- Recall that for the statistical error bound for the LASSO estimates with exactly sparse model, we need the following condition:

$$(*) \qquad \frac{\|X\Delta\|_2^2}{2n} \geq \kappa \|\Delta\|_2^2 \quad \forall \Delta \in \mathcal{C} = \{\Delta \in \mathbb{R}^d : \|\Delta_{S^c}\|_1 \leq 3\|\Delta\|_1\}$$

where $\kappa > 0$.

- In general, one would not expect $(*)$ to hold for arbitrary X; e.g., fix $\Delta \in \mathcal{C}$ and construct X s.t. $\Delta \in \text{null}(X)$. Thus, some assumptions on X are needed.
- Let us assume that the rows of X are iid $\mathcal{N}(0, I)$. Our goal is to prove

<u>Theorem</u>: With probability $\geq 1 - c_1 \exp(-c_2 n)$ for some $c_1, c_2 > 0$,

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \frac{1}{4}\|\Delta\|_2 - 9\sqrt{\frac{\log d}{n}}\|\Delta\|_1 \qquad \forall \Delta \in \mathbb{R}^d.$$

Note that the theorem trivially holds for $\Delta = 0$. Thus, without loss of generality, we may assume $\|\Delta\|_2 = 1$.

- The proof of the theorem consists of 3 main steps:

<u>Step 1</u>: For a given $r > 0$, define

$$V(r) = \{\Delta \in \mathbb{R}^d : \|\Delta\|_2 = 1, \|\Delta\|_1 \leq r\}.$$

<u>Proposition 1</u>:

$$\mathbb{E}\left[\inf_{\Delta \in V(r)} \frac{\|X\Delta\|_2}{\sqrt{n}}\right] \geq 3\left[\frac{1}{4} - \sqrt{\frac{\log d}{n}}\, r\right].$$

whenever $V(r) \neq \phi$.

Step 2: We show that the random variable $Q(r, X)$ is concentrated around its mean, where

$$Q(r, X) = \inf_{\Delta \in V(r)} \frac{\|X\Delta\|_2}{\sqrt{n}}.$$

Proposition 2: Let $r > 0$ be s.t. $V(r) \neq \phi$. Then,

$$\Pr\left[\left| Q(r,X) - \mathbb{E}[Q(r,X)] \right| \geq \tfrac{1}{2} t(r) \right]$$

$$\leq 2\exp(-nt^2(r)/8), \text{ where}$$

$$t(r) = \frac{1}{4} + 3\sqrt{\frac{\log d}{n}}\, r.$$

Step 3: The results of Propositions 1 and 2 show that w/ probability at least $1 - 2\exp(-n\, t^2(r)/8)$,

$$Q(r, X) = \inf_{\Delta \in V(r)} \frac{\|X\Delta\|_2}{\sqrt{n}} \geq \mathbb{E}[Q(r,X)] - \frac{1}{2} t(r)$$

$$\geq 1 - \frac{3}{2} t(r) = \frac{5}{8} - \frac{9}{2}\sqrt{\frac{\log d}{n}}\, r.$$

Here, we need $\|\Delta\|_1 \leq r$, where r is fixed. However, we need the above to hold for all r. This is the goal of this step.

- Probability tools used
  - (i) Comparison inequality for Gaussian process
  - (ii) Concentration of measure for Lipschitz functions of Gaussians
  - (iii) "Peeling argument" from empirical process theory

Proof of Proposition 1

- The quantity of interest here is

$$\tilde{Q}(r,X) = \inf_{\Delta \in V(r)} \|X\Delta\|_2 = \inf_{\Delta \in V(r)} \sup_{u \in S^{n-1}} u^T X \Delta$$

where $S^{n-1} = \{ u \in \mathbb{R}^n : \|u\|_2 = 1 \}$.

- Note that for each $(u, \Delta) \in S^{n-1} \times V(r)$,

$$Y_{u,\Delta} = u^T X \Delta \quad \text{is a zero-mean Gaussian RV}$$

- To get a lower bound on $\mathbb{E}[\tilde{Q}(r,X)]$, or equivalently, an upper bound on $\mathbb{E}[-\tilde{Q}(r,X)] = \mathbb{E}\left[\sup_{\Delta \in V(r)} \inf_{u \in S^{n-1}} u^T X \Delta\right]$, a powerful idea is to construct another family of Gaussian RVs $\{Z_{u,\Delta}\}$ s.t. $\mathbb{E}\left[\sup_{\Delta \in V(r)} \inf_{u \in S^{n-1}} Z_{u,\Delta}\right]$ is easy to compute and is related to $\mathbb{E}\left[\sup_{\Delta \in V(r)} \inf_{u \in S^{n-1}} Y_{u,\Delta}\right]$

This is the content of Gordon's inequality:

<u>Fact</u> (<u>Gordon's Inequality</u>)

Let U, V be arbitrary index sets. Consider two families $\{Y_{u,v}\}$ and $\{Z_{u,v}\}$ of zero-mean Gaussian RVs. Suppose that

$$\sigma(Y_{u,v} - Y_{u',v'}) \le \sigma(Z_{u,v} - Z_{u',v'}) \quad \forall (u,v), (u',v') \in U \times V,$$

and

$$\sigma(Y_{u,v} - Y_{u,v'}) = \sigma(Z_{u,v} - Z_{u,v'}) \quad \forall\, u \in U,\ v, v' \in V,$$

where $\sigma(\cdot)$ denotes the standard deviation of the argument. Then,

$$\mathbb{E}\left[\sup_{u \in U} \inf_{v \in V} Y_{u,v}\right] \le \mathbb{E}\left[\sup_{u \in U} \inf_{v \in V} Z_{u,v}\right].$$

- To apply Gordon's inequality, let us first compute $\sigma(Y_{u,\Delta} - Y_{u',\Delta'})$, where $Y_{u,\Delta} = u^T X \Delta$, $(u,\Delta) \in S^{n-1} \times V(r)$, and see what properties are needed for $\{Z_{u,\Delta}\}$.

By definition,

$$\sigma^2(Y_{u,\Delta} - Y_{u',\Delta'}) = \mathbb{E}\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{d} X_{ij}(u_i\Delta_j - u_i'\Delta_j')\right)^2\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{d}(u_i\Delta_j - u_i'\Delta_j')^2 \quad [\because X_{ij} \sim \mathcal{N}(0,1)]$$

$$= \| u\Delta^T - u'\Delta'^T \|_F^2 .$$

Observe that

$$\| u\Delta^T - u'\Delta'^T\|_F^2 = \| (u-u')\Delta^T + u'(\Delta-\Delta')^T\|_F^2$$

$$= \|\Delta\|_2^2 \|u-u'\|_2^2 + \|u'\|_2^2 \|\Delta-\Delta'\|_2^2$$

$$+ 2\,(u^Tu' - \|u'\|_2^2)(\|\Delta\|_2^2 - \Delta^T\Delta')$$

$$= \|u-u'\|_2^2 + \|\Delta-\Delta'\|_2^2 - 2(\|u'\|_2^2 - u^Tu')(\|\Delta\|_2^2 - \Delta^T\Delta')$$

(since $\|\Delta\|_2^2 = \|u'\|_2^2 = 1$)

$$\leq \|u-u'\|_2^2 + \|\Delta-\Delta'\|_2^2$$

(by Cauchy-Schwarz, w/ equality holding if $u=u'$ or $\Delta=\Delta'$).

This suggests that we should define

$$Z_{u,\Delta} = g_1^T u + g_2^T \Delta$$

where $g_1 \sim \mathcal{N}(0, I_n)$, $g_2 \sim \mathcal{N}(0, I_d)$ are independent.

Indeed, we have $Z_{u,\Delta} \sim \mathcal{N}(0, \|u\|_2^2 + \|\Delta\|_2^2)$ and

$$\sigma^2(Z_{u,\Delta} - Z_{u',\Delta'}) = \|u-u'\|_2^2 + \|\Delta-\Delta'\|_2^2 .$$

By Gordon's inequality, with $U = V(r)$ and $V = S^{n-1}$, we have

$$\mathbb{E}\left[\sup_{\Delta\in V(r)} \inf_{u\in S^{n-1}} u^T X \Delta\right] \leq \mathbb{E}\left[\sup_{\Delta\in V(r)} \inf_{u\in S^{n-1}} (g_1^T u + g_2^T \Delta)\right]$$

$$= \mathbb{E}\left[\inf_{u\in S^{n-1}} g_1^T u\right] + \mathbb{E}\left[\sup_{\Delta\in V(r)} g_2^T \Delta\right]$$

$$\leq -\mathbb{E}\left[\|g_1\|_2\right] + \mathbb{E}\left[\sup_{\Delta\in V(r)} \|g_2\|_\infty \cdot \|\Delta\|_1\right] \qquad \text{— } (**)$$

The problem now reduces to computing $\mathbb{E}[\|g_1\|_2]$ and $\mathbb{E}[\|g_2\|_\infty]$.

<u>Claim 1:</u> $\mathbb{E}[\|g_2\|_\infty] \leq 3\sqrt{\log d}$.

<u>Proof:</u> Since $\|g_2\|_\infty$ is a non-negative RV, we have

$$\mathbb{E}[\|g_2\|_\infty] = \int_0^\infty \Pr[\|g_2\|_\infty > t]\, dt$$

$$\leq \delta + \int_\delta^\infty \left[1 - \Pr[\|g_2\|_\infty \leq t]\right] dt \quad \text{for any } \delta > 0$$

Note that

$$\Pr[\|g_2\|_\infty \leq t] = (\Pr[|g| \leq t])^d \quad \text{where } g \sim N(0,1), \ (\text{since } g_2 \sim N(0, I_d))$$

$$= [1 - \Pr[|g| > t]]^d \geq 1 - d\Pr[|g| > t].$$

Hence,

$$\mathbb{E}[\|g_2\|_\infty] \leq \delta + d \cdot \int_\delta^\infty \Pr[|g| > t]\, dt.$$

Using the estimate

$$\Pr[|g| > t] \leq \sqrt{\frac{2}{\pi}} \cdot \frac{1}{t} \cdot \exp(-t^2/2) \quad \forall t > 0 \quad (\text{Exercise}),$$

we get

$$\mathbb{E}[\|g_2\|_\infty] \leq \delta + d \cdot \sqrt{\frac{2}{\pi}} \int_\delta^\infty \frac{1}{t} \exp(-t^2/2)\, dt \leq \delta + \frac{d}{\delta}\sqrt{\frac{2}{\pi}} \int_\delta^\infty \exp(-t^2/2)\, dt$$

$$\leq \delta + \frac{d}{\delta^2}\sqrt{\frac{2}{\pi}}\exp(-\delta^2/2).$$

Taking $\delta = \sqrt{2\log d}$, we get

$$\mathbb{E}[\|g_2\|_\infty] \leq \sqrt{2\log d} + \sqrt{\frac{2}{\pi}}\cdot\frac{1}{2\log d} \leq 3\sqrt{\log d}.$$

<u>Claim 2</u> (<u>Exercise</u>) $\mathbb{E}[\|g_1\|_2] \geq \frac{3}{4}\sqrt{n}$ for, say, $n \geq 10$.

Using (**) and Claim 1 and Claim 2, since $\|\Delta\|_1 \leq r$ $\forall \Delta \in V(r)$,

$$\mathbb{E}[-\tilde{Q}(r,X)] \leq 3r\sqrt{\log d} - \frac{3}{4}\sqrt{n},$$

which implies that

$$\inf_{\Delta \in V(r)} \frac{\|X\Delta\|_2}{\sqrt{n}} \geq 3\left[\frac{1}{4} - \sqrt{\frac{\log d}{n}}\cdot r\right].$$

This finishes the proof of Proposition 1.

Proof of Proposition 2

We will make use of the following classic fact:

Fact (Concentration of Measure for Lipschitz Functions of Gaussian RVs)

Let $g \sim \mathcal{N}(0, I_m)$ and $F : \mathbb{R}^m \to \mathbb{R}$ be a Lipschitz function w/ constant $L > 0$; i.e.,

$$|F(x) - F(y)| \leq L \cdot \|x - y\|_2 .$$

Then, for all $t \geq 0$,

$$\Pr\Big[\, |F(g) - \mathbb{E}[F(g)]| \geq t \Big] \leq 2 \exp\left(-t^2/2L^2\right)$$

Our goal now is to show that $Q(r, X)$ is Lipschitz in X and determine the Lipschitz constant. Towards that end, we compute

$$\sqrt{n}\left[Q(r, X) - Q(r, Y) \right] = \inf_{\Delta \in V(r)} \|X\Delta\|_2 - \inf_{\Delta \in V(r)} \|Y\Delta\|_2$$

$\leq \inf_{\Delta \in V(r)} \|X\Delta\|_2 - \|Y\hat{\Delta}\|_2$ where $\hat{\Delta} = \operatorname{argmin}_{\Delta \in V(r)} \|Y\Delta\|_2$, which exists because $V(r)$ is compact and $\Delta \mapsto \|Y\Delta\|_2$ is continuous

$\leq \|X\hat{\Delta}\|_2 - \|Y\hat{\Delta}\|_2$

$\leq \sup_{\Delta \in V(r)} \|(X - Y)\Delta\|_2$

$\leq \|X - Y\| \cdot \sup_{\Delta \in V(r)} \|\Delta\|_2$ where $\|\cdot\|$ denotes the spectral norm; i.e., $\|A\| = \sup_{\|u\|_2 = 1} \|Au\|_2$

$\leq \|X - Y\|_F$ Since $\|A\| \leq \|A\|_F$ and $\|\Delta\|_2 = 1 \ \ \forall \Delta \in V(r)$

It follows that

$$\left| Q(r,X) - Q(r,Y) \right| \le \frac{1}{\sqrt{n}} \| X - Y \|_F \, ;$$

i.e., $X \mapsto Q(r,X)$ is $\frac{1}{\sqrt{n}}$ - Lipschitz.

Consequently, by the Fact, we have

$$\Pr\left[\left| Q(r,X) - \mathbb{E}[Q(r,X)] \right| \geq \frac{1}{2} t(r) \right]$$

$$\le 2 \exp\left(- n t^2(r)/8 \right), \quad \text{where } t(r) = \frac{1}{4} + 3\sqrt{\frac{\log d}{n}}\, r.$$

- Finally, we execute Step 3. Recall that for a fixed $r>0$, we have

$$Q(r,X) \geq 1 - \frac{3}{2} t(r) = \frac{5}{8} - \frac{9}{2}\sqrt{\frac{\log d}{n}}\, r$$

with high probability. To get a bound that is uniform in r, we use a device called the peeling argument.

<u>Proposition 3</u>: Let $A \subseteq \mathbb{R}^d$ be non-empty, $f: \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ and $h: \mathbb{R}^d \to \mathbb{R}_+$ be given, and $g: \mathbb{R} \to \mathbb{R}$ be non-negative and strictly increasing. Consider the optimization problem $\sup_{h(v) \le r,\, v \in A} f(v, X)$, where X is a random vector.

Suppose $g(r) \geq \mu$ for all $r \geq 0$ and $\exists\, c>0$ s.t. for all $r>0$,

$$\Pr\left[\sup_{h(v) \le r,\, v \in A} f(v,X) \geq g(r) \right] \le 2 \exp\left(-c \cdot a \cdot g^2(r) \right)$$

for some $a > 0$.

Then,

$$\Pr[\mathcal{E}] \leq \frac{2\exp(-4c\cdot a\cdot\mu^2)}{1-\exp(-4c\cdot a\cdot\mu^2)}$$

where

$$\mathcal{E} = \{\exists v \in A \text{ s.t. } f(v,X) \geq 2g(h(v))\}.$$

- Let us first see how to apply Proposition 3. Take

$$A = S^{d-1} = \{\Delta \in \mathbb{R}^d : \|\Delta\|_2 = 1\}.$$

$$h(\Delta) = \|\Delta\|_1, \qquad f(\Delta, X) = 1 - \frac{\|X\Delta\|_2}{\sqrt{n}},$$

$$g(r) = \frac{3}{2}t(r).$$

Then, we have $g(r) \geq 3/8 = \mu$. Moreover,

$$\Pr\left[\sup_{h(\Delta)\leq r,\, \Delta\in A} f(\Delta,X) \geq g(r)\right]$$

$$= \Pr\left[\sup_{\Delta\in V(r)}\left(1 - \frac{\|X\Delta\|_2}{\sqrt{n}}\right) \geq \frac{3}{2}t(r)\right]$$

$$= \Pr\left[Q(r,X) \leq 1 - \frac{3}{2}t(r)\right]$$

$$\leq 2\exp(-nt^2(r)/8) = 2\exp(-n g^2(r)/18)$$

It follows that $\exists c_1, c_2 > 0$ such that

$\Pr[\mathcal{E}] \leq c_1 \exp(-c_2 n)$, where

$$\mathcal{E} = \left\{\exists \Delta \in \mathbb{R}^d : \|\Delta\|_2 = 1,\ \frac{\|X\Delta\|_2}{\sqrt{n}} \leq 1 - 3t(\|\Delta\|_1) = \frac{1}{4} - 9\sqrt{\frac{\log d}{n}}\cdot\|\Delta\|_1\right\},$$

which implies w/ probability at least $1 - c_1 \exp(-c_2 n)$, event $\mathcal{E}^c$ occurs; i.e.,

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq \frac{1}{4} - 9\sqrt{\frac{\log d}{n}}\|\Delta\|_1 \qquad \forall \Delta \in \mathbb{R}^d \text{ with } \|\Delta\|_2 = 1.$$

This establishes the theorem.

- It now remains to prove Proposition 3.

Proof: Since $g(r) \geq \mu$ for all $r \geq 0$, define

$$A_m = \{ v \in A : 2^{m-1}\mu \leq g(h(v)) \leq 2^m \mu \}$$

for $m = 1, 2, \dots$. For a given X, if $v \in A$ is such that $f(v,X) \geq 2g(h(v))$, then $v \in A_m$ for some m.

$\Rightarrow$ by the Union bound,

$$\Pr[\mathcal{E}] \leq \sum_{m \geq 1} \Pr\left[\exists v \in A_m \text{ s.t. } f(v,X) \geq 2g(h(v)) \right]$$

By definition, if $v \in A_m$ and $f(v,X) \geq 2g(h(v))$, then

$$f(v,X) \geq 2(2^{m-1}\mu) = 2^m \mu.$$

Also, $g(h(v)) \leq 2^m\mu$. Since g is strictly increasing, this implies $h(v) \leq g^{-1}(2^m\mu)$. Hence,

$$\Pr[\mathcal{E}] \leq \sum_{m \geq 1} \Pr\left[\sup_{\substack{h(v) \leq g^{-1}(2^m\mu) \\ v \in A}} f(v,X) \geq 2^m \mu \right]$$

$$\leq 2 \sum_{m \geq 1} \exp\left[-c \cdot a \cdot (g(g^{-1}(2^m\mu)))^2 \right]$$

$$= 2 \sum_{m \geq 1} \exp\left[-c \cdot a \cdot 2^{2m} \cdot \mu^2\right]$$

$$\leq 2 \sum_{m \geq 1} \exp\left(-4c \cdot a \cdot \mu^2 \cdot m\right)$$

$$= \frac{2 \exp(-4 \cdot c \cdot a \cdot \mu^2)}{1 - \exp(-4 \cdot c \cdot a \cdot \mu^2)} \qquad \text{(geometric sum)}$$

Discussion

① We have only dealt with the case where each row of X is iid $\mathcal{N}(0, I)$. The analysis can be extended to the case where each row of X is iid $\mathcal{N}(0, \Sigma)$ for some $\Sigma \succ 0$.

② Generalizations to cases where entries of X are not necessarily Gaussian are possible.