Penalizations L^1 and L^2

1 Warm-up

Consider a model given by

$$Y = X\theta_\star + \varepsilon\,,$$

where $X \in \mathbb{R}^{n\times d}$ and $\varepsilon \sim \mathcal{N}(0, \sigma_\star^2 I_n)$. The Ridge estimator is defined for all $\lambda > 0$ by:

$$\widehat{\theta}_\lambda \in \mathrm{Argmin}_{\theta\in\mathbb{R}^d}\mathcal{L}(\theta) \quad \text{with} \quad \mathcal{L}(\theta) = \frac{1}{n}\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2\,.$$

For all $\lambda > 0$, the excess risk is given by

$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_\lambda) - \mathsf{R}(\theta_\star)\right] = \lambda^2\theta_\star^\top\left(\frac{1}{n}X^\top X + \lambda I_d\right)^{-2}\frac{1}{n}X^\top X\theta_\star + \frac{\sigma_\star^2}{n}\mathrm{Trace}\left((n^{-1}X^\top X)^2(n^{-1}X^\top X + \lambda I_d)^{-2}\right)\,.$$

1. Prove that
$$\mathbb{E}\left[\mathsf{R}(\widehat{\theta}_\lambda) - \mathsf{R}(\theta_\star)\right] \leqslant \frac{\lambda}{2}\|\theta_\star\|_2^2 + \frac{\sigma_\star^2}{2n^2\lambda}\mathrm{Trace}\left(X^\top X\right)\,.$$

2. Propose an "optimal" value for λ and compute the associated excess risk upper bound.

2 Elastic-Net

Consider a model given by

$$Y = X\theta_\star + \varepsilon\,,$$

where $X \in \mathbb{R}^{n\times d}$ and $\varepsilon \sim \mathcal{N}(0, \sigma_\star^2 I_n)$. The Elastic-Net estimator involves both L^1 and L^2 penalties. It is defined for all $\lambda, \mu > 0$ by:

$$\widehat{\theta}_{\lambda,\mu} \in \mathrm{Argmin}_{\theta\in\mathbb{R}^d}\mathcal{L}(\theta) \quad \text{with} \quad \mathcal{L}(\theta) = \|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2 + \mu\|\theta\|_1\,.$$

In the following, we assume that for all $1 \le j \le d$, the j-th column of X satisfies $\|\mathbf{X}_j\|_2 = 1$.

1. For all $1 \le j \le d$ provide the partial derivative of $\mathcal{L}$ with respect to θ_j for $\theta_j \neq 0$.

2. Provide an expression of the answer of the first question with $R_j(\theta) = \mathbf{X}_j^\top(Y - \sum_{k\neq j}\theta_k\mathbf{X}_k)$.

3. Assume that θ_k, $1 \le k \neq j \le d$ are fixed and assume that the minimum of $\theta_j \mapsto \mathcal{L}(\theta)$ is reached at a $\theta_j^\star \neq 0$. Prove that the sign of $\theta_j^\star$ is the same as the sign of R_j and conclude.

4. Provide an algorithm to obtain an approximation of $\widehat{\theta}_{\lambda,\mu}$.