Let us now apply the results in the previous week to some concrete applications.

## Example 1 (LASSO with exactly sparse models)

Consider the standard linear model

$$y = X\theta^* + w, \quad y, w \in \mathbb{R}^n; \quad X \in \mathbb{R}^{n \times d}; \quad \theta^* \in \mathbb{R}^d.$$

We assume that the ground-truth $\theta^*$ is exactly sparse with sparsity $s$. The LASSO estimator takes the form

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \underbrace{\frac{1}{2n}\|y - X\theta\|_2^2}_{\mathcal{L}(\theta)} + \lambda \underbrace{\|\theta\|_1}_{R(\theta)} \right\}$$

By taking $S$ to be the support of $\theta^*$ (in particular, $|S| = s$), $\mathcal{C}$ takes the form

$$\mathcal{C} = \left\{ \Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1 \right\}.$$

Moreover, to show that $\mathcal{L}$ satisfies the RSC property, it suffices to prove that

$$\frac{\|X\Delta\|_2^2}{n} \geq \kappa \|\Delta\|_2^2 \qquad \forall \Delta \in \mathcal{C} \qquad \text{— (*)}$$

for some $\kappa > 0$. Before we do that, let us use (*) to obtain a bound on the statistical error of the LASSO estimate:

**Proposition 1:** Suppose that

(i) $X$ satisfies (*);

(ii) $\frac{\|X_j\|_2}{\sqrt{n}} \leq 1 \quad \forall j$, where $X_j$ is the $j^{\text{th}}$ column of $X$;

(iii) $w$ is sub-Gaussian with parameter $\sigma > 0$; i.e., $w$ has zero mean and for any fixed $v \in \mathbb{R}^n$ with $\|v\|_2 = 1$,

$$\Pr\left[|v^T w| \geq t\right] \leq 2\exp\left(-t^2/2\sigma^2\right) \quad \forall t > 0;$$

(iv) $\theta^*$ is $s$-sparse.

Then, by taking $\lambda = 4\sigma\sqrt{\frac{\log d}{n}}$ in the LASSO estimator, we have

$$\|\hat{\theta} - \theta^*\|_2^2 \leq 36 \frac{\sigma^2}{K^2} \cdot \frac{s \log d}{n}$$

with probability at least $1 - 2/d$.

Proof: Consider the model subspace

$$\mathcal{M} \equiv \mathcal{M}(S) = \bar{\mathcal{M}}(S) = \{\theta \in \mathbb{R}^d : \theta_j = 0 \ \forall j \notin S\},$$

where $S = \text{supp}(\theta^*)$. Then, we have $\theta^*_{\mathcal{M}^\perp} = 0$. By our general statistical error bound result (Theorem 1 in Week 2's notes), we need to bound $\Psi^2(\bar{\mathcal{M}})$ and $\lambda$. For the former, we have

$$\Psi^2(\bar{\mathcal{M}}) = \sup_{u \in \bar{\mathcal{M}} \setminus \{0\}} \frac{\|u\|_1^2}{\|u\|_2^2} = s$$

For the latter, recall that we need $\lambda \geq 2\mathcal{R}^*(\nabla \mathcal{L}(\theta^*))$. Since

$$\nabla \mathcal{L}(\theta^*) = \frac{1}{n} X^T(X\theta^* - y) = -\frac{1}{n} X^T w \quad \text{and} \quad \mathcal{R}^*(v) = \|v\|_\infty,$$

we need to bound $\|X^T w\|_\infty$. Towards that end, we utilize our statistical assumptions on the model. Specifically, we have

$$\Pr\left[\left|\frac{X_j^T w}{n}\right| \geq t\right] = \Pr\left[\left|\left(\frac{X_j}{\sqrt{n}}\right)^T w\right| \geq \sqrt{n}\, t\right]$$

$$\leq 2\exp\left(-t^2 n / 2\sigma^2\right) \quad \text{for } j = 1, \cdots, d.$$

Hence, by the union bound,

$$\Pr\left[\left\|\frac{X^T w}{n}\right\|_\infty \geq t\right] \leq 2d \cdot \exp\left(-t^2 n / 2\sigma^2\right).$$

Upon setting $t^2 = \frac{4\sigma^2 \log d}{n}$, the RHS equals $\frac{2}{d}$. It follows that with probability at least $1 - 2/d$, we have $\lambda \geq \frac{2}{n}\|X^T w\|_\infty$. □

Now, let us return to (*). Key to proving it is the following result:

Theorem 1: Suppose that the rows of $X$ are iid $\mathcal{N}(0,\Sigma)$ with $\Sigma \succ 0$. Then, there exist constants $c_1, c_2 > 0$ such that

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Sigma^{1/2}\Delta\|_2 - 9\rho(\Sigma)\sqrt{\frac{\log d}{n}}\|\Delta\|_1 \qquad \forall \Delta \in \mathbb{R}^d \qquad \text{—(**)}$$

with probability at least $1 - c_1 \exp(-c_2 n)$, where

$$\rho^2(\Sigma) = \max_{1 \leq j \leq d} \Sigma_{jj}.$$

Implications and Remarks

1) Note that $\forall \Delta \in \mathcal{C}$,

$$\|\Delta\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \leq 4\|\Delta_S\|_1 \leq 4\sqrt{s}\|\Delta_S\|_2 \leq 4\sqrt{s}\|\Delta\|_2.$$

Thus, if we take $\Sigma = I$ for simplicity, then Theorem 1 yields

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Delta\|_2 - 36\sqrt{\frac{s\log d}{n}}\|\Delta\|_2$$

In particular, as long as

$$k = \frac{1}{4} - 36\sqrt{\frac{s\log d}{n}} > 0 \iff n > 144^2 s \log d,$$

we have (*) holds.

2) Theorem 1 essentially says that the nullspace of $X$ cannot have vectors that are too sparse.

3) The subtlety in proving (**) is that it has to hold for all $\Delta \in \mathbb{R}^d$, not just a single $\Delta$. This requires tools from random matrix theory and empirical process theory.

## Example 2 (LASSO with weakly sparse models)

Let us now consider the setup of Example 1, except that $\theta^*$ is not necessarily exactly sparse but can be well approximated by a sparse vector. To formalize this, one way is to stipulate that

$$\theta^* \in B_q(R_q) \triangleq \left\{ \theta \in \mathbb{R}^d : \|\theta\|_q^q \le R_q \right\}, \quad q \in [0,1]$$

When $q=0$, $\theta^*$ is exactly sparse with sparsity $R_0$. For $q \in (0,1]$, the ball $B_q(R_q)$ imposes a certain decay rate on the ordered absolute values of $\theta^*$.

In this weakly sparse case, $\mathcal{C}$ takes the form

$$\mathcal{C} = \left\{ \Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1 + 4\|\theta^*_{S^c}\|_1 \right\}, \quad S \subseteq \{1, \cdots, d\}$$

Note that $\mathcal{C}$ contains a ball centered at the origin. Hence, it is impossible to have a $\kappa > 0$ such that

$$\frac{\|X\Delta\|_2^2}{2n} \ge \kappa \|\Delta\|_2^2 \qquad \forall \Delta \in \mathcal{C}.$$

This is why in the definition of RSC, we need the extra tolerance term $\tau(\cdot)$.

Similar to Proposition 1, we have the following statistical error bound for the weakly sparse case:

**Proposition 2:** Suppose that

(i) There exist constants $\kappa_1, \kappa_2 > 0$ such that

$$\frac{\|X\Delta\|_2^2}{n} \ge \kappa_1 \|\Delta\|_2^2 - \kappa_2 \frac{\log d}{n} \|\Delta\|_1^2 \qquad \forall \Delta \in \mathbb{R}^d, \quad \text{---(***)}$$

(ii) $\frac{\|X_j\|_2}{\sqrt{n}} \le 1 \quad \forall j$, where $X_j$ is the $j^{\text{th}}$ column of $X$;

(iii) $w$ is sub-Gaussian with parameter $\sigma > 0$;

(iv) $\theta^* \in B_q(R_q)$, where $\sqrt{R_q}\left(\frac{\log d}{n}\right)^{\frac{1}{2} - \frac{q}{4}} \le 1$.

Then, by taking $\lambda = 4\sigma\sqrt{\frac{\log d}{n}}$ in the LASSO estimator, we have

$$\|\hat{\theta} - \theta^*\|_2^2 \leq \left[17 + 128\frac{K_2}{K_1}\left(\frac{4\sigma}{K_1}\right)^{-8}\right]\cdot\left(\frac{16\sigma^2}{K_1^2}\right)^{1-8/2} R_8 \cdot \left(\frac{\log d}{n}\right)^{1-8/2}$$

with probability at least $1 - 2/d$.

Proof: For a threshold $\eta > 0$, define

$$S_\eta = \{ j : |\theta_j^*| > \eta \}.$$

and consider the model subspace

$$\mathcal{M} \equiv \mathcal{M}(S_\eta) = \bar{\mathcal{M}}(S_\eta) = \{ \theta \in \mathbb{R}^d : \theta_j = 0 \ \forall j \notin S_\eta \}.$$

Claim 2: Under the assumptions of Proposition 2 and the assumption that $n \geq 64\frac{K_2}{K_1}|S_\eta|\log d$, the following RSC property of $\theta \mapsto \mathcal{L}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$ holds:

$$\mathcal{L}(\theta^* + \Delta) \geq \mathcal{L}(\theta^*) + \nabla\mathcal{L}(\theta^*)^T\Delta + \frac{K_1}{2}\|\Delta\|_2^2 - \overbrace{32K_2\frac{\log d}{n}\|\theta^*_{S_\eta^c}\|_1^2}^{\tau^2(\theta^*)}$$

$$\forall \Delta \in \mathbb{C}.$$

Assuming the claim, by invoking our general statistical error bound result (Theorem 1 in Week 2's notes), we get

$$\|\hat{\theta} - \theta^*\|_2^2 \leq \frac{9\lambda^2}{K_1^2}|S_\eta| + \frac{4}{K_1}\left(32K_2\frac{\log d}{n}\|\theta^*_{S_\eta^c}\|_1^2 + 2\lambda\|\theta^*_{S_\eta^c}\|_1\right), \quad -(\Delta)$$

where we use the facts that $\Psi^2(\bar{\mathcal{M}}) = |S_\eta|$ and $\lambda \geq \frac{2}{n}\|X^T w\|_\infty$ with probability at least $1 - 2/d$ by our calculation in the proof of Proposition 1. To obtain the final result, it remains to bound $|S_\eta|$ and $\|\theta^*_{S_\eta^c}\|_1$.

First, observe that

$$R_8 \geq \|\theta^*\|_8^8 \geq \sum_{j \in S_\eta} |\theta_j^*|^8 \geq \eta^8|S_\eta|.$$

It follows that

$$|S_\eta| \leq \eta^{-8} R_8 \quad \text{for any } \eta > 0.$$

Next, we have

$$\|\theta^*_{S_\eta^c}\|_1 = \sum_{j \in S_\eta^c}|\theta_j^*| = \sum_{j \in S_\eta^c}|\theta_j^*|^8|\theta_j^*|^{1-8} \leq R_8\,\eta^{1-8}.$$

Substituting these into $(\Delta)$ and setting $\eta = \frac{\lambda}{K_1}$ yields

$$\|\hat{\theta} - \theta^*\|_2^2 \le \frac{9\lambda^2}{K_1^2}\eta^{-q} R_q + \frac{8\lambda}{K_1}\cdot\eta^{1-q} R_q + 128\,\frac{K_2}{K_1}\,\frac{\log d}{n}\,\eta^{2-2q} R_q^2$$

$$= 17\left(\frac{\lambda^2}{K_1^2}\right)^{1-q/2} R_q + 128\,\frac{K_2}{K_1}\left(\frac{\lambda^2}{K_1^2}\right)^{1-q/2} R_q \cdot \left[R_q \frac{\log d}{n}\left(\frac{\lambda}{K_1}\right)^{-q}\right]$$

By our choice of $R_q$ and $\lambda$, we have

$$R_q \frac{\log d}{n}\left(\frac{\lambda}{K_1}\right)^{-q} = \left(\frac{4\sigma}{K_1}\right)^{-q} \cdot \underbrace{R_q\left(\frac{\log d}{n}\right)^{1-q/2}}_{\le 1} \le \left(\frac{4\sigma}{K_1}\right)^{-q}$$

It follows that

$$\|\hat{\theta} - \theta^*\|_2^2 \le \left[17 + 128\,\frac{K_2}{K_1}\left(\frac{4\sigma}{K_1}\right)^{-q}\right] R_q \left(\frac{\lambda^2}{K_1^2}\right)^{1-q/2}$$

$$= \left[17 + 128\,\frac{K_2}{K_1}\left(\frac{4\sigma}{K_1}\right)^{-q}\right] \cdot \left(\frac{16\sigma^2}{K_1^2}\right)^{1-q/2} \cdot R_q \cdot \left(\frac{\log d}{n}\right)^{1-q/2}. \quad \square$$

Proof of Claim 2: For any $\Delta \in \mathcal{C}$, we have

$$\|\Delta\|_1 \le \|\Delta_{S_\eta}\|_1 + \|\Delta_{S_\eta^c}\|_1 \le 4\|\Delta_{S_\eta}\|_1 + 4\|\theta^*_{S_\eta^c}\|_1.$$

Hence, using the inequality $(a+b)^2 \le 2(a^2+b^2)$ $\forall a, b \ge 0$, we get

$$\|\Delta\|_1^2 \le 32\left(\|\Delta_{S_\eta}\|_1^2 + \|\theta^*_{S_\eta^c}\|_1^2\right)$$

$$\le 32\left(|S_\eta| \cdot \|\Delta_{S_\eta}\|_2^2 + \|\theta^*_{S_\eta^c}\|_1^2\right)$$

$$\le 32|S_\eta| \cdot \|\Delta\|_2^2 + 32\|\theta^*_{S_\eta^c}\|_1^2.$$

This, together with (***), yields

$$\frac{\|X\Delta\|_2^2}{n} \ge \left(K_1 - 32K_2\frac{\log d}{n}|S_\eta|\right)\|\Delta\|_2^2 - 32K_2\frac{\log d}{n}\|\theta^*_{S_\eta^c}\|_1^2 \quad \forall \Delta \in \mathcal{C}.$$

Our choice of $n$ implies that $K_1 - 32K_2\frac{\log d}{n}|S_\eta| \ge \frac{K_1}{2}$. This completes the proof. $\square$