightarrow \sum_{i=1}^{n} e_i = 0$

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 $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ be the SVD, where $\mathbf{D} = \text{diag}(d_1, \ldots, d_p)$, $\mathbf{U} \in \mathbb{R}^{n \times p}$ 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

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

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

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

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

Using the SVD identities $X^T X = V D^2 V^T$ and $X^T y = V D U^T y$,

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

Since $\lambda I = V(\lambda I) V^T$ and Using $I = V V^T$

$$V D^2 V^T + \lambda I = V(D^2 + \lambda I) V^T. \longrightarrow$$

Since V is orthogonal $\longrightarrow$ Using $(VAV^T)^{-1} = V A^{-1} V^T$

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

because $V^T V = I$

□

(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_j by: $\frac{d_j}{d_j^2 + \lambda}$

when d_j is small, this factor becomes very small, so ridge heavily shrinks components along poorly-identified directions. Small d_j correspond to near-collinearity and an ill-conditioned $X^T X$, where OLS would amplify noise and produce unstable (high-variance) coefficients. Adding λI 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, $\mathbf{u}_j^\top \mathbf{y}$, and $\mathbf{v}_j$).

From the previous result

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

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

where u_j and v_j are the j-columns of U and V, respectively.

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

Problem 2: Understanding LASSO Sparsity (Geometric View)

Recall that LASSO solves:

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

(a) 1D warm-up (very short). In the 1D case with a single parameter β, 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 λ makes the minimizer more likely to be exactly $\hat{\beta} = 0$.

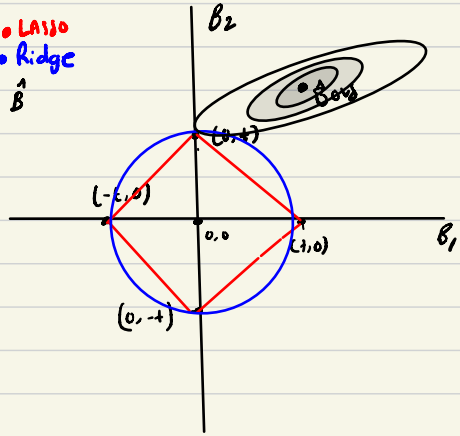
The squared-error term is a smooth parabola in β, while $\lambda|\beta|$ is a V-shaped function with a sharp corner at $\beta = 0$. As λ increases, the V-penalty becomes more influential and pulls the minimizer toward the corner. Because of the corner (non-differentiability) at zero, the optimum can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

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

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

- Sketch the feasible region $|\beta_1| + |\beta_2| \le t$ (a diamond in the (β_1, β_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 $\beta_1^2 + \beta_2^2 \le t$ (a circle). Why does Ridge typically *not* set coefficients exactly to zero?

Geometry of Ridge vs Lasso



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 ($\beta_2 = 0$) which yields sparsity. In contrast, "ridge" uses the L2 ball $\beta_1^2 + \beta_2^2 \leq t$, which is smooth (a circle) so tangency typically happens off the axes and doesn't force coefficients to be exactly 0.

the ellipses represent the contours of the least square error function.

- $\beta_1^2 + \beta_2^2 \leq t$
- $|\beta_1| + |\beta_2| \leq t$