Statistics 210B Lecture 20 Notes

Daniel Raban

April 5, 2022

1 Restricted Eigenvalue Condition for Gaussian Random Matrices

1.1 Recap: Noisy, sparse linear estimation and the restricted eigenvalue condition

Let's continue our analysis of noisy, sparse linear regression. Our model is $y = X\theta^* + w \in \mathbb{R}^n$, where

$$w \in \mathbb{R}^n, \qquad X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times d}, \qquad \theta^* \in \mathbb{R}^d, \qquad |S(\theta^*)| \le s.$$

We looked at the λ formulation of the LASSO problem, where

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_1.$$

We also looked at the 1-norm constrained and error-constrained formulations of the problem. We defined the $\mathbb{C}_\alpha$ cone

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

Using this cone, we defined the restricted eigenvalue condition for efficient bounds on estimation.

Definition 1.1. $X \sim \mathrm{RE}(S, (\kappa, \alpha))$ if

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

We proved the following result.

Theorem 1.1. *Assume that* $\mathrm{RE}(s,(\kappa,3))$. *With a proper choice of hyperparameter, we have*

$$\|\widehat{\theta}-\theta^*\|_2 \lesssim \frac{1}{\kappa}\sqrt{s}\left\|\frac{X^\top w}{n}\right\|_\infty \lesssim \sigma\sqrt{\frac{s\log d}{n}}.$$

Now we would like to answer the question: when does RE hold?

1.2 Restricted eigenvalue condition for Gaussian random matrices

Theorem 1.2. *Let* $X_i \overset{\text{iid}}{\sim} N(0,\Sigma)$, *where* $\Sigma \in S_+^{d\times d}$. *There exist universal constants* $c_1<1<c_2$ *such that*

$$\frac{\|X\Delta\|_2^2}{n} \geq c_1\|\sqrt{\Sigma}\Delta\|_2^2 - c_2\rho^2(\Sigma)\frac{\log d}{n}\|\Delta\|_1^2 \qquad \forall\Delta\in\mathbb{R}^d$$

with probability at least $1-\frac{e^{-n/32}}{1-e^{n/32}}$. *Here,* $\rho^2(\Sigma)=\max_{i\in[d]}\Sigma_{i,i}$.

We think of this as a generalized RE condition. Let's show that this implies $\mathrm{RE}(S,(\kappa,3))$ for every S with cardinality $\leq s$. For all $\Delta\in\mathbb{C}_3(S)$, we want to show that $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$. Given the inequality $\|\Delta\|_1^2\leq 4s\|\Delta\|_2^2$, we can lower bound the right hand side in the theorem:

$$\begin{aligned} c_1\|\sqrt{\Sigma}\Delta\|_2^2 - c_2\rho^2(\Sigma)\frac{\log d}{n}\|\Delta\|_1^2 &\geq c_1\lambda_{\min}(\Sigma)\|\Delta\|_2^2 - c_2\rho^2(\Sigma)\frac{\log d}{n}4s\|\Delta\|_2^2 \\ &= \underbrace{\left(c_1\lambda_{\min}(\Sigma) - 4c_2\rho^2(\Sigma)\frac{s\log d}{n}\right)}\|\Delta\|_2^2 \end{aligned}$$

If $n\geq s\log d\frac{8c_2}{c_1}\frac{\rho^2(\Sigma)}{\lambda_{\min}(\Sigma)}$, we have the inequality $4c_2\rho^2(\Sigma)\frac{s\log d}{n}\leq\frac{c_1}{2}\lambda_{\min}(\Sigma)$. We can use it to lower bound the bracketed part.

$$\geq \frac{1}{2}c\lambda_{\min}(\Sigma)\|\Delta\|_2^2.$$

Proof. Let's prove the theorem in the case where $\Sigma=I_d$, so $X_i\overset{\text{iid}}{\sim}N(0,I_d)$. Our goal is the inequality

$$\frac{\|X\Delta\|_2^2}{n} + c_2'\frac{\log d}{n}\|\Delta\|_1^2 \geq c_1'\|\Delta\|_2^2 \qquad \forall\Delta\in\mathbb{R}^d.$$

Call $\|X\Delta\|_2^2$ the "X norm of Δ." We want to relate this to the 1-norm and 2 norm of Δ. A sufficient condition is to have

$$\frac{\|X\Delta\|_2}{\sqrt{n}} + c_2\sqrt{\frac{\log d}{n}}\|\Delta\|_1 \geq c_1\|\Delta\|_2 \qquad \forall\Delta\in\mathbb{R}^d$$

because if $a, b > 0$, then $a + b \leq c \implies a^2 + b^2 \leq c^2$. This inequality is invariant to scaling Δ, so it is sufficient to show that

$$\frac{\|X\Delta\|_2}{\sqrt{n}} + c_2\sqrt{\frac{\log d}{n}}\|\Delta\|_1 \geq c_1 \qquad \forall \|\Delta\|_2 = 1.$$

So we want to check that

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq c_1 - c_2\sqrt{\frac{\log d}{n}}\|\Delta\|_1 \quad \forall \|\Delta\|_2 = 1.$$

It is sufficient to show this for all Δ with bounded 1-norm:

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq c_1 - c_2\sqrt{\frac{\log d}{n}}r \quad \forall \|\Delta\|_2 = 1, \|\Delta\|_1 \leq r$$

for all $r > 0$. This means we can show that

$$\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\|X\Delta\|_2}{\sqrt{n}} \geq c_1 - c_2\sqrt{\frac{\log d}{n}}r \qquad \forall r > 0.$$

The intuition is that we want to apply the Gaussian comparison inequality, for which we need a $\|X\Delta\|_2$ on the left hand side and no Δ dependence on the right hand side. We have 3 steps:

Step 1: Expectation bound for fixed $r > 0$ (Gaussian comparison inequality)

$$\mathbb{E}\left[\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\|X\Delta\|_2}{\sqrt{n}}\right] \geq c_1 - c_2\sqrt{\frac{\log d}{n}}r$$

Step 2: Concentration for fixed $r > 0$ (Gaussian concentration)

$$G_r = \left\{\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\|X\Delta\|_2}{\sqrt{n}} \geq c_1 - c_2\sqrt{\frac{\log d}{n}}r\right\}$$

occurs with high probability.

Step 3: Union bound over $r > 0$ (Peeling argument): Let $G = \bigcap_{r>0} G_r$, so that $G^c = \bigcup_{r>0} G_r^c$. Then we can calculate

$$\mathbb{P}(G^c) \leq \sum_{r>0} \mathbb{P}(G_r^c).$$

We need to discretize the sum to get a bound that works.

We provide the rest of the proof in lemmas. □

Lemma 1.1 (Gaussian comparison)**.** *There exist constants c_1, c_2 such that*

$$\mathbb{E}\left[\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \le r} \frac{\|X\Delta\|_2}{\sqrt{n}}\right] \ge c_1 - c_2\sqrt{\frac{\log d}{n}} r$$

Proof. By the variational representation of the norm,

$$\mathbb{E}\left[\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \le r} \frac{\|X\Delta\|_2}{\sqrt{n}}\right] = \mathbb{E}\left[\inf_{\Delta \in S^{d-1}(1) \cap B_1(r)} \sup_{u \in S^{n-1}} \frac{\langle u, X\Delta\rangle}{n}\right].$$

By Gordon's inequality,

$$\mathbb{E}\left[\inf_{\Delta \in S} \sup_{u \in T} \langle u, X\Delta\rangle\right] \ge \mathbb{E}\left[\inf_{\Delta \in S} \sup_{u \in T} \langle h, \Delta\rangle + \langle g, u\rangle\right],$$

for any S, T, where $X_{i,j}, g_i, h_i \overset{\text{iid}}{\sim} N(0,1)$. So we get

$$\begin{aligned}
\mathbb{E}\left[\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \le r} \frac{\|X\Delta\|_2}{\sqrt{n}}\right] &\ge \mathbb{E}\left[\inf_{\Delta \in S^{d-1}(1) \cap B_1(r)} \sup_{\|u\|_2=1} \frac{\langle h, \Delta\rangle}{\sqrt{n}} + \frac{\langle g, u\rangle}{\sqrt{n}}\right] \\
&= \mathbb{E}\left[\inf_{\Delta} \frac{\langle h, \Delta\rangle}{\sqrt{n}} + \sup_{\|u\|_2=1} \frac{\langle g, u\rangle}{\sqrt{n}}\right] \\
&= \mathbb{E}\left[\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \le r} \frac{\langle h, \Delta\rangle}{\sqrt{n}}\right] + \mathbb{E}\left[\sup_{\|u\|_2=1} \frac{\langle g, u\rangle}{\sqrt{n}}\right]
\end{aligned}$$

Since $\mathbb{E}[\|g_2\|^2/n] = 1$, the expectation of the square root will be close to 1. We have the lower bound $\mathbb{E}[\|g\|_2/\sqrt{n}] \ge 1/4$. The first term on the other hand, can be expresed as $-\mathbb{E}\left[\sup_{\|\Delta\|_2=1, \|\Delta\|_1 \le r} \frac{\langle -h, \Delta\rangle}{\sqrt{n}}\right] \ge -\mathbb{E}\left[\sup_{\|\Delta\|_1 \le r} \frac{\langle -h, \Delta\rangle}{\sqrt{n}}\right] = -\mathbb{E}\left[\frac{\|-h\|_\infty}{\sqrt{n}}\right] r \ge -2\sqrt{\frac{\log d}{n}} r$. So we get

$$\ge \frac{1}{4} - 2\sqrt{\frac{\log d}{n}} r.$$ □

Lemma 1.2 (Concentration)**.** *Let $X_{i,j} \overset{\text{iid}}{\sim} N(0,1)$. The the event*

$$G_r = \left\{\inf_{\|\Delta\|_2=1, \|\Delta\|_1 \le r} \frac{\|X\Delta\|_2}{\sqrt{n}} \ge c_1 - c_2\sqrt{\frac{\log d}{n}} r\right\}$$

occurs with high probability.

Proof. Define the function

$$f(X) = \inf_{\|\Delta\|_2=1, \Delta \in S} \frac{\|X\Delta\|_2}{\sqrt{2}}.$$

We want to show that f is Lipschitz for the Frobenius norm, so we can use the Gaussian concentration lemma. Define $\Delta^* = \arg\min \|X_2\Delta\|_2/\sqrt{n}$. Then

$$\begin{aligned} f(X_1) - f(X_2) &\leq \frac{\|X_1\Delta^*\|_1}{\sqrt{n}} - \frac{\|X_2\Delta^*\|_2}{\sqrt{n}} \\ &\leq \frac{\|(X_1 - X_2)\Delta^*\|_1}{\sqrt{n}} \\ &\leq \frac{\|X_1 - X_2\|_{\text{op}}\|\Delta^*\|_1}{\sqrt{n}} \\ &\leq \frac{\|X_1 - X_2\|_F}{\sqrt{n}} \end{aligned}$$

This means that f is $\frac{1}{\sqrt{n}}$-Lipschitz in $\|X\|_F$, so $f(X)$ is $\text{sG}(1/\sqrt{n})$. Then

$$\mathbb{P}(f(X) \leq E[f(X)] - t) \leq e^{-nt^2/2},$$

so

$$G_r := \left\{ \inf_{\|\Delta\|_2=1, \|\Delta\|_1 \leq r} \frac{\|X\Delta\|_2}{\sqrt{n}} \geq c_1 - c_2\sqrt{\frac{\log d}{n}} r \right\}$$

occurs with high probability. □

Lemma 1.3 (Peeling argument)**.** *Let the bad event be*

$$G^c = \left\{ \exists \Delta, \|\Delta\|_2 = 1 \; s.t. \frac{\|X\Delta\|_2}{\sqrt{n}} \leq c_1 - c_2\sqrt{\frac{\log d}{n}} \|\Delta\|_1 \right\}.$$

then $G^c \subseteq \bigcup_{m=m_{\min}}^{m_{\max}} G^c_{2^{m+1}}$, *so* $\mathbb{P}(G^c) \leq \sum_{m=m_{\min}}^{m_{\max}} \mathbb{P}(G^c_{2^{m+1}})$.

Proof. Note that $\|\Delta\|_2 \leq \|\Delta\|_1 \leq \sqrt{d}\|\Delta\|_2$, so we get $1 \leq \|\Delta\|_1 \leq \sqrt{d}$. We discretize the interval in the log scale:

$$[1, \sqrt{d}] = \bigcup_{m=0}^{m_{\max}} [2^m, 2^{m+1}), \qquad m_{\max} = \log_2(\sqrt{d}) \approx \log d.$$

The we can write

$$G^c \subseteq \bigcup_{m=m_{\min}}^{m_{\max}} \left\{ \exists \Delta, \|\Delta\|_2 = 1, 2^m \leq \|\Delta\|_1 \leq 2^{m+1} \text{ s.t.} \frac{\|X\Delta\|_2}{\sqrt{n}} \leq c_1 - c_2\sqrt{\frac{\log d}{n}} 2^m \right\}$$

$$\subseteq \underbrace{\left\{ \inf_{\|\Delta\|_2=1, \|\Delta_1 \le 2^{m+1}} \frac{\|X\Delta\|_2}{\sqrt{n}} \le c_1 - \frac{c_2}{2}\sqrt{\frac{\log d}{n}} \right\}}_{G^c_{2^{m+1}}}.$$

So we have shown that $G^c \subseteq \bigcup_{m=m_{\min}}^{m_{\max}} G^c_{2^{m+1}}$. □

1.3 LASSO oracle inequality

We have shown that we can efficiently bound the approximation error of θ^* if θ^* is sparse. But what if θ^* is not exactly sparse but is instead approximately sparse? That is, what if $\theta^*_{S^c} \neq 0$ but $\|\theta^*_{S^c}\|_1$ is small?

Definition 1.2. We say that an estimator $\widehat{\theta}$ satisfies an **oracle inequality** with respect to the risk R, set Θ, and model $\{\mathbb{P}_\theta : \theta \in \Theta^*\}$ ($\Theta \subseteq \Theta^*$), if there exist constants c and $\varepsilon_n(\mathbb{P}_{\theta^*}, \Theta)$ such that for any $\theta^* \in \Theta^*$, then

$$R(\widehat{\theta}; \theta^*) \le c \underbrace{\inf_{\theta \in \Theta} R(\theta; \theta^*)}_{\text{approx. error/oracle risk}} + \underbrace{\varepsilon_n(\mathbb{P}_{\theta^*}, \Theta)}_{\text{statistical error}}.$$

We hope that c is not too large and that ε_n is small. If $\theta^* \in \Theta$, then

$$\inf_{\theta \in \Theta} R(\theta; \theta^*) = 0.$$

Let $\Theta = \{\Delta \mathbb{R}^d : \|\Delta\|_0 \le s\}$ be the set of s-sparse vectors and let $R(\theta; \theta^*) = \|\theta - \theta^*\|_2$. Then if θ^* is s-sparse, $\inf_{\theta \in \Theta} R(\theta; \theta^*) = 0$. If θ^* is not s-sparse, then

$$\inf_{\theta \in \Theta} R(\theta, \theta^*) > 0.$$

We use our generalized RE condition:

$$\frac{\|X\Delta\|_2^2}{n} \ge c_1 \|\sqrt{\Sigma}\Delta\|_2^2 - c_2 \rho^2(\Sigma) \frac{\log d}{n} \|\Delta\|_1^2, \qquad \forall \Delta \in \mathbb{R}^d.$$

Theorem 1.3 (LASSO oracle inequality)**.** *Assume the generalized RE condition holds for $X \in \mathbb{R}^{n \times d}$. Let $\widehat{\theta}$ be solution to the λ formulation of LASSO with $\lambda_n \ge 2\|\frac{X^\top w}{n}\|_\infty$. Then for any S with $|S| \le \frac{c_1}{64 c_2} \frac{\overline{\kappa}}{\rho^2(\Sigma)} \frac{n}{\log d}$ (where $\overline{\kappa} = \lambda_{\min}(\Sigma)$,*

$$\|\widehat{\theta} - \theta^*\|_2^2 \le \underbrace{\frac{144}{c_1^2} \frac{\lambda_n^2}{\overline{\kappa}^2} |S|}_{\text{statistical error} \lesssim \sigma^2 \frac{s \log d}{n}} + \underbrace{\frac{16}{c_1} \frac{\lambda_n}{\overline{\kappa}} \|\theta^*_{S^c}\|_1 + \frac{32 c_2}{c_1} \frac{\rho^2(\Sigma)}{\overline{\kappa}} \frac{\log d}{n} \|\theta^*_{S^c}\|_1^2}_{\text{approx. error/oracle risk} \lesssim \varepsilon_n + \varepsilon_n^2},$$

where $\varepsilon_n = \sqrt{\frac{\log d}{n}} \|\theta^*_{S^c}\|_1$.

Proof. This this a deterministic inequality, so the proof is to derive a basic inequality and then use some algebra. The proof is in the textbook. □