# SUPPLEMENT TO "ASYMPTOTIC NORMALITY AND OPTIMALITIES IN ESTIMATION OF LARGE GAUSSIAN GRAPHICAL MODEL"

By Zhao Ren [*,‡], Tingni Sun[§], Cun-Hui Zhang[†,¶] and Harrison H. Zhou[*,‡]

*Yale University*[‡], *University of Pennsylvania*[§] *and Rutgers University*[¶]

In this supplement we collect proofs for proving auxiliary lemmas.

**1. Proof of Lemma 5.** The proof of this Lemma is similar to that of Lemma 2 in Section 8.1. But for the latent variables case in both algebraic analysis and probabilistic analysis we need to replace $\beta_i$, $\theta_{ij}^{ora}$, $\sigma^{ora}$, $\epsilon_i$ and $\nu$ by $\beta_i^S$, $\theta_{ij}^{ora,S}$, $\sigma^{ora,S}$, $\epsilon_i^S$ and $\nu^S$ respectively, and subsequently define

$$I_1 = \left\{\nu^S \leq \sigma^{ora,S}\lambda\frac{\xi-1}{\xi+1}(1-\tau)\right\}, \tag{102}$$

$$I_3 = \left\{\sigma^{ora,S} \in \left[\sqrt{1/(2M)}, \sqrt{2M}\right]\right\}, \tag{103}$$

where $\sigma^{ora,S} = \left(\theta_{ii}^{ora,S}\right)^{1/2} = \frac{\|\epsilon_i^S\|}{\sqrt{n}}$ and $\nu^S = \left\|\mathbf{Y}^T\epsilon_i^S/n\right\|_\infty$, while the definitions of $I_2$ and $I_4$ are the same as before in Equations (87) and (89). As in Section 8.1 we define $E = \cap_{i=1}^4 I_i$ and need to show that $\mathbb{P}\{E\} \geq 1-(1+o(1))p^{-\delta+1}$. We will need only to show that

$$\mathbb{P}\{I_1^c\} \leq o\left(p^{-\delta+1}\right), \text{ and } \mathbb{P}\{I_3^c\} \leq o(p^{-\delta}).$$

The arguments for $\mathbb{P}\{I_2^c\}$ and $\mathbb{P}\{I_4^c\}$ are identical to those of the non-latent variables case in Section 8.1.2.

It is relatively easy to establish the probabilistic bound for $I_3$. It is a consequence of the following two bounds,

$$\mathbb{P}\left\{(\sigma^{ora})^2 \notin [3/(4M), 5M/4]\right\} \leq \mathbb{P}\left\{\left|\frac{(\sigma^{ora})^2}{\theta_{ii}} - 1\right| \geq \frac{1}{4}\right\} = \mathbb{P}\left\{\left|\frac{\chi^2_{(n)}}{n} - 1\right| \geq \frac{1}{4}\right\} = o(p^{-\delta}), \tag{104}$$

which follows from Equation (93), and

$$\mathbb{P}\left\{\left|(\sigma^{ora})^2 - \left(\sigma^{ora,S}\right)^2\right| > 1/(4M)\right\} = o(p^{-\delta}) \tag{105}$$

[*]The research of Zhao Ren and Harrison H. Zhou was supported in part by NSF Career Award DMS-0645676 and NSF FRG Grant DMS-0854975.

[†]The research of Cun-Hui Zhang was supported in part by the NSF Grants DMS 0906420, DMS-11-06753 and NSA Grant H98230-11-1-0205.

which follows two bounds for $D_1$ (70) and $D_2$ (71).

Now we establish the probabilistic bound for $I_1$, i.e., $\mathbb{P}\{I_1^c\} \le o\left(p^{-\delta+1}\right)$. Write the event $I_1^c$ as follows,

$$\begin{aligned} & \nu^S - \nu + \nu \\ > & \left[\sigma^{ora,S}\lambda\frac{\xi-1}{\xi+1}(1-\tau) - \sigma^{ora}\sqrt{\frac{2\delta\log p}{n}}\right] + \sigma^{ora}\sqrt{\frac{2\delta\log p}{n}} \\ = & \left[\sigma^{ora,S}\left(\lambda\frac{\xi-1}{\xi+1}(1-\tau) - \sqrt{\frac{2\delta\log p}{n}}\right) + \left(\sigma^{ora,S} - \sigma^{ora}\right)\sqrt{\frac{2\delta\log p}{n}}\right] + \sigma^{ora}\sqrt{\frac{2\delta\log p}{n}}. \end{aligned}$$

Set $\xi = 6/\varepsilon + 5$ such that

$$\lambda\frac{\xi-1}{\xi+1}(1-\tau) > (1+\varepsilon/2)\sqrt{\frac{2\delta\log p}{n}}$$

for $\lambda = (1+\varepsilon)\sqrt{\frac{2\delta\log p}{n}}$ on $I_2 \cap I_3$. Then probabilistic bound for $I_1$ is a consequence the following two bounds,

$$\mathbb{P}\left\{\nu > \sigma^{ora}\sqrt{\frac{2\delta\log p}{n}}\right\} \le O\left(p^{-\delta+1}/\sqrt{\log p}\right), \tag{106}$$

which follows from Equation (100) or Proposition 5, and

$$\mathbb{P}\left\{\nu^S - \nu > \frac{\varepsilon}{2}\sigma^{ora,S}\sqrt{\frac{2\delta\log p}{n}} + \left(\sigma^{ora,S} - \sigma^{ora}\right)\sqrt{\frac{2\delta\log p}{n}}\right\} = o\left(p^{-\delta}\right), \tag{107}$$

which is established as follows. From Equations (103), $D_1$ (70) and $D_2$ (71), we have

$$\frac{\varepsilon}{2}\sigma^{ora,S} - \left|\left(\sigma^{ora,S}\right) - \left(\sigma^{ora}\right)\right| \ge \frac{\varepsilon}{2}\sqrt{\frac{1}{2M}} - O\left(\frac{1}{p} + \sqrt{\frac{\log p}{np}}\right) \ge \frac{\varepsilon}{4}\sqrt{\frac{1}{2M}}$$

with probability $1 - o\left(p^{-\delta}\right)$. It is then enough to show

$$\mathbb{P}\left\{\nu^S - \nu > \frac{\varepsilon}{4}\sqrt{\frac{1}{2M}}\sqrt{\frac{2\delta\log p}{n}}\right\} = o\left(p^{-\delta}\right) \tag{108}$$

to establish Equation (107). On the event $I_4$ we have

$$\begin{aligned} \left|\nu^S - \nu\right| & \le \max_{k\in A^c}\left|\left|\frac{\mathbf{Y}_k^T\epsilon_i^S}{n}\right| - \left|\frac{\mathbf{Y}_k^T\epsilon_i}{n}\right|\right| \le \max_{k\in A^c}\left|\frac{\mathbf{Y}_k^T\mathbf{X}_{A^c}\beta_i^L}{n}\right| \\ & = \max_{k\in A^c}\left|\frac{\sqrt{n}}{\|\mathbf{X}_k^T\|}\frac{\mathbf{X}_k^T\mathbf{X}_{A^c}\beta_i^L}{n}\right| \le \sqrt{2M}\max_{k\in A^c}\left|\frac{1}{n}\sum_{g=1}^n X_k^{(g)}\left(X_{A^c}^{(g)}\right)^T\beta_i^L\right|, \end{aligned} \tag{109}$$

where $X^{(g)}$ denotes the $g$th sample. Note that for each $k \in A^c$ the term $\frac{1}{n}\sum_{g=1}^{n} X_k^{(g)}\left(X_{A^c}^{(g)}\right)^T \beta_i^L$ in (109) is an average of $n$ i.i.d. random variables with

$$\left|\mathbb{E}X_k^{(g)}\left(X_{A^c}^{(g)}\right)^T \beta_i^L\right| \leq \left\|\Sigma_{XX}\beta_i^L\right\|_2 \leq \|\Sigma_{XX}\|_2 \left\|\beta_i^L\right\|_2 \leq \frac{CM}{\sqrt{p}},$$

$$Var\left(X_k^{(g)}\left(X_{A^c}^{(g)}\right)^T \beta_i^L\right) \leq \frac{C}{p}.$$

From the classical large deviations bound in Theorem 2.8 of Petrov (1995), there exist some uniform constant $c_1, c_2 > 0$ such that

$$\mathbb{P}\left\{\left|\frac{\sqrt{p}}{n}\sum_{g=1}^{n}\left[X_k^{(g)}\left(X_{A^c}^{(g)}\right)^T \beta_i^L - \mathbb{E}\left(X_k^{(g)}\left(X_{A^c}^{(g)}\right)^T \beta_i^L\right)\right]\right| > t\right\} \leq 2\exp\left(-nt^2/c_2\right) \quad \text{for } 0 < t < c_1.$$

then by setting $t = \sqrt{\frac{2\delta c_2 \log p}{n}}$, with probability $1 - o\left(p^{-\delta}\right)$, we have

$$\left|\nu^S - \nu\right| \leq \sqrt{2M}\left[\frac{CM}{\sqrt{p}} + \frac{1}{\sqrt{p}}\sqrt{\frac{2\delta c_2 \log p}{n}}\right] = o\left(\sqrt{\frac{2\delta \log p}{n}}\right).$$

where the last inequality follows from Equation (45). Therefore we have shown Equation (108), i.e., with probability $1 - o\left(p^{-\delta}\right)$,

$$\nu^S - \nu = o\left(\sqrt{\frac{2\delta \log p}{n}}\right) < \frac{\varepsilon}{4}\sqrt{\frac{1}{2M}}\sqrt{\frac{2\delta \log p}{n}},$$

which together with Equation (106), imply the probabilistic bound for $I_1$.

**2. Proof of Proposition 1.** From the KKT conditions (80) we have

$$\left\|\frac{\mathbf{Y}^T\mathbf{Y}}{n}\left(\hat{d}(\mu) - d^{true}\right)\right\|_\infty = \left\|\frac{\mathbf{Y}^T\left(\mathbf{X}_i - \mathbf{Y}\hat{d}(\mu)\right)}{n} - \frac{\mathbf{Y}^T\left(\mathbf{X}_i - \mathbf{Y}d^{true}\right)}{n}\right\|_\infty \leq \mu + \nu, \tag{110}$$

where $\nu$ is defined in Equation (83). We also have

$$\begin{aligned}
& \frac{1}{n}\left\|\mathbf{Y}\left(d^{true} - \hat{d}(\mu)\right)\right\|^2 \\
= \; & \frac{\left(d^{true} - \hat{d}(\mu)\right)^T\left(\mathbf{Y}^T\left(\mathbf{X}_i - \mathbf{Y}\hat{d}(\mu)\right) - \mathbf{Y}^T\left(\mathbf{X}_i - \mathbf{Y}d^{true}\right)\right)}{n} \\
\leq \; & \mu\left(\left\|d^{true}\right\|_1 - \left\|\hat{d}(\mu)\right\|_1\right) + \nu\left\|d^{true} - \hat{d}(\mu)\right\|_1 \qquad (111) \\
\leq \; & (\mu + \nu)\left\|\left(d^{true} - \hat{d}(\mu)\right)_T\right\|_1 + 2\mu\left\|\left(d^{true}\right)_{T^c}\right\|_1 - (\mu - \nu)\left\|\left(d^{true} - \hat{d}(\mu)\right)_{T^c}\right\|_1,
\end{aligned}$$

where the first inequality follows from the KKT conditions (80). Then on the event $\left\{\nu \leq \mu\frac{\xi-1}{\xi+1}\right\}$ we have

(112)
$$\frac{1}{n}\left\|\mathbf{Y}\left(d^{true}-\hat{d}\left(\mu\right)\right)\right\|^{2} \leq 2\mu\left(\frac{\xi\left\|\left(d^{true}-\hat{d}\left(\mu\right)\right)_{T}\right\|_{1}}{\xi+1}+\left\|\left(d^{true}\right)_{T^{c}}\right\|_{1}-\frac{\left\|\left(d^{true}-\hat{d}\left(\mu\right)\right)_{T^{c}}\right\|_{1}}{\xi+1}\right).$$

If we could show that $\hat{d}\left(\mu\right)-d^{true}\in\mathcal{C}\left(\frac{\xi+\zeta}{1-\zeta},T\right)$, then by the definition (81) and inequality (110) we would obtain

$$\left\|\hat{d}\left(\mu\right)-d^{true}\right\|_{1} \leq \frac{\left(\nu+\mu\right)\left|T\right|}{CIF_{1}\left(\frac{\xi+\zeta}{1-\zeta},T,\mathbf{Y}\right)}. \tag{113}$$

Suppose that

$$\left\|\hat{d}\left(\mu\right)-d^{true}\right\|_{1} \geq \frac{1+\xi}{\zeta}\left\|\left(d^{true}\right)_{T^{c}}\right\|_{1}, \tag{114}$$

then the inequality (112) becomes

$$\frac{\left\|\mathbf{Y}\left(d^{true}-\hat{d}\left(\mu\right)\right)\right\|^{2}}{n} \leq \frac{2\mu}{\xi+1}\left(\left(\xi+\zeta\right)\left\|\left(d^{true}-\hat{d}\left(\mu\right)\right)_{T}\right\|_{1}-\left(1-\zeta\right)\left\|\left(d^{true}-\hat{d}\left(\mu\right)\right)_{T^{c}}\right\|_{1}\right).$$

Thus we have

$$\hat{d}\left(\mu\right)-d^{true}\in\mathcal{C}\left(\frac{\xi+\zeta}{1-\zeta},T\right).$$

Combining this fact under the condition (114) with (113), we obtain the first desired inequality (90)

$$\left\|\hat{d}\left(\mu\right)-d^{true}\right\|_{1} \leq \max\left\{\frac{1+\xi}{\zeta}\left\|\left(d^{true}\right)_{T^{c}}\right\|_{1},\frac{\left(\nu+\mu\right)\left|T\right|}{CIF_{1}\left(\frac{\xi+\zeta}{1-\zeta},T,\mathbf{Y}\right)}\right\}.$$

We complete our proof by letting $\zeta=1/2$ and noting that (111) implies the second desired inequality (91).

**3. Proof of Proposition 2.** For $\tau$ defined in Equation (84), we need to show that $\hat{\sigma}\geq\sigma^{ora}\left(1-\tau\right)$ and $\hat{\sigma}\leq\sigma^{ora}\left(1+\tau\right)$ on the event $\left\{\nu\leq\sigma^{ora}\lambda\frac{\xi-1}{\xi+1}\left(1-\tau\right)\right\}$. Let $\hat{d}\left(\sigma\lambda\right)$ be the solution of (79) as a function of $\sigma$, then

$$\frac{\partial}{\partial\sigma}L_{\lambda}\left(\hat{d}\left(\sigma\lambda\right),\sigma\right)=\frac{1}{2}-\frac{\left\|\mathbf{X}_{i}-\mathbf{Y}\hat{d}\left(\sigma\lambda\right)\right\|^{2}}{2n\sigma^{2}} \tag{115}$$

since $\left\{\frac{\partial}{\partial d}L_{\lambda}\left(d,\sigma\right)|_{d=\hat{d}(\sigma\lambda)}\right\}_{k}=0$ for all $\hat{d}_{k}\left(\sigma\lambda\right)\neq0$, and $\left\{\frac{\partial}{\partial\sigma}\hat{d}\left(\sigma\lambda\right)\right\}_{k}=0$ for all $\hat{d}_{k}\left(\sigma\lambda\right)=0$ which follows from the fact that $\left\{k:\hat{d}_{k}\left(\sigma\lambda\right)=0\right\}$ is unchanged in a neighborhood of $\sigma$ for almost all $\sigma$. Equation (115) plays a key in the proof.

**(1)**. To show that $\hat{\sigma} \geq \sigma^{ora}(1-\tau)$ it's enough to show

$$\frac{\partial}{\partial \sigma} L_\lambda\left(\hat{d}(\sigma\lambda), \sigma\right)|_{\sigma=t_1} \leq 0.$$

where $t_1 = \sigma^{ora}(1-\tau)$, due to the strict convexity of the objective function $L_\lambda(d,\sigma)$ in $\sigma$. Equation (115) implies that

$$\begin{aligned} 2t_1^2 \frac{\partial}{\partial \sigma} L_\lambda\left(\hat{d}(\sigma\lambda), \sigma\right)|_{\sigma=t_1} &= t_1^2 - \frac{\left\|\mathbf{X}_m - \mathbf{Y}\hat{d}(t_1\lambda)\right\|^2}{n} \\ &\leq t_1^2 - \frac{\left\|\mathbf{X}_m - \mathbf{Y}d^{true} + \mathbf{Y}\left(\hat{d}(t_1\lambda) - d^{true}\right)\right\|^2}{n} \\ &\leq t_1^2 - (\sigma^{ora})^2 + 2\left(d^{true} - \hat{d}(t_1\lambda)\right)^T \frac{\mathbf{Y}^T(\mathbf{X}_m - \mathbf{Y}d^{true})}{n} \\ &\leq 2t_1(t_1 - \sigma^{ora}) + 2\nu\left\|d^{true} - \hat{d}(t_1\lambda)\right\|_1. \end{aligned} \tag{116}$$

From Equation (90) in Proposition 1, on the event $\left\{\nu \leq t_1\lambda\frac{\xi-1}{\xi+1}\right\} = \left\{\nu/\sigma^{ora} < \lambda\frac{\xi-1}{\xi+1}(1-\tau)\right\}$ we have

$$\left\|\hat{d}(t_1\lambda) - d^{true}\right\|_1 \leq \max\left\{2(1+\xi)\left\|\left(d^{true}\right)_{T^c}\right\|_1, \frac{(\nu + t_1\lambda)|T|}{CIF_1(2\xi+1, T, \mathbf{Y})}\right\}.$$

then

$$\begin{aligned} & 2t_1^2 \frac{\partial}{\partial \sigma} L_\lambda\left(\hat{d}(\sigma\lambda), \sigma\right)|_{\sigma=t_1} \\ \leq\ & 2t_1(t_1 - \sigma^{ora}) + 2t_1\lambda \max\left\{2(1+\xi)\left\|\left(d^{true}\right)_{T^c}\right\|_1, \frac{(\nu + t_1\lambda)|T|}{CIF_1(2\xi+1, T, \mathbf{Y})}\right\} \\ \leq\ & 2t_1\left[-\tau\sigma^{ora} + \lambda \max\left\{2(1+\xi)\left\|\left(d^{true}\right)_{T^c}\right\|_1, \frac{2\sigma^{ora}\lambda|T|}{CIF_1(2\xi+1, T, \mathbf{Y})}\right\}\right] \\ =\ & 2t_1\sigma^{ora}\left[-\tau + \lambda \max\left\{\frac{2(1+\xi)}{\sigma^{ora}}\left\|\left(d^{true}\right)_{T^c}\right\|_1, \frac{2\lambda|T|}{CIF_1(2\xi+1, T, \mathbf{Y})}\right\}\right] < 0 \end{aligned}$$

where last inequality is from the definition of $\tau$.

**(2)**. Let $t_2 = \sigma^{ora}(1+\tau)$.To show the other side $\hat{\sigma} \leq \sigma^{ora}(1+\tau)$ it is enough to show

$$\frac{\partial}{\partial \sigma} L_\lambda\left(\hat{d}(\sigma\lambda), \sigma\right)|_{\sigma=t_2} \geq 0.$$

Equation (115) implies that on the event $\left\{\nu \leq t_2\lambda\frac{\xi-1}{\xi+1}\right\} = \left\{\nu/\sigma^{ora} < \lambda\frac{\xi-1}{\xi+1}(1+\tau)\ \right\}$ we have

$$
\begin{aligned}
& 2t_2^2 \frac{\partial}{\partial \sigma} L_\lambda \left(\hat{d}\left(\sigma\lambda\right), \sigma\right) |_{\sigma=t_2} \\
= \;& t_2^2 - \frac{\left\|\mathbf{X}_m - \mathbf{Y}\hat{d}\left(t_2\lambda\right)\right\|^2}{n} \\
= \;& t_2^2 - \left(\sigma^{ora}\right)^2 + \left(\sigma^{ora}\right)^2 - \frac{\left\|\mathbf{X}_m - \mathbf{Y}\hat{d}\left(t_2\lambda\right)\right\|^2}{n} \\
= \;& t_2^2 - \left(\sigma^{ora}\right)^2 + \frac{\left\|\mathbf{X}_m - \mathbf{Y}d^{true}\right\|^2 - \left\|\mathbf{X}_m - \mathbf{Y}\hat{d}\left(t_2\lambda\right)\right\|^2}{n} \\
= \;& t_2^2 - \left(\sigma^{ora}\right)^2 + \frac{\left(\hat{d}\left(t_2\lambda\right) - d^{true}\right)^T \mathbf{Y}^T \left(\mathbf{X}_m - \mathbf{Y}d^{true} + \mathbf{X}_m - \mathbf{Y}\hat{d}\left(t_2\lambda\right)\right)}{n} \\
\geq \;& t_2^2 - \left(\sigma^{ora}\right)^2 - \left\|\hat{d}\left(t_2\lambda\right) - d^{true}\right\|_1 \left(\nu + t_2\lambda\right).
\end{aligned}
$$

Equation (90) and the fact $1 + \tau \leq 2$ imply

$$
\begin{aligned}
& 2t_2^2 \frac{\partial}{\partial \sigma} L_\lambda \left(\hat{d}\left(\sigma\lambda\right), \sigma\right) |_{\sigma=t_2} \\
\geq \;& \left(t_2 + \sigma^{ora}\right) \sigma^{ora} \tau - \max\left\{2\left(1+\xi\right)\left(\nu + t_2\lambda\right)\left\|\left(d^{true}\right)_{T^c}\right\|_1, \frac{\left(\nu + t_2\lambda\right)^2 |T|}{CIF_1\left(2\xi+1, T, \mathbf{Y}\right)}\right\} \\
\geq \;& \left(\sigma^{ora}\right)^2 \left(\left(2+\tau\right)\tau - \max\left\{\frac{2\left(1+\xi\right)\left(2\lambda\left(1+\tau\right)\right)}{\sigma^{ora}}\left\|\left(d^{true}\right)_{T^c}\right\|_1, \frac{8\left(1+\tau\right)\lambda^2 |T|}{CIF_1\left(2\xi+1, T, \mathbf{Y}\right)}\right\}\right) \\
\geq \;& \left(\sigma^{ora}\right)^2 \tau,
\end{aligned}
$$

where last inequality is from the definition of $\tau$.

**4. Proof of Proposition 4.** This Proposition essentially follows from the shifting inequality Proposition 5 in Ye and Zhang (2010). We will give a brief proof using results and notations in that paper.

Define the generalized version of $l_q$ cone invertibility factor (81),

$$
CIF'_{q,l}\left(\alpha, K, \mathbf{Y}\right) = \inf\left\{\frac{|K|^{1/q}\left\|\frac{\mathbf{Y}^T\mathbf{Y}}{n}u\right\|_\infty}{\left\|u_A\right\|_q} : u \in \mathcal{C}\left(\alpha, K\right), u \neq 0, |A \backslash K| \leq l\right\}.
$$

When $q = 1$ and $l = p$, $CIF'_{q,l}\left(\alpha, K, \mathbf{Y}\right) = CIF'_{1,p}\left(\alpha, K, \mathbf{Y}\right) = CIF_1\left(\alpha, K, \mathbf{Y}\right)$. By Equations (17), (18) and (20) of Ye and Zhang (2010) we have

$$
\begin{aligned}
CIF_1\left(\alpha, K, \mathbf{Y}\right) &= CIF'_{1,p}\left(\alpha, K, \mathbf{Y}\right) \geq \frac{CIF'_{2,l}\left(\alpha, K, \mathbf{Y}\right)}{C_{1,2}\left(\alpha, \frac{|K|}{l}\right)} \\
&\geq \frac{\phi^*_{2,l}\left(\alpha, K, \mathbf{Y}\right)}{C_{1,2}\left(\alpha, \frac{|K|}{l}\right)} \geq \frac{\tilde{\phi}^*_{2,l}\left(\alpha, K, \mathbf{Y}\right)}{C_{1,2}\left(\alpha, \frac{|K|}{l}\right)\left(\left(1+\alpha\right) \wedge \sqrt{1 + \frac{l}{|K|}}\right)}
\end{aligned}
$$

where $C_{1,2}\left(\alpha, \frac{|K|}{l}\right)$, $\phi^*_{2,l}(\alpha, K, \mathbf{Y})$ and $\tilde{\phi}^*_{2,l}(\alpha, K, \mathbf{Y})$ are defined on page 3523-3524 of Ye and Zhang (2010). From the definition of $\tilde{\phi}^*_{2,l}(\alpha, K, \mathbf{Y})$ in Equation (20) of Ye and Zhang (2010), setting $r = 2$ (thus $a_r = 1/4$ on page 3523) we have

$$\tilde{\phi}^*_{2,l}(\alpha, K, \mathbf{Y}) \geq 1 - \pi^-_{l+k}(\mathbf{Y}) - \alpha\sqrt{\frac{k}{4l}}\theta_{4l,k+l}(\mathbf{Y}).$$

Since $C_{1,2}\left(\alpha, \frac{|K|}{l}\right) = 1 + \alpha$, then

$$CIF_1(\alpha, K, \mathbf{Y}) \geq \frac{1}{(1+\alpha)\left((1+\alpha) \wedge \sqrt{1+\frac{l}{k}}\right)}\left(1 - \pi^-_{l+k}(\mathbf{Y}) - \alpha\sqrt{\frac{k}{4l}}\theta_{4l,k+l}(\mathbf{Y})\right).$$

**5. Proof of Lemma 1.** Most of this proof is the same as that of Lemma 2. Hence we only emphasize the differences here and provide details whenever necessary. The proof also consists of algebraic analysis and probabilistic analysis. Recall that the main idea in the proof of Lemma 2 is to show the events $\cap_{i=1}^4 I_i$ defined in (86)-(89) occur with high probability and whenever they hold, Proposition 2 and Proposition 1 establish the desired results. Since we decrease the penalty term $\lambda^{new}$, the event $I_1$ is no longer valid with high probability. Now we need to redefine an appropriate event $I_1^{new}$ and show that it occurs with high probability and on $I_1^{new}$, similar properties like Proposition 2 and Proposition 1 also hold.

Recall $\nu = \|h\|_\infty$ with $h := \mathbf{Y}^T\epsilon_m/n$ defined in (83). Similar as (82), we define the index set $T = \{k \in A^c, |d_k^{true}| \geq \lambda^{new}\}$ of $d^{true}$ with large coordinates. Now by using smaller $\lambda^{new}$, although $\nu$ is no longer smaller than $\sigma^{ora}\lambda^{new}\frac{\xi-1}{\xi+1}(1-\tau)$ in $I_1$ w.h.p., most coordinates of $h$ still hold. Define a random index set

$$\hat{T}_0 = \left\{k \in A^c, |h_k| \geq \sigma^{ora}\lambda^{new}\frac{\xi-1}{\xi+1}(1-\tau)\right\}.$$

The new event can be defined as

$$I_1^{new} = \left\{\nu \leq \sigma^{ora}\frac{\lambda^{new}}{1-t}\frac{\xi-1}{\xi+1}(1-\tau)\right\} \cup \left\{\left|\hat{T}_0\right| \leq C_u s_{\max}\right\},$$

where $C_u$ is some universal constant. Note that $\lambda^{new} \geq (1-t)\lambda$ by our assumption $s_{\max} = O(p^t)$, thus the probabilistic analysis for $I_1$ in Lemma 2 immediately implies that the first component of $I_1^{new}$ holds w.h.p.. The second component holds w.h.p. can be shown by large deviation result of order statistics, where the first component is also can be seen as a special case. (See, e.g. Reiss (1989)). Therefore, we briefly showed that $\mathbb{P}\{(I_1^{new})^c\} \leq \left(p^{-\delta+1}/\sqrt{\log p}\right)$. We also need to modify the event $I_2$ a little bit as we not only care about index $T$ but also index $\hat{T}_0$ which are out of the bound.

$$I_2^{new} = \left\{ CIF_1 \left( \xi + 2 + \frac{\xi - 1}{1 - t}, \hat{T}_0 \cup T, \mathbf{Y} \right) \geq C > 0 \right\}.$$

The probabilistic analysis of $I_2$ in Lemma 2 implies that there is no difference by using $I_2^{new}$ since the probability bound for $I_2$ is universal for all index sets with cardinality less than $s + C_u s_{\max} = o\left(\frac{n}{\log p}\right)$. We don't change events $I_3$ and $I_4$. Thus we finish the probabilistic analysis.

Now it's enough for us to show that on $\cap_{i=1}^{2} I_i^{new} \cap_{i=3}^{4} I_i$, the desired results hold. It's not hard to see in the proof of Proposition 2 that as long as a similar property like Proposition 1 holds (we will provide details and prove this key result in a minute), Proposition 2 is still valid when we replace $I_1$ by $I_1^{new}$ in the assumption. The only thing we need to show is the following Proposition 6. Note on $\cap_{i=1}^{2} I_i^{new} \cap_{i=3}^{4} I_i$, Proposition 2 is valid and hence the assumption of the following Proposition with $\mu = \hat{\sigma}\lambda^{new}$ is also satisfied. We then apply this Proposition 6 again to finish the algebraic analysis and hence complete our proof.

PROPOSITION 6. *For any $\xi > 1$, on the event $\left\{ \nu \leq \frac{\mu}{1-t} \frac{\xi-1}{\xi+1} \right\} \cup \left\{ \left| \hat{T}_1 \right| \leq C_u s_{\max} \right\}$ with $\hat{T}_1 = \left\{ k \in A^c, |h_k| \geq \mu \frac{\xi-1}{\xi+1} \right\}$, we have*

$$\left\| \hat{d}(\mu) - d^{true} \right\|_1 \leq \max \left\{ (2 + 2\xi) \left\| \left( d^{true} \right)_{T^c} \right\|_1, \frac{(\nu + \mu) \left| T \cup \hat{T}_1 \right|}{CIF_1} \right\}, \tag{117}$$

$$\frac{1}{n} \left\| \mathbf{Y} \left( d^{true} - \hat{d}(\mu) \right) \right\|^2 \leq (\nu + \mu) \left\| \hat{d}(\mu) - d^{true} \right\|_1, \tag{118}$$

*where $CIF_1$ above is short for $CIF_1 \left( \xi + 2 + \frac{\xi-1}{1-t}, T \cup \hat{T}_1, \mathbf{Y} \right)$.*

The proof is a modification of that for Proposition 1. We still have equation (110) $\left\| \frac{\mathbf{Y}^T \mathbf{Y}}{n} \left( \hat{d}(\mu) - d^{true} \right) \right\|_\infty \leq \mu + \nu$. Define $\Delta(\mu) := d^{true} - \hat{d}(\mu)$. The equation (112) needs to be modified as follows,

$$\frac{\|\mathbf{Y}\Delta(\mu)\|^2}{n} = \frac{\Delta^T(\mu) \left( \mathbf{Y}^T \left( \mathbf{X}_i - \mathbf{Y}\hat{d}(\mu) \right) - \mathbf{Y}^T \left( \mathbf{X}_i - \mathbf{Y} d^{true} \right) \right)}{n}$$

$$\leq \mu \left( \left\| d^{true} \right\|_1 - \left\| \hat{d}(\mu) \right\|_1 \right) + \nu \left\| \Delta(\mu)_{\hat{T}_1} \right\|_1 + \mu \frac{\xi - 1}{\xi + 1} \left\| \Delta(\mu)_{\hat{T}_1^c} \right\|_1 \tag{119}$$

$$\leq \left( \mu + \frac{\mu}{1 - t} \frac{\xi - 1}{\xi + 1} \right) \left\| \Delta(\mu)_{T \cup \hat{T}_1} \right\|_1 + 2\mu \left\| \left( d^{true} \right)_{T^c} \right\|_1 - \left( \mu - \mu \frac{\xi - 1}{\xi + 1} \right) \left\| \Delta(\mu)_{(T \cup \hat{T}_1)^c} \right\|_1. \tag{120}$$

where the first inequality follows from the KKT conditions (80) and our assumed event. The remaining part is the same as that in Proposition 1. Suppose $\|\Delta(\mu)\|_1 \geq 2(1+\xi)\left\|(d^{true})_{T^c}\right\|_1$, then inequality (120) becomes

$$\left(\left(\xi+2+\frac{\xi-1}{1-t}\right)\left\|\Delta(\mu)_{T\cup\hat{T}_1}\right\|_1-\left\|\Delta(\mu)_{\left(T\cup\hat{T}_1\right)^c}\right\|_1\right)\geq 0.$$

Thus we have

$$d^{true}-\hat{d}(\mu)=\Delta(\mu)\in\mathcal{C}\left(\xi+2+\frac{\xi-1}{1-t},T\cup\hat{T}_1\right).$$

Combining this fact with equation (110), we obtain the first desired inequality (117). We complete our proof by noting that (119) implies the second desired inequality (118).

Now we show that $\lambda^{new}$ can be replaced by its finite sample version $\lambda^{new}_{finite}$. As we have seen, the analysis of event $I_1$ is the key result. All we need to show is that $\mathbb{P}\left\{|h_k|>\lambda^{new}_{finite}\right\}\leq O\left(p^{-\delta}\right)$, where $\frac{\sqrt{n-1}h_k}{\sqrt{1-h_k^2}}$ follows an $t$ distribution with $n-1$ degrees of freedom. Since $h_k$ is an increasing function of $\frac{\sqrt{n-1}h_k}{\sqrt{1-h_k^2}}$ on $\mathbb{R}^+$, we can take the quantile of $t_{(n-1)}$ distribution rather than use the concentration inequality above.

## References.


PETROV, V. V. (1995). *Limit theorems of probability theory.* Oxford Science Publications.

REISS, R. D. (1989). *Approximate distributions of order statistics: with applications to nonparametric statistics.* Springer-Verlag New York.

YE, F. and ZHANG, C. H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the $\ell_q$ loss in $\ell_r$ ball. *Journal of Machine Learning Research* **11** 3481-3502.



DEPARTMENT OF STATISTICS
YALE UNIVERSITY
NEW HAVEN, CONNECTICUT 06511
USA
E-MAIL: zhao.ren@yale.edu
huibin.zhou@yale.edu

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: tingni@wharton.upenn.edu

DEPARTMENT OF STATISTICS AND BIOSTATISTICS
HILL CENTER, BUSCH CAMPUS
RUTGERS UNIVERSITY
PISCATAWAY, NEW JERSEY 08854
USA
E-MAIL: cunhui@stat.rutgers.edu