LASSO and Instrumental Variable

Dingyi Li, Yuchen Xu, Ritwik Sadhu, Yixin Shen, Yen-Lin Chen

November 2019

In this report, we will pave the steps toward proving the LASSO oracle inequality, Theorem 7.19 of the book and use Gaussian case to obtain a better bound in its application on Instrumental Variable. We will derive 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 $y = \mathbf{X}\theta^* + w$.

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

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

1 Theorem 7.13

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 eigenvalue 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 $\mathbb{C}_\alpha(S) = \left\{\Delta \in \mathbf{R}^d | \|\Delta_{S^C}\|_1 \leq \alpha \|\Delta_S\|_1\right\}$ 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 we summarize it briefly. $\hat{\theta}$ 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.

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

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

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

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

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

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

□

Theorem 7.13 bases on the restricted eigenvalue condition of $\mathbf{X}$. Let's consider the general 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 $\frac{1}{n}\|\mathbf{X}\theta\|_2^2$.

2 Theorem 7.16

Theorem 7.16. *Consider a random matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ with iid rows $x_i \in \mathbb{R}^d$ 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 $2\rho(\mathbf{\Sigma})\sqrt{\frac{\log d}{n}}\|\theta\|_1 \geq 0$.

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

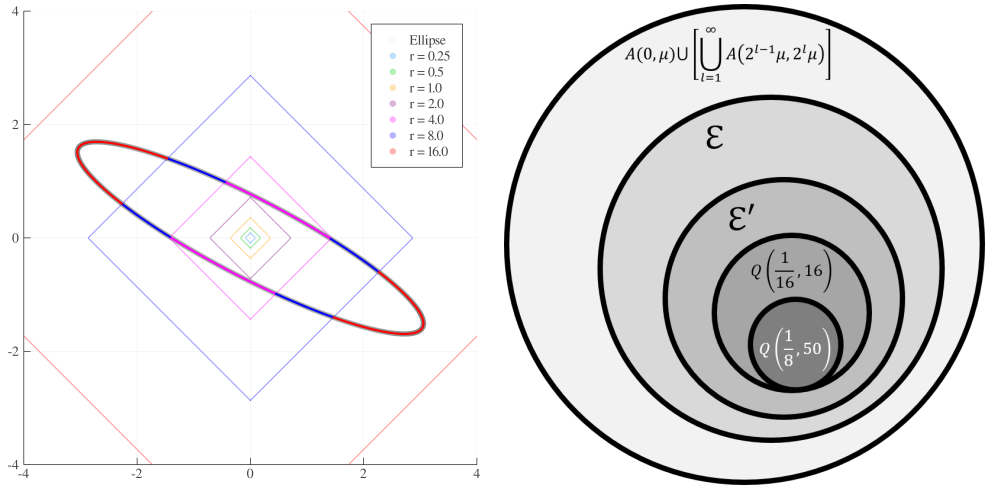

Figure 1: (Left) The two-dimensional illustration of the set $\mathbb{K}(r_l, r_u)$ with increasing r_l and r_u. (Right) The diagram of all the event sets discussed.

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 (left). 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, we will derive the general values of μ in the following proof.

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

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

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

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

   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 $l \in \mathbb{N} \Longrightarrow 2\rho(\mathbf{\Sigma})\sqrt{\frac{\log d}{n}} \left\|\theta\right\|_1 \geq 2^{l-1}\mu$.

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

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

   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 .

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

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

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

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

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

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

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

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,

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

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

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

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

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

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

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

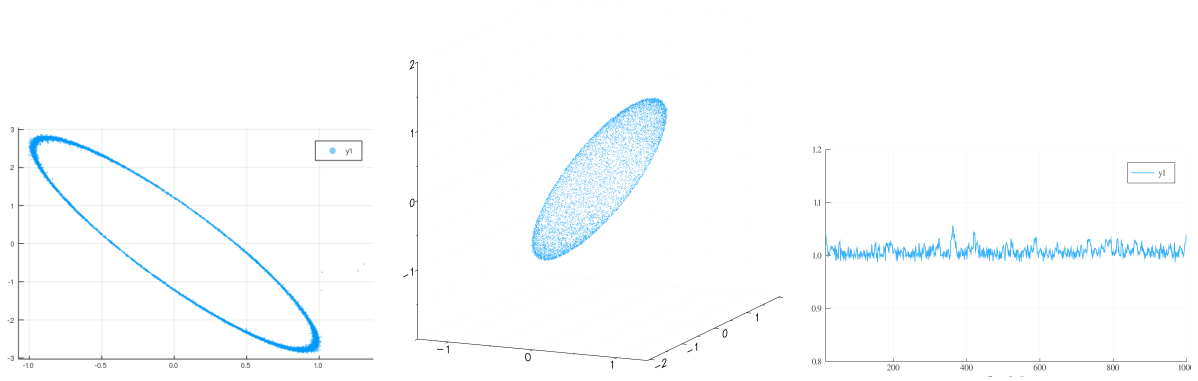

Figure 2: The demonstration of sampling from a high-dimensional ellipse surface using random-walk Metropolis-Hastings algorithm. (Left) The two-dimensional ellipse. (Middle) The three-dimensional case. (Right) The 500-dimensional with x- and y-axis being the sample number and $x'\mathbf{\Sigma}x$ value, which is close to 1 by definition.

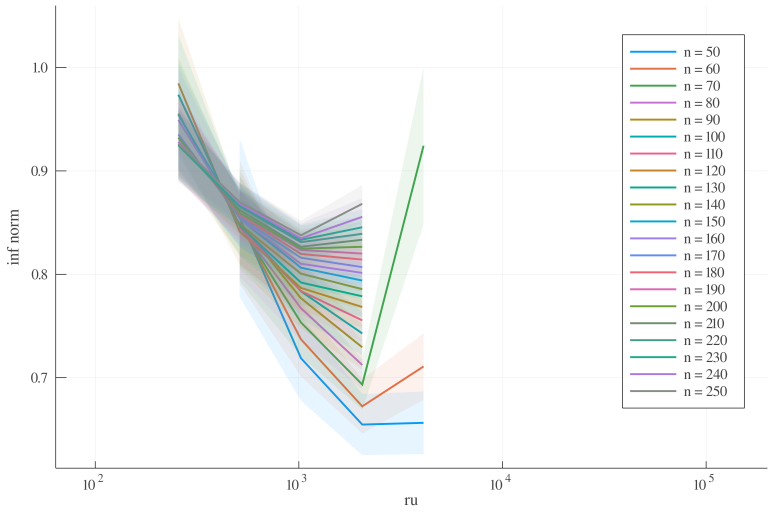

Figure 3: The simulation result for the *inf-norm* using all the combinations of n and r_u. See the main text for detailed explanation.

It is helpful to visualize the sets of the event discussed and used so far. Fig. 1 (right)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 $\|\mathbf{X}\theta\|_2 \geq 0$.

2.1 Simulation Studies

We are 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. We demonstrate the performance of the random-walk M-H algorithm in Fig. 2.

We 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, we 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 between these two trade-offs depending on the design of the matrix $\mathbf{\Sigma}$ but the event will still be 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 we refer to it as *inf-norm*. The *inf-norm* for our simulation is shown in Fig. 3 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 $\mathbf{\Sigma}$, we found that only when $r_u \in [128, 4096]$ does the corresponding $l1$-ball intersect with the ellipse. However, for such r_u, $P\left[\mathcal{A}(r_l, r_u)\right] = 0 \leq \epsilon$ 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.

3 Theorem 7.19

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*

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

where the cardinality of S *satisfies*

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

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

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

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

Now use Theorem 7.16.

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

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

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

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

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

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

   Combining both cases by summing up both upper bounds,

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

which completes the proof. □

4 Bounds on prediction error

We call $X(\hat{\theta} - \theta^*)$ as the prediction error, since it measures how our predicted mean for y compares with it's true mean.

Theorem 7.20. *Let:*

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

as usual. Consider the Lagrangian LASSO:

$$\min_{\theta} \frac{1}{2n}\|y - X\theta\|^2 + \lambda_n\|\theta\|_1$$

with a strictly positive regularization parameter $\lambda_n \geq 2\|\frac{X^T w}{n}\|_\infty$. *Let* $\hat{\theta}$ *be any optimal solution, and put* $\hat{\Delta} = \hat{\theta} - \theta^*$. *Then,*

(a) $\frac{\|X\hat{\Delta}\|_2^2}{n} \leq 12\|\theta^*\|_1\lambda_n$

(b) If further θ^* *is supported on a subset* S *of cardinality* s, *and* X *satisfies the* $RE(\kappa; 3)$ *condition over* S, *we have:*

$$\frac{\|X\hat{\Delta}\|_2^2}{n} \leq \frac{9}{\kappa}s\lambda_n^2$$

Proof. (a) Note that, since $\hat{\theta}$ is an optimal solution to the Lagrangian LASSO,

$$\begin{aligned} &\frac{1}{2n}\|y - \mathbf{X}\hat{\theta}\|_2^2 + \lambda_n\|\hat{\theta}\|_1 \leq \frac{1}{2n}\|y - \mathbf{X}\theta^*\|_2^2 + \lambda_n\|\theta^*\|_1 \\ \iff &\frac{1}{2n}\|w - \mathbf{X}(\hat{\theta} - \theta^*)\|_2^2 \leq \frac{1}{2n}\|w\|_2^2 + \lambda_n\left(\|\theta^*\|_1 - \|\hat{\theta}\|_1\right) \\ \iff &\frac{1}{n}\|\mathbf{X}\hat{\Delta}\|_2^2 \leq 2\frac{w^T\mathbf{X}\hat{\Delta}}{n} + 2\lambda_n\left(\|\theta^*\|_1 - \|\hat{\theta}\|_1\right) \end{aligned}$$

Now, by Hölder's Inequality,

$$\left|\frac{w^T X\hat{\Delta}}{n}\right| \leq \left\|\frac{X^T w}{n}\right\|_\infty \|\hat{\Delta}\|_1$$

Which, since $\lambda_n \geq 2\|\frac{X^T w}{n}\|_\infty$, implies:

$$\left|\frac{w^T X\hat{\Delta}}{n}\right| \leq \frac{\lambda_n}{2}\|\hat{\Delta}\|_1 \leq \frac{\lambda_n}{2}\left\{\|\theta^*\|_1 + \|\hat{\theta}\|_1\right\}$$

where the second step follows by triangle inequality. Taken together:

$$\begin{aligned} 0 \leq \frac{1}{n}\|\mathbf{X}\hat{\Delta}\|_2^2 \leq \lambda_n\{\|\theta^*\|_1 + \|\hat{\theta}\|_1\} + 2\lambda_n\left\{\|\theta^*\|_1 - \|\hat{\theta}\|_1\right\} \\ = \lambda_n\{3\|\theta^*\|_1 - \|\hat{\theta}\|_1\} \end{aligned}$$

i.e., $\lambda_n\{3\|\theta^*\|_1 - \|\hat{\theta}\|_1\} \geq 0$. Since $\lambda_n > 0$ by assumption, the result follows.

(b) From the previous part:

$$\frac{1}{n}\|\mathbf{X}\hat{\Delta}\|_2^2 \leq \frac{1}{n}w^T X\hat{\Delta} + \lambda_n\{\|\theta^*\|_1 - \|\hat{\theta}\|_1\}$$

Now, if θ^* is supported on S, we have:

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

where the subscript S indicates the vector restricted to coefficients in S. Substituting this in the last equation we get:

$$\frac{1}{n}\|\mathbf{X}\hat{\Delta}\|_2^2 \leq 2\frac{w^T X\hat{\Delta}}{n} + 2\lambda_n\{\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1\}$$

But as in part (a), we'd have, by Hölder's inequality:

$$\left|2\frac{w^T X\hat{\Delta}}{n}\right| \leq \lambda_n\|\hat{\Delta}\|_1 = \lambda_n\{\|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1\}$$

Taken together:

$$\begin{aligned}\frac{1}{n}\|\mathbf{X}\hat{\Delta}\|_2^2 &\leq \lambda_n\left\{\|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{S^c}\|_1\right\} + 2\lambda_n\{\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1\} \\ &= \lambda_n\{3\|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{S^c}\|_1\} \\ &\leq 3\lambda_n\|\hat{\Delta}_S\|_1 \leq 3\sqrt{s}\lambda_n\|\hat{\Delta}_S\|_2 \qquad \text{[Cauchy-Schwarz]} \\ &\leq 3\sqrt{s}\lambda_n\|\hat{\Delta}\|_2\end{aligned}$$

From the above calculation we also have $3\|\hat{\Delta}_S\|_2 - \|\hat{\Delta}_{S^c}\| \geq 0$, i.e., $\hat{\Delta} \in \mathcal{C}_3(S)$. Hence, since X follows $RE(\kappa; 3)$ property over S, we also have:

$$\frac{\|X\hat{\Delta}\|_2^2}{n} \geq \kappa\|\hat{\Delta}\|_2^2 \implies \|\hat{\Delta}\|_2 \geq \frac{\|X\hat{\Delta}\|_2}{\sqrt{n\kappa}}$$

which, when substituted into the preceding inequality gives:

$$\frac{1}{n}\|\mathbf{X}\hat{\Delta}\|_2^2 \leq 3\sqrt{s}\lambda_n\frac{\|X\hat{\Delta}\|_2}{\sqrt{n\kappa}} \implies \frac{\|X\hat{\Delta}\|_2}{\sqrt{n}} \leq \frac{3\sqrt{s}\lambda_n}{\sqrt{\kappa}}$$

which, when squared, gives us the requisite result.

□

4.1 Example: Classical Linear Regression with Sub-Gaussian Errors

Consider now a fixed design model:

$$y = X\theta + w$$

where the errors w_i are 0-mean sub-Gaussian with variance proxy σ^2. Let the true parameter θ^* belong to a sparse subset S with cardinality s, and suppose the $n \times d$ design matrix X satisfies $REP(\kappa; 3)$ over S, and is also C-column normalized, i.e.

$$\max_{j=1}^{d}\frac{\|X_j\|_2}{\sqrt{n}} \leq C \tag{50}$$

Then, $X_j'w$ is 0 mean subGaussian with variance proxy $\sigma^2\|X_j\|_2^2$, and we have:

$$\begin{aligned}
\mathbb{E}\left[\exp\left(s\left\|\frac{X'w}{n}\right\|_\infty\right)\right] &= \mathbb{E}\left[\exp\left(s\max_{j=1}^{d}\frac{|X_j'w|}{n}\right)\right] \\
&\leq \sum_{j=1}^{d}\mathbb{E}\left[\exp\left(s\frac{|X_j'w|}{n}\right)\right] \\
&\leq \sum_{j=1}^{d} 2\exp\left(\frac{s^2\|X_j\|_2^2\sigma^2}{2n^2}\right) \qquad [X_j'w \sim subG(\sigma^2\|X_j\|_2^2)] \\
&\leq \sum_{j=1}^{d} 2\exp\left(\frac{s^2nC^2\sigma^2}{2n^2}\right) \qquad [\text{ By (50)}] \\
&\leq 2d\exp\left(\frac{s^2C^2\sigma^2}{2n}\right)
\end{aligned}$$

Now, by Markov's inequality on $\exp\left(s\|\frac{X'w}{n}\|_\infty\right)$, we have:

$$\mathbb{P}\left[\|\frac{X'w}{n}\|_\infty > C\sigma\left(\sqrt{\frac{2\log d}{n}}+\delta\right)\right] \leq \frac{\mathbb{E}\left[\exp\left(s\left\|\frac{X'w}{n}\right\|_\infty\right)\right]}{\exp\left(sC\sigma\left(\sqrt{\frac{2\log d}{n}}+\delta\right)\right)} \tag{51}$$

$$\leq 2\exp\left(\frac{s^2C^2\sigma^2}{2n}+\log d - sC\sigma\left(\sqrt{\frac{2\log d}{n}}+\delta\right)\right) \tag{52}$$

$$= 2\exp\left(\left(\frac{sC\sigma}{\sqrt{2n}}-\sqrt{\log d}\right)^2 - \delta sC\sigma\right) \tag{53}$$

Put $z = sC\sigma$, $g(z) = \left(\frac{z}{\sqrt{2n}}-\sqrt{\log d}\right)^2 - \delta z$. Then, setting $g'(z) = 0$ and solving, we get:

$$\begin{aligned}
2\left(\frac{z}{\sqrt{2n}}-\sqrt{\log d}\right)/\sqrt{2n} - \delta &= 0 \\
\implies z &= \sqrt{2n\log d} + n\delta
\end{aligned}$$

Substituting this into (53), we get:

$$\begin{aligned}
\mathbb{P}\left[\left\|\frac{X'w}{n}\right\|_\infty > C\sigma\left(\sqrt{\frac{2\log d}{n}}+\delta\right)\right] &\leq 2\exp\left(-\frac{n\delta^2}{2}-\delta\sqrt{2n\log d}\right) \\
&\leq 2\exp\left(-\frac{n\delta^2}{2}\right)
\end{aligned}$$

Hence, setting $\lambda_n = 2C\sigma\left(\sqrt{\frac{2\log d}{n}}+\delta\right)$, we'd get:

$$\mathbb{P}\left[\lambda_n \geq 2\left\|\frac{X'w}{n}\right\|_\infty\right] \geq 1 - 2\exp\left(-\frac{n\delta^2}{2}\right)$$

Also, on the above event, by Theorem 7.13(a), we'd have:

$$\|\hat{\theta}-\theta^*\|_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda_n = \frac{6C\sigma}{\kappa}\sqrt{s}\left\{\sqrt{\frac{2\log d}{n}}+\delta\right\}$$

which gives us a bound on the estimation error with high probability. On the other hand, we can also get the following (standardized) prediction error bound on the above event, based on our assumptions on the design matrix:

$$\frac{\|X\hat{\Delta}\|_2^2}{n} \leq \frac{9}{\kappa} s\lambda_n^2 = \frac{36}{\kappa} C^2\sigma^2 s \left(\sqrt{\frac{2\log d}{n}} + \delta\right)^2 \leq \frac{72}{\kappa} C^2\sigma^2 \left(\frac{2s\log d}{n} + s\delta^2\right)$$

Thus, the convergence rate for the standardized prediction error can be much faster than that for the estimation error.

4.2 Simulations on Example 7.14

Let $S = \{1, \ldots s\}$. Consider the following X matrix:

$$X = \left[\sqrt{n}P_{n,s}, \frac{1}{n(d-s)} J_{n,d-s}\right]$$

Then,

$$X'X = \begin{bmatrix} nI_s & 0 \\ 0 & \frac{1}{d-s} J_{d-s} \end{bmatrix}$$

We have: $\|\hat{\Delta}_{S^c}\|_2 \leq \|\hat{\Delta}_{S^c}\|_1 \leq \sqrt{s}\|\hat{\Delta}_{S^c}\|_2$, which would imply, for any $\hat{\Delta} \in C_3(S)$:

$$\begin{aligned} \|\hat{\Delta}\|_2^2 &= \|\hat{\Delta}_S\|_2^2 + \|\hat{\Delta}_{S^c}\|_2^2 \\ &\leq \|\hat{\Delta}_S\|_2^2 + \|\hat{\Delta}_{S^c}\|_1^2 \\ &\leq \|\hat{\Delta}_S\|_2^2 + 9\|\hat{\Delta}_S\|_1^2 \\ &\leq \|\hat{\Delta}_S\|_2^2 + 9s\|\hat{\Delta}_S\|_2^2 = (1+9s)\|\hat{\Delta}_S\|_2^2 \end{aligned}$$

Hence, $\|X\hat{\Delta}\|_2^2/n \geq \|\hat{\Delta}_S\|_2^2 \geq \frac{1}{1+9s}\|\hat{\Delta}\|_2^2$, i.e., X satisfies $REP(\frac{1}{1+9s}; 3)$ over S. It is also, obviously, 1 column normalized for $n > s$. Also, fix the error variance to be 0.1, and take delta to be 0.1. We then have, for $\lambda_n = 0.2\left(\sqrt{\frac{2\log d}{n}} + 0.1\right)$

$$\frac{\|X\hat{\Delta}\|_2^2}{n} \leq 0.72(1+9s)\left(\frac{2s\log d}{n} + 0.01s\right)$$

$$\|\hat{\Delta}\|_2^2 \leq 0.6(1+9s)\left(\sqrt{\frac{2s\log d}{n}} + 0.1\sqrt{s}\right)$$

. We carry out the simulations with Gaussian errors for $s = 5, \ldots, 20$ in intervals of 5 and $n = 2000, \ldots, 6000$ in intervals of 1000. The generated plots are given below:

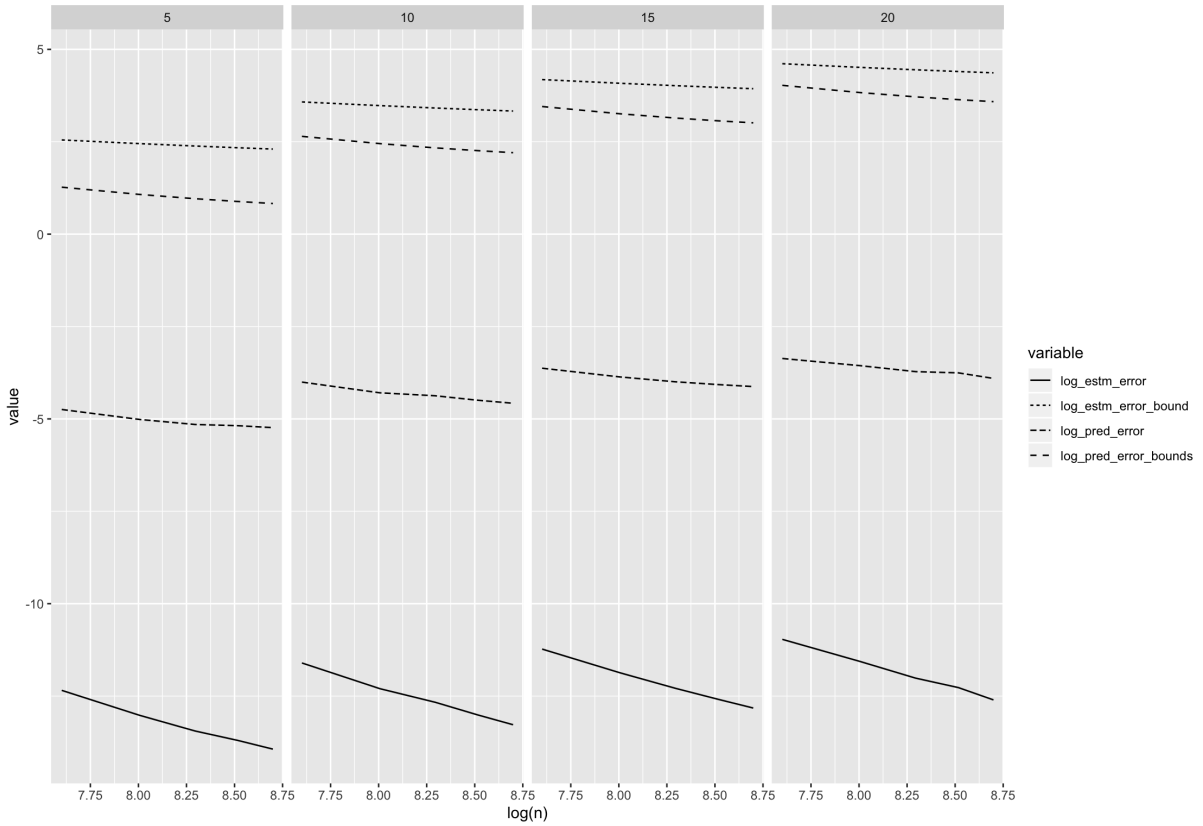

Figure 4: Plot of sample errors and theoretical bounds in log scale for different values of sparsity dimension s

5 Theorem 7.21

Assumptions needed:

1. *Lower eigenvalue:* The smallest eigenvalue of the sample covariance submatrix indexed by S is bounded below:

$$\gamma_{\min}(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n}) \geq c_{\min} > 0.$$

2. *Mutual incoherence:* There exists $\alpha \in [0, 1)$ such that

$$\max_{j \in S^c} \|(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T X_j\|_1 \leq \alpha.$$

And the theorem of consistency for variable selection is stated as:

Theorem 7.21. *For an S-sparse linear regression model for which the design matrix satisfies the above two assumptions. Then for any choice of regularization parameter such that*

$$\lambda_n \geq \frac{2}{1-\alpha}\|\mathbf{X}_{S^c}^T\Pi_{S^\perp}(\mathbf{X})\frac{w}{n}\|_\infty,$$

the Lagrangian Lasso (7.18) has the following properties:

(a) Uniqueness: There is a unique optimal solution $\hat{\theta}$.

(b) No false inclusion: This solution has its support set $\hat{S}$ contained within the true support set S.

(c) ℓ_∞-bounds: The error $\hat{\theta} - \theta^*$ satisfies

$$\|\hat{\theta}_S - \theta^*_S\|_\infty \leq \underbrace{\|(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n})^{-1}\mathbf{X}_S^T\frac{w}{n}\|_\infty + \|(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n})^{-1}\|_\infty\lambda_n}_{=:B(\lambda_n;\mathbf{X})}.$$

(d) No false exclusion: The Lasso includes all indices $i \in S$ such that $|\theta^*_i| > B(\lambda_n; \mathbf{X})$ and hence is variable selection consistent if $\min_{i\in S}|\theta^*_i| > B(\lambda_n; \mathbf{X})$.

Proof. Using the subgradient, the zero-subgradient condition is

$$\frac{1}{n}\mathbf{X}^T(\mathbf{X}\hat{\theta} - y) + \lambda_n\hat{z} = 0. \tag{54}$$

The proof is based on a constructive procedure, known as the *primal-dual witness method* (PDW construction):

1. Set $\hat{\theta}_{S^c} = 0$.
2. Determinr $(\hat{\theta}_S, \hat{z}_S) \in \mathbb{R}^S \times \mathbb{R}^S$ by solving

$$\hat{\theta}_S \in \arg\min_{\theta_S\in\mathbb{R}^S}\{\underbrace{\frac{1}{2n}\|y - \mathbf{X}_S\theta_S\|_2^2}_{=:f(\theta_S)} + \lambda_n\|\theta_S\|_1\}, \tag{55}$$

   and then choosing $\hat{z}_S \in \partial\|\hat{\theta}_S\|_1$ such that $\nabla f(\theta_S)|_{\theta_S=\hat{\theta}_S} + \lambda_n\hat{z}_S = 0$.
3. Solve for $\hat{z}_{S^c} \in \mathbb{R}^{d-s}$ based on equation (54) and check whether $\|\hat{z}_{S^c}\|_\infty < 1$.

Then using the fact that $\hat{\theta}_{S^c} = \theta^*_{S^c} = 0$, we rewrite the equation (54) as

$$\frac{1}{n}\begin{pmatrix}\mathbf{X}_S^T\mathbf{X}_S & \mathbf{X}_S^T\mathbf{X}_{S^c}\\ \mathbf{X}_{S^c}^T\mathbf{X}_S & \mathbf{X}_{S^c}^T\mathbf{X}_{S^c}\end{pmatrix}\begin{pmatrix}\hat{\theta}_S - \theta^*_S\\ 0\end{pmatrix} - \frac{1}{n}\begin{pmatrix}\mathbf{X}_S^Tw\\ \mathbf{X}_{S^c}^Tw\end{pmatrix} + \lambda_n\begin{pmatrix}\hat{z}_S\\ \hat{z}_{S^c}\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix}, \tag{56}$$

here we use the fact that $y = \mathbf{X}_S\theta^*_S + w$. Here we need the following lemma:

Lemma I. *f the lower eigenvalue assumption holds, then success of PDW construction implies $(\hat{\theta}_S, 0) \in \mathbb{R}^d$ is the unique optimal solution.*

Proof of lemma: Assume $\hat{\theta} = (\hat{\theta}_S, 0)$ is an optimal solution with associated subgradient $\hat{z}$ and by definition we have $\langle\hat{z}, \hat{\theta}\rangle = \|\hat{\theta}\|_1$. Suppose we have another optimal solution $\tilde{\theta}$ then we are guaranteed that $F(\hat{\theta}) + \lambda_n\langle\hat{z}, \hat{\theta}\rangle = F(\tilde{\theta}) + \lambda_n\|\tilde{\theta}\|$ where $F(\theta) = \frac{1}{2n}\|y - \mathbf{X}\theta\|_2^2$, hence using the fact that $\lambda_n\hat{z} = -\nabla F(\hat{\theta})$ and rearranging a little bit,

$$F(\hat{\theta}) + \langle\nabla F(\hat{\theta}), \tilde{\theta} - \hat{\theta}\rangle - F(\tilde{\theta}) = \lambda_n(\|\tilde{\theta}\| - \langle\hat{z}, \tilde{\theta}\rangle).$$

By convexity of F we have the left-hand side is negative and then $\|\tilde{\theta}\|_1 \leq \langle\hat{z}, \tilde{\theta}\rangle$. But we also have $\langle\hat{z}, \tilde{\theta}\rangle \leq \|\hat{z}\|_\infty\|\tilde{\theta}\|_1$ we must have $\|\tilde{\theta}\|_1 = \langle\hat{z}, \tilde{\theta}\rangle$ which implies $\tilde{\theta}_j = 0$ if $j \in S^c$. In conclusion the optimal solution must be supported on S and hence can be obtained from solving the PDW construction. In addition when the lower eigenvalue assumption holds the problem (55) is strictly convex and hence minimizer is unique.□

Now we have proved the lemma and in order to prove (a) and (b) in the theorem it suffices to show that $\hat{z}_{S^c}$ in step 3 satisfies $\|\hat{z}_{S^c}\| < 1$. From 56 we have

$$\hat{z}_{S^c} = -\frac{1}{\lambda_n n}\mathbf{X}_{S_c}^T\mathbf{X}_S(\hat{\theta}_S - \theta^*_S) + \mathbf{X}_{S^c}^T(\frac{w}{\lambda_n n}), \tag{57}$$

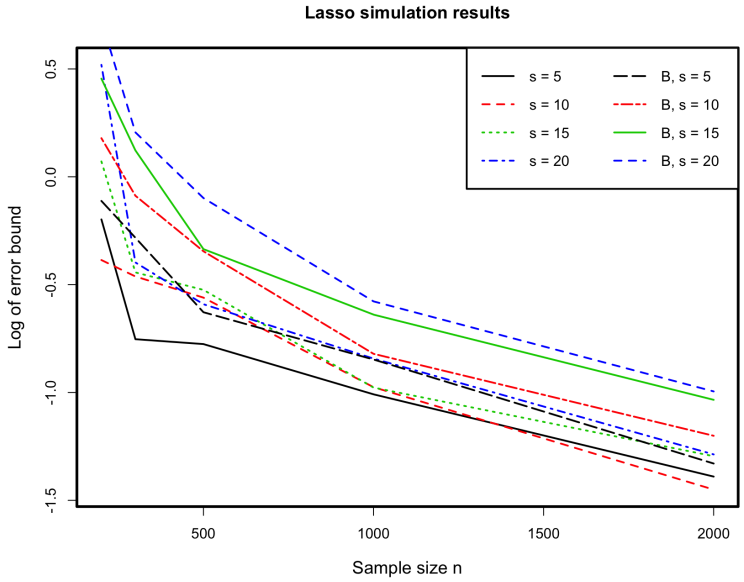

Figure 5: The simulation results for the variable selection recovery, in terms of $\log(\|\hat{\theta} - \theta^*\|_\infty)$. Here the legends starting with letter '**B**' represent the plot for oracle bound $B(\lambda_n; \mathbf{X})$ for each case.

and

$$\hat{\theta}_S - \theta^*_S = (\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T w - \lambda_n(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\hat{z}_S. \tag{58}$$

Doing substitution we get

$$\hat{z}_{S^c} = \underbrace{\mathbf{X}_{S_c}^T\mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\hat{z}_S}_{\mu} + \underbrace{\mathbf{X}_{S^c}^T[I - \mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\mathbf{X}_S^T](\frac{w}{\lambda_n n})}_{V_{S^c}}, \tag{59}$$

then by triangular inequality and the mutual incoherence assumption we have that $\|\mu\|_\infty \leq \alpha$ and by the choice of λ_n we have $\|V_{S^c}\| \leq \frac{1-\alpha}{2}$, hence $\|\hat{z}_{S^c}\| \leq \frac{1+\alpha}{2} < 1$.

In order to show the ℓ_∞ bound, we refer to 58 and get that

$$\|\hat{\theta}_S - \theta^*_S\|_\infty \leq \|(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n})^{-1}\mathbf{X}_S^T\frac{w}{n}\|_\infty + \|(\frac{\mathbf{X}_S^T\mathbf{X}_S}{n})^{-1}\|_\infty\lambda_n, \tag{60}$$

and it completes the proof. □

5.1 Simulation

In this part we include the simulation that satisfies all the assumptions and make all non-zero coefficients greater than $B(\lambda_n; \mathbf{X})$. Then by running the built-in package for Lasso we could analyze the performance by comparing θ^* and $\hat{\theta}$. Here we design the simulations for different triples of $(n, d = 10n, s)$ and get the plot.

6 Instrumental Variable

6.1 Definition

Consider the simplest classical homoskedastic IV model:

$$y_t = x_t'\beta + u_t \tag{61}$$

$$x_t = z_t'\pi + v_t, \tag{62}$$

where y_t is one-dimensional, x_t is $n \times 1$, z_t is $k \times 1$ and $Eu_t z_t = 0$, one observes i.i.d. data $\{y_t, x_t, z_t\}$ (WLLN).

With the data, we want to estimate Equation (61). However, different with classical models, u_t and v_t are correlated, where $(u_t, v_t)' \sim N(0, [\sigma_u^2, \sigma_{uv}; \sigma_{uv}, \sigma_v^2])$, and thus, x_t is an endogenous regressor. Therefore, OLS provides inconsistent estimators with the violation of the classical assumption.

Assume that $n \leq k$. z_t is exogenous, if it also relevant ($Z'X$ has rank n), then it can serve as instrument and the model is identified. Under homoskedastic, the usual TSLS from efficient GMM will be,

$$\hat{\beta}_{TSLS} = (X' P_Z X)^{-1} X' P_Z Y \tag{63}$$

where $P_Z = Z(Z'Z)^{-1}Z'$.

Plug in $Y = X\beta + u$ into Equation (63) with WLLN . We can present consistency.

6.2 Weak IV

For the weak IV $\pi = 0$ here, we consider a simple case. We assume that x_t is one dimensional. Moreover, for simplicity, we assume $\epsilon_t = \beta v_t + u_t$ and $(\epsilon_t, v_t) \sim N(0, [\sigma_1^2, \rho\sigma_1\sigma_2; \rho\sigma_1\sigma_2, \sigma_2^2])$. We can derive, $\sigma_1 = \beta^2\sigma_v^2 + \sigma_u^2 + 2\beta\sigma_u v$, $\rho\sigma_1\sigma_2 = \beta\sigma_v^2 + \sigma_{uv}$ and $\sigma_2^2 = \sigma_v^2$. We can further assume, $\epsilon_t = \sigma_1(\rho v_t/\sigma_2 + \sqrt{1-\rho^2}v_0)$, where $v_0 \sim N(0,1)$ and v_0 is independent of v_2.

Then,

$$\begin{aligned}
\hat{\beta}_{TSLS} =& \frac{X' P_Z Y}{X' P_Z X} \\
=& \frac{(\pi' Z' + v') P_Z (\beta Z\pi + \epsilon)}{(\pi' Z' + v') P_Z (Z\pi + v)} \\
=& \beta + \frac{\pi' Z' (\sigma_{uv} v/\sigma_v^2 + \sigma_1\sqrt{1-\rho^2}v_0) + v_2' P_Z (\sigma_{uv} v_2/\sigma_v^2 + \sigma_1\sqrt{1-\rho^2}v_0)}{(\pi' Z' + v') P_Z (Z\pi + v)}
\end{aligned}$$

If $\pi = 0$ (weak instrument), we can get,

$$\begin{aligned}
\hat{\beta}_{TSLS} =& \rho\sigma_1/\sigma_2 + \sigma_1\sqrt{1-\rho^2}\frac{v_2' P_Z v_0}{v_2' P_Z v_2} \\
=& \beta + \sigma_{uv}/\sigma_v^2 + \sigma_1\sqrt{1-\rho^2}\frac{v_2' P_Z v_0}{v_2' P_Z v_2}
\end{aligned}$$

By the same way, we can derive the OLS estimator when $\pi = 0$ (an easier way is noticing that $P_Z = I$),

$$\hat{\beta}_{OLS} = \beta + \sigma_{uv}/\sigma_v^2 + \sigma_1\sqrt{1-\rho^2}\frac{v_2' v_0}{v_2' v_2}$$

Neither 2SLS or OLS provides consistent results.

7 LASSO Methods for Gaussian IVs

Belloni, A., Chernozhukov, V. and Hansen, C., 2011. Lasso methods for Gaussian instrumental variables models.

We use a similar model here as before:

$$y_t = x_t\beta_1 + w_t'\beta_2 + u_t \tag{64}$$
$$x_t = z_t'\pi + v_t, \tag{65}$$

x_t is an one dimensional endogenous variable. w_t is $k \times 1$.The sample size is n. z_t is $d \times 1$ where $d >> n$ and the cardinality of its support $s := |S(\pi)| << n$. We assume $s^2 log^2 d = o(n)$.

Theorem 1 (Generic Result on Optimal IV Estimation) Under our setting, assume the eigenvalues of $Q_n = E[A_i A_i']$, where $A_t = (z_t'\pi, w_t')'$, are bounded away from zero and from above uniformly in n. Let $\hat{D}_t = z_t'\hat{\pi}$ be a generic sparsity-based estimator of optimal instruments $D_t = z_t'\pi$ that obeys as n grows

$$||z_t'\hat{\pi} - z_t'\pi||_{2,n} + ||\mathbb{G}_n(z_t u_t)||_\infty ||\hat{\pi} - \pi||_1 = o_p(1). \tag{66}$$

Then the IV estimator based on $\hat{A}_t = (z_t'\hat{\pi}, w_t')'$ is consistent and asymptotically efficient.

Proof

Step 0. Recall that $A_t = (D_t, w_t')'$ are instruments and $d_t = (x_t, w_t')'$ are regressors for $t = 1, ..., n$. The condition that $\mathbb{E}_n[A_t A_t'] = Q_n$ has eigenvalues bounded from above uniformly in n implies that

$$\mathbb{E}_n[D_t^2] + \mathbb{E}_n[||w_t||^2] = \mathbb{E}_n[||A_t||^2] = trace(Q_n) \lesssim (1 + k)$$

is bounded from above uniformly in n.

Also, we have $\mathbb{E}_n[A_t u_t] \sim N(0, \sigma_u^2 Q_n/n)$ and $\mathbb{E}_n[A_t v_t] \sim N(0, \sigma_v^2 Q_n/n)$ so that, with the tail bounds,

$$\begin{aligned}
||\mathbb{E}_n[d_t u_t]|| &\leq |\mathbb{E}_n[v_t u_t]| + ||\mathbb{E}_n[A_t u_t]|| \lesssim_P \sigma_{uv} + \sigma_u\sqrt{(1+k)/n} \\
||\mathbb{E}_n[A_t v_t]|| &= |\mathbb{E}_n[D_t v_t]|^2 + ||\mathbb{E}_n[w_t v_t]||^2 \lesssim_P \sigma_v^2(1+k)/n \\
||d_t||_{2,n} &\leq ||v_t||_{2,n} + ||A_t||_{2,n} \lesssim_P \sigma_v + \sqrt{1+k}
\end{aligned}$$

Step 1. We have that by $E[u + t|A_t] = 0$

$$\begin{aligned}
\sqrt{n}(\hat{\beta} - \beta) =& \mathbb{E}_n[\hat{A}_t d_t']^{-1}\sqrt{n}\mathbb{E}_n[\hat{A}_t u_t] \\
=& \mathbb{E}_n[\hat{A}_t d_t']^{-1}\mathbb{G}_n[\hat{A}_t u_t] \\
=& (\mathbb{E}_n[A_t d_t'] + o_P(1))^{-1}(\mathbb{G}_n[A_t u_t] + o_P(1))
\end{aligned}$$

where by Steps 2 and 3 below:

$$\mathbb{E}_n[\hat{A}_t d_t'] = \mathbb{E}_n[A_t d_t'] + o_P(1) \tag{67}$$
$$\mathbb{G}_n[\hat{A}_t u_t] = \mathbb{G}_n[A_t u_t] + o_P(1) \tag{68}$$

Thus, since $\mathbb{E}_n[D_t v_t] = o_P(1)$ and $\mathbb{E}_n[w_t v_t) = o_P(1)$ by Step 0, note that $\mathbb{E}_n[A_t d_t'] = Q_n + o_P(1)$. Moreover, $Var(\mathbb{G}_n[A_t u_t]) = \sigma_u^2 Q_n$ has engenvalues bounded away from zero and bounded from above uniformly in n. Therefore,

$$\sqrt{n}(\hat{\beta} - \beta) = Q_n^{-1}\mathbb{G}_n[A_t u_t] + o_P(1)$$

and $Q_n^{-1}\mathbb{G}_n[A_t u_t]$ is a vector distributed as normal with mean zero and covariance $\sigma_u^2 Q_n^{-1}$.

Step 2. To show (67), note that $\hat{A}_t - A_t = (\hat{D}_t - D_t, 0')'$. Thus,

$$\begin{aligned} ||\mathbb{E}_n[(\hat{A}_t - A_t)d_i']|| = ||\mathbb{E}_n[(\hat{D}_t - D_t)d_i']|| \leq & \quad \mathbb{E}_n[||(\hat{D}_t - D_t)||\,||d_i||] \\ \leq & \quad \sqrt{\mathbb{E}_n[||(\hat{D}_t - D_t)||^2]\mathbb{E}_n[||d_t||^2]} \\ = & \quad ||(\hat{D}_t - D_t)||_{2,n}||d_t||_{2,n} \\ \lesssim_P & \quad ||(\hat{D}_t - D_t)||_{2,n} = o_P(1) \end{aligned}$$

by step 0, and the assumption.

Step 3. To show (68), note that

$$\begin{aligned} ||\mathbb{G}_n[(\hat{A}_t - A_t)u_t]|| &= ||\mathbb{G}_n[(\hat{D}_t - D_t)u_t]|| \\ &= ||\mathbb{G}_n[z_t'(\hat{\pi} - \pi)u_t]|| \\ &= ||\sum_{j=1}^{d} \mathbb{G}_n(z_{tj}u_t)'(\hat{\pi}_j - \pi_j)|| \\ &\leq ||\mathbb{G}_n(z_t u_t)||_\infty ||\hat{\pi} - \pi||_1 = o_P(1) \end{aligned}$$

We can also show the consistency of the variance estimator in the homoscedastic case with $\hat{\sigma}_u^2$ and $\hat{Q}_n = \mathbb{E}_n[\hat{A}_t\hat{A}_t']$.

Theorem 2 (Generic Result on Optimal IV Estimation) Under our setting, assume the eigenvalues of $Q_n = E[A_i A_i']$, where $A_t = (z_t'\pi, w_t')'$, are bounded away from zero and from above uniformly in n. RE holds for $E_n[z_t z_t']$, $s^2 log^2 d = o(n)$. Let $\hat{D}_i = z_t'\hat{\pi}$ where $\hat{\pi}$ is the LASSO. Then the IV estimator based on $\hat{A}_t = (z_t'\hat{\pi}, w_t')'$ is consistent and asymptotically efficient.

Proof

First note that by a union bound and tail properties of Gaussian random variables,

$$||\mathbb{G}_n(z_t u_t)||_\infty \lesssim_P \sigma_u \sqrt{log\, d}$$

since $u_t \sim N(0, \sigma_u^2)$ and $\mathbb{E}_n[z_{tj}^2] = 1$ for $j = 1, ..., d$. Under the condition $s^2 log^2 d = o(n)$, we have

$$\begin{aligned} ||\hat{D}_t - D_t||_{2,n} &\lesssim_P \sigma_v \sqrt{\frac{slog(d)}{n}}, \\ ||\hat{\pi}_t - \pi_t||_2 &\lesssim_P \sigma_v \sqrt{\frac{slog(d)}{n}}, \\ ||\hat{\pi}_t - \pi_t||_1 &\lesssim_P \sigma_v \sqrt{\frac{s^2 log(d)}{n}}, \end{aligned}$$

where the last inequality comes from $||\hat{\pi}_t - \pi_t||_1 \leq \sqrt{||\hat{\pi}_t - \pi_t||_0}||\hat{\pi}_t - \pi_t||_2 \lesssim_P \sqrt{s}||\hat{\pi}_t - \pi_t||_2$.

By using theorem 1, we get the result.

8 A more general family of estimators

We've already studied the class of sparse linear models and the associated use of l1-regularization, and these are special cases of a more general family of estimators, based

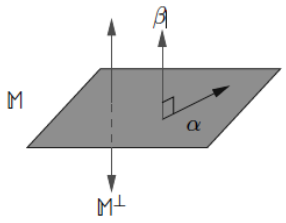

Figure 9.6 In the ideal case, decomposability is defined in terms of a subspace pair $(\mathbb{M}, \mathbb{M}^{\perp})$. For any $\alpha \in \mathbb{M}$ and $\beta \in \mathbb{M}^{\perp}$, the regularizer should decompose as $\Phi(\alpha+\beta) = \Phi(\alpha) + \Phi(\beta)$.

on combining a cost function with a regularizer. Minimizing such a function gives an estimation method known as an M-estimator.

The first ingredient of a general M-estimator is a cost function $L_n : \Omega \times Z^n \to \mathbb{R}$, where the value Ln($\theta$; Z_1^n) provides a measure of the fit of parameter θ to the data Z_1^n, where $Z_1^n = (Z_1, \ldots, Z_n)$ is a collection of n samples. Each sample Z_i takes values in some space **Z**, and is drawn independently according to some distribution **P**. Then, combine the empirical cost function with a regularizer or penalty function $\Phi : \Omega \to \mathbb{R}$, we have the M-estimator:

$$\hat{\theta} = \arg\min_{\theta \in \Omega} L_n(\theta; Z_1^n) + \lambda_n \Phi(\theta),$$

where $\lambda_n > 0$ is a user-defined regularization weight. Later on, for convenience, denote $L_n(\theta; Z_1^n)$ as $L_n(\theta)$. Here, we can see that the lasso estimator mentioned above is a special case of this generalized estimator.

8.1 Bounding the estimation error

Similarly, as what we did in the previous sections, we also need to consider bounding the estimation error $\hat{\theta}$ - θ^*. Because of the time, I will just state some of the useful properties and then prove a theorem about the estimation error. The notion of a decomposable regularizer is defined in terms of a pair of subspaces $M \subset M$ of R^d. The role of the model subspace M is to capture the constraints specified by the model; The orthogonal complement of the space M, namely the set $\overline{M}^{\perp} := \{v \in R^d | < u, v >= 0 \text{ for all } u \in \overline{M}\}$ is referred to as the perturbation subspace, representing deviations away from the model subspace M.

Definition 9.9 Given a pair of subspaces M $\subset \overline{M}$, a norm-based regularizer Φ is decomposable with respect to (M,$M^{\perp}$) if $\Phi(\alpha + \beta) = \Phi(\alpha) + \Phi(\beta)$ for all $\alpha \in M$ and $\beta \in M^{\perp}$.

Define error vector $\hat{\Delta} := \hat{\theta} - \theta^*$, and we will show that decomposability—in conjunction with a suitable choice for the regularization weight λ_n - ensures that the error $\hat{\Delta}$ must lie in a very restricted set. In order to specify a "suitable" choice of regularization parameter λ_n, we need to define the notion of the dual norm associated with our regularizer. Given any norm $\Phi : R^d \to R$, its dual norm is defined in a variational manner as $\Phi^*(v) :=\sup_{\Phi(u)\leq 1} < u, v >$. Our choice of regularization parameter is specified in terms of the random vector $\nabla L_n(\theta^*)$ —the gradient of the empirical cost evaluated at θ^*, also referred to as the score function. Under ideal circumstances, we expect that the score function will not be too large, and we measure its fluctuations in terms of the dual norm, thereby defining the "good event"

$$G(\lambda_n) := \{\Phi^*(\nabla L_n(\theta^*)) \leq \lambda_n/2\}$$

Then, we can have a Proposition:

Proposition 9.13 Let $L_n : \Omega \to \mathbb{R}$ be a convex function, let the regularizer $\Phi : \Omega \to [0, \infty)$ be a norm, and consider a subspace pair $(\mathrm{M}, \overline{M}^{\perp})$ over which Φ is decomposable. Then conditioned on the event $\mathrm{G}(\lambda_n)$, the error $\hat{\Delta} = \hat{\theta} - \theta^*$ belongs to the set $C_{\theta^*}(M, \overline{M}^{\perp}) := \{\Delta \in \Omega | \Phi(\Delta_{\overline{M}^{\perp}}) \leq 3\Phi(\Delta_{\overline{M}}) + 4\Phi(\theta^*_{M^{\perp}})\}$.

When the subspaces $(\mathrm{M}, M^{\perp})$ and parameter θ^* are clear from the context, we adopt the shorthand notation C. We begin by describing the notion of restricted strong convexity, which is defined by the Taylor-series expansion. Given any differentiable cost function, we can use the gradient to form the first-order Taylor approximation, which then defines the first-order Taylor-series error

$$E_n(\Delta) := L_n(\theta^* + \Delta) - L_n(\theta^*) - < \nabla L_n(\theta^*), \Delta > .$$

We use this to define the notion of restricted strong convexity:

Definition 9.15 For a given norm $\|\cdot\|$ and regularizer $\Phi(\cdot)$, the cost function satisfies a restricted strong convexity (RSC) condition with radius $R > 0$, curvature $\kappa > 0$ and tolerance τ_n^2 if $E_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|^2 - \tau_n^2 \Phi^2(\Delta)$ for all $\Delta \in B(R)$.

Note that the set B(R) is the unit ball defined by the given norm $\|\cdot\|$. In our applications of RSC, the norm $\|\cdot\|$ will be derived from an inner product on the space Ω.

Definition 9.18 (Subspace Lipschitz constant) For any subspace S of R^d, the subspace Lipschitz constant with respect to the pair $(\Phi, ||\cdot||)$ is given by

$$\Psi(S) := \sup_{u \in S \setminus \{0\}} \Phi(u)/||u||$$

This quantity is the Lipschitz constant of the regularizer with respect to the error norm, but as restricted to the subspace S. It corresponds to the worst-case price of translating between the Φ- and $||\cdot||$-norms for any vector in S.

Then we are able to have a general result that holds under the restricted strong convexity condition:

Theorem 9.19 (Bounds for general models) Under conditions

(A1) the cost function is convex, and satisfies the local RSC condition (9.38) with curvature κ, radius R and tolerance τ_n^2 with respect to an inner-product induced norm $||\cdot||$, and

(A2) there is a pair of subspaces $M \subset \overline{M}$ such that the regularizer decomposes over $(\mathrm{M}, \overline{M}^{\perp})$.

Consider the regularized M-estimator (9.3) conditioned on the event $\mathrm{G}(\lambda_n)$,

(a) Any optimal solution satisfies the bound

$$\Phi(\hat{\theta} - \theta^*) \leq 4\left\{\Psi(\overline{M})||\hat{\theta} - \theta^*|| + \Phi(\theta^*_{M^{\perp}})\right\}$$

(b) For any subspace pair $(M, M^{\perp})$ such that $\tau_n^2 \Psi^2(\overline{M}) \leq \kappa/64$ and $\varepsilon_n(\overline{M}, M^{\perp}) \leq R$, we have $||\hat{\theta} - \theta^*||^2 \leq \varepsilon_n^2(\overline{M}, M^{\perp})$. where

$$\varepsilon_n^2(\overline{\mathbb{M}}, \mathbb{M}^{\perp}) := \underbrace{9\,\frac{\lambda_n^2}{\kappa^2}\,\Psi^2(\overline{\mathbb{M}})}_{\text{estimation error}} + \underbrace{\frac{8}{\kappa}\left\{\lambda_n \Phi(\theta^*_{\mathbb{M}^{\perp}}) + 16\tau_n^2 \Phi^2(\theta^*_{\mathbb{M}^{\perp}})\right\}}_{\text{approximation error}},$$