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

   Proof in lecture notes.

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

   Proof in lecture notes.

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 $\theta_j \neq 0$.

   Note that for all $\theta \in \mathbb{R}^d$,
$$\nabla_\theta(\|Y - X\theta\|_2^2 + \lambda\|\theta\|_2^2) = 2X^\top X\theta - 2X^\top Y + 2\lambda\theta = 2X^\top\left(\sum_{k=1}^d \theta_k\mathbf{X}_k - Y\right) + 2\lambda\theta\,.$$

   Therefore, for $1 \le j \le d$ such that $\theta_j \neq 0$,
$$\partial_j\mathcal{L}(\theta) = 2\mathbf{X}_j^\top\left(\sum_{k=1}^d \theta_k\mathbf{X}_k - Y\right) + 2\lambda\theta_j + \mu\,\mathrm{sign}(\theta_j)\,.$$

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

Since $\|\mathbf{X}_j\|_2 = 1$, *for* $1 \leq j \leq d$ *such that* $\theta_j \neq 0$,

$$\begin{aligned}\partial_j \mathcal{L}(\theta) &= 2\theta_j - 2R_j(\theta) + 2\lambda\theta_j + \mu \mathrm{sign}(\theta_j)\\ &= 2\left((1+\lambda)\theta_j - R_j(\theta) + \frac{\mu}{2}\mathrm{sign}(\theta_j)\right)\,.\end{aligned}$$

3. Assume that θ_k, $1 \leq k \neq j \leq d$ 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 signe of R_j and conclude.

If the minimum of $\theta_j \mapsto \mathcal{L}(\theta)$ *is reached at some* $\theta_j^\star \neq 0$ *it means that* $\partial_j\mathcal{L}((\theta_1,\ldots,\theta_{j-1},\theta_j^\star,\theta_{j+1},\ldots,\theta_d)) = 0$. *Since*

$$\partial_j\mathcal{L}((\theta_1,\ldots,\theta_{j-1},\theta_j^\star,\theta_{j+1},\ldots,\theta_d)) = 2\left((1+\lambda)\theta_j^\star - R_j(\theta) + \frac{\mu}{2}\mathrm{sign}(\theta_j^\star)\right)\,,$$

$\theta_j^\star$ *and* $R_j(\theta)$ *have the same sign. Indeed, if* $\theta_j^\star > 0$ *and* $R_j(\theta) \leq 0$ *then* $\partial_j\mathcal{L}((\theta_1,\ldots,\theta_{j-1},\theta_j^\star,\theta_{j+1},\ldots,\theta_d)) > 0$ *and* $\theta_j^\star < 0$ *and* $R_j(\theta) \geq 0$ *then* $\partial_j\mathcal{L}((\theta_1,\ldots,\theta_{j-1},\theta_j^\star,\theta_{j+1},\ldots,\theta_d)) < 0$. *Therefore,*

$$\begin{aligned}\theta_j^\star &= \frac{R_j(\theta)}{1+\lambda}\left(1 - \frac{\mu\mathrm{sign}(\theta_j^\star)}{2R_j(\theta)}\right)\,,\\ &= \frac{R_j(\theta)}{1+\lambda}\left(1 - \frac{\mu}{2|R_j(\theta)|}\right)\,.\end{aligned}$$

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

The estimator $\widehat{\theta}_{\lambda,\mu}$ *can be approximated recursively coordinate by coordinate. Starting from a random vector, at each iteration, a coordinate* $1 \leq j \leq d$ *is chosen at random and we update* θ_j, *keeping all other coordinates fixed.*

- *Compute* $R_j(\theta)$.
- *If* $1 - \mu/(2|R_j(\theta)|) > 0$ *set* $\theta_j = \frac{R_j(\theta)}{1+\lambda}\left(1 - \frac{\mu}{2|R_j(\theta)|}\right)$.
- *If* $1 - \mu/(2|R_j(\theta)|) \leq 0$ *set* $\theta_j = 0$.