Lecture 3

July 1, 2025

Restricted null space property (RNS)

- Define

$$\mathbb{C}(S) := \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}.$$

Theorem

The following two are equivalent:

- *For any $\theta^* \in \mathbb{R}^d$ with support $\subset S$, the basis pursuit program applied to the data $(X, y = X\theta^*)$ has unique solution $\hat{\theta} = \theta^*$.*
- *The restricted null space (RNS) property holds, i.e.,*

$$\mathbb{C}(S) \cap \ker(X) = \{0\}.$$

Recall that $\ker(X) = \{\theta \in \mathbb{R}^d : X\theta = 0\}$.

Sufficient conditions for restricted nullspace

- $[d] = \{1, \ldots, d\}$.

Sufficient conditions for restricted nullspace

- $[d] = \{1, \ldots, d\}$.
- For a matrix $X \in \mathbb{R}^{n\times}$ let X_j be its jth column for $j \in [d]$.

Sufficient conditions for restricted nullspace

- $[d] = \{1, \ldots, d\}$.
- For a matrix $X \in \mathbb{R}^{n\times}$ let X_j be its jth column for $j \in [d]$.
- The pairwise incoherence of X is defined as

$$\delta_{PW}(X) = \max_{i,j\in[d]} \left| \frac{\langle X_i, X_j \rangle}{n} - 1_{\{i=j\}} \right|.$$

Sufficient conditions for restricted nullspace

- $[d] = \{1, \ldots, d\}$.
- For a matrix $X \in \mathbb{R}^{n\times}$ let X_j be its jth column for $j \in [d]$.
- The pairwise incoherence of X is defined as

$$\delta_{PW}(X) \;=\; \max_{i,j\in[d]} \left| \frac{\langle X_i, X_j \rangle}{n} - 1_{\{i=j\}} \right|.$$

- If $\delta_{PW}(X)$ is small then

$$\langle X_i/\sqrt{n}, X_j/\sqrt{n} \rangle \;\approx\; \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{otherwise.} \end{cases}$$

- Alternative form: $X^\top X$ is the Gram matrix of X.

- Alternative form: $X^\top X$ is the Gram matrix of X.
- $(X^\top X)_{i,j} = \langle X_i, X_j \rangle$, and

$$\delta_{PW}(X) = \|\frac{X^\top X}{n} - I_d\|_\infty$$

where $\|\cdot\|_\infty$ is the vector ℓ_∞ norm of a matrix.

- Alternative form: $X^\top X$ is the Gram matrix of X.
- $(X^\top X)_{i,j} = \langle X_i, X_j \rangle$, and

$$\delta_{PW}(X) = \|\frac{X^\top X}{n} - I_d\|_\infty$$

  where $\|\cdot\|_\infty$ is the vector ℓ_∞ norm of a matrix.
- **Proposition.** (Uniform) restricted nullspace holds for all S with $|S| \leq s$ if

$$\delta_{PW}(X) \leq \frac{1}{3s}.$$

- **Proof.** For a vector $\theta \in \mathbb{R}^d$ and set $S \subset \{1, \ldots, d\}$, let $\theta_S \in \mathbb{R}^d$ be given as

$$(\theta_S)_i = \begin{cases} \theta_i & \text{if } \; i \in S \\ 0 & \text{Otherwise.} \end{cases}$$

- **Proof.** For a vector $\theta \in \mathbb{R}^d$ and set $S \subset \{1, \ldots, d\}$, let $\theta_S \in \mathbb{R}^d$ be given as

$$(\theta_S)_i = \begin{cases} \theta_i & \text{if } \ i \in S \\ 0 & \text{Otherwise.} \end{cases}$$

Notice that

$$\left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 \;=\; \theta_S^\top \frac{X^\top X}{n} \theta_S$$

- **Proof.** For a vector $\theta \in \mathbb{R}^d$ and set $S \subset \{1, \ldots, d\}$, let $\theta_S \in \mathbb{R}^d$ be given as

$$(\theta_S)_i = \begin{cases} \theta_i & \text{if } i \in S \\ 0 & \text{Otherwise.} \end{cases}$$

Notice that

$$\begin{aligned} \left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \theta_S^\top \frac{X^\top X}{n} \theta_S \\ &= \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_S + \|\theta_S\|_2^2 \end{aligned}$$

- **Proof.** For a vector $\theta \in \mathbb{R}^d$ and set $S \subset \{1, \ldots, d\}$, let $\theta_S \in \mathbb{R}^d$ be given as

$$(\theta_S)_i = \begin{cases} \theta_i & \text{if } \; i \in S \\ 0 & \text{Otherwise.} \end{cases}$$

Notice that

$$\begin{aligned} \left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \theta_S^\top \frac{X^\top X}{n} \theta_S \\ &= \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_S + \|\theta_S\|_2^2 \\ &\geq -\left\| \frac{X^\top X}{n} - I_d \right\|_\infty \cdot \|\theta_S\|_1^2 + \|\theta_S\|_2^2 \end{aligned}$$

- **Proof.** For a vector $\theta \in \mathbb{R}^d$ and set $S \subset \{1, \ldots, d\}$, let $\theta_S \in \mathbb{R}^d$ be given as

$$(\theta_S)_i = \begin{cases} \theta_i & \text{if } \; i \in S \\ 0 & \text{Otherwise.} \end{cases}$$

Notice that

$$\begin{aligned}
\left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \theta_S^\top \frac{X^\top X}{n} \theta_S \\
&= \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_S + \|\theta_S\|_2^2 \\
&\geq -\left\| \frac{X^\top X}{n} - I_d \right\|_\infty \cdot \|\theta_S\|_1^2 + \|\theta_S\|_2^2 \\
&\geq -\delta_{PW}(X) \cdot \|\theta_S\|_1^2 + \|\theta_S\|_2^2
\end{aligned}$$

- **Proof.** For a vector $\theta \in \mathbb{R}^d$ and set $S \subset \{1, \ldots, d\}$, let $\theta_S \in \mathbb{R}^d$ be given as

$$(\theta_S)_i = \begin{cases} \theta_i & \text{if } \ i \in S \\ 0 & \text{Otherwise.} \end{cases}$$

Notice that

$$\begin{aligned}
\left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \theta_S^\top \frac{X^\top X}{n} \theta_S \\
&= \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_S + \|\theta_S\|_2^2 \\
&\geq -\left\| \frac{X^\top X}{n} - I_d \right\|_\infty \cdot \|\theta_S\|_1^2 + \|\theta_S\|_2^2 \\
&\geq -\delta_{PW}(X) \cdot \|\theta_S\|_1^2 + \|\theta_S\|_2^2 \\
&\geq -\delta_{PW}(X) \cdot s\|\theta_S\|_2^2 + \|\theta_S\|_2^2
\end{aligned}$$

- **Proof.** For a vector $\theta \in \mathbb{R}^d$ and set $S \subset \{1, \ldots, d\}$, let $\theta_S \in \mathbb{R}^d$ be given as

$$(\theta_S)_i = \begin{cases} \theta_i & \text{if } i \in S \\ 0 & \text{Otherwise.} \end{cases}$$

Notice that

$$\begin{aligned} \left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \theta_S^\top \frac{X^\top X}{n} \theta_S \\ &= \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_S + \|\theta_S\|_2^2 \\ &\geq -\left\| \frac{X^\top X}{n} - I_d \right\|_\infty \cdot \|\theta_S\|_1^2 + \|\theta_S\|_2^2 \\ &\geq -\delta_{PW}(X) \cdot \|\theta_S\|_1^2 + \|\theta_S\|_2^2 \\ &\geq -\delta_{PW}(X) \cdot s\|\theta_S\|_2^2 + \|\theta_S\|_2^2 \\ &\geq \|\theta_S\|_2^2 (1 - s\delta_{PW}(X)) \end{aligned}$$

- **Proof.** For a vector $\theta \in \mathbb{R}^d$ and set $S \subset \{1, \ldots, d\}$, let $\theta_S \in \mathbb{R}^d$ be given as

$$(\theta_S)_i = \begin{cases} \theta_i & \text{if } \ i \in S \\ 0 & \text{Otherwise.} \end{cases}$$

Notice that

$$\begin{aligned} \left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \theta_S^\top \frac{X^\top X}{n} \theta_S \\ &= \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_S + \|\theta_S\|_2^2 \\ &\geq -\left\| \frac{X^\top X}{n} - I_d \right\|_\infty \cdot \|\theta_S\|_1^2 + \|\theta_S\|_2^2 \\ &\geq -\delta_{PW}(X) \cdot \|\theta_S\|_1^2 + \|\theta_S\|_2^2 \\ &\geq -\delta_{PW}(X) \cdot s\|\theta_S\|_2^2 + \|\theta_S\|_2^2 \\ &\geq \|\theta_S\|_2^2 (1 - s\delta_{PW}(X)) \end{aligned}$$

where the first inequality follows from the ineqaulity $u^T M v \leq \|M\|_\infty \|u\|_1 \|v\|_1$, and the third from the inequality $\|\theta_S\|_1 \leq \sqrt{s}\|\theta_S\|_2$.

- On the other hand, if $X\theta = 0$ then $X(\theta_S + \theta_{S^c}) = 0$ or $X\theta_S = -X\theta_{S^c}$.

- On the other hand, if $X\theta = 0$ then $X(\theta_S + \theta_{S^c}) = 0$ or $X\theta_S = -X\theta_{S^c}$. Thus,

$$\left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 = \left| \left\langle \frac{X\theta_S}{\sqrt{n}}, \frac{-X\theta_{S^c}}{\sqrt{n}} \right\rangle \right|$$

- On the other hand, if $X\theta = 0$ then $X(\theta_S + \theta_{S^c}) = 0$ or $X\theta_S = -X\theta_{S^c}$. Thus,

$$\begin{aligned}
\left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \left| \left\langle \frac{X\theta_S}{\sqrt{n}}, \frac{-X\theta_{S^c}}{\sqrt{n}} \right\rangle \right| \\
&= \left| \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_{S^c} + \theta_S^\top \theta_{S^c} \right|
\end{aligned}$$

- On the other hand, if $X\theta = 0$ then $X(\theta_S + \theta_{S^c}) = 0$ or $X\theta_S = -X\theta_{S^c}$. Thus,

$$
\begin{aligned}
\left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \left| \left\langle \frac{X\theta_S}{\sqrt{n}}, \frac{-X\theta_{S^c}}{\sqrt{n}} \right\rangle \right| \\
&= \left| \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_{S^c} + \theta_S^\top \theta_{S^c} \right| \\
&= \left| \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_{S^c} \right|
\end{aligned}
$$

- On the other hand, if $X\theta = 0$ then $X(\theta_S + \theta_{S^c}) = 0$ or $X\theta_S = -X\theta_{S^c}$. Thus,

$$
\begin{aligned}
\left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \left| \left\langle \frac{X\theta_S}{\sqrt{n}}, \frac{-X\theta_{Sc}}{\sqrt{n}} \right\rangle \right| \\
&= \left| \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_{S^c} + \theta_S^\top \theta_{S^c} \right| \\
&= \left| \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_{S^c} \right| \\
&\leq \delta_{PW}(X) \cdot \|\theta_S\|_1 \cdot \|\theta_{S^c}\|_1
\end{aligned}
$$

- On the other hand, if $X\theta = 0$ then $X(\theta_S + \theta_{S^c}) = 0$ or $X\theta_S = -X\theta_{S^c}$. Thus,

$$
\begin{aligned}
\left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \left| \left\langle \frac{X\theta_S}{\sqrt{n}}, \frac{-X\theta_{S^c}}{\sqrt{n}} \right\rangle \right| \\
&= \left| \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_{S^c} + \theta_S^\top \theta_{S^c} \right| \\
&= \left| \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_{S^c} \right| \\
&\leq \delta_{PW}(X) \cdot \|\theta_S\|_1 \cdot \|\theta_{S^c}\|_1 \\
&\leq \delta_{PW}(X) \cdot \sqrt{s} \|\theta_S\|_2 \cdot \|\theta_{S^c}\|_1
\end{aligned}
$$

- On the other hand, if $X\theta = 0$ then $X(\theta_S + \theta_{S^c}) = 0$ or $X\theta_S = -X\theta_{S^c}$. Thus,

$$
\begin{aligned}
\left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \left| \left\langle \frac{X\theta_S}{\sqrt{n}}, \frac{-X\theta_{S^c}}{\sqrt{n}} \right\rangle \right| \\
&= \left| \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_{S^c} + \theta_S^\top \theta_{S^c} \right| \\
&= \left| \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_{S^c} \right| \\
&\leq \delta_{PW}(X) \cdot \|\theta_S\|_1 \cdot \|\theta_{S^c}\|_1 \\
&\leq \delta_{PW}(X) \cdot \sqrt{s} \|\theta_S\|_2 \cdot \|\theta_{S^c}\|_1
\end{aligned}
$$

- Therefore

$$
\|\theta_S\|_2^2 (1 - s\delta_{PW}(X)) \leq \left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 \leq \delta_{PW}(X) \cdot \sqrt{s} \|\theta_S\|_2 \cdot \|\theta_{S^c}\|_1 .
$$

- On the other hand, if $X\theta = 0$ then $X(\theta_S + \theta_{S^c}) = 0$ or $X\theta_S = -X\theta_{S^c}$. Thus,

$$
\begin{aligned}
\left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 &= \left| \left\langle \frac{X\theta_S}{\sqrt{n}}, \frac{-X\theta_{S^c}}{\sqrt{n}} \right\rangle \right| \\
&= \left| \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_{S^c} + \theta_S^\top \theta_{S^c} \right| \\
&= \left| \theta_S^\top \left(\frac{X^\top X}{n} - I_d \right) \theta_{S^c} \right| \\
&\leq \delta_{PW}(X) \cdot \|\theta_S\|_1 \cdot \|\theta_{S^c}\|_1 \\
&\leq \delta_{PW}(X) \cdot \sqrt{s} \|\theta_S\|_2 \cdot \|\theta_{S^c}\|_1
\end{aligned}
$$

- Therefore

$$
\|\theta_S\|_2^2 (1 - s\delta_{PW}(X)) \leq \left\| \frac{X\theta_S}{\sqrt{n}} \right\|_2^2 \leq \delta_{PW}(X) \cdot \sqrt{s} \|\theta_S\|_2 \cdot \|\theta_{S^c}\|_1 .
$$

- Hence,

$$
\frac{1}{\sqrt{s}} \|\theta_S\|_1 \leq \|\theta_S\|_2 \leq \frac{\sqrt{s}\,\delta_{PW}(X)}{(1 - s\delta_{PW}(X))} \|\theta_{S^c}\|_1
$$

- And so

$$\|\theta_S\|_1 \;\leq\; \frac{s\delta_{PW}(X)}{(1 - s\delta_{PW}(X))}\|\theta_{S^c}\|_1.$$

- And so
$$\|\theta_S\|_1 \;\leq\; \frac{s\delta_{PW}(X)}{(1 - s\delta_{PW}(X))}\|\theta_{S^c}\|_1.$$

- Thus, we have proven that if $\delta_{PW}(X) \;\leq\; \frac{1}{3s}$ then $\theta \in \mathrm{Ker}(X)$ implies
$$\|\theta_S\|_1 \;\leq\; \frac{1}{2}\|\theta_{S^c}\|_1,$$

- And so
$$\|\theta_S\|_1 \;\leq\; \frac{s\delta_{PW}(X)}{(1 - s\delta_{PW}(X))}\|\theta_{S^c}\|_1.$$

- Thus, we have proven that if $\delta_{PW}(X) \;\leq\; \frac{1}{3s}$ then $\theta \in \mathrm{Ker}(X)$ implies
$$\|\theta_S\|_1 \;\leq\; \frac{1}{2}\|\theta_{S^c}\|_1,$$
which means $\mathbb{C}(S) \cap \mathrm{Ker}(X) = \{0\}$, since
$$\mathbb{C}(S) \;:=\; \{\Delta \in \mathbb{R}^d \;:\; \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}$$

- And so

$$\|\theta_S\|_1 \;\leq\; \frac{s\delta_{PW}(X)}{(1 - s\delta_{PW}(X))}\|\theta_{S^c}\|_1.$$

- Thus, we have proven that if $\delta_{PW}(X) \;\leq\; \frac{1}{3s}$ then $\theta \in \mathrm{Ker}(X)$ implies

$$\|\theta_S\|_1 \;\leq\; \frac{1}{2}\|\theta_{S^c}\|_1,$$

which means $\mathbb{C}(S) \cap \mathrm{Ker}(X) = \{0\}$, since

$$\mathbb{C}(S) \;:=\; \{\Delta \in \mathbb{R}^d \;:\; \|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1\}$$

and so $\Delta \in \mathbb{C}(S) \cap \mathrm{Ker}(X)$ would imply $\|\Delta_{S^c}\|_1 \leq \|\Delta_S\|_1$ and $\|\Delta_S\|_1 \leq \|\Delta_{S^c}\|_1/2$.

- **Definition.** $X \in \mathbb{R}^{n \times d}$ satisfies a restricted isometry property (RIP) of order s with constant $\delta_s(X) > 0$ if

$$\left\|\frac{X_S^\top X_S}{n} - I\right\|_{op} \le \delta_s(X), \quad \text{for all} \quad S \ \text{ with } |S| \le s.$$

- **Definition.** $X \in \mathbb{R}^{n \times d}$ satisfies a restricted isometry property (RIP) of order s with constant $\delta_s(X) > 0$ if

$$\|\frac{X_S^\top X_S}{n} - I\|_{op} \leq \delta_s(X), \quad \text{for all} \quad S \text{ with } |S| \leq s.$$

- PW incoherence is close to RIP with $s = 2$.

- **Definition.** $X \in \mathbb{R}^{n\times d}$ satisfies a restricted isometry property (RIP) of order s with constant $\delta_s(X) > 0$ if

$$\|\frac{X_S^\top X_S}{n} - I\|_{op} \le \delta_s(X), \quad \text{for all} \quad S \text{ with } |S| \le s.$$

- PW incoherence is close to RIP with $s = 2$.
- In general, for any $s \ge 2$, it holds that

$$\delta_{PW}(X) \le \delta_s(X) \le s\delta_{PW}(X).$$

RIP gives sufficient conditions:

- **Proposition (HDS Prop. 7.2).** (Uniform) restricted null space holds for all S with $|S| \le s$ if

$$\delta_{2s}(X) \le \frac{1}{3}.$$

RIP gives sufficient conditions:

- **Proposition (HDS Prop. 7.2).** (Uniform) restricted null space holds for all S with $|S| \leq s$ if

$$\delta_{2s}(X) \leq \frac{1}{3}.$$

- Compare this with the condition $\delta_{PW}(X) \leq \frac{1}{3s}$.

RIP gives sufficient conditions:

- **Proposition (HDS Prop. 7.2).** (Uniform) restricted null space holds for all S with $|S| \leq s$ if

$$\delta_{2s}(X) \leq \frac{1}{3}.$$

- Compare this with the condition $\delta_{PW}(X) \leq \frac{1}{3s}$.
- Consider a sub-Gaussian matrix X with i.i.d. entries and $\mathbb{E}(X_{ij}) = 0$ and $\mathbb{E}(X_{ij}^2) = 1$ (Exercise 7.7):
  - We have that

$$n \gtrsim s^2 \log d \implies \delta_{PW}(X) \leq \tfrac{1}{3s}, \; w.h.p.$$

RIP gives sufficient conditions:

- **Proposition (HDS Prop. 7.2).** (Uniform) restricted null space holds for all S with $|S| \leq s$ if

$$\delta_{2s}(X) \leq \frac{1}{3}.$$

- Compare this with the condition $\delta_{PW}(X) \leq \frac{1}{3s}$.
- Consider a sub-Gaussian matrix X with i.i.d. entries and $\mathbb{E}(X_{ij}) = 0$ and $\mathbb{E}(X_{ij}^2) = 1$ (Exercise 7.7):
  - We have that

$$n \gtrsim s^2 \log d \implies \delta_{PW}(X) \leq \tfrac{1}{3s}, \ w.h.p.$$
$$n \gtrsim s \log(\tfrac{ed}{s}) \implies \delta_s(X) \leq \tfrac{1}{3}, \ w.h.p.$$

RIP gives sufficient conditions:

- **Proposition (HDS Prop. 7.2).** (Uniform) restricted null space holds for all S with $|S| \leq s$ if

$$\delta_{2s}(X) \leq \frac{1}{3}.$$

- Compare this with the condition $\delta_{PW}(X) \leq \frac{1}{3s}$.
- Consider a sub-Gaussian matrix X with i.i.d. entries and $\mathbb{E}(X_{ij}) = 0$ and $\mathbb{E}(X_{ij}^2) = 1$ (Exercise 7.7):
  - We have that

$$\begin{aligned} n \gtrsim s^2 \log d &\implies \delta_{PW}(X) \leq \tfrac{1}{3s}, \ w.h.p. \\ n \gtrsim s \log(\tfrac{ed}{s}) &\implies \delta_s(X) \leq \tfrac{1}{3}, \ w.h.p. \end{aligned}$$

  - Sample complexity requirement for RIP is milder.

Noisy sparse regression

- Recall the model
$$y = X\theta^* + w$$
where θ^* is the parameter of interest and w is a vector of noise errors.

Noisy sparse regression

- Recall the model
$$y = X\theta^* + w$$
where θ^* is the parameter of interest and w is a vector of noise errors.
- A very popular estimator is the ℓ_1-regularized least squares:
$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right\}. \tag{1}$$

Noisy sparse regression

- Recall the model
$$y = X\theta^* + w$$
where θ^* is the parameter of interest and w is a vector of noise errors.
- A very popular estimator is the ℓ_1-regularized least squares:
$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right\}. \tag{1}$$
- The idea: minimizing ℓ_1 norm leads to sparse solutions.

Noisy sparse regression

- Recall the model

$$y = X\theta^* + w$$

  where θ^* is the parameter of interest and w is a vector of noise errors.
- A very popular estimator is the ℓ_1-regularized least squares:

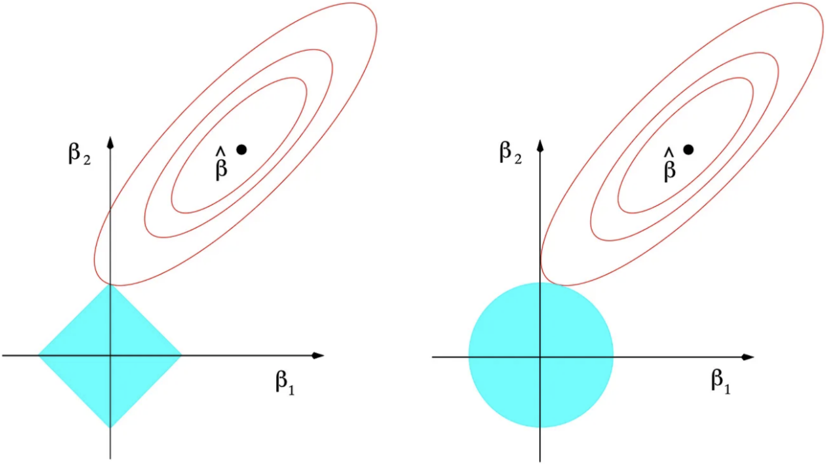
$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right\}. \tag{1}$$

- The idea: minimizing ℓ_1 norm leads to sparse solutions.
- (1) is a convex program; global solution can be obtained efficiently.

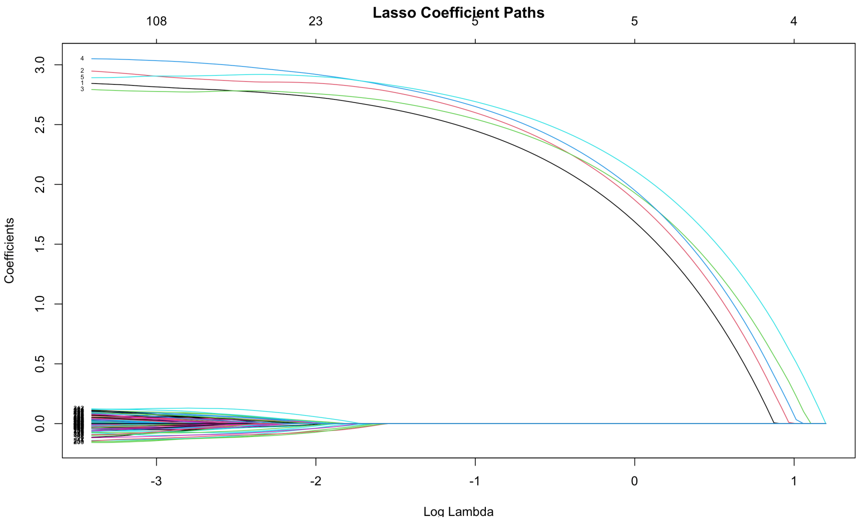
Noisy sparse regression

- Recall the model
$$y = X\theta^* + w$$
where θ^* is the parameter of interest and w is a vector of noise errors.
- A very popular estimator is the ℓ_1-regularized least squares:
$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda \|\theta\|_1 \right\}. \tag{1}$$
- The idea: minimizing ℓ_1 norm leads to sparse solutions.
- (1) is a convex program; global solution can be obtained efficiently.
- Other options: constrained form of lasso
$$\min_{\|\theta\|_1 \leq R} \frac{1}{2n} \|y - X\theta\|_2^2$$

Noisy sparse regression

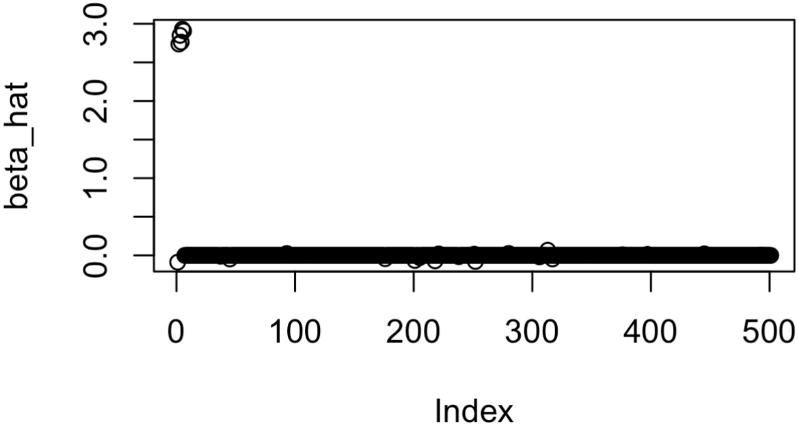


```r
library(glmnet)

# Simulate data
set.seed(42)
n <- 200  # number of observations
p <- 500   # number of predictors

X <- matrix(rnorm(n * p), nrow = n, ncol = p)
beta <- c(rep(3, 5), rep(0, p - 5))  # sparse true coefficients
y <- X %*% beta + rnorm(n)

# Fit Lasso regression (alpha=1 for Lasso)
lasso_fit <- glmnet(X, y, alpha = 1)

# Plot solution paths
plot(lasso_fit, xvar = "lambda", label = TRUE)
title("Lasso Coefficient Paths")

# Cross-validation to select best lambda
cv_fit <- cv.glmnet(X, y, alpha = 1)
plot(cv_fit)
title("Cross-Validation Error")
```

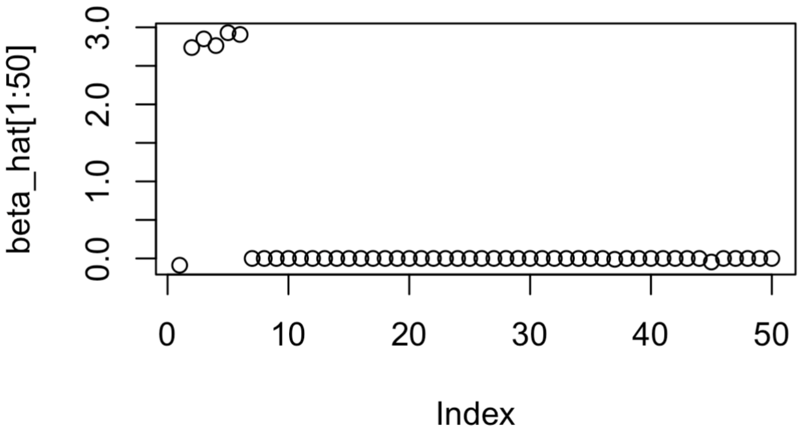


Lasso Coefficient Paths
108
23
5
5
4
Coefficients
3.0
2.5
2.0
1.5
1.0
0.5
0.0
4
2
5
1
3
-3
-2
-1
0
1
Log Lambda

```r
# Cross-validation to select best lambda
cv_fit <- cv.glmnet(X, y, alpha = 1)
plot(cv_fit)
title("Cross-Validation Error")

# Best lambda value
best_lambda <- cv_fit$lambda.min
cat("Best lambda:", best_lambda, "\n")

# Coefficients at best lambda
beta_hat = coef(cv_fit, s = "lambda.min")
plot(beta_hat)
plot(beta_hat[1:50])
```



beta_hat
0.0
1.0
2.0
3.0
0
100
200
300
400
500
Index


beta_hat[1:50]
0.0
1.0
2.0
3.0
0
10
20
30
40
50
Index

Restricted eigenvalue condition

- For a constant $\alpha \geq 1$, we define

$$\mathbb{C}_\alpha(S) := \left\{ \Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1 \right\}.$$

Restricted eigenvalue condition

- For a constant $\alpha \geq 1$, we define

$$\mathbb{C}_\alpha(S) := \left\{ \Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1 \right\}.$$

- **Definition.** A matrix X satisfies the restricted eigenvalue (RE) condition over S with parameters (κ, α) if

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \kappa \|\Delta\|_2^2 \quad \forall\, \Delta \in \mathbb{C}_\alpha(S).$$

Restricted eigenvalue condition

- For a constant $\alpha \geq 1$, we define

$$\mathbb{C}_\alpha(S) := \left\{ \Delta \in \mathbb{R}^d \,:\, \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1 \right\}.$$

- **Definition.** A matrix X satisfies the restricted eigenvalue (RE) condition over S with parameters (κ, α) if

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \kappa \|\Delta\|_2^2 \quad \forall\, \Delta \in \mathbb{C}_\alpha(S).$$

Recall that RNS corresponds to $\mathbb{C}_1(S) \cap \ker(X) = \{0\}$. Thus,

$$\frac{1}{n}\|X\Delta\|_2^2 > 0$$

for all $\Delta \in \mathbb{C}_1(S)\backslash\{0\}$.

Deviation bounds under RE

Theorem

Assume that $y = X\theta^* + w$*, where* $X \in \mathbb{R}^{n\times d}$*,* $\theta^* \in \mathbb{R}^d$ *and*

Deviation bounds under RE

Theorem

Assume that $y = X\theta^* + w$*, where* $X \in \mathbb{R}^{n\times d}$*,* $\theta^* \in \mathbb{R}^d$ *and*

- θ^* *is supported on* $S \subset [d]$ *with* $|S| \leq s$*.*

Deviation bounds under RE

Theorem

Assume that $y = X\theta^* + w$, where $X \in \mathbb{R}^{n \times d}$, $\theta^* \in \mathbb{R}^d$ and

- *θ^* is supported on $S \subset [d]$ with $|S| \leq s$.*
- *X satisfies $RE(\kappa, 3)$ over S.*

Deviation bounds under RE

Theorem

Assume that $y = X\theta^* + w$, *where* $X \in \mathbb{R}^{n\times d}$, $\theta^* \in \mathbb{R}^d$ *and*

- θ^* *is supported on* $S \subset [d]$ *with* $|S| \le s$.
- X *satisfies RE*$(\kappa, 3)$ *over* S.

Let us define $z = X^\top w/n$. *Then we have the following:*

Deviation bounds under RE

Theorem

Assume that $y = X\theta^* + w$*, where* $X \in \mathbb{R}^{n \times d}$, $\theta^* \in \mathbb{R}^d$ *and*

- θ^* *is supported on* $S \subset [d]$ *with* $|S| \leq s$.
- X *satisfies* $RE(\kappa, 3)$ *over* S.

Let us define $z = X^\top w/n$*. Then we have the following:*

- *Any solution of Lasso (1) with* $\lambda \geq 2\|z\|_\infty$ *satisfies*

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda.$$

Deviation bounds under RE

Theorem

Assume that $y = X\theta^* + w$, where $X \in \mathbb{R}^{n \times d}$, $\theta^* \in \mathbb{R}^d$ and

- *θ^* is supported on $S \subset [d]$ with $|S| \leq s$.*
- *X satisfies RE$(\kappa, 3)$ over S.*

Let us define $z = X^\top w/n$. Then we have the following:

- *Any solution of Lasso (1) with $\lambda \geq 2\|z\|_\infty$ satisfies*

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda.$$

- *Any solution of constrained Lasso with $R = \|\theta^*\|_1$ satisfies*

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{4}{\kappa}\sqrt{s}\|z\|_\infty.$$

Example (fixed design regression)

- Assume that $y = X\theta^* + w$, where $w \sim N(0, \sigma^2 I_n)$.

Example (fixed design regression)

- Assume that $y = X\theta^* + w$, where $w \sim N(0, \sigma^2 I_n)$.
- $X \in \mathbb{R}^{n \times d}$ fixed and satisfying RE condition and normalization

$$\max_{j=1,\dots,d} \frac{\|X_j\|}{\sqrt{n}} \leq C$$

where X_j is the jth column of X.

Example (fixed design regression)

- Assume that $y = X\theta^* + w$, where $w \sim N(0, \sigma^2 I_n)$.
- $X \in \mathbb{R}^{n \times d}$ fixed and satisfying RE condition and normalization

$$\max_{j=1,\dots,d} \frac{\|X_j\|}{\sqrt{n}} \leq C$$

  where X_j is the jth column of X.
- Recall $z = X^\top w/n$.

Example (fixed design regression)

- Assume that $y = X\theta^* + w$, where $w \sim N(0, \sigma^2 I_n)$.
- $X \in \mathbb{R}^{n \times d}$ fixed and satisfying RE condition and normalization
$$\max_{j=1,\dots,d} \frac{\|X_j\|}{\sqrt{n}} \leq C$$
where X_j is the jth column of X.
- Recall $z = X^\top w/n$.
- It is easy to show that w.p. $\geq 1 - 2\exp(-n\delta^2/2)$
$$\|z\|_\infty \leq C\sigma \left(\sqrt{\frac{2\log d}{n}} + \delta \right)$$

Example (fixed design regression)

- Assume that $y = X\theta^* + w$, where $w \sim N(0, \sigma^2 I_n)$.
- $X \in \mathbb{R}^{n\times d}$ fixed and satisfying RE condition and normalization

$$\max_{j=1,\ldots,d} \frac{\|X_j\|}{\sqrt{n}} \leq C$$

  where X_j is the jth column of X.
- Recall $z = X^\top w/n$.
- It is easy to show that w.p. $\geq 1 - 2\exp(-n\delta^2/2)$

$$\|z\|_\infty \leq C\sigma \left(\sqrt{\frac{2\log d}{n}} + \delta \right)$$

- Thus, setting $\lambda = 2C\sigma \left(\sqrt{\frac{2\log d}{n}} + \delta \right)$, Lasso solution satisfies

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{6C\sigma}{\kappa}\sqrt{s} \left(\frac{2\log d}{n} + \delta \right)$$

  w.p at least $1 - 2\exp(-n\delta^2/2)$.

- Taking $\delta = \sqrt{2 \log d / n}$, we have

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \sigma \sqrt{\frac{s \log d}{n}}$$

w.h.p. (i.e., at least $1 - 2d^{-1}$).

- Taking $\delta = \sqrt{2 \log d/n}$, we have

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \sigma \sqrt{\frac{s \log d}{n}}$$

  w.h.p. (i.e., at least $1 - 2d^{-1}$).
- This is the typical high-dimensional scaling in sparse problems.

- Taking $\delta = \sqrt{2 \log d / n}$, we have

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \sigma \sqrt{\frac{s \log d}{n}}$$

  w.h.p. (i.e., at least $1 - 2d^{-1}$).
- This is the typical high-dimensional scaling in sparse problems.
- Had we known the support S in advance, our rate would be (w.h.p.)

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \sigma \sqrt{\frac{s}{n}}$$

- Taking $\delta = \sqrt{2 \log d / n}$, we have

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \sigma \sqrt{\frac{s \log d}{n}}$$

  w.h.p. (i.e., at least $1 - 2d^{-1}$).
- This is the typical high-dimensional scaling in sparse problems.
- Had we known the support S in advance, our rate would be (w.h.p.)

$$\|\hat{\theta} - \theta^*\|_2 \lesssim \sigma \sqrt{\frac{s}{n}}$$

- The $\log d$ factor is the price for not knowing the support.

Proof of Theorem

- Let us simplify the loss $L(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$.

Proof of Theorem

- Let us simplify the loss $L(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$.
- Setting $\Delta = \theta - \theta^*$,

$$L(\theta) = \frac{1}{2n}\|X(\theta - \theta^*) - w\|_2^2$$

Proof of Theorem

- Let us simplify the loss $L(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$.
- Setting $\Delta = \theta - \theta^*$,

$$
\begin{aligned}
L(\theta) &= \frac{1}{2n}\|X(\theta - \theta^*) - w\|_2^2 \\
&= \frac{1}{2n}\|X\Delta - w\|_2^2
\end{aligned}
$$

Proof of Theorem

- Let us simplify the loss $L(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$.
- Setting $\Delta = \theta - \theta^*$,

$$
\begin{aligned}
L(\theta) &= \frac{1}{2n}\|X(\theta - \theta^*) - w\|_2^2 \\
&= \frac{1}{2n}\|X\Delta - w\|_2^2 \\
&= \frac{1}{2n}\|X\Delta\|_2^2 - \frac{1}{n}\langle X\Delta, w\rangle + \text{const.}
\end{aligned}
$$

Proof of Theorem

- Let us simplify the loss $L(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$.
- Setting $\Delta = \theta - \theta^*$,

$$\begin{aligned} L(\theta) &= \frac{1}{2n}\|X(\theta - \theta^*) - w\|_2^2 \\ &= \frac{1}{2n}\|X\Delta - w\|_2^2 \\ &= \frac{1}{2n}\|X\Delta\|_2^2 - \frac{1}{n}\langle X\Delta, w\rangle + \text{const.} \\ &= \frac{1}{2n}\|X\Delta\|_2^2 - \frac{1}{n}\langle \Delta, X^\top w\rangle + \text{const.} \end{aligned}$$

Proof of Theorem

- Let us simplify the loss $L(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$.
- Setting $\Delta = \theta - \theta^*$,

$$
\begin{aligned}
L(\theta) &= \frac{1}{2n}\|X(\theta - \theta^*) - w\|_2^2 \\
&= \frac{1}{2n}\|X\Delta - w\|_2^2 \\
&= \frac{1}{2n}\|X\Delta\|_2^2 - \frac{1}{n}\langle X\Delta, w\rangle + \text{const.} \\
&= \frac{1}{2n}\|X\Delta\|_2^2 - \frac{1}{n}\langle \Delta, X^\top w\rangle + \text{const.} \\
&= \frac{1}{2n}\|X\Delta\|_2^2 - \langle \Delta, z\rangle + \text{const.}
\end{aligned}
$$

where $z = X^\top w/n$.

Proof of Theorem

- Let us simplify the loss $L(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$.
- Setting $\Delta = \theta - \theta^*$,

$$
\begin{aligned}
L(\theta) &= \frac{1}{2n}\|X(\theta - \theta^*) - w\|_2^2 \\
&= \frac{1}{2n}\|X\Delta - w\|_2^2 \\
&= \frac{1}{2n}\|X\Delta\|_2^2 - \frac{1}{n}\langle X\Delta, w\rangle + \text{const.} \\
&= \frac{1}{2n}\|X\Delta\|_2^2 - \frac{1}{n}\langle \Delta, X^\top w\rangle + \text{const.} \\
&= \frac{1}{2n}\|X\Delta\|_2^2 - \langle \Delta, z\rangle + \text{const.}
\end{aligned}
$$

where $z = X^\top w/n$. Hence,

$$L(\theta) - L(\theta^*) := \frac{1}{2n}\|X\Delta\|^2 - \langle \Delta, z\rangle.$$

Proof (constrained version)

- By optimality of $\hat{\theta}$ and feasibility of θ^*,

$$L(\hat{\theta}) \leq L(\theta^*).$$

Proof (constrained version)

- By optimality of $\hat{\theta}$ and feasibility of θ^*,
$$L(\hat{\theta}) \leq L(\theta^*).$$

- Error vector $\hat{\Delta} = \hat{\theta} - \theta^*$ satisfies the basic inequality
$$\frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \langle \hat{\Delta}, z \rangle$$

Proof (constrained version)

- By optimality of $\hat{\theta}$ and feasibility of θ^*,

$$L(\hat{\theta}) \leq L(\theta^*).$$

- Error vector $\hat{\Delta} = \hat{\theta} - \theta^*$ satisfies the basic inequality

$$\frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \langle \hat{\Delta}, z \rangle$$

- Using Holder inequality

$$\frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \|z\|_\infty \|\hat{\Delta}\|_1.$$

Proof (constrained version)

- By optimality of $\hat{\theta}$ and feasibility of θ^*,

$$L(\hat{\theta}) \leq L(\theta^*).$$

- Error vector $\hat{\Delta} = \hat{\theta} - \theta^*$ satisfies the basic inequality

$$\frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \langle \hat{\Delta}, z \rangle$$

- Using Holder inequality

$$\frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \|z\|_\infty \|\hat{\Delta}\|_1.$$

- Since $\|\hat{\theta}\|_1 \leq \|\theta^*\|_1$, we have that $\hat{\Delta} = \hat{\theta} - \theta^* \in \mathbb{C}_1(S)$:

$$\|\theta^*_S\|_1 = \|\theta^*\|_1$$

Proof (constrained version)

- By optimality of $\hat{\theta}$ and feasibility of θ^*,

$$L(\hat{\theta}) \leq L(\theta^*).$$

- Error vector $\hat{\Delta} = \hat{\theta} - \theta^*$ satisfies the basic inequality

$$\frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \langle \hat{\Delta}, z \rangle$$

- Using Holder inequality

$$\frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \|z\|_\infty \|\hat{\Delta}\|_1.$$

- Since $\|\hat{\theta}\|_1 \leq \|\theta^*\|_1$, we have that $\hat{\Delta} = \hat{\theta} - \theta^* \in \mathbb{C}_1(S)$:

$$\|\theta^*_S\|_1 = \|\theta^*\|_1 \geq \|\theta^* + \hat{\Delta}\|_1$$

Proof (constrained version)

- By optimality of $\hat{\theta}$ and feasibility of θ^*,

$$L(\hat{\theta}) \leq L(\theta^*).$$

- Error vector $\hat{\Delta} = \hat{\theta} - \theta^*$ satisfies the basic inequality

$$\frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \langle \hat{\Delta}, z \rangle$$

- Using Holder inequality

$$\frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \|z\|_\infty \|\hat{\Delta}\|_1.$$

- Since $\|\hat{\theta}\|_1 \leq \|\theta^*\|_1$, we have that $\hat{\Delta} = \hat{\theta} - \theta^* \in \mathbb{C}_1(S)$:

$$\|\theta^*_S\|_1 \; = \; \|\theta^*\|_1 \; \geq \; \|\theta^* + \hat{\Delta}\|_1 \; = \; \|\theta^*_S + \hat{\Delta}_S\|_1 + \|\theta^*_{S^c} + \hat{\Delta}_{S^c}\|_1$$

Proof (constrained version)

- By optimality of $\hat{\theta}$ and feasibility of θ^*,

$$L(\hat{\theta}) \leq L(\theta^*).$$

- Error vector $\hat{\Delta} = \hat{\theta} - \theta^*$ satisfies the basic inequality

$$\frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \langle \hat{\Delta}, z \rangle$$

- Using Holder inequality

$$\frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \|z\|_\infty \|\hat{\Delta}\|_1.$$

- Since $\|\hat{\theta}\|_1 \leq \|\theta^*\|_1$, we have that $\hat{\Delta} = \hat{\theta} - \theta^* \in \mathbb{C}_1(S)$:

$$\|\theta^*_S\|_1 = \|\theta^*\|_1 \geq \|\theta^* + \hat{\Delta}\|_1 = \|\theta^*_S + \hat{\Delta}_S\|_1 + \|\theta^*_{S^c} + \hat{\Delta}_{S^c}\|_1$$
$$= \|\theta^*_S + \hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1$$

Proof (constrained version)

- By optimality of $\hat{\theta}$ and feasibility of θ^*,

$$L(\hat{\theta}) \leq L(\theta^*).$$

- Error vector $\hat{\Delta} = \hat{\theta} - \theta^*$ satisfies the basic inequality

$$\frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \langle \hat{\Delta}, z \rangle$$

- Using Holder inequality

$$\frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \|z\|_\infty \|\hat{\Delta}\|_1.$$

- Since $\|\hat{\theta}\|_1 \leq \|\theta^*\|_1$, we have that $\hat{\Delta} = \hat{\theta} - \theta^* \in \mathbb{C}_1(S)$:

$$\begin{aligned} \|\theta^*_S\|_1 \; = \; \|\theta^*\|_1 \; \geq \; \|\theta^* + \hat{\Delta}\|_1 \; = \; \|\theta^*_S + \hat{\Delta}_S\|_1 + \|\theta^*_{S^c} + \hat{\Delta}_{S^c}\|_1 \\ = \; \|\theta^*_S + \hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \; \geq \; \|\theta^*_S\|_1 \; - \; \|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1 \end{aligned}$$

- Since $\hat{\Delta} = \hat{\theta} - \theta^* \in \mathbb{C}_1(S)$,

$$\|\hat{\Delta}\|_1 = \|\hat{\Delta}_{S^c}\|_1 + \|\hat{\Delta}_S\|_1 \leq 2\|\hat{\Delta}_S\|_1 \leq 2\sqrt{s}\|\hat{\Delta}\|_2.$$

- Since $\hat{\Delta} = \hat{\theta} - \theta^* \in \mathbb{C}_1(S)$,

$$\|\hat{\Delta}\|_1 = \|\hat{\Delta}_{S^c}\|_1 + \|\hat{\Delta}_S\|_1 \leq 2\|\hat{\Delta}_S\|_1 \leq 2\sqrt{s}\|\hat{\Delta}\|_2.$$

- Combined with RE condition ($\hat{\Delta} \in \mathbb{C}_1(S) \subset \mathbb{C}_3(S)$ as well)

$$\frac{1}{2}\kappa\|\hat{\Delta}\|_2^2 \leq \frac{1}{2n}\|X\hat{\Delta}\|^2 \leq \|z\|_\infty\|\hat{\Delta}\|_1 \leq \|z\|_\infty 2\sqrt{s}\|\hat{\Delta}\|_2$$

which gives the desired result.

Proof (Lagrangian version)

- Let
$$L(\theta) := \frac{1}{2n}\|y - X\theta|_2^2 + \lambda\|\theta\|_1$$
be the regularized loss.

Proof (Lagrangian version)

- Let
$$L(\theta) := \frac{1}{2n}\|y - X\theta|_2^2 + \lambda\|\theta\|_1$$
be the regularized loss.
- Basic inequality is
$$L(\hat{\theta}) \leq L(\theta^*).$$

Proof (Lagrangian version)

- Let
$$L(\theta) := \frac{1}{2n}\|y - X\theta|_2^2 + \lambda\|\theta\|_1$$
be the regularized loss.
- Basic inequality is
$$L(\hat{\theta}) \le L(\theta^*).$$
- Rearranging
$$\frac{1}{2n}\|X\hat{\Delta}\|_2^2 \le \langle z, \hat{\Delta}\rangle + \lambda(\|\theta^*\|_1 - \|\hat{\theta}\|_1)$$

Proof (Lagrangian version)

- Let
$$L(\theta) := \frac{1}{2n}\|y - X\theta|_2^2 + \lambda\|\theta\|_1$$
be the regularized loss.
- Basic inequality is
$$L(\hat{\theta}) \leq L(\theta^*).$$
- Rearranging
$$\frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \langle z, \hat{\Delta}\rangle + \lambda(\|\theta^*\|_1 - \|\hat{\theta}\|_1)$$
- We have
$$\|\theta^*\|_1 - \|\hat{\theta}\|_1 = \|\theta^*_S\|_1 + \|\theta^*_{S^c}\|_1 - \|\theta^*_S + \hat{\Delta}_S\|_1 - \|\theta^*_{S^c} + \hat{\Delta}_{S^c}\|_1$$

Proof (Lagrangian version)

- Let
$$L(\theta) := \frac{1}{2n}\|y - X\theta|_2^2 + \lambda\|\theta\|_1$$
be the regularized loss.
- Basic inequality is
$$L(\hat{\theta}) \leq L(\theta^*).$$
- Rearranging
$$\frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \langle z, \hat{\Delta}\rangle + \lambda(\|\theta^*\|_1 - \|\hat{\theta}\|_1)$$
- We have
$$\begin{aligned} \|\theta^*\|_1 - \|\hat{\theta}\|_1 &= \|\theta^*_S\|_1 + \|\theta^*_{S^c}\|_1 - \|\theta^*_S + \hat{\Delta}_S\|_1 - \|\theta^*_{S^c} + \hat{\Delta}_{S^c}\|_1 \\ &= \|\theta^*_S\|_1 - \|\theta^*_S + \hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1 \end{aligned}$$

Proof (Lagrangian version)

- Let
$$L(\theta) := \frac{1}{2n}\|y - X\theta|_2^2 + \lambda\|\theta\|_1$$
be the regularized loss.
- Basic inequality is
$$L(\hat{\theta}) \leq L(\theta^*).$$
- Rearranging
$$\frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \langle z, \hat{\Delta}\rangle + \lambda(\|\theta^*\|_1 - \|\hat{\theta}\|_1)$$
- We have
$$\begin{aligned} \|\theta^*\|_1 - \|\hat{\theta}\|_1 &= \|\theta^*_S\|_1 + \|\theta^*_{S^c}\|_1 - \|\theta^*_S + \hat{\Delta}_S\|_1 - \|\theta^*_{S^c} + \hat{\Delta}_{S^c}\|_1 \\ &= \|\theta^*_S\|_1 - \|\theta^*_S + \hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1 \\ &\leq \|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1 \end{aligned}$$

- And so

$$\frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \langle z, \hat{\Delta}\rangle + \lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1).$$

- And so

$$\frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \langle z, \hat{\Delta}\rangle + \lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1).$$

- This implies, since $\lambda \geq 2\|z\|_\infty$,

$$\frac{1}{n}\|X\hat{\Delta}\|_2^2 \quad \leq \quad 2\langle z, \hat{\Delta}\rangle + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1)$$

- And so

$$\frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \langle z, \hat{\Delta}\rangle + \lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1).$$

- This implies, since $\lambda \geq 2\|z\|_\infty$,

$$\begin{aligned} \tfrac{1}{n}\|X\hat{\Delta}\|_2^2 &\leq 2\langle z, \hat{\Delta}\rangle + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\ &\leq 2\|z\|_\infty\|\hat{\Delta}\|_1 + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \end{aligned}$$

- And so

$$\frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \langle z, \hat{\Delta}\rangle + \lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1).$$

- This implies, since $\lambda \geq 2\|z\|_\infty$,

$$\begin{aligned}
\tfrac{1}{n}\|X\hat{\Delta}\|_2^2 &\leq 2\langle z, \hat{\Delta}\rangle + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\
&\leq 2\|z\|_\infty\|\hat{\Delta}\|_1 + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\
&\leq \lambda\|\hat{\Delta}\|_1 + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1)
\end{aligned}$$

- And so

$$\frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \langle z, \hat{\Delta}\rangle + \lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1).$$

- This implies, since $\lambda \geq 2\|z\|_\infty$,

$$\begin{aligned}
\tfrac{1}{n}\|X\hat{\Delta}\|_2^2 &\leq 2\langle z, \hat{\Delta}\rangle + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\
&\leq 2\|z\|_\infty\|\hat{\Delta}\|_1 + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\
&\leq \lambda\|\hat{\Delta}\|_1 + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\
&= \lambda(\|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1) + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1)
\end{aligned}$$

- And so

$$\frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \langle z, \hat{\Delta}\rangle + \lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1).$$

- This implies, since $\lambda \geq 2\|z\|_\infty$,

$$\begin{aligned}
\tfrac{1}{n}\|X\hat{\Delta}\|_2^2 &\leq 2\langle z, \hat{\Delta}\rangle + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\
&\leq 2\|z\|_\infty\|\hat{\Delta}\|_1 + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\
&\leq \lambda\|\hat{\Delta}\|_1 + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\
&= \lambda(\|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1) + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\
&= \lambda(3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1)
\end{aligned}$$

- And so

$$\frac{1}{2n}\|X\hat{\Delta}\|_2^2 \leq \langle z, \hat{\Delta}\rangle + \lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1).$$

- This implies, since $\lambda \geq 2\|z\|_\infty$,

$$\begin{aligned}
\frac{1}{n}\|X\hat{\Delta}\|_2^2 &\leq 2\langle z, \hat{\Delta}\rangle + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\
&\leq 2\|z\|_\infty\|\hat{\Delta}\|_1 + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\
&\leq \lambda\|\hat{\Delta}\|_1 + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\
&= \lambda(\|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1) + 2\lambda(\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1) \\
&= \lambda(3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1)
\end{aligned}$$

- We obtain that $\hat{\Delta} \in \mathbb{C}_3(S)$ ($\|\hat{\Delta}_{S^c}\|_1 \leq 3\|\hat{\Delta}_S\|_1$) and the rest of the proof follows.

Key condition: Restricted eigenvalue condition

- For a constant $\alpha \geq 1$, we define

$$\mathbb{C}_\alpha(S) := \left\{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq \alpha\|\Delta_S\|_1\right\}.$$

- **Definition.** A matrix X satisfies the restricted eigenvalue (RE) condition over S with parameters (κ, α) if

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2 \quad \forall\, \Delta \in \mathbb{C}_\alpha(S).$$

  Recall that RNS corresponds to $\mathbb{C}_1(S) \cap \ker(X) = \{0\}$. Thus,

$$\frac{1}{n}\|X\Delta\|_2^2 > 0$$

  for all $\Delta \in \mathbb{C}_1(S)\backslash\{0\}$.

Deviation bounds under RE

Theorem

Assume that $y = X\theta^* + w$, *where* $X \in \mathbb{R}^{n\times d}$, $\theta^* \in \mathbb{R}^d$ *and*

- θ^* *is supported on* $S \subset [d]$ *with* $|S| \leq s$.
- X *satisfies* $RE(\kappa, 3)$ *over* S.

Let us define $z = X^\top w/n$. *Then we have the following:*

- *Any solution of Lasso (1) with* $\lambda \geq 2\|z\|_\infty$ *satisfies*

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda.$$

- *Any solution of constrained Lasso with* $R = \|\theta^*\|_1$ *satisfies*

$$\|\hat{\theta} - \theta^*\|_2 \leq \frac{4}{\kappa}\sqrt{s}\|z\|_\infty.$$

RE condition for anisotropic design

- For a PSD matrix Σ, let $\rho^2(\Sigma) = \max_{i,i} \Sigma_{ii}$.

Theorem

Let $X \in \mathbb{R}^{n \times d}$ withrows i.i.d. from $N(0, \Sigma)$. Then, there exist universal constants $0 < c_1 < 1 < c_2$ such that

$$\frac{1}{n}\|X\theta\|_2^2 \geq c_1\|\sqrt{\Sigma}\theta\|_2^2 - c_2\rho^2(\Sigma)\frac{\log d}{n}\|\theta\|_1^2, \quad \forall \theta \in \mathbb{R}^d \tag{2}$$

with probability at least $1 - e^{-n/32}/(1 - e^{-n/32})$.

RE condition for anisotropic design

- For a PSD matrix Σ, let $\rho^2(\Sigma) = \max_{i,i} \Sigma_{ii}$.

Theorem

Let $X \in \mathbb{R}^{n \times d}$ withrows i.i.d. from $N(0, \Sigma)$. Then, there exist universal constants $0 < c_1 < 1 < c_2$ such that

$$\frac{1}{n}\|X\theta\|_2^2 \geq c_1\|\sqrt{\Sigma}\theta\|_2^2 - c_2\rho^2(\Sigma)\frac{\log d}{n}\|\theta\|_1^2, \quad \forall \theta \in \mathbb{R}^d \tag{2}$$

with probability at least $1 - e^{-n/32}/(1 - e^{-n/32})$.

- (2) implies RE condition over $\mathbb{C}_3(S)$ uniformly over all subsets of cardinality

$$|S| \leq \frac{c_1}{32c_2}\frac{\gamma_{\min}(\Sigma)}{\rho^2(\Sigma)}\frac{n}{\log d}.$$

RE condition for anisotropic design

- For a PSD matrix Σ, let $\rho^2(\Sigma) = \max_{i,i} \Sigma_{ii}$.

Theorem

Let $X \in \mathbb{R}^{n \times d}$ withrows i.i.d. from $N(0, \Sigma)$. Then, there exist universal constants $0 < c_1 < 1 < c_2$ such that

$$\frac{1}{n}\|X\theta\|_2^2 \geq c_1\|\sqrt{\Sigma}\theta\|_2^2 - c_2\rho^2(\Sigma)\frac{\log d}{n}\|\theta\|_1^2, \quad \forall \theta \in \mathbb{R}^d \tag{2}$$

with probability at least $1 - e^{-n/32}/(1 - e^{-n/32})$.

- (2) implies RE condition over $\mathbb{C}_3(S)$ uniformly over all subsets of cardinality

$$|S| \leq \frac{c_1}{32c_2}\frac{\gamma_{\min}(\Sigma)}{\rho^2(\Sigma)}\frac{n}{\log d}.$$

- In other words, $n \gtrsim s\log d \implies$ RE over $\mathbb{C}_3(S)$ for all $|S| \leq s$.