Statistical Learning Theory Group 3 Project

Yen-Lin Chen*
Part of the Work of Group 3

November 27, 2019

This short report is part of the work of Group 3. In this report, I will pave the steps toward proving the LASSO oracle inequality, Theorem 7.19 of the book. I will focus mainly on Theorem 7.16 and Lemma 7.24 because with both, the LASSO oracle inequality becomes straightforward. We consider the Lagrangian LASSO setting with the model $y = \mathbf{X}\theta^* + w$.

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

$\mathbf{X} \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$ and $w \in \mathbb{R}^n$. The aim is to upper bound the quantity $\|\hat{\theta} - \theta^*\|_2$.

Theorem 7.13. *Under the following 2 assumptions:*

1. *sparse θ^* is supported on $S \subseteq \{1, 2, \ldots, d\}$ with $|S| = s$.*
2. *$\mathbf{X}$ satisfies restricted eigenvale condition over S with parameter (κ, α).*

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

where $\mathbf{C}_\alpha(S) = \left\{\Delta \in \mathbf{R}^d | \|\Delta_{S^C}\|_1 \leq \alpha \|\Delta_S\|_1\right\}$ and S^C denotes the complementary set of S.

If $\lambda_n \geq 2\|\frac{\mathbf{X}'w}{n}\|_\infty$, $\hat{\theta}$ satisfies the bound:

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

Proof. The proof was done in class so I summarize it briefly. Since $\hat{\theta}$ is optimal in Eq. (1).

$$\frac{1}{2n}\left\|y - \mathbf{X}\hat{\theta}\right\|_2^2 + \lambda_n \left\|\hat{\theta}\right\|_1 \leq \frac{1}{2n}\|y - \mathbf{X}\theta^*\|_2^2 + \lambda_n \|\theta^*\|_1 \tag{4}$$

$$\iff \frac{1}{2n}\left\|w - \mathbf{X}(\hat{\theta} - \theta^*)\right\|_2^2 \leq \frac{1}{2n}\|w\|_2^2 + \lambda_n \left(\|\theta^*\|_1 - \left\|\hat{\theta}\right\|_1\right) \tag{5}$$

$$\iff \frac{1}{n}\left\|\mathbf{X}\hat{\Delta}\right\|_2^2 \leq 2\frac{w'\mathbf{X}\hat{\Delta}}{n} + 2\lambda_n \left(\|\theta^*\|_1 - \left\|\hat{\theta}\right\|_1\right) \tag{6}$$

where $\hat{\Delta} = \hat{\theta} - \theta^*$. We use the following three properties.

$$\kappa\|\hat{\Delta}\|_2^2 \leq \frac{1}{n}\left\|\mathbf{X}\hat{\Delta}\right\|_2^2 \tag{7}$$

*Department of Applied and Engineering Physics

$$\frac{w'\mathbf{X}\hat{\Delta}}{n} = \left(\frac{\hat{\Delta}\mathbf{X}}{n}\right)' w \leq \left\|\frac{\mathbf{X}'w}{n}\right\|_\infty \left\|\hat{\Delta}\right\|_1 \leq \frac{\lambda_n}{2} \tag{8}$$

$$\begin{aligned}\|\theta^*\|_1 - \left\|\hat{\theta}\right\|_1 &= \|\theta^*_S\|_1 - \left\|\theta^* + \hat{\Delta}\right\|_1 \\ &= \|\theta^*_S\|_1 - \left\|\theta^*_S + \bar{\Delta}_S\right\|_1 - \left\|\hat{\Delta}_{S^C}\right\|_1 \\ &\leq \left\|\hat{\Delta}_S\right\|_1 - \left\|\hat{\Delta}_{S^C}\right\|_1\end{aligned} \tag{9}$$

Therefore, by plugging the above inequalities into Eq. (6),

$$\begin{aligned}\kappa\|\hat{\Delta}\|_2^2 &\leq \lambda_n \left\|\hat{\Delta}\right\|_1 + 2\lambda_n \left(\left\|\hat{\Delta}_S\right\|_1 - \left\|\hat{\Delta}_{S^C}\right\|_1\right) \\ &\leq \lambda_n \left(3\left\|\hat{\Delta}_S\right\|_1 - \left\|\hat{\Delta}_{S^C}\right\|_1\right) \\ &\leq 3\lambda_n \left\|\hat{\Delta}_S\right\|_1 \\ &\leq 3\lambda_n\sqrt{s}\left\|\hat{\Delta}\right\|_2\end{aligned} \tag{10}$$

$$\|\hat{\Delta}\|_2 = \|\hat{\theta} - \theta^*\|_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda_n \tag{11}$$

□

Theorem 7.13 bases on the restricted eigenvalue condition of $\mathbf{X}$. Let's consider the case where $\mathbf{X}$ is now random and the aim is again to upper bound the quantity $\|\hat{\theta} - \theta^*\|_2$ with high probability, where $\hat{\theta}$ is the solution for the Lagrangian LASSO equation, i.e. Eq. (1). To this end, we need the following property for the quantity $\frac{1}{n}\|\mathbf{X}\theta\|_2^2$.

Theorem 7.16. *Consider a random matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ with iid rows $x_i \in \mathbf{R}^d$ from the normal distribution $\mathcal{N}(0, \mathbf{\Sigma})$. Then there are universal constants $c_1 < 1 < c_2$ such that*

$$\frac{1}{n}\|\mathbf{X}\theta\|_2^2 \geq c_1 \left\|\sqrt{\mathbf{\Sigma}}\theta\right\|_2^2 - c_2\rho^2(\mathbf{\Sigma})\frac{\log d}{n}\|\theta\|_1 \quad \forall\theta \in \mathbb{R}^d \tag{12}$$

with probability at least $1 - \frac{e^{-n/32}}{1-e^{-n/32}}$.

proof of Theorem 7.16. By re-scaling of the vector θ, it suffices to prove the result on the ellipse:

$$\mathbb{S}^{d-1}(\mathbf{\Sigma}) = \left\{\theta \in \mathbb{R}^d \middle| \left\|\sqrt{\mathbf{\Sigma}\theta}\right\|_2 = 1\right\} \tag{13}$$

To obtain the "$\geq$" in Eq. (12), it is equivalent to upper bound the probability of the "$\leq$" event:

$$\mathcal{Q}(c_1, c_2) = \left\{\mathbf{X} \in \mathbb{R}^{n\times d} \middle| \frac{1}{n}\|\mathbf{X}\theta\|_2^2 \leq c_1 \left\|\sqrt{\mathbf{\Sigma}}\theta\right\|_2^2 - c_2\rho^2(\mathbf{\Sigma})\frac{\log d}{n}\|\theta\|_1^2\right\} \tag{14}$$

For all $\theta \in \mathbb{S}^{d-1}(\mathbf{\Sigma})$, define the "bad" event as the following:

$$\mathcal{E} = \left\{\mathbf{X} \in \mathbb{R}^{n\times d} \middle| \inf_{\theta\in\mathbb{S}^{d-1}(\mathbf{\Sigma})} \frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}} \leq \frac{1}{4} - 4\rho(\mathbf{\Sigma})\sqrt{\frac{\log d}{n}}\|\theta\|_1\right\} \tag{15}$$

$\forall\mathbf{X}$ satisfying the events in $\mathcal{E}$, the following holds because $2\rho(\mathbf{\Sigma})\sqrt{\frac{\log d}{n}}\|\theta\|_1 \geq 0$.

$$\inf_{\theta\in\mathbb{S}^{d-1}(\mathbf{\Sigma})} \frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}} \leq \frac{1}{4} + 4\rho(\mathbf{\Sigma})\sqrt{\frac{\log d}{n}}\|\theta\|_1 \tag{16}$$

Define another event as

$$\mathcal{E}' = \left\{ \mathbf{X} \in \mathbb{R}^{n\times d} \middle| \inf_{\theta \in \mathbb{S}^{d-1}(\mathbf{\Sigma})} \frac{\|\mathbf{X}\theta\|_2^2}{n} \leq \frac{1}{16} - 16\rho^2(\mathbf{\Sigma}) \frac{\log d}{n} \|\theta\|_1^2 \right\} \tag{17}$$

$\forall \theta \in \mathbb{S}^{d-1}(\mathbf{\Sigma})$, $\left\| \sqrt{\mathbf{\Sigma}}\theta \right\|_2 = 1$. $\mathcal{Q}\left(\frac{1}{16}, 16\right) \subseteq \mathcal{E}' \subseteq \mathcal{E}$. Notice that under "good" events $\mathcal{E}^C$, the event $\mathcal{Q}\left(\frac{1}{16}, 16\right)^C$ occurs for sure. Now, the aim is to upper bound the probability $P(\mathcal{E})$.

For a pair of radii $0 \leq r_l < r_u$, define the set $\mathbb{K}(r_l, r_u)$

$$\mathbb{K}(r_l, r_u) = \left\{ \theta \in \mathbb{S}^{d-1}(\mathbf{\Sigma}) \middle| 2\rho(\mathbf{\Sigma}) \sqrt{\frac{\log d}{n}} \|\theta\|_1 \in [r_l, r_u] \right\} \tag{18}$$

and the corresponding bad event $\mathcal{A}$

$$\mathcal{A}(r_l, r_u) = \left\{ \inf_{\theta \in \mathbb{K}(r_l, r_u)} \frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}} \leq \frac{1}{2} - 2r_u \right\} \tag{19}$$

We will need the following lemma. Lemma 7.24 uses $\mathbb{K}(r_l, r_u)$ and $\mathcal{A}(r_l, r_u)$ to upper bound the probability $P(\mathcal{E})$.

Lemma 7.24. *For any pair of radii $0 \leq r_l < r_u$, we have*

$$P\left[\mathcal{A}(r_l, r_u)\right] \leq e^{-\frac{n}{32}} e^{-\frac{n}{2} r_u^2} \tag{20}$$

Furthermore, for some constant $\mu \geq \frac{1}{8}$, we have

$$\mathcal{E} \subseteq \mathcal{A}(0, \mu) \cup \left(\bigcup_{l=1}^{\infty} \mathcal{A}(2^{l-1}\mu, 2^l \mu) \right) \tag{21}$$

The intuition is to partition the ellipse, $\theta \in \mathbb{S}^{d-1}(\mathbf{\Sigma})$ into infinitely many disjoint subsets $\mathbb{K}(r_l, r_u)$ and obtain the probability of "bad" events $\mathcal{A}(r_l, r_u)$ within each subset. As r_u grows: $(r_l, r_u) = (0, \mu) \to (\mu, 2\mu) \to (2\mu, 4\mu) \to \ldots$ we will have fewer and fewer "bad" events. The two-dimensional illustration of the set $\mathbb{K}(r_l, r_u)$ with increasing r_u is shown in Fig. 1. It is apparent that the set $\mathbb{K}(r_l, r_u)$ is empty for most pairs of the radii. In other words, the d-dimensional $l1$-ball only intersects with the ellipse for a specific choice of its radius. Note that by the construction of $\mathcal{A}(r_l, r_u)$ in Eq. (19) , $\mathcal{A}(r_l, r_u) = \phi$ for all $r_u > \frac{1}{4}$. In the textbook, it says $\mu = \frac{1}{4}$ for the sake of proving Theorem 7.16 (with minor errors in the original proof) but $\mu = \frac{1}{4}$ is not general for Lemma 7.24. Here, I will derive the general values of μ in the following proof.

proof of Lemma 7.24. I'll start with the proof of Eq. (21) by considering the following two cases.

1. $\theta \in \mathbb{K}(r_l = 0, r_u = \mu) \Longrightarrow 2\rho(\mathbf{\Sigma}) \sqrt{\frac{\log d}{n}} \|\theta\|_1 \leq \mu$.

   Therefore, if θ certifies the event $\mathcal{E}$, I want to determine μ such that $\mathcal{A}(0, \mu)$ is certified for sure. That is to say,

$$\begin{aligned} \frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}} &\leq \frac{1}{4} - 4\rho(\mathbf{\Sigma}) \sqrt{\frac{\log d}{n}} \|\theta\|_1 \\ &\leq \frac{1}{4} \\ &\leq \frac{1}{2} - 2\mu \quad \text{ensuring} \quad \mathcal{A}(0, \mu) \end{aligned} \tag{22}$$

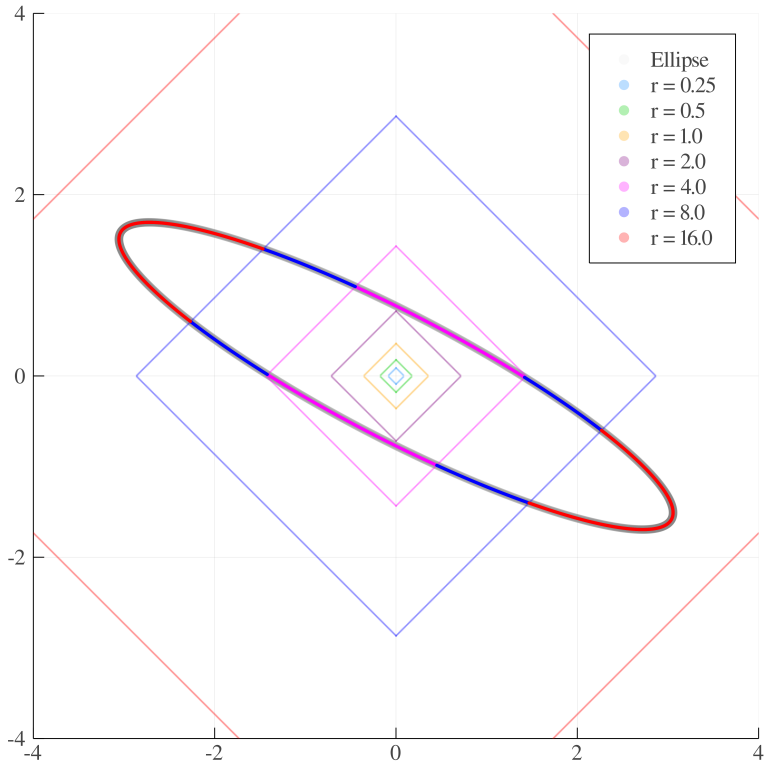

Figure 1: The two-dimensional illustration of the set $\mathbb{K}(r_l, r_u)$ with increasing r_l and r_u.

Therefore, we have $\mu \geq \frac{1}{8}$.

2. $\theta \in \mathbb{K}(r_l = 2^{l-1}\mu, r_u = 2^l\mu)$ for some $l \in \mathbb{N} \Longrightarrow 2\rho(\mathbf{\Sigma})\sqrt{\frac{\log d}{n}}\,\|\theta\|_1 \geq 2^{l-1}\mu$.

If θ certifies the event $\mathcal{E}$, i.e.

$$\begin{aligned} \frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}} &\leq \frac{1}{4} - 4\rho(\mathbf{\Sigma})\sqrt{\frac{\log d}{n}}\,\|\theta\|_1 \\ &\leq \frac{1}{4} - 2\left(2^{l-1}\mu\right) = \frac{1}{4} - 2^l\mu \\ &\leq 2\left(\frac{1}{4} - 2^l\mu\right) = \frac{1}{2} - 2r_u \end{aligned} \tag{23}$$

Therefore, $\mathcal{A}(r_l = 2^{l-1}\mu, r_u = 2^l\mu)$ occurs for sure. Notice that this this case, the value of μ is totally irrelevant, as long as it is positive.

Combining case 1 and 2, the proof of Eq. (21) is complete. Let's now focus on constructing the tail bound for the probability of event $\mathcal{A}(r_l, r_u)$. By the construction of $\mathcal{A}$, it is equivalent to upper bound the following quantity .

$$T(r_l, r_u) = -\inf_{\theta \in \mathbb{K}(r_l, r_u)} \frac{\|\mathbf{X}\theta\|_2}{\sqrt{n}} \tag{24}$$

For any vector $v \in \mathbb{R}^n$, the l_2-norm $\|v\|_2$ can be written as the following

$$\|v\|_2 = \sup_{u \in \mathbb{S}^{n-1}} \langle u, v \rangle \tag{25}$$

where $\mathbb{S}^{n-1}$ is the ellipse in $\mathbb{R}^n$. Therefore,

$$T(r_l, r_u) = -\inf_{\theta \in \mathbb{K}(r_l, r_u)} \left[\sup_{u \in \mathbb{S}^{n-1}} \frac{\langle u, \mathbf{X}\theta \rangle}{\sqrt{n}} \right] = \sup_{\theta \in \mathbb{K}(r_l, r_u)} \left[\inf_{u \in \mathbb{S}^{n-1}} \frac{\langle u, \mathbf{X}\theta \rangle}{\sqrt{n}} \right] \tag{26}$$

Rewrite $\mathbf{X} = \mathbf{W}\sqrt{\mathbf{\Sigma}}$ where $\mathbf{W} \in \mathbb{R}^{n \times d}$ is a standard Gaussian matrix. Moreover, with $v = \sqrt{\mathbf{\Sigma}}\theta$, we have $\mathbf{X}\theta = \mathbf{W}\sqrt{\mathbf{\Sigma}}\theta = \mathbf{W}v$.

$$T(r_l, r_u) = \sup_{v \in \tilde{\mathbb{K}}(r_l, r_u)} \left[\inf_{u \in \mathbb{S}^{n-1}} \frac{\langle u, \mathbf{W}v \rangle}{\sqrt{n}} \right] = \sup_{v \in \tilde{\mathbb{K}}(r_l, r_u)} \left[\inf_{u \in \mathbb{S}^{n-1}} Z_{u,v} \right] \tag{27}$$

where $Z_{u,v} = \frac{\langle u, \mathbf{W}v \rangle}{\sqrt{n}}$ and this operation transforms the ellipse set of θ in to a ball set of v, i.e.

$$\tilde{\mathbb{K}}(r_l, r_u) = \left\{ v \in \mathbb{R}^d \middle| 2\rho(\mathbf{\Sigma}) \sqrt{\frac{\log d}{n}} \left\| \mathbf{\Sigma}^{-\frac{1}{2}} v \right\|_1 \in [r_l, r_u] \right\} \tag{28}$$

Note that after the transformation, $u \in \mathbb{S}^{n-1} \subseteq \mathbb{R}^n$ and $v \in \mathbb{S}^{d-1} \subseteq \mathbb{R}^d$. Therefore $Z_{u,v} \sim \mathcal{N}(0, n^{-1})$, which is useful in designing another random variable with larger variance to upper bound Eq. (27). Let $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^d$ with iid elements from $\mathcal{N}(0,1)$, construct two random variables as the following

$$Y'_u = \frac{\langle g, u \rangle}{\sqrt{n}} \quad Y_{u,v} = \frac{\langle g, u \rangle}{\sqrt{n}} + \frac{\langle h, v \rangle}{\sqrt{n}} \tag{29}$$

with $\mathrm{var}(Z_{u,v}) \leq \mathrm{var}(Y'_u) \leq \mathrm{var}(Y_{u,v})$. With Gordon's inequality,

$$\begin{aligned}
E\left\{ \sup_{v \in \tilde{\mathbb{K}}(r_l, r_u)} \left[\inf_{u \in \mathbb{S}^{n-1}} Z_{u,v} \right] \right\} &\leq E\left\{ \sup_{v \in \tilde{\mathbb{K}}(r_l, r_u)} \left[\inf_{u \in \mathbb{S}^{n-1}} Y'_u \right] \right\} \\
&\leq E\left\{ \sup_{v \in \tilde{\mathbb{K}}(r_l, r_u)} \left[\inf_{u \in \mathbb{S}^{n-1}} Y_{u,v} \right] \right\} \\
&= E\left\{ \sup_{v \in \tilde{\mathbb{K}}(r_l, r_u)} \left[\inf_{u \in \mathbb{S}^{n-1}} \frac{\langle g, u \rangle}{\sqrt{n}} \right] \right\} + E\left\{ \sup_{v \in \tilde{\mathbb{K}}(r_l, r_u)} \left[\inf_{u \in \mathbb{S}^{n-1}} \frac{\langle h, v \rangle}{\sqrt{n}} \right] \right\} \\
&= E\left(\inf_{u \in \mathbb{S}^{n-1}} \frac{\langle g, u \rangle}{\sqrt{n}} \right) + E\left(\sup_{v \in \tilde{\mathbb{K}}(r_l, r_u)} \frac{\langle h, v \rangle}{\sqrt{n}} \right) \\
&= -E\left(\frac{\|g\|_2}{\sqrt{n}} \right) + E\left(\sup_{v \in \tilde{\mathbb{K}}(r_l, r_u)} \frac{\langle h, v \rangle}{\sqrt{n}} \right) \\
&= -E\left(\frac{\|g\|_2}{\sqrt{n}} \right) + E\left(\sup_{\theta \in \mathbb{K}(r_l, r_u)} \frac{\langle h, \sqrt{\mathbf{\Sigma}}\theta \rangle}{\sqrt{n}} \right) \\
&= -E\left(\frac{\|g\|_2}{\sqrt{n}} \right) + E\left(\sup_{\theta \in \mathbb{K}(r_l, r_u)} \frac{\langle \sqrt{\mathbf{\Sigma}}h, \theta \rangle}{\sqrt{n}} \right)
\end{aligned} \tag{30}$$

Since the random vector g is drawn iid from $\mathcal{N}(0,1)$,

$$\begin{aligned} E\left(\frac{\|g\|_2}{\sqrt{n}}\right) &= \frac{1}{\sqrt{n}} E\left(\sqrt{\sum_{i=1}^{n} g_i^2}\right) \\ &\geq \frac{1}{n} E\left(\sum_{i=1}^{n} |g_i|\right) \\ &\geq \frac{1}{n} \sum_{i=1}^{n} E(|g_i|) = E(|g_i|) \\ &= 2\int_0^\infty \frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}} dx = \sqrt{\frac{2}{\pi}} \end{aligned} \tag{31}$$

The last term in Eq. (30) can also be bound.

$$\begin{aligned} E\left(\sup_{\theta\in\mathbb{K}(r_l,r_u)} \frac{\langle\sqrt{\mathbf{\Sigma}}h,\theta\rangle}{\sqrt{n}}\right) &\leq E\left(\sup_{\theta\in\mathbb{K}(r_l,r_u)} \|\theta\|_1 \frac{\left\|\sqrt{\mathbf{\Sigma}}h\right\|_\infty}{\sqrt{n}}\right) \\ &\leq E\left(\frac{\left\|\sqrt{\mathbf{\Sigma}}h\right\|_\infty}{\sqrt{n}}\right)\left(\sup_{\theta\in\mathbb{K}(r_l,r_u)} \|\theta\|_1\right) \\ &\leq \left[2\rho(\mathbf{\Sigma})\sqrt{\frac{\log d}{n}}\right]\left(\frac{r_u}{2\rho(\mathbf{\Sigma})\sqrt{(\log d)/n}}\right) \\ &= r_u \end{aligned} \tag{32}$$

Finally, by combining Eq. (30), (31) and (32),

$$E\left[T(r_l,r_u)\right] \leq -\sqrt{\frac{2}{\pi}} + r_u \tag{33}$$

With the upper tail bound in Theorem 2.26,

$$P\left\{T(r_l,r_u) \geq E\left[T(r_l,r_u)\right] + \delta\right\} \leq e^{-n\delta^2/2} \tag{34}$$

$$\begin{aligned} P\left[T(r_l,r_u) \geq \left(-\sqrt{\frac{2}{\pi}} + r_u\right) + \delta\right] &\leq e^{-n\delta^2/2} \\ P\left[T(r_l,r_u) \geq \left(-\sqrt{\frac{2}{\pi}} + r_u\right) + \left(\sqrt{\frac{2}{\pi}} - \frac{1}{2} + r_u\right)\right] &\leq e^{-n\left(\sqrt{\frac{2}{\pi}} - \frac{1}{2} + r_u\right)^2/2} \\ &\leq e^{-\frac{n}{2}\left(\sqrt{\frac{2}{\pi}} - \frac{1}{2}\right)^2} e^{-nr_u^2/2} \\ &\leq e^{-\frac{n}{32}} e^{-\frac{n}{2}r_u^2} \end{aligned} \tag{35}$$

Therefore, by plugging the definition of $T(r_l,r_u)$ and flip the sign, the proof of Lemma 7.24 is complete.

$$P\left[\mathcal{A}(r_l,r_u)\right] \leq e^{-\frac{n}{32}} e^{-\frac{n}{2}r_u^2} \tag{36}$$

□

Let's now use Lemma 7.24 to continue the proof of Theorem 7.16.

proof of Theorem 7.16 (continued). Now with Lemma 7.24 and setting $\mu = \frac{1}{4}$, we have

$$\begin{aligned} P(\mathcal{E}) &\leq P\left[\mathcal{A}(0,\mu)\right] + \sum_{l=1}^{\infty} P\left[\mathcal{A}(2^{l-1}\mu, 2^l\mu)\right] \\ &\leq e^{-\frac{n}{32}} e^{-\frac{n}{2}\mu^2} + \sum_{l=1}^{\infty} e^{-\frac{n}{32}} e^{-\frac{n}{2}2^{2l}\mu^2} \\ &= e^{-\frac{n}{32}} \sum_{l=0}^{\infty} e^{-\frac{n}{2}2^{2l}\mu^2} \\ &\leq e^{-\frac{n}{32}} \sum_{l=0}^{\infty} e^{-nl\mu^2} \\ &= e^{-\frac{n}{32}} \frac{1}{1-e^{-n\mu^2}} = \frac{e^{-\frac{n}{32}}}{1-e^{-\frac{n}{16}}} \\ &\leq \frac{e^{-\frac{n}{32}}}{1-e^{-\frac{n}{32}}} \end{aligned} \tag{37}$$

Combining with Eq. (15) and (17),

$$\begin{aligned} \frac{e^{-\frac{n}{32}}}{1-e^{-\frac{n}{32}}} &\geq P(\mathcal{E}) \\ &\geq P[\mathcal{E}'] = P\left\{\mathbf{X} \in \mathbb{R}^{n\times d} \middle| \inf_{\theta\in\mathbb{S}^{d-1}(\mathbf{\Sigma})} \frac{\|\mathbf{X}\theta\|_2^2}{n} \leq \frac{1}{16} - 16\rho^2(\mathbf{\Sigma})\frac{\log d}{n}\|\theta\|_1^2\right\} \\ &\geq P\left\{\mathbf{X} \in \mathbb{R}^{n\times d} \middle| \inf_{\theta\in\mathbb{S}^{d-1}(\mathbf{\Sigma})} \frac{\|\mathbf{X}\theta\|_2^2}{n} \leq \frac{1}{8} - 32\rho^2(\mathbf{\Sigma})\frac{\log d}{n}\|\theta\|_1^2\right\} \\ &\geq P\left\{\mathbf{X} \in \mathbb{R}^{n\times d} \middle| \inf_{\theta\in\mathbb{S}^{d-1}(\mathbf{\Sigma})} \frac{\|\mathbf{X}\theta\|_2^2}{n} \leq \frac{1}{8} - 50\rho^2(\mathbf{\Sigma})\frac{\log d}{n}\|\theta\|_1^2\right\} \\ &\geq P\left[Q(\frac{1}{8}, 50)\right] \end{aligned} \tag{38}$$

Therefore, with $c_1 = \frac{1}{8}$ and $c_2 = 50$, the bound holds. □

It is helpful to visualize the sets of the event discussed and used so far. Fig. 2 shows the relationship among them. The ultimate goal was to upper bound the probability of $\mathcal{Q}(c_1, c_2)$ whose size is determined by constants c_1 and c_2. For $(c_1, c_2) = \left(\frac{1}{16}, 16\right)$, the event set $\mathcal{Q}$ is the subset of the "bad" event $\mathcal{E}$. Lemma 7.24 further shows that $\mathcal{E}$ is the subset of $\mathcal{A}(0,\mu) \cup \left(\bigcup_{l=1}^{\infty} \mathcal{A}(2^{l-1}\mu, 2^l\mu)\right)$ whose probability is upper bound using the tail bound property. Closer look into Lemma 7.24 suggests that this bound is very loose because for $r_u > \frac{1}{4}$, $P(\mathcal{A}) = 0$ because $\|\mathbf{X}\theta\|_2 \geq 0$.

The context is now set up for showing the LASSO Oracle Inequality. Note that there is a minor error in the original proof (in Eq. (7.36)) but the result remains.

Theorem 7.19. *Under the condition of Theorem 7.16 and consider the Lagrangian LASSO equation, Eq. (1) with $\lambda_n \geq 2\|\frac{\mathbf{X}'w}{n}\|_\infty$. For any $\theta^* \in \mathbb{R}^d$ and optimal solution $\hat{\theta}$ satisfies the bound*

$$\left\|\hat{\theta} - \theta^*\right\|_2^2 \leq \frac{144\lambda_n^2}{c_1^2\bar{\kappa}^2}|S| + \frac{16\lambda_n}{c_1\bar{\kappa}}\left\|\theta^*_{S^C}\right\|_1 + \frac{32c_2\rho^2(\mathbf{\Sigma})}{c_1\bar{\kappa}}\frac{\log d}{n}\left\|\theta^*_{S^C}\right\|_1 \tag{39}$$

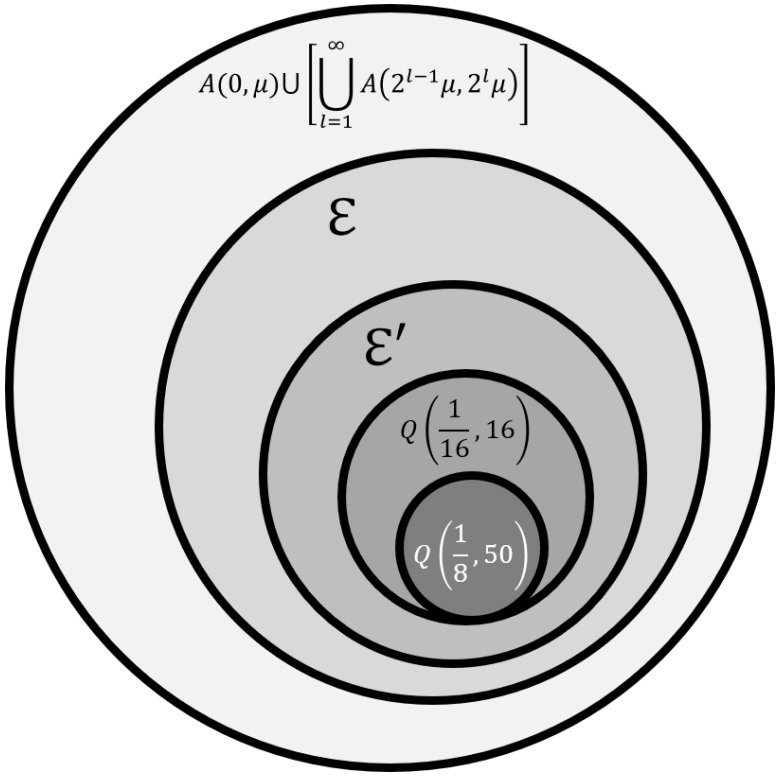

Figure 2: The diagram of all the event sets discussed.

where the cardinality of S satisfies

$$|S| \leq \frac{c_1 \bar{\kappa}}{64 c_2 \rho^2(\mathbf{\Sigma})} \frac{\log d}{n} \tag{40}$$

proof of Theorem 7.19. Since Theorem 7.16 involves the $l1$-norm, the first goal is to obtain the bound for $\|\theta\|_1$. From the first inequality in Eq. (10),

$$\begin{aligned} 0 &\leq \lambda_n \left[3 \left\| \hat{\Delta}_S \right\|_1 - \left\| \hat{\Delta}_{S^C} \right\|_1 + 2 \left\| \theta^*_{S^C} \right\|_1 \right] \\ &= \lambda_n \left[3 \left\| \hat{\Delta}_S \right\|_1 - \left(\left\| \hat{\Delta}_S \right\|_1 - \left\| \hat{\Delta} \right\|_1 \right) + 2 \left\| \theta^*_{S^C} \right\|_1 \right] \\ &\leq \lambda_n \left[4 \left\| \hat{\Delta}_S \right\|_1 - \left\| \hat{\Delta} \right\|_1 + 2 \left\| \theta^*_{S^C} \right\|_1 \right] \end{aligned} \tag{41}$$

$$\begin{aligned} \left\| \hat{\Delta} \right\|_1^2 &\leq \left(4 \left\| \hat{\Delta}_S \right\|_1 + 2 \left\| \theta^*_{S^C} \right\|_1 \right)^2 \\ &\leq \left(4\sqrt{|S|} \left\| \hat{\Delta}_S \right\|_2 + 2 \left\| \theta^*_{S^C} \right\|_1 \right)^2 \\ &\leq \left(4\sqrt{|S|} \left\| \hat{\Delta} \right\|_2 + 2 \left\| \theta^*_{S^C} \right\|_1 \right)^2 \\ &\leq (1^2 + 1^2) \left[\left(4\sqrt{|S|} \left\| \hat{\Delta} \right\|_2 \right)^2 + \left(2 \left\| \theta^*_{S^C} \right\|_1 \right)^2 \right] \\ &= 32|S| \left\| \hat{\Delta} \right\|_2^2 + 8 \left\| \theta^*_{S^C} \right\|_1^2 \end{aligned} \tag{42}$$

Now use Theorem 7.16.

$$\begin{aligned}\frac{1}{n}\left\|\mathbf{X}\hat{\Delta}\right\|_2^2 &\geq c_1\left\|\sqrt{\mathbf{\Sigma}}\hat{\Delta}\right\|_2^2 - c_2\rho^2(\mathbf{\Sigma})\frac{\log d}{n}\left\|\hat{\Delta}\right\|_1 \\ &\geq c_1\bar{\kappa}\left\|\hat{\Delta}\right\|_2^2 - c_2\rho^2(\mathbf{\Sigma})\frac{\log d}{n}\left(32|S|\left\|\hat{\Delta}\right\|_2^2 + 8\left\|\theta^*_{S^C}\right\|_1^2\right) \\ &= \left(c_1\bar{\kappa} - 32c_2\rho^2(\mathbf{\Sigma})\frac{\log d}{n}|S|\right)\left\|\hat{\Delta}\right\|_2^2 - 8c_2\rho^2(\mathbf{\Sigma})\frac{\log d}{n}\left\|\theta^*_{S^C}\right\|_1^2\end{aligned} \tag{43}$$

where $\bar{\kappa}$ is the minimum eigenvalue of matrix $\mathbf{\Sigma}$. Using the constraint in the cardinality from Eq. (40),

$$\frac{1}{n}\left\|\mathbf{X}\hat{\Delta}\right\|_2^2 \geq \frac{1}{2}c_1\bar{\kappa}\left\|\hat{\Delta}\right\|_2^2 - 8c_2\rho^2(\mathbf{\Sigma})\frac{\log d}{n}\left\|\theta^*_{S^C}\right\|_1^2 \tag{44}$$

Now it is left to compare two norms: $\left\|\hat{\Delta}\right\|_2^2$ and $\left\|\theta^*_{S^C}\right\|_1^2$.

1. Let $\frac{1}{4}c_1\bar{\kappa}\left\|\hat{\Delta}\right\|_2^2 \geq 8c_2\rho^2(\mathbf{\Sigma})\frac{\log d}{n}\left\|\theta^*_{S^C}\right\|_1^2$.

$$\begin{aligned}\frac{1}{4}c_1\bar{\kappa}\left\|\hat{\Delta}\right\|_2^2 &\leq \frac{1}{n}\left\|\mathbf{X}\hat{\Delta}\right\|_2^2 \\ &\leq \lambda_n\left[3\sqrt{|S|}\left\|\hat{\Delta}\right\|_2 + 2\left\|\theta^*_{S^C}\right\|\right]\end{aligned} \tag{45}$$

   Solving for $\left\|\hat{\Delta}\right\|_2$:

$$0 \leq \left\|\hat{\Delta}\right\|_2 \leq \frac{1}{2}\left[\frac{12\lambda_n\sqrt{|S|}}{c_1\bar{\kappa}} + \sqrt{\frac{144\lambda_n^2|S|}{c_1^2\bar{\kappa}^2} + \frac{32\lambda_n\left\|\theta^*_{S^C}\right\|_1}{c_1\bar{\kappa}}}\right] \tag{46}$$

$$\begin{aligned}0 \leq \left\|\hat{\Delta}\right\|_2^2 &\leq \frac{1}{4}\left[\frac{12\lambda_n\sqrt{|S|}}{c_1\bar{\kappa}} + \sqrt{\frac{144\lambda_n^2|S|}{c_1^2\bar{\kappa}^2} + \frac{32\lambda_n\left\|\theta^*_{S^C}\right\|_1}{c_1\bar{\kappa}}}\right]^2 \\ &\leq \frac{1}{4}(1^2+1^2)\left[\frac{288\lambda_n^2|S|}{c_1^2\bar{\kappa}^2} + \frac{32\lambda_n\left\|\theta^*_{S^C}\right\|_1}{c_1\bar{\kappa}}\right] \\ &= \frac{144\lambda_n^2}{c_1^2\bar{\kappa}^2}|S| + \frac{16\lambda_n}{c_1\bar{\kappa}}\left\|\theta^*_{S^C}\right\|_1\end{aligned} \tag{47}$$

2. Otherwise, $\frac{1}{4}c_1\bar{\kappa}\left\|\hat{\Delta}\right\|_2^2 < 8c_2\rho^2(\mathbf{\Sigma})\frac{\log d}{n}\left\|\theta^*_{S^C}\right\|_1^2$.

$$\left\|\hat{\Delta}\right\|_2^2 < \frac{32c_2\rho^2(\mathbf{\Sigma})}{c_1\bar{\kappa}}\frac{\log d}{n}\left\|\theta^*_{S^C}\right\|_1 \tag{48}$$

   Combining both cases by summing up both upper bounds,

$$\left\|\hat{\Delta}\right\|_2^2 = \left\|\hat{\theta} - \theta^*\right\|_2^2 \leq \frac{144\lambda_n^2}{c_1^2\bar{\kappa}^2}|S| + \frac{16\lambda_n}{c_1\bar{\kappa}}\left\|\theta^*_{S^C}\right\|_1 + \frac{32c_2\rho^2(\mathbf{\Sigma})}{c_1\bar{\kappa}}\frac{\log d}{n}\left\|\theta^*_{S^C}\right\|_1 \tag{49}$$

which completes the proof.

□

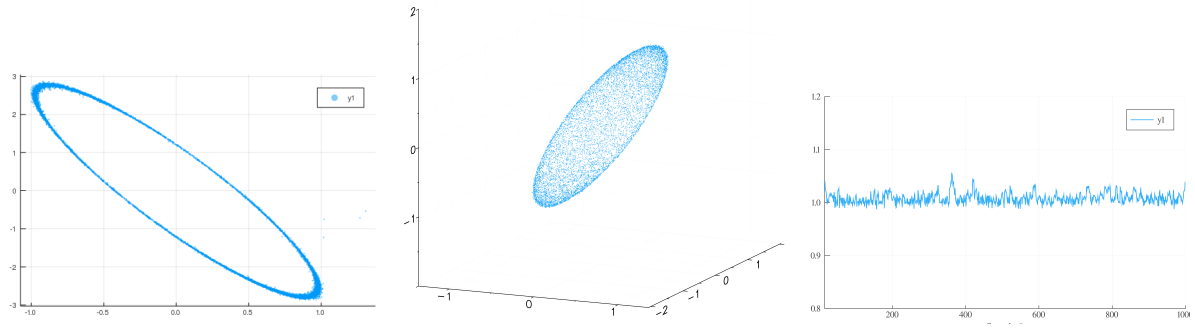

Figure 3: The demonstration of sampling from a high-dimensional ellipse surface using random-walk Metropolis-Hastings algorithm. (Left) The two-dimensional ellipse. (Middle) The three-dimensional case. (Right) The 500-dimensional with x- and y-axis being the sample number and $x'\mathbf{\Sigma}x$ value, which is close to 1 by definition.

Simulation Studies

I am interested in running simulations to check and visualize the bound in Eq. (20), i.e. $P\left[\mathcal{A}(r_l, r_u)\right] \leq e^{-\frac{n}{32}} e^{-\frac{n}{2}r_u^2}$. The major challenge here is to sample θ from the space $\theta \in \mathbb{K}(r_l, r_u)$, or equivalently, to sample from an ellipse surface in high dimensional space: $\mathbb{S}^{d-1}(\mathbf{\Sigma}) = \left\{\theta \in \mathbb{R}^d \middle| \theta'\mathbf{\Sigma}\theta = 1\right\}$. This challenge can be tackled by the random-walk Metropolis-Hastings algorithm described as follows.

1. Generate the $(k+1)^{\text{th}}$ sample x_{k+1} from x_k.

$$x_{k+1} \sim \mathcal{N}(x_k, \sigma^2 I_d) \tag{50}$$

2. Calculate the un-normalized probability from the symmetric kernel which decays super fast with the distance between x_{k+1} and the ellipse surface.

$$p(x_{k+1}|x_k) \propto e^{-k|x'_{k+1}\mathbf{\Sigma}x_{k+1}-1|} \tag{51}$$

3. Compute the acceptance rate α.

$$\alpha = \min\left(1, \frac{e^{-k|x'_{k+1}\mathbf{\Sigma}x_{k+1}-1|}}{e^{-k|x'_k\mathbf{\Sigma}x_k-1|}}\right) \tag{52}$$

4. Generate a random variable $u \sim U[0,1]$. If $u < \alpha$, accept x_{k+1}.

The samples $x_1, x_2, \ldots, x_N$ will be on the surface of this high dimensional ellipse. I demonstrate the performance of the random-walk M-H algorithm in Fig. 3. With this sampling tool, the goal is to numerically examine the bound for $P\left[\mathcal{A}(r_l, r_u)\right]$.

I ran the simulation using 25,000 samples on the 500-dimensional ellipse surface for $r_u \in \left\{\frac{1}{4}, \frac{1}{2}, 1, \ldots, 18024\right\}$ and the number of rows in $\mathbf{X}$, $n \in \{50, 60, 70, \ldots, 250\}$. The sparsity setup is satisfied for $d > n$. For each combination of radii pair $(\frac{1}{2}r_u, r_u)$ and n, I didn't find any event of $\mathcal{A}(r_l, r_u)$. The reason is the following. The event $\mathcal{A}(r_l, r_u)$ requires small r_u, i.e. $r_u \leq \frac{1}{4}$ but under such condition, the set $\mathbb{K}(r_l, r_u)$ is empty. As r_u increases, the set $\mathbb{K}(r_l, r_u)$ is non-empty but the event $\mathcal{A}(r_l, r_u)$ becomes impossible because $\|\mathbf{X}\theta\|_2 \geq 0$. There might be a sweet spot

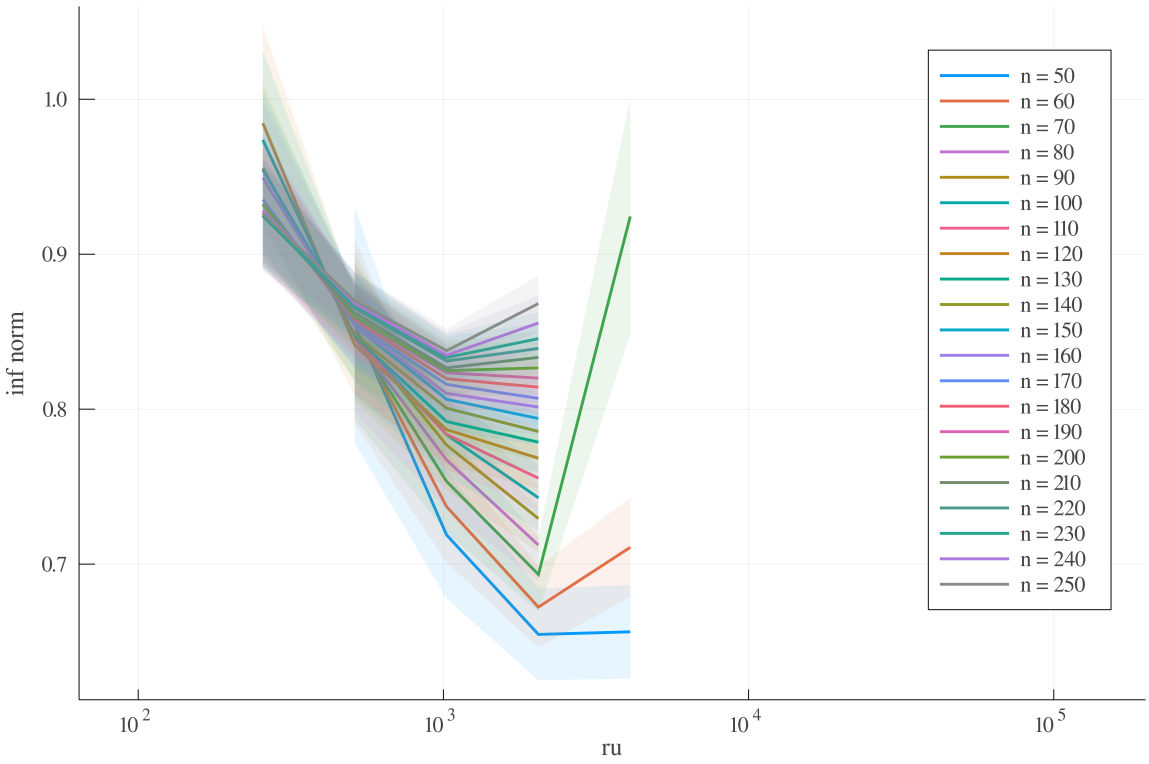

Figure 4: The simulation result for the *inf-norm* using all the combinations of n andr_u. See the main text for detailed explanation.

between these two trade-offs depending on the design of the matrix $\mathbf{\Sigma}$ but the event is extremely rare if not impossible.

Essentially, the aim for Theorem 7.16 and Lemma 7.24 is to investigate the quantity $\inf_\theta \frac{1}{\sqrt{n}} \|\mathbf{X}\theta\|_2$ which I refer to it as *inf-norm*. The *inf-norm* for my simulation is shown in Fig. 4 for all the n tested. The shaded regions represent the standard deviations from 1000 generated random matrices $\mathbf{X}$. Notice that if r_u is too large or too small, the set $\mathbb{K}(r_l, r_u)$ is empty and there is no statistics to report. For this specific matrix $\mathbf{\Sigma}$, I found that only when $r_u \in [128, 4096]$ does the corresponding $l1$-ball intersect with the ellipse. However, for such r_u, $P[\mathcal{A}(r_l, r_u)] = 0 \leq \epsilon$ for an arbitrary $\epsilon \geq 0$. Therefore, the bounds in Theorem 7.16 is very loose and better bound (but still loose) can achieved by picking c_1 and c_2 more carefully.