Supplementary Materials for "A Novel Convex Clustering Method for High-Dimensional Data Using Semi-Proximal ADMM"

HUANGYUE CHEN[1], LINGCHEN KONG[1], YAN LI[2]

[1] Department of Applied Mathematics, Beijing Jiaotong University, Beijing 100044, China

[2] School of Insurance and Economics, University of International Business and Economics, Beijing 100029, China

hychen@bjtu.edu.cn, lchkong@bjtu.edu.cn, liyan2010@uibe.edu.cn

July 8, 2020

In this part, we provide proofs of main theoretical results of the sparse group lasso convex clustering (SGLCC). Before that, we shall introduce some notations and establish some useful properties to facilitate the proofs of Theorems. We first define $\widetilde{H}_\ell = I_p \otimes (e_{\ell_1} - e_{\ell_2})^T \in \mathbb{R}^{p\times np}$ is a submatrix of $\widetilde{H} = (\widetilde{H}_1^T, \widetilde{H}_2^T, \cdots, \widetilde{H}_{|\Theta|}^T)^T$ and provide some results of $H, \widetilde{H}$ in Proposition 0.1.

Proposition 0.1 *According to the definitions of H and $\widetilde{H}$, we have following results.*
(1) rank$(H) = n-1$ and rank$(\widetilde{H}) = (n-1)p$.
(2) $\Lambda_{\min}(H) = \Lambda_{\max}(H) = \sqrt{n}$ and $\Lambda_{\min}(\widetilde{H}) = \Lambda_{\max}(\widetilde{H}) = \sqrt{n}$, where $\Lambda_{\min}(\cdot)$ and $\Lambda_{\max}(\cdot)$ are the minimum nonzero singular value and the maximum singular value of matric, respectively.
(3) The SGLCC model is equivalent to

$$\min_{x\in\mathbb{R}^{np}} \left\{ \frac{1}{2}\|x-a\|_2^2 + \gamma_1 \sum_{\ell\in\Theta} \omega_\ell \|\widetilde{H}_\ell x\|_q + \gamma_2[(1-\alpha)\sum_{j=1}^{p} u_j \|E_j x\|_2 + \alpha\|x\|_1] \right\}, \tag{1}$$

where $E_j = e_j^T \otimes I_n \in \mathbb{R}^{n\times np}$ and $a = vec(A)$, and its optimal solution is denoted as $\hat{x} = (\hat{x}_1^T, \hat{x}_2^T, \cdots, \hat{x}_p^T)^T$.

The following lemma provides a boundary for probability of the quadratic forms of independent sub-Gaussian random variables.

Lemma 0.1 *([1]) Let ϵ be independent sub-Gaussian random variables with mean 0 and variance δ^2. Let M be a symmetric matrix. Then there exists some positive constants b_1, b_2 such that,*

$$P(\epsilon^T M \epsilon > t + \delta^2 tr(M)) \leq exp\left\{ -\min\left(\frac{b_1 t^2}{\delta^4\|M\|_F}, \frac{b_2 t}{\delta^2\|M\|_2} \right) \right\}, \text{ for any } t > 0.$$

For the sake of simplicity of our analysis, we apply the reformulate technique as in [2]. Note that $\text{rank}(\widetilde{H}) = (n-1)p$. Let $\widetilde{H} = U\Sigma V_f^T$ be the singular value decomposition of $\widetilde{H}$, where $U \in \mathbb{R}^{|\Theta|p\times(n-1)p}$ such that $U^TU = I_{(n-1)p}$, $\Sigma \in \mathbb{R}^{(n-1)p\times(n-1)p}$ is a diagonal matrix, and $V_f \in \mathbb{R}^{np\times(n-1)p}$ such that $V_f^TV_f = I_{(n-1)p}$. Construct a matrix $V_g \in \mathbb{R}^{np\times p}$ such that $V = [V_f, V_g] \in \mathbb{R}^{np\times np}$ is orthogonal matrix, that is $V^TV = I_{np}$, and note that $V_f^TV_g = 0$. Moreover, let $f = V_f^Tx \in \mathbb{R}^{(n-1)p}$ and $g = V_g^Tx \in \mathbb{R}^p$, and thus $x = V_ff + V_gg$. Similarly, we denote $f^* = V_f^Tx^* \in \mathbb{R}^{(n-1)p}$ and $g^* = V_g^Tx^* \in \mathbb{R}^p$, and thus $x^* = V_ff^* + V_gg^*$.

Hence, based on these notations above and $\omega_\ell = 1$, the model (1) is equivalent to

$$\begin{aligned}\min_{f,g}\ & \frac{1}{2}\|a - V_ff - V_gg\|_2^2 + \gamma_1\sum_{\ell\in\Theta}\|G_\ell f\|_q + \gamma_2(1-\alpha)\sum_{j=1}^p u_j\|E_j(V_ff + V_gg)\|_2 \\ & + \gamma_2\alpha\|V_ff + V_gg\|_1,\end{aligned} \tag{2}$$

where G_ℓ is a submatrix of $G = (G_1; G_2; \cdots; G_{|\Theta|})$ such that $G = U\Sigma$. Note that $rank(G) = (n-1)p$, then there exists pseudo-inverse $G^+ \in \mathbb{R}^{(n-1)p\times|\Theta|p}$ such that $G^+G = I_{(n-1)p}$. Let $(\hat{f}, \hat{g})$ be the solution to (2). Then $\hat{f} = V_f^T\hat{x}, \hat{g} = V_g^T\hat{x}$ and thus $\hat{x} = V_f\hat{f} + V_g\hat{g}$.

Proof of Theorem 3.1: By the definition of $(\hat{f}, \hat{g})$, we obtain that

$$\begin{aligned}&\frac{1}{2}\|a - V_f\hat{f} - V_g\hat{g}\|_2^2 + \gamma_1\|G\hat{f}\|_1 + \gamma_2(1-\alpha)\sum_{j=1}^p u_j\|E_j(V_f\hat{f} + V_g\hat{g})\|_2 + \gamma_2\alpha\|V_f\hat{f} + V_g\hat{g}\|_1 \\ &\leq \frac{1}{2}\|a - V_ff^* - V_gg^*\|_2^2 + \gamma_1\|Gf^*\|_1 + \gamma_2(1-\alpha)\sum_{j=1}^p u_j\|E_j(V_ff^* + V_gg^*)\|_2 + \\ &\gamma_2\alpha\|V_ff^* + V_gg^*\|_1.\end{aligned}$$

Using elementary relations $\frac{1}{2}(\|\eta_1\|^2 - \|\eta_2\|^2) = \langle\eta_1 - \eta_2, \eta_2\rangle + \frac{1}{2}||\eta_1 - \eta_2||^2$, we further obtain

$$\begin{aligned}\frac{1}{2}\|V_f(\hat{f} - f^*) + V_g(\hat{g} - g^*)\|_2^2 \leq\ & \epsilon^T[V_f(\hat{f} - f^*) + V_g(\hat{g} - g^*)] + \gamma_1(\|Gf^*\|_1 - \|G\hat{f}\|_1) \\ & + \gamma_2(1-\alpha)\sum_{j=1}^p u_j[\|E_j(V_ff^* + V_gg^*)\|_2 - \|E_j(V_f\hat{f} + V_g\hat{g})\|_2] \\ & + \gamma_2\alpha[\|V_ff^* + V_gg^*\|_1 - \|V_f\hat{f} + V_g\hat{g}\|_1].\end{aligned} \tag{3}$$

Next, we shall build up the relationship between $\hat{g}$ and g_0. By the optimality condition of (2), we have $-V_g^T(a - V_f\hat{f} - V_g\hat{g}) + \beta = 0$. This, together with $\epsilon = a - x^* = a - V_ff^* - V_gg^*$, implies

$$\hat{g} - g^* = V_g^T\epsilon - \beta,$$

where $\beta = \beta_1 + \beta_2$, β_1 and β_2 is subgradient of the third term and fourth term in (2), respectively.

We shall estimate the first term in the right-hand side of (3). Then

$$\begin{aligned}|\epsilon^T[V_f(\hat{f}-f^*)+V_g(\hat{g}-g^*)]| &= |\epsilon^T[V_f(\hat{f}-f^*)+V_g(V_g^T\epsilon-\beta)]| \\ &\le |\epsilon^T V_g V_g^T \epsilon| + |\epsilon^T V_g \beta| + |\epsilon^T V_f(\hat{f}-f^*)| \\ &\le |\epsilon^T V_g V_g^T \epsilon| + |\epsilon^T V_g \beta| + \|\epsilon^T V_f G^+\|_\infty \|G(\hat{f}-f^*)\|_1,\end{aligned}$$

where the first inequality follows from Cauchy-Schwarz inequality, and the last inequality follows from $G^+G = I$ and the Holder's inequality. We now establish bounds for three terms in the right-hand side of the inequality that hold with high probability.

Bounds for $|\epsilon^T V_g V_g^T \epsilon|$ and $\|\epsilon^T V_f G^+\|_\infty$. Using Lemma 0.1 and condition **A1**, we obtain that

$$P\left(|\epsilon^T V_g V_g^T \epsilon| \ge \delta^2\left[p+\sqrt{p\log(np)}\right]\right) \le \exp\left\{-\min\left(b_1\log(np), b_2\sqrt{p\log(np)}\right)\right\} \tag{4}$$

and

$$P\left(\|\epsilon^T V_f G^+\|_\infty \ge 2\delta\sqrt{\frac{\log(p|\Theta|)}{n}}\right) \le \frac{2}{p|\Theta|}, \tag{5}$$

see Lemma 6 in [3] for detailed derivation.

Bound for $|\epsilon^T V_g \beta|$. Note that $\beta_1 = \gamma_2(1-\alpha)\sum_{j=1}^p u_j \frac{V_g^T E_j^T E_j(V_f\hat{f}+V_g\hat{g})}{\|E_j(V_f\hat{f}+V_g\hat{g})\|_2}$ when $\hat{x} \neq 0$, and thus

$$\begin{aligned}\|\beta_1\|_2 &= \left\|\gamma_2(1-\alpha)\sum_{j=1}^p u_j \frac{V_g^T E_j^T E_j(V_f\hat{f}+V_g\hat{g})}{\|E_j(V_f\hat{f}+V_g\hat{g})\|_2}\right\|_2 \\ &\le \gamma_2(1-\alpha)\sum_{j=1}^p u_j \|V_g^T\|_2 \|E_j^T\|_2 = \gamma_2(1-\alpha)\|u\|_1.\end{aligned}$$

Moreover, setting $h(g) := \gamma_2\alpha\|V_f\hat{f}+V_g g\|_1$, we know that

$$\begin{aligned}|h(g_1)-h(g_2)| &= \gamma_2\alpha\left|\|V_f\hat{f}+V_g g_1\|_1 - \|V_f\hat{f}+V_g g_2\|_1\right| \\ &\le \gamma_2\alpha\|V_g(g_1-g_2)\|_1 \le \gamma_2\alpha\|V_g\|_2\|g_1-g_2\|_2 = \gamma_2\alpha\|g_1-g_2\|_2,\end{aligned}$$

which implies function $h(g)$ is Lipschitz continuous with constant $\gamma_2\alpha$. Hence, $||\beta_2||_2 \le \gamma_2\alpha$ by Proposition 2.47 in [4]. Next, we have

$$\|V_g\beta\|_2 \le \|V_g\|_2\|\beta\|_2 = \|\beta\|_2 \le \gamma_2(1-\alpha)\|u\|_1 + \gamma_2\alpha.$$

We know that $\epsilon^T V_g \beta$ is a sub-Gaussian random variable with mean zero and variance $\|V_g\beta\|_2^2\delta^2$ by Condition **A1**. Using Chebyshev's inequality, we have

$$P\left(|\epsilon^T V_g \beta| > t\right) \le \frac{\|V_g\beta\|_2^2\delta^2}{t^2} \le \frac{(\gamma_2(1-\alpha)\|u\|_1+\gamma_2\alpha)^2\delta^2}{t^2}.$$

Picking $t = \sqrt{np}$, we obtain from the above inequality that

$$P\left(|\epsilon^T V_g \beta| > \sqrt{np}\right) \leq \frac{(\gamma_2(1-\alpha)\|u\|_1 + \gamma_2\alpha)^2\delta^2}{np}. \tag{6}$$

Together with (4) and (5), we further obtain that

$$\begin{aligned}&\epsilon^T V_g V_g^T \epsilon + |\epsilon^T V_g \beta| + \|\epsilon^T V_f G^+\|_\infty \|G(\hat{f} - f^*)\|_1 \geq \\ &\delta^2 \left[p + \sqrt{p\log(np)}\right] + \sqrt{np} + 2\delta\sqrt{\frac{\log(p|\Theta|)}{n}}\|G(\hat{f} - f^*)\|_1\end{aligned}$$

holds with probability at most c_4, where

$$c_4 = \exp\left\{-\min\left(b_1\log(np), b_2\sqrt{p\log(np)}\right)\right\} + \frac{2}{p|\Theta|} + \frac{(\gamma_2(1-\alpha)\|u\|_1 + \gamma_2\alpha)^2\delta^2}{np}.$$

Hence,

$$|\epsilon^T[V_f(\hat{f} - f^*) + V_g(\hat{g} - g^*)]| \leq \delta^2\left[p + \sqrt{p\log(np)}\right] + \sqrt{np} + 2\delta\sqrt{\frac{\log(p|\Theta|)}{n}}\|G(\hat{f} - f^*)\|_1 \tag{7}$$

holds with probability at least $1 - c_4$.

Furthermore, it is clear that

$$\begin{aligned}&\gamma_2(1-\alpha)\sum_{j=1}^p u_j[\|E_j(V_f f^* + V_g g^*)\|_2 - \|E_j(V_f \hat{f} + V_g \hat{g})\|_2] \\ &= \gamma_2(1-\alpha)\sum_{j=1}^p u_j(\|x_j^*\|_2 - \|\hat{x}_j\|_2) \leq \frac{\gamma_2(1-\alpha)}{2}(\|u\|_2^2 + \|x^* - \hat{x}\|_2^2),\end{aligned} \tag{8}$$

where the last inequality follows from triangle inequality and $\eta_1\eta_2 \leq \frac{1}{2}(\eta_1^2 + \eta_2^2)$. Similarly, the last term in (3) can be estimated as below:

$$\gamma_2\alpha[\|V_f f^* + V_g g^*\|_1 - \|V_f \hat{f} + V_g \hat{g}\|_1] = \gamma_2\alpha(\|x^*\|_1 - \|\hat{x}\|_1) \leq \gamma_2\alpha\|x^*\|_1. \tag{9}$$

Substituting the inequality (7-9) into (3) and letting $\gamma_1 > 2\delta\sqrt{\frac{\log(p|\Theta|)}{n}}$, we know

$$\begin{aligned}\frac{1}{2}\|V_f(\hat{f} - f^*) + V_g(\hat{g} - g^*)\|_2^2 \leq &\ \gamma_1\|G(\hat{f} - f^*)\|_1 + \gamma_1(\|Gf^*\|_1 - \|G\hat{f}\|_1) \\ &+ \frac{\gamma_2(1-\alpha)}{2}(\|u\|_2^2 + \|x^* - \hat{x}\|_2^2) + \gamma_2\alpha\|x^*\|_1 \\ &+ \delta^2\left[p + \sqrt{p\log(np)}\right] + \sqrt{np}.\end{aligned}$$

holds with probability at least $1 - c_4$. Dividing both sides by $np/2$, we further obtain that

$$\begin{aligned}
\frac{1+\gamma_2(\alpha-1)}{np}\|\hat{x}-x^*\|_2^2 &\leq \frac{4\gamma_1}{np}\|Gf^*\|_1 + \frac{\gamma_2(1-\alpha)}{np}\|u\|_2^2 + \frac{2\gamma_2\alpha}{np}\|x^*\|_1 \\
&\quad + 2\delta^2\left[\frac{1}{n} + \sqrt{\frac{log(np)}{n^2p}}\right] + \frac{2}{\sqrt{np}} \\
&\leq \frac{4\gamma_1}{np}\|\widetilde{H}x^*\|_1 + \frac{\gamma_2(1-\alpha)}{np}\|u\|_2^2 + \frac{2\gamma_2\alpha}{np}\|x^*\|_1 \\
&\quad + 2\delta^2\left[\frac{1}{n} + \sqrt{\frac{\log(np)}{n^2p}}\right] + \frac{2}{\sqrt{np}}
\end{aligned}$$

holds with probability at least $1 - c_4$.

Proof of Theorem 3.2: Let us prove Theorem 3.2 along the lines of Theorem 3.1. Since $(\hat{f}, \hat{g})$ is the global minimizer of (2), we have

$$\begin{aligned}
&\frac{1}{2}\|a - V_f\hat{f} - V_g\hat{g}\|_2^2 + \gamma_1\sum_{\ell\in\Theta}\|G_\ell\hat{f}\|_2 + \gamma_2(1-\alpha)\sum_{j=1}^{p}u_j\|E_j(V_f\hat{f}+V_g\hat{g})\|_2 + \gamma_2\alpha\|V_f\hat{f}+V_g\hat{g}\|_1 \\
&\leq \frac{1}{2}\|a - V_ff^* - V_gg^*\|_2^2 + \gamma_1\sum_{\ell\in\Theta}\|G_\ell f^*\|_2 + \gamma_2(1-\alpha)\sum_{j=1}^{p}u_j\|E_j(V_ff^*+V_gg^*)\|_2 \\
&+ \gamma_2\alpha\|V_ff^*+V_gg^*\|_1
\end{aligned}$$

After some simple manipulations we further obtain that

$$\begin{aligned}
\frac{1}{2}\|V_f(\hat{f}-f^*) + V_g(\hat{g}-g^*)\|_2^2 &\leq \epsilon^T[V_f(\hat{f}-f^*) + V_g(\hat{g}-g^*)] + \gamma_1\sum_{\ell\in\Theta}\Big(\|G_\ell f^*\|_2 - \|G_\ell\hat{f}\|_2\Big) \\
&\quad + \gamma_2(1-\alpha)\sum_{j=1}^{p}u_j[\|E_j(V_ff^*+V_gg^*)\|_2 - \|E_j(V_f\hat{f}+V_g\hat{g})\|_2] \qquad (10)\\
&\quad + \gamma_2\alpha[\|V_ff^*+V_gg^*\|_1 - \|V_f\hat{f}+V_g\hat{g}\|_1].
\end{aligned}$$

Similarly, we obtain that $\hat{g} - g^* = V_g^T\epsilon - \beta$ by following the same arguments in Theorem 3.1. Thus,

$$\begin{aligned}
|\epsilon^T[V_f(\hat{f}-f^*) + V_g(\hat{g}-g^*)]| &= |\epsilon^TV_gV_g^T\epsilon - \epsilon^TV_g\beta + \epsilon^TV_f(\hat{f}-f^*)| \\
&\leq |\epsilon^TV_gV_g^T\epsilon| + \max_{\ell\in\Theta}\|\epsilon^TV_fG_\ell^+\|_2\sum_{\ell\in\Theta}\|G_\ell(\hat{f}-f^*)\|_2 + |\epsilon^TV_g\beta|,
\end{aligned}$$

where the last inequality follows from triangle inequality and Cauchy-Schwarz inequality. Next, we shall establish boundedness for the three terms on the right-hand side of the inequality.

Bounds for $|\epsilon^TV_gV_g^T\epsilon|$ and $|\epsilon^TV_g\beta|$. The boundedness of $|\epsilon^TV_gV_g^T\epsilon|$ and $|\epsilon^TV_g\beta|$ are established in (4) and (6), respectively.

Bound for $\max_{\ell\in\Theta}\|\epsilon^TV_fG_\ell^+\|_2$. We first notice that $\epsilon^TV_fG_\ell^+ \in \mathbb{R}^p$. Hence, we obtain that

$\|\epsilon^T V_f G_\ell^+\|_2 \leq \sqrt{p}\|\epsilon^T V_f G_\ell^+\|_\infty$, which implies

$$\max_{\ell\in\Theta} \|\epsilon^T V_f G_\ell^+\|_2 \leq \sqrt{p} \max_{\ell\in\Theta} \|\epsilon^T V_f G_\ell^+\|_\infty \leq \sqrt{p}\|\epsilon^T V_f G^+\|_\infty.$$

Thus,

$$P\left(\max_{\ell\in\Theta} \|\epsilon^T V_f G_\ell^+\|_2 \geq 2\delta\sqrt{\frac{p\log(p|\Theta|)}{n}}\right) \leq P\left(\sqrt{p}\|\epsilon^T V_f G^+\|_\infty \geq 2\delta\sqrt{\frac{p\log(p|\Theta|)}{n}}\right) \leq \frac{2}{p|\Theta|},$$

where the last inequality follows from (5). Together with (4) and (6), we obtain that

$$\begin{aligned}\left|\epsilon^T[V_f(\hat{f}-f^*)+V_g(\hat{g}-g^*)]\right| \leq 2\delta\sqrt{\frac{p\log(p|\Theta|)}{n}} \sum_{\ell\in\Theta} \|G_\ell(\hat{f}-f^*)\|_2 \\ +\delta^2\left[p+\sqrt{p\log(np)}\right]+\sqrt{np}\end{aligned} \tag{11}$$

holds with probability at least $1-c_4$.

Substituting the inequality (8-9) and (11) into (10), and letting $\gamma_1 > 2\delta\sqrt{\frac{p\log(p|\Theta|)}{n}}$, we obtain that the following relation holds with propability at least $1-c_4$,

$$\begin{aligned}\frac{1}{2}\|V_f(\hat{f}-f^*)+V_g(\hat{g}-g^*)\|_2^2 \leq \gamma_1 \sum_{\ell\in\Theta} \|G_\ell(\hat{f}-f^*)\|_2 + \gamma_1 \sum_{\ell\in\Theta}\left(\|G_\ell f^*\|_2 - \|G_\ell \hat{f}\|_2\right) \\ +\frac{\gamma_2(1-\alpha)}{2}(\|u\|_2^2+\|x^*-\hat{x}\|_2^2)+\gamma_2\alpha\|x^*\|_1 \\ +\delta^2\left[p+\sqrt{p\log(np)}\right]+\sqrt{np},\end{aligned}$$

Further, we know

$$\begin{aligned}\frac{1+\gamma_2(\alpha-1)}{np}\|\hat{x}-x^*\|_2^2 &\leq \frac{4\gamma_1}{np}\sum_{\ell\in\Theta}\|G_\ell f^*\|_2 + \frac{\gamma_2(1-\alpha)}{np}\|u\|_2^2 + \frac{2\gamma_2\alpha}{np}\|x^*\|_1 \\ &+\frac{2\delta^2}{np}\left[p+\sqrt{p\log(np)}\right]+\frac{2}{\sqrt{np}} \\ &\leq \frac{4\gamma_1}{np}\sum_{\ell\in\Theta}\|\widetilde{H}_\ell x^*\|_2 + \frac{\gamma_2(1-\alpha)}{np}\|u\|_2^2 + \frac{2\gamma_2\alpha}{np}\|x^*\|_1 \\ &+2\delta^2\left[\frac{1}{n}+\sqrt{\frac{\log(np)}{n^2p}}\right]+\frac{2}{\sqrt{np}}\end{aligned}$$

holds with probability at least $1-c_4$.

In the proofs of Theorems 3.3-3.4, we only need to verify that $P(||\hat{x}_j||_2=0)\to 1$ holds for an element $j\in\mathcal{I}^c$. Without loss of generality, let $p\in\mathcal{I}^c$, we will prove that $P(||\hat{x}_p||_2=0)\to 1$, the similar arguments apply to another elements in set $\mathcal{I}^c$.

Proof of Theorem 3.3: Suppose by contradiction that $||\hat{x}_p||_2 \neq 0$. The optimality condition of optimization problem (1) with $q = 1$ at $\hat{x}_p$ be given by

$$\hat{x}_p - a_p + \gamma_1 \sum_{\ell \in \Theta} \text{sign}((e_{\ell_1} - e_{\ell_2})^T \hat{x}_p)(e_{\ell_1} - e_{\ell_2}) + \gamma_2(1-\alpha)u_p \frac{\hat{x}_p}{||\hat{x}_p||_2} + \gamma_2 \alpha \text{sign}(\hat{x}_p) = 0$$

which implies that

$$\frac{1}{\sqrt{n}}(\hat{x}_p - x_p^*) - \frac{1}{\sqrt{n}}\epsilon_p + \frac{\gamma_1}{\sqrt{n}} \sum_{\ell \in \Theta} \text{sign}((e_{\ell_1} - e_{\ell_2})^T \hat{x}_p)(e_{\ell_1} - e_{\ell_2}) + \frac{1}{\sqrt{n}}\gamma_2(1-\alpha)u_p \frac{\hat{x}_p}{||\hat{x}_p||_2} + \frac{1}{\sqrt{n}}\gamma_2 \alpha \text{sign}(\hat{x}_p) = 0.$$

Based on Theorem 3.1 and its discussions, we know that $\frac{1}{n}||\hat{x}_p - x_p^*||_2^2 = o_P(1)$, that is, the first term is of the order $o_P(1)$. The term $\frac{1}{\sqrt{n}}\epsilon_p$ is of the order $o_P(1)$ because the ϵ_p is sub-Gaussian random variable. The fourth term and last term are of the order $o_P(1)$ by

$$\gamma_2(1-\alpha)u_p \frac{||\hat{x}_p||_2}{\sqrt{n}||\hat{x}_p||_2} = \frac{\gamma_2(1-\alpha)u_p}{\sqrt{n}} \to 0 \text{ and } \frac{\gamma_2 \alpha ||\text{sign}(\hat{x}_p)||_2}{\sqrt{n}} \to 0,$$

respectively.

Without loss of generality, we assume the first component of $\hat{x}_p$ is nonzero. Hence, the first component of $\sum_{\ell \in \Theta} \text{sign}((e_{\ell_1} - e_{\ell_2})^T \hat{x}_p)(e_{\ell_1} - e_{\ell_2})$ is of the order $O_P(n)$. We know from $\gamma_1 \geq 2\delta\sqrt{\frac{\log(p|\Theta|)}{n}}$ that the third term diverges to infinity, which contradicts with the optimality condition. Therefore, $\hat{x}_p = 0$ holds with probability tending to one.

Proof of Theorem 3.4: Let us prove Theorem 3.4 along the lines of Theorem 3.3. Suppose that $||\hat{x}_p||_2 \neq 0$. The optimality condition of optimization problem (1) with $q = 2$ at $\hat{x}_p$ be given by

$$\hat{x}_p - a_p + \gamma_1 \sum_{\ell \in \Theta} \frac{e_{\ell_1} - e_{\ell_2}}{||\widetilde{J}_\ell \hat{x}||_2} + \gamma_2(1-\alpha)u_p \frac{\hat{x}_p}{||\hat{x}_p||_2} + \gamma_2 \alpha \text{sign}(\hat{x}_p) = 0$$

which implies that

$$\frac{1}{\sqrt{n}}(\hat{x}_p - x_p^*) - \frac{1}{\sqrt{n}}\epsilon_p + \frac{\gamma_1}{\sqrt{n}} \sum_{\ell \in \Theta} \frac{e_{\ell_1} - e_{\ell_2}}{||\widetilde{J}_\ell \hat{x}||_2} + \frac{1}{\sqrt{n}}\gamma_2(1-\alpha)u_p \frac{\hat{x}_p}{||\hat{x}_p||_2} + \frac{1}{\sqrt{n}}\gamma_2 \alpha \text{sign}(\hat{x}_p) = 0.$$

We only need to discuss the third term because it is different with the case when $q = 1$. Note that all entries of $\hat{x}$ cannot the same, and thus

$$0 < ||\widetilde{H}_\ell \hat{x}||_2 \leq ||\widetilde{H}_\ell||_F ||\hat{x}||_2 = \sqrt{2p}||\hat{x}||_2.$$

We also know $1/||\hat{x}||_2 = O_P(\sqrt{np})$, and then,

$$\left(\sum_{\ell \in \Theta} \frac{e_{\ell_1} - e_{\ell_2}}{||\widetilde{J}_\ell \hat{x}||_2} \right)_1 = \frac{n}{||\widetilde{J}_\ell \hat{x}||_2} = O_P(n^{\frac{3}{2}}).$$

Further, We know from $\gamma_1 \geq 2\delta\sqrt{\frac{p\log(p|\Theta|)}{n}}$ that the third term diverges to infinity, which contradicts with the optimality condition. Hence, $\hat{x}_p = 0$ holds with probability tending to one.

References

[1] D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

[2] J. Liu, L. Yuan, and J. P. Ye. Guaranteed sparse recovery under linear transformation. *Proceedings of the 30th International Conference on Machine Learning, Atlanta, USA*, 28(3):91–99, 2013.

[3] K. M. Tan and D. Witten. Statistical properties of convex clustering. *Electronic Journal of Statistics*, 9(2):2324–2347, 2015.

[4] B. S. Mordukhovich and N. M. Nam. *An Easy Path to Convex Analysis and Applications*. Morgan & Claypool Publishers, St. Louis, 2014.