Appendix

A. Proof of Proposition 1

As mentioned in the statement, $\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = \widetilde{\boldsymbol{\alpha}} + \widetilde{\boldsymbol{\beta}} - \bar{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}} + \widetilde{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}} = \boldsymbol{\Delta} + \boldsymbol{\Gamma}$. If (i) either $[\boldsymbol{\Delta}]_j$ or $[\boldsymbol{\Gamma}]_j$ is zero, or (ii) $\mathrm{sign}([\boldsymbol{\Delta}]_j) = \mathrm{sign}([\boldsymbol{\Gamma}]_j)$, then $([\boldsymbol{\Delta}]_j + [\boldsymbol{\Gamma}]_j)^2 \geq ([\boldsymbol{\Delta}]_j)^2 + ([\boldsymbol{\Gamma}]_j)^2$. Therefore, if this happens for every j, the inequality (10) holds. When either $[\boldsymbol{\alpha}^*]_j \neq 0$ or $[\boldsymbol{\beta}^*]_j \neq 0$ holds, $[\boldsymbol{\Delta}]_j = 0$ or $[\boldsymbol{\Gamma}]_j = 0$ is guaranteed by construction (rule 2 or 3 above). If both $[\boldsymbol{\alpha}^*]_j$ and $[\boldsymbol{\alpha}^*]_j$ are zero (in case of rule 1), the following lemma ensures that $[\widetilde{\boldsymbol{\alpha}}]_j$ and $[\widetilde{\boldsymbol{\beta}}]_j$ (and therefore $[\boldsymbol{\Delta}]_j$ and $[\boldsymbol{\Gamma}]_j$ because $[\bar{\boldsymbol{\alpha}}]_j = [\bar{\boldsymbol{\beta}}]_j = 0$ in this case) have same signs.

Lemma 1. *The signs of $\widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\beta}}$ are always consistent whenever both are not zeros: sign$([\widetilde{\boldsymbol{\alpha}}]_j)$ = sign$([\widetilde{\boldsymbol{\beta}}]_j)$ for all j such that $[\widetilde{\boldsymbol{\alpha}}]_j \neq 0$ and $[\widetilde{\boldsymbol{\beta}}]_j \neq 0$.*

The proof of Lemma 1 is trivial. Suppose that $[\widetilde{\boldsymbol{\alpha}}]_j$ and $[\widetilde{\boldsymbol{\beta}}]_j$ have opposite signs; say $[\widetilde{\boldsymbol{\alpha}}]_j > 0$ and $[\widetilde{\boldsymbol{\beta}}]_j < 0$. Then $[\widetilde{\boldsymbol{\alpha}}]_j - \epsilon$ and $[\widetilde{\boldsymbol{\beta}}]_j + \epsilon$ with arbitrary small positive ϵ and all others fixed, will have the same loss by $\mathcal{L}(\cdot)$, but smaller values in the regularizers, which violates the stationary condition of local minimum.

Since the sign consistency is always guaranteed (or at least one of them is zero), the decomposability in the statement trivially holds.

Showing (11) also comes from the definitions of support sets. For any j such that $j \notin \boldsymbol{s}^*$ but $j \in U$, we set $\bar{\boldsymbol{\alpha}}_j := \widetilde{\boldsymbol{\alpha}}_j$ and hence $\boldsymbol{\Delta}_j = 0$ by definition. Therefore, the projection does not make any difference. The equality holds for $\boldsymbol{\Delta}_{\bar{\boldsymbol{s}}}$ since $\boldsymbol{s}^* \subseteq \bar{\boldsymbol{s}} \subseteq U$. The same reasoning holds for $\boldsymbol{\Gamma}$ as well.

B. Proofs for ℓ_2 Error Bounds

B.1. Proof of Theorem 1

Recall that we get $(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}})$ by the transformation $\mathcal{T}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*; \widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}})$. For notational simplicity, we use $\boldsymbol{\Lambda}$ to denote the error vector on our estimation: $\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$. We also define the individual error vectors as $\boldsymbol{\Delta} := \widetilde{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}$ and $\boldsymbol{\Gamma} := \widetilde{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}$. Note that $\boldsymbol{\Lambda} = \boldsymbol{\Delta} + \boldsymbol{\Gamma}$ since $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\alpha}} + \widetilde{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\alpha}} + \bar{\boldsymbol{\beta}}$ by definitions.

We first show that under (RSC), $\|\boldsymbol{\Lambda}\|_2 \leq 1$ is guaranteed, and hence the first inequality (12) for the case of $\|\boldsymbol{\Lambda}\|_2 \leq 1$ only matters. Toward this, given all the assumptions in the statement, suppose that $\|\boldsymbol{\Lambda}\|_2 \geq 1$.

Then setting $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ as $(\boldsymbol{\theta}^*, \boldsymbol{\Lambda})$ in (13), we obtain

$$\kappa_2 \|\boldsymbol{\Lambda}\|_2 - \tau_2 \|\boldsymbol{\Lambda}\|_{\boldsymbol{\eta}} \leq \left\langle \nabla \mathcal{L}(\boldsymbol{\theta}^* + \boldsymbol{\Lambda}) - \nabla \mathcal{L}(\boldsymbol{\theta}^*), \boldsymbol{\Lambda} \right\rangle . \tag{19}$$

At the same time, the stationary condition of (14) ensures that any local optimum satisfies

$$\left\langle \nabla \mathcal{L}(\widetilde{\boldsymbol{\theta}}) + \nabla \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}}, \boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}} \right\rangle \geq 0 \tag{20}$$

for any feasible $\boldsymbol{\theta}$. This is more general than the first-order stationary condition; for some local optima, the first-order stationary condition does not hold due to the inequality constraint of program (14). (see (Loh & Wainwright, 2014; 2015) for more details.) Since $\boldsymbol{\theta}^*$ is feasible by assumption, setting $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ yields

$$\left\langle \nabla \mathcal{L}(\widetilde{\boldsymbol{\theta}}) + \nabla \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}}, \boldsymbol{\Lambda} \right\rangle \leq 0 . \tag{21}$$

Combining (19) and (21) yields

$$\kappa_2 \|\boldsymbol{\Lambda}\|_2 - \tau_2 \|\boldsymbol{\Lambda}\|_{\boldsymbol{\eta}} \leq \left\langle -\nabla \mathcal{L}(\boldsymbol{\theta}^*) - \nabla \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}}, \boldsymbol{\Lambda} \right\rangle .$$

By Hölder's inequality,

$$\begin{aligned} \left\langle -\nabla \mathcal{L}(\boldsymbol{\theta}^*) - \nabla \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}}, \boldsymbol{\Lambda} \right\rangle &\leq \left\| \nabla \mathcal{L}(\boldsymbol{\theta}^*) + \nabla \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} \right\|_{\boldsymbol{\eta}}^* \|\boldsymbol{\Lambda}\|_{\boldsymbol{\eta}} \\ &\leq \left(\left\| \nabla \mathcal{L}(\boldsymbol{\theta}^*) \right\|_{\boldsymbol{\eta}}^* + \bar{\eta} \right) \|\boldsymbol{\Lambda}\|_{\boldsymbol{\eta}} \end{aligned} \tag{22}$$

where we utilize Lemma 3 with defining $\bar{\eta} := \max\{\frac{\lambda_1}{\eta_1}, \frac{\lambda_2}{\eta_2}\}$ in the second inequality. Moreover, since $\left\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\right\|_{\boldsymbol{\eta}}^* = \max\left(\frac{\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty}{\eta_1}, \frac{\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_{\infty,a^*}}{\eta_2}\right) \leq \max\left(\frac{\lambda_1}{4\eta_1}, \frac{\lambda_2}{4\eta_2}\right)$ by assumption, we obtain

$$\kappa_2\|\boldsymbol{\Lambda}\|_2 - \tau_2\|\boldsymbol{\Lambda}\|_{\boldsymbol{\eta}} \leq \frac{5\bar{\eta}}{4}\|\boldsymbol{\Lambda}\|_{\boldsymbol{\eta}}\,. \tag{23}$$

Rearranging (23), we have

$$\|\boldsymbol{\Lambda}\|_2 \leq \frac{\|\boldsymbol{\Lambda}\|_{\boldsymbol{\eta}}}{\kappa_2}\Big(\frac{5\bar{\eta}}{4} + \tau_2\Big) \overset{(i)}{\leq} \frac{2r}{\kappa_2}\Big(\frac{5\bar{\eta}}{4} + \tau_2\Big) \overset{(ii)}{\leq} 1$$

where (i) follows the fact $\|\boldsymbol{\Lambda}\|_{\boldsymbol{\eta}} \leq \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\eta}} + \|\boldsymbol{\theta}^*\|_{\boldsymbol{\eta}} \leq 2r$ (since $\boldsymbol{\theta}^*$ is feasible) and (ii) holds under the conditions that $r \leq \frac{\kappa_2}{4\tau_2}$ and $r\bar{\eta} \leq \frac{\kappa_2}{5}$.

From now we revisit the stationary condition of the problem (14) and the RSC condition (12) for $\|\boldsymbol{\Lambda}\|_2 \leq 1$. Recall that $\bar{\boldsymbol{s}}$ is the union of $supp(\boldsymbol{\alpha}^*)$ and $supp(\bar{\boldsymbol{\alpha}})$, and $\boldsymbol{\Delta}_{\bar{\boldsymbol{s}}}$ is the projection of $\boldsymbol{\Delta}$ onto the space w.r.t. $\bar{\boldsymbol{s}}$ (simply meaning $[\boldsymbol{\Delta}_{\bar{\boldsymbol{s}}}]_j = 0$ if $j \notin \bar{\boldsymbol{s}}$). $\bar{\boldsymbol{b}}$ was similarly defined as $supp(\boldsymbol{\beta}^*) \cup supp(\bar{\boldsymbol{\beta}})$. Rearranging the stationary condition (21), we obtain

$$\begin{aligned}
\Big\langle \nabla\mathcal{L}(\widetilde{\boldsymbol{\theta}})\,,\, \boldsymbol{\Lambda}\Big\rangle &\leq -\Big\langle \nabla\|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}}\,,\, \widetilde{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}\Big\rangle \overset{(i)}{\leq} \|\bar{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} - \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} \\
&\overset{(ii)}{\leq} \lambda_1\|\bar{\boldsymbol{\alpha}}\|_1 + \lambda_2\|\bar{\boldsymbol{\beta}}\|_{1,a} - \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} \overset{(iii)}{\leq} \lambda_1\|\bar{\boldsymbol{\alpha}}\|_1 + \lambda_2\|\bar{\boldsymbol{\beta}}\|_{1,a} - \lambda_1\|\widetilde{\boldsymbol{\alpha}}\|_1 - \lambda_2\|\widetilde{\boldsymbol{\beta}}\|_{1,a} \\
&= \lambda_1\Big(\|\bar{\boldsymbol{\alpha}}\|_1 + \big\|[\boldsymbol{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1 - \big\|[\boldsymbol{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1 - \|\widetilde{\boldsymbol{\alpha}}\|_1\Big) \\
&\quad + \lambda_2\Big(\|\bar{\boldsymbol{\beta}}\|_{1,a} + \big\|[\boldsymbol{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a} - \big\|[\boldsymbol{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a} - \|\widetilde{\boldsymbol{\beta}}\|_{1,a}\Big)
\end{aligned}$$

where we use (i) the convexity and (iii) the triangular inequality of $\|\cdot\|_{\boldsymbol{\lambda}}$ norm, and we obtain the inequality (ii) since $\|\bar{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}}$ has the infimal sum of two regularizers by definition. Hence,

$$\begin{aligned}
&\Big\langle \nabla\mathcal{L}(\widetilde{\boldsymbol{\theta}})\,,\, \boldsymbol{\Lambda}\Big\rangle \\
&\leq \lambda_1\Big(\|\bar{\boldsymbol{\alpha}}\|_1 + \big\|[\boldsymbol{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1 - \big\|[\boldsymbol{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1 - \|\widetilde{\boldsymbol{\alpha}}\|_1\Big) \\
&\quad + \lambda_2\Big(\|\bar{\boldsymbol{\beta}}\|_{1,a} + \big\|[\boldsymbol{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a} - \big\|[\boldsymbol{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a} - \|\widetilde{\boldsymbol{\beta}}\|_{1,a}\Big) \\
&\overset{(i)}{=} \lambda_1\Big(\big\|\bar{\boldsymbol{\alpha}} + [\boldsymbol{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1 - \big\|[\boldsymbol{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1 - \|\widetilde{\boldsymbol{\alpha}}\|_1\Big) \\
&\quad + \lambda_2\Big(\big\|\bar{\boldsymbol{\beta}} + [\boldsymbol{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a} - \big\|[\boldsymbol{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a} - \|\widetilde{\boldsymbol{\beta}}\|_{1,a}\Big) \\
&\overset{(ii)}{\leq} \lambda_1\Big(\big\|\bar{\boldsymbol{\alpha}} + [\boldsymbol{\Delta}]_{\bar{\boldsymbol{s}}^c} + [\boldsymbol{\Delta}]_{\bar{\boldsymbol{s}}}\big\|_1 + \big\|[\boldsymbol{\Delta}]_{\bar{\boldsymbol{s}}}\big\|_1 - \big\|[\boldsymbol{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1 - \|\widetilde{\boldsymbol{\alpha}}\|_1\Big) \\
&\quad + \lambda_2\Big(\big\|\bar{\boldsymbol{\beta}} + [\boldsymbol{\Gamma}]_{\bar{\boldsymbol{b}}^c} + [\boldsymbol{\Gamma}]_{\bar{\boldsymbol{b}}}\big\|_{1,a} + \big\|[\boldsymbol{\Gamma}]_{\bar{\boldsymbol{b}}}\big\|_{1,a} - \big\|[\boldsymbol{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a} - \|\widetilde{\boldsymbol{\beta}}\|_{1,a}\Big) \\
&= \lambda_1\Big(\big\|[\boldsymbol{\Delta}]_{\bar{\boldsymbol{s}}}\big\|_1 - \big\|[\boldsymbol{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1\Big) + \lambda_2\Big(\big\|[\boldsymbol{\Gamma}]_{\bar{\boldsymbol{b}}}\big\|_{1,a} - \big\|[\boldsymbol{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a}\Big)
\end{aligned} \tag{24}$$

where (i) and (ii) hold by the decomposability and triangular inequality of norms, respectively.

We also compute the upper bound for term $|\langle\nabla\mathcal{L}(\bar{\boldsymbol{\theta}})\,,\, \boldsymbol{\Lambda}\rangle|$:

$$\begin{aligned}
-\Big\langle \nabla\mathcal{L}(\bar{\boldsymbol{\theta}})\,,\, \boldsymbol{\Lambda}\Big\rangle &= -\Big\langle \nabla\mathcal{L}(\boldsymbol{\theta}^*)\,,\, \boldsymbol{\Lambda}\Big\rangle \overset{(i)}{=} -\langle\nabla\mathcal{L}(\boldsymbol{\theta}^*), \boldsymbol{\Delta}\rangle - \langle\nabla\mathcal{L}(\boldsymbol{\theta}^*), \boldsymbol{\Gamma}\rangle \\
&\overset{(ii)}{\leq} \|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty\,\|\boldsymbol{\Delta}\|_1 + \|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_{\infty,a^*}\,\|\boldsymbol{\Gamma}\|_{1,a}
\end{aligned} \tag{25}$$

since (i) holds from the fact $\boldsymbol{\Lambda} = \boldsymbol{\Delta} + \boldsymbol{\Gamma}$ by definition, and (ii) does by the two standard Hölder's inequalities.

We now combine (12) with (24) and (25), and obtain

$$\begin{aligned}
&\kappa_1\|\mathbf{\Lambda}\|_2^2 - \tau_1\|\mathbf{\Lambda}\|_{\boldsymbol{\eta}}^2 \leq \left\langle \nabla\mathcal{L}(\bar{\boldsymbol{\theta}}+\mathbf{\Lambda}) - \nabla\mathcal{L}(\bar{\boldsymbol{\theta}}), \mathbf{\Lambda}\right\rangle \\
&\leq \|\nabla\mathcal{L}(\bar{\boldsymbol{\theta}})\|_\infty \|\mathbf{\Delta}\|_1 + \|\nabla\mathcal{L}(\bar{\boldsymbol{\theta}})\|_{\infty,a^*}\|\mathbf{\Gamma}\|_{1,a} \\
&\quad + \lambda_1\Big(\big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}}\big\|_1 - \big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1\Big) + \lambda_2\Big(\big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}}\big\|_{1,a} - \big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a}\Big).
\end{aligned}$$

Moreover, since $\|\mathbf{\Lambda}\|_{\boldsymbol{\eta}} \leq \eta_1\|\mathbf{\Delta}\|_1 + \eta_2\|\mathbf{\Gamma}\|_{1,a}$ by definition of dirty regularizer $\|\cdot\|_{\boldsymbol{\eta}}$, the above inequality can be rearranged as

$$\begin{aligned}
\kappa_1\|\mathbf{\Lambda}\|_2^2 &\leq \Big(\|\nabla\mathcal{L}(\bar{\boldsymbol{\theta}})\|_\infty + 2\tau_1\eta_1 r\Big)\|\mathbf{\Delta}\|_1 + \Big(\|\nabla\mathcal{L}(\bar{\boldsymbol{\theta}})\|_{\infty,a^*} + 2\tau_1\eta_2 r\Big)\|\mathbf{\Gamma}\|_{1,a} \\
&\quad + \lambda_1\Big(\big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}}\big\|_1 - \big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1\Big) + \lambda_2\Big(\big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}}\big\|_{1,a} - \big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a}\Big) \\
&\leq \frac{3\lambda_1}{2}\big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}}\big\|_1 + \frac{3\lambda_2}{2}\big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}}\big\|_{1,a}.
\end{aligned}$$

Note that in the last inequality, we utilize the assumption on the setting λ_1 and λ_2 in the statement (that is, $4\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty \leq \lambda_1$ and $8\tau_1\eta_1 r \leq \lambda_1$ for $\mathbf{\Delta}$ case) and the fact that $\mathbf{\Delta} = [\mathbf{\Delta}]_{\bar{\boldsymbol{s}}} + [\mathbf{\Delta}]_{\bar{\boldsymbol{s}}^c}$. Also note that we dropped minus terms at the end.

Now, in order to relate the ℓ_1 and ℓ_2 norms of projected error vector, $[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}}$, we need to consider the sparsity level of $\bar{\boldsymbol{\alpha}}$. During the construction of $\bar{\boldsymbol{\alpha}}$, the sparsity level of $\bar{\boldsymbol{\alpha}}$ might be possibly greater than that of $\boldsymbol{\alpha}^*$. However, by (11) of Proposition 1, we have $\mathbf{\Delta}_{\boldsymbol{s}^*} = \mathbf{\Delta}_{\bar{\boldsymbol{s}}}$ as well as $\mathbf{\Gamma}_{\boldsymbol{b}^*} = \mathbf{\Gamma}_{\bar{\boldsymbol{b}}}$. Hence,

$$\begin{aligned}
\kappa_1\|\mathbf{\Lambda}\|_2^2 &\leq \frac{3\lambda_1}{2}\sqrt{s}\big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}}\big\|_2 + \frac{3\lambda_2}{2}\sqrt{s_{\mathcal{G}}}\big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}}\big\|_2 \\
&\leq \frac{3\lambda_1}{2}\sqrt{s}\|\mathbf{\Delta}\|_2 + \frac{3\lambda_2}{2}\sqrt{s_{\mathcal{G}}}\|\mathbf{\Gamma}\|_2 \leq \frac{3\bar{\lambda}}{2}\Big(\|\mathbf{\Delta}\|_2 + \|\mathbf{\Gamma}\|_2\Big) \qquad (26)
\end{aligned}$$

where $\bar{\lambda} := \max\{\lambda_1\sqrt{s}\,,\, \lambda_2\sqrt{s_{\mathcal{G}}}\}$.

Combining (26) with the result in (10) of Proposition 1, we have

$$\frac{\kappa_1}{2}\Big(\|\mathbf{\Delta}\|_2 + \|\mathbf{\Gamma}\|_2\Big)^2 \leq \kappa_1\Big(\|\mathbf{\Delta}\|_2^2 + \|\mathbf{\Gamma}\|_2^2\Big) \leq \kappa_1\|\mathbf{\Lambda}\|_2^2 \leq \frac{3\bar{\lambda}}{2}\Big(\|\mathbf{\Delta}\|_2 + \|\mathbf{\Gamma}\|_2\Big),$$

and finally the upper bound of ℓ_2 error can be computed as stated:

$$\|\mathbf{\Lambda}\|_2 \leq \|\mathbf{\Delta}\|_2 + \|\mathbf{\Gamma}\|_2 \leq \frac{3}{\kappa_1}\bar{\lambda},$$

which completes the proof.

B.2. Proof of Corollary 1

We first show the loss function of (3) satisfy RSC condition in the following proposition.

Proposition 2. *Consider a design matrix $X \in \mathbb{R}^{n\times p}$ whose rows are independently sampled from $N(0,\Sigma)$. Then, with probability at least $1 - c_1\exp(-c_2 n)$ for some constants c_1 and c_2,*

$$\frac{1}{n}\|X\boldsymbol{\theta}\|_2^2 \geq \kappa_1'\|\boldsymbol{\theta}\|_2^2 - \tau_1'\|\boldsymbol{\theta}\|_{\boldsymbol{\eta}'}^2 \qquad \textit{for any } \boldsymbol{\theta} \in \mathbb{R}^p. \qquad (27)$$

where $\eta_1' = \sqrt{\frac{\log p}{n}}$ and $\eta_2' = \frac{\mathbb{E}(\|\varepsilon\|_{\infty,a^*})}{\sqrt{n}}$ letting a standard normal vector ε for $\boldsymbol{\eta}'$, and κ_1' and τ_1' are some constants depending only on Σ.

As discussed in (Negahban et al., 2012), for the case of $a = 2$, $\frac{\mathbb{E}(\|\varepsilon\|_{\infty,2})}{\sqrt{n}} \leq \left\{\sqrt{\frac{m}{n}} + \sqrt{\frac{3\log q}{n}}\right\}$ by the standard tail bound of χ^2 variables.

The next step in order to appeal to Theorem 1, is to set the proper regularization parameters. Under the conditions in the statement, (Negahban et al., 2012) show that

$$4\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty = 4\Big\|\frac{1}{n}X^\top w\Big\|_\infty \leq 8\sigma\sqrt{\frac{\log p}{n}}$$

with probability at least $1 - c_1\exp(-c_2 n\lambda_1^2)$. Moreover,

$$4\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_{\infty,a^*} = 4\max_{t=1,2,\ldots,q}\Big\|\frac{1}{n}X_{G_t}^\top w\Big\|_{a^*} \leq 8\sigma\left\{\frac{m^{1-1/a}}{\sqrt{n}} + \sqrt{\frac{\log q}{n}}\right\}$$

with probability at least $1 - 2\exp(-2\log q)$.

Since $\|\boldsymbol{\theta}\|_{\boldsymbol{\eta}'}^2 \leq \frac{3}{64}\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}^2$ for the values of $\boldsymbol{\eta}'$ and $\boldsymbol{\lambda}$ specified, (12) holds with $\kappa_1 = \kappa_1'$ and $\tau_1 = \frac{3}{64}\tau_1'$ by (27). In addition, (13) holds with $\kappa_2 = \kappa_1$ and $\tau_2 = \frac{3r}{32}\tau_1'$, by Lemma 8.

Therefore, combining all pieces, for a constant $r := \min\left\{\sqrt{\frac{8\kappa_1}{3\tau_1'}}, \frac{\kappa_1}{5}, \frac{8}{3\tau_1'}\right\}$, we can guarantee that

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \frac{24\sigma}{\kappa_1}\max\left\{\sqrt{\frac{s\log p}{n}}, \sqrt{\frac{s_{\mathcal{G}}m}{n}} + \sqrt{\frac{s_{\mathcal{G}}\log q}{n}}\right\}.$$

B.3. Proof of Proposition 2

The proof of Proposition 2 is the simple extension of proofs given by (Raskutti et al., 2010; Negahban et al., 2012): with probability greater than $1 - c_1\exp(\boldsymbol{\lambda}_2 n)$ for some constants c_1 and c_2,

$$\frac{\|X\boldsymbol{\theta}\|_2}{\sqrt{n}} \geq \frac{\lambda_{\min}(\Sigma^{1/2})}{4}\|\boldsymbol{\theta}\|_2 - 3\frac{\mathbb{E}\big(\|w\|_{\boldsymbol{\eta}}^*\big)}{\sqrt{n}}\|\boldsymbol{\theta}\|_{\boldsymbol{\eta}} \qquad \text{for all } \boldsymbol{\theta}\in\mathbb{R}^p \tag{28}$$

where $w \sim N(0,\Sigma)$, $\lambda_{\min}(\Sigma^{1/2})$ is the minimum eigenvalue of $\Sigma^{1/2}$. Here we use $\|\cdot\|_{\boldsymbol{\eta}}$ and its dual norm to arrive at the statement rather than ℓ_1 (and ℓ_∞) as in (Raskutti et al., 2010) or $\|\cdot\|_{1,a}$ (and its dual) in (Negahban et al., 2012). In particular, (Raskutti et al., 2010) obtains the result in terms of ℓ_1 from the fact $\frac{\mathbb{E}(\|w\|_\infty)}{\sqrt{n}} \leq 3(\max_i\Sigma_{ii})\sqrt{\frac{\log p}{n}}$ while (Negahban et al., 2012) does for group lasso case from $\frac{\mathbb{E}(\|w\|_{\infty,a^*})}{\sqrt{n}} \leq \max_{t=1,\ldots,N_{\mathcal{G}}}|||(\Sigma^{1/2})_{\mathcal{G}_t}|||_{a^*}\frac{\mathbb{E}(\|\varepsilon\|_{\infty,a^*})}{\sqrt{n}}$ where ε is a standard normal vector, and $|||A|||_{a^*} := \max_{\|\boldsymbol{\theta}\|_{a^*}=1}\|A\boldsymbol{\theta}\|_{a^*}$.

Recalling $\mathbb{E}\big(\|w\|_{\boldsymbol{\eta}}^*\big) = \mathbb{E}\big(\max\{\frac{\|w\|_\infty}{\eta_1}, \frac{\|w\|_{\infty,a^*}}{\eta_2}\}\big)$,

$$\mathbb{E}\big(\|w\|_{\boldsymbol{\eta}}^*\big) \leq \mathbb{E}\left(\frac{\|w\|_\infty}{\eta_1} + \frac{\|w\|_{\infty,a^*}}{\eta_2}\right) = \mathbb{E}\left(\frac{\|w\|_\infty}{\eta_1}\right) + \mathbb{E}\left(\frac{\|w\|_{\infty,a^*}}{\eta_2}\right) \tag{29}$$

since both $\frac{\|w\|_\infty}{\eta_1}$ and $\frac{\|w\|_{\infty,a^*}}{\eta_2}$ are always greater than or equal to zero. Setting $\eta_1 = \sqrt{\frac{\log p}{n}}$ and $\eta_2 = \frac{\mathbb{E}(\|\varepsilon\|_{\infty,a^*})}{\sqrt{n}}$ for $\boldsymbol{\eta}$, we have

$$\mathbb{E}\big(\|w\|_{\boldsymbol{\eta}}^*\big) \leq \Big(3(\max_i\Sigma_{ii}) + \max_{t=1,\ldots,N_{\mathcal{G}}}|||(\Sigma^{1/2})_{\mathcal{G}_t}|||_{a^*}\Big)\sqrt{n} \tag{30}$$

and we can establish the bound

$$\frac{\|X\boldsymbol{\theta}\|_2}{\sqrt{n}} \geq \frac{\lambda_{\min}(\Sigma^{1/2})}{4}\|\boldsymbol{\theta}\|_2 - 3\Big(3(\max_i\Sigma_{ii}) + \max_{t=1,\ldots,N_{\mathcal{G}}}|||(\Sigma^{1/2})_{\mathcal{G}_t}|||_{a^*}\Big)\|\boldsymbol{\theta}\|_{\boldsymbol{\eta}}\,. \tag{31}$$

Finally, (27) can be obtained with constants $\kappa_1 = \frac{\lambda_{\min}(\Sigma)}{32}$ and $\tau_1 = 9\Big(3(\max_i\Sigma_{ii}) + \max_{t=1,\ldots,N_{\mathcal{G}}}|||(\Sigma^{1/2})_{\mathcal{G}_t}|||_{a^*}\Big)^2$ from the fact that $a \geq c - b$ implies $2(a^2+b^2) \geq (a+b)^2 \geq c^2$ for positive real a, b and c.

B.4. Proof of Corollary 2

As shown in (Loh & Wainwright, 2015), it can be easily shown that the RSC condition (12) holds with $\kappa_1 = (\|\Theta^*\|_2+1)^{-2}$ and $\tau_1 = 0$. Hence by Lemma 7, (13) also holds with $\kappa_2 = \kappa_1$ and $\tau_2 = 0$. Hence, the condition on r in Theorem 1 is reduced to $r \leq \frac{1}{5(\|\Theta^*\|_2+1)^2}$. Moreover, since $\nabla\mathcal{L}(\boldsymbol{\theta}^*) = \widehat{\Sigma} - (\Theta^*)^{-1}$, the choices of regularization parameters satisfy the condition of Theorem 1, and its error bound holds.

C. Proofs for Support Set Recovery Guarantees

Before providing the actual proof of Theorem 2, we briefly review the primal-dual witness (PDW) proof technique (Wainwright, 2009; Jalali et al., 2010; Yang et al., 2015; Loh & Wainwright, 2014) for our setting:

(i) Solve the *restricted* problem

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^U,\, \|\boldsymbol{\theta}\|_{\boldsymbol{\eta}}\leq r}{\text{minimize}} \; \mathcal{L}(\boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta};\boldsymbol{\lambda}) \tag{32}$$

where $U := supp(\boldsymbol{\alpha}^*) \cup supp(\boldsymbol{\beta}^*)$ as defined earlier. We set $\widehat{\boldsymbol{\theta}}_U$ as the local minimum of this problem.

(ii) Define $\widehat{\boldsymbol{z}}_1 \in \left[\nabla\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}\right]_U$ and $q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) := \|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}} - \mathcal{R}(\boldsymbol{\theta};\boldsymbol{\lambda})$. Choose $\widehat{\boldsymbol{z}}_2$ such that

$$\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}}) - \nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}) + \widehat{\boldsymbol{z}} = \mathbf{0} \tag{33}$$

where $\widehat{\boldsymbol{z}} := (\widehat{\boldsymbol{z}}_1, \widehat{\boldsymbol{z}}_2)$, $\widehat{\boldsymbol{\theta}} := (\widehat{\boldsymbol{\theta}}_U, \mathbf{0})$ and $[\nabla\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}]_j = [\nabla\mathcal{R}(\boldsymbol{\theta};\boldsymbol{\lambda})]_j = [\widehat{\boldsymbol{z}}_2]_j$ for $j \notin U$ so that $[\nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})]_j = 0$ at $\widehat{\boldsymbol{\theta}}_j = 0$. Establish strict dual feasibility of $\|\widehat{\boldsymbol{z}}_2\|^*_{\boldsymbol{\lambda}} \leq 1-\delta$ for some $\delta \in (0,1]$.

(iii) Show that all stationary points of (17) are supported on U.

C.1. Proof of Theorem 2

Since $\widehat{\boldsymbol{\theta}}$ has the same sparsity pattern with $\boldsymbol{\theta}^*$, we can begin with very similar analysis as the proof of Theorem 1 by handling $\widehat{\boldsymbol{\theta}}$ as $\boldsymbol{\theta}^*$. Hence, in this proof we re-define $(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}) := \mathcal{T}(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}; \widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}})$ where $(\widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}})$ is the (local) minimizer of non-convex dirty regularizer such that $\widetilde{\boldsymbol{\alpha}} + \widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\theta}}$. Therefore, by construction $\bar{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}$. We also re-use the notation $\boldsymbol{\Lambda}$, $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$ to represent again $\boldsymbol{\Lambda} := \widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}$, $\widetilde{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}$ respectively but for newly defined $\bar{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{\beta}}$.

We begin with the first-order stationary condition of (17):

$$\left\langle \nabla\mathcal{L}(\widetilde{\boldsymbol{\theta}}) + \nabla\mathcal{R}(\widetilde{\boldsymbol{\theta}};\boldsymbol{\lambda}), \boldsymbol{\Lambda} \right\rangle \leq 0\,, \tag{34}$$

which is in the form discussed in (21).

As before, we first show that $\|\boldsymbol{\Lambda}\|_2 \leq 1$ under (RSC). In order to show by contradiction, suppose that $\|\boldsymbol{\Lambda}\|_2 \geq 1$.

Then setting $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (\widehat{\boldsymbol{\theta}}, \boldsymbol{\Lambda})$ in (13), we obtain

$$\kappa_2\|\boldsymbol{\Lambda}\|_2 - \tau_2\|\boldsymbol{\Lambda}\|_{\boldsymbol{\eta}} \leq \left\langle \nabla\mathcal{L}(\widehat{\boldsymbol{\theta}} + \boldsymbol{\Lambda}) - \nabla\mathcal{L}(\widehat{\boldsymbol{\theta}}), \boldsymbol{\Lambda} \right\rangle. \tag{35}$$

Combining (34) and (35) yields

$$\kappa_2\|\boldsymbol{\Lambda}\|_2 - \tau_2\|\boldsymbol{\Lambda}\|_{\boldsymbol{\eta}} \leq \left\langle -\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}}) - \nabla\mathcal{R}(\widetilde{\boldsymbol{\theta}};\boldsymbol{\lambda}), \boldsymbol{\Lambda} \right\rangle. \tag{36}$$

Since we set $\widehat{\boldsymbol{z}}$ to satisfy (33),

$$\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}}) = \nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}) - \widehat{\boldsymbol{z}} = \nabla\|\widehat{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} - \nabla\mathcal{R}(\widehat{\boldsymbol{\theta}};\boldsymbol{\lambda}) - \widehat{\boldsymbol{z}} = -\nabla\mathcal{R}(\widehat{\boldsymbol{\theta}};\boldsymbol{\lambda})\,,$$

and hence by Hölder's inequality and Lemma 4,

$$\begin{aligned}\kappa_2\|\boldsymbol{\Lambda}\|_2 - \tau_2\|\boldsymbol{\Lambda}\|_{\boldsymbol{\eta}} &\leq \left\langle \nabla\mathcal{R}(\widehat{\boldsymbol{\theta}};\boldsymbol{\lambda}) - \nabla\mathcal{R}(\widetilde{\boldsymbol{\theta}};\boldsymbol{\lambda}), \boldsymbol{\Lambda} \right\rangle \\ &\leq \left(\left\|\nabla\mathcal{R}(\widehat{\boldsymbol{\theta}};\boldsymbol{\lambda})\right\|^*_{\boldsymbol{\eta}} + \left\|\nabla\mathcal{R}(\widetilde{\boldsymbol{\theta}};\boldsymbol{\lambda})\right\|^*_{\boldsymbol{\eta}} \right) \|\boldsymbol{\Lambda}\|_{\boldsymbol{\eta}} \leq 2\bar{\eta}\|\boldsymbol{\Lambda}\|_{\boldsymbol{\eta}}\end{aligned}$$

where $\bar{\eta} := \max\{\frac{\lambda_1}{\eta_1}, \frac{\lambda_2}{\eta_2}\}$. Since $\|\mathbf{\Lambda}\|_{\boldsymbol{\lambda}} = \|\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} \leq 2r$ by the constraint of (17),

$$\kappa_2 \|\mathbf{\Lambda}\|_2 \leq (\tau_2 + 2\bar{\eta}) \|\mathbf{\Lambda}\|_{\boldsymbol{\eta}} \leq 2r(\tau_2 + 2\bar{\eta}).$$

Therefore, if $2r(\tau_2 + 2\bar{\eta}) \leq 1$ as assumed, we should have $\|\mathbf{\Lambda}\|_2 \leq 1$.

Now we focus on the RSC condition in (12):

$$\kappa_1 \|\mathbf{\Lambda}\|_2^2 - \tau_1 \|\mathbf{\Lambda}\|_{\boldsymbol{\eta}}^2 \leq \left\langle \nabla \mathcal{L}(\widetilde{\boldsymbol{\theta}}) - \nabla \mathcal{L}(\widehat{\boldsymbol{\theta}}), \mathbf{\Lambda} \right\rangle. \tag{37}$$

Let $\bar{\mathcal{L}}(\boldsymbol{\theta}) := \mathcal{L}(\boldsymbol{\theta}) - q_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$. (Recalling $q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) := \|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}} - \mathcal{R}(\boldsymbol{\theta}; \boldsymbol{\lambda})$, $\bar{\mathcal{L}}(\boldsymbol{\theta})$ actually can be rewritten as $\mathcal{L}(\boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta}; \boldsymbol{\lambda}) - \|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}$). Then, from (37),

$$\kappa_1 \|\mathbf{\Lambda}\|_2^2 - \tau_1 \|\mathbf{\Lambda}\|_{\boldsymbol{\eta}}^2 \leq \left\langle \nabla \bar{\mathcal{L}}(\widetilde{\boldsymbol{\theta}}) - \nabla \bar{\mathcal{L}}(\widehat{\boldsymbol{\theta}}), \mathbf{\Lambda} \right\rangle + \left\langle \nabla q_{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\theta}}) - \nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}), \mathbf{\Lambda} \right\rangle. \tag{38}$$

In order to upper-bound the second term in the RHS of (38), we apply the mean-value theorem:

$$\nabla q_{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\theta}}) - \nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}) = \nabla^2 q_{\boldsymbol{\lambda}}(\underline{\boldsymbol{\theta}})(\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}})$$

where $\underline{\boldsymbol{\theta}}$ is a parameter vector on the line between $\widetilde{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}$. By the Cauchy-Schwarz inequality

$$\left| \left\langle \nabla q_{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\theta}}) - \nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}), \mathbf{\Lambda} \right\rangle \right| \leq \left\| \nabla^2 q_{\boldsymbol{\lambda}}(\underline{\boldsymbol{\theta}}) \right\|_2 \left\| \widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}} \right\|_2^2 \tag{39}$$

where $\|\cdot\|_2$ denotes the spectral norm of the matrix.

Combining (38) and (39) with the assumption $\left\| \nabla^2 q_{\boldsymbol{\lambda}}(\underline{\boldsymbol{\theta}}) \right\|_2 \leq \mu$ (which holds for non-convex regularizers considered in this paper, as shown in Lemma 5) yields

$$(\kappa_1 - \mu) \|\mathbf{\Lambda}\|_2^2 - \tau_1 \|\mathbf{\Lambda}\|_{\boldsymbol{\eta}}^2 \leq \left\langle \nabla \bar{\mathcal{L}}(\widetilde{\boldsymbol{\theta}}) - \nabla \bar{\mathcal{L}}(\widehat{\boldsymbol{\theta}}), \mathbf{\Lambda} \right\rangle. \tag{40}$$

At the same time, we represent the stationary condition (34) using the notation $q_{\boldsymbol{\lambda}}(\widetilde{\boldsymbol{\theta}})$:

$$0 \leq \left\langle \nabla \bar{\mathcal{L}}(\widetilde{\boldsymbol{\theta}}), \widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \right\rangle + \left\langle \widetilde{\boldsymbol{z}}, \widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \right\rangle \tag{41}$$

where $\widetilde{\boldsymbol{z}} \in \partial \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}}$. From our construction of $\widehat{\boldsymbol{z}}$ in (33), we have $\left\langle \nabla \bar{\mathcal{L}}(\widehat{\boldsymbol{\theta}}) + \widehat{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}} \right\rangle = 0$, and when combined with (41) this implies

$$\begin{aligned} 0 &\leq \left\langle \nabla \bar{\mathcal{L}}(\widehat{\boldsymbol{\theta}}) - \nabla \bar{\mathcal{L}}(\widetilde{\boldsymbol{\theta}}), \widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}} \right\rangle - \left\langle \widehat{\boldsymbol{z}}, \widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}} \right\rangle - \left\langle \widetilde{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}} \right\rangle \\ &= \left\langle \nabla \bar{\mathcal{L}}(\widehat{\boldsymbol{\theta}}) - \nabla \bar{\mathcal{L}}(\widetilde{\boldsymbol{\theta}}), \widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}} \right\rangle + \left\langle \widehat{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}} \right\rangle - \|\widehat{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} + \left\langle \widetilde{\boldsymbol{z}}, \widehat{\boldsymbol{\theta}} \right\rangle - \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}}. \end{aligned} \tag{42}$$

With (40), this inequality implies

$$\begin{aligned} (\kappa_1 - \mu) \|\mathbf{\Lambda}\|_2^2 - \tau_1 \|\mathbf{\Lambda}\|_{\boldsymbol{\eta}}^2 &\leq \left\langle \widehat{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}} \right\rangle - \|\widehat{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} + \left\langle \widetilde{\boldsymbol{z}}, \widehat{\boldsymbol{\theta}} \right\rangle - \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} \\ &\leq \left\langle \widehat{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}} \right\rangle - \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} \end{aligned} \tag{43}$$

where we use the fact $\left\langle \widetilde{\boldsymbol{z}}, \widehat{\boldsymbol{\theta}} \right\rangle \leq \|\widetilde{\boldsymbol{z}}\|_{\boldsymbol{\lambda}}^* \|\widehat{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} \leq \|\widehat{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}}$ by Lemma 3.

Assume for now that

$$\|\mathbf{\Lambda}\|_{\boldsymbol{\eta}} \leq \max\{\lambda_1 \sqrt{s}, \lambda_2 \sqrt{s_{\mathcal{G}}}\} \max\left\{ \frac{\eta_1}{\lambda_1}, \frac{\eta_2}{\lambda_2} \right\} \sqrt{2}\Big(\frac{4}{\delta} + 2\Big) \|\mathbf{\Lambda}\|_2 \tag{44}$$

for some $\delta \in (0, 1]$. Then, defining $C := \max\{\lambda_1 \sqrt{s}, \lambda_2 \sqrt{s_{\mathcal{G}}}\} \max\left\{ \frac{\eta_1}{\lambda_1}, \frac{\eta_2}{\lambda_2} \right\} \sqrt{2}\Big(\frac{4}{\delta} + 2\Big)$ so that $\|\mathbf{\Lambda}\|_{\boldsymbol{\eta}} \leq C \|\mathbf{\Lambda}\|_2$, we have

$$(\kappa_1 - \mu) \|\mathbf{\Lambda}\|_2^2 - \tau_1 \|\mathbf{\Lambda}\|_{\boldsymbol{\eta}}^2 \geq \kappa_1 \|\mathbf{\Lambda}\|_2^2 - \tau_1 C^2 \|\mathbf{\Lambda}\|_2^2. \tag{45}$$

Therefore, as long as $\kappa_1 > \tau_1 C^2$,

$$0 \leq (\kappa_1 - \mu)\|\mathbf{\Lambda}\|_2^2 - \tau_1\|\mathbf{\Lambda}\|_{\boldsymbol{\eta}}^2 \leq \langle \widehat{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}} \rangle - \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} \tag{46}$$

implying $\|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} \leq \langle \widehat{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}} \rangle$. Actually, since $\langle \widehat{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}} \rangle \leq \|\widehat{\boldsymbol{z}}\|_{\boldsymbol{\lambda}}^* \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} \leq \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}}$ by Hölder's inequality and Lemma 3, it should hold that $\|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} = \langle \widehat{\boldsymbol{z}}, \widetilde{\boldsymbol{\theta}} \rangle$. As discussed in previous works (Wainwright, 2009; Jalali et al., 2010; Yang et al., 2015; Loh & Wainwright, 2014) using PDW approach, this equality guarantees that for all $j \notin U$, $\widetilde{\boldsymbol{\theta}}_j = 0$ under the strict dual feasibility.

The remaining procedure is to show (44), which is proved in the following lemma.

Lemma 2. *Suppose* $\|\widehat{\boldsymbol{z}}_{U^c}\|_{\boldsymbol{\lambda}}^* \leq 1 - \delta$ *for some* $\delta \in (0, 1]$, *and* $2r\tau_1\eta_1 \leq \frac{\delta}{2}\lambda_1$ *and* $2r\tau_1\eta_2 \leq \frac{\delta}{2}\lambda_2$. *Then*

$$\|\mathbf{\Lambda}\|_{\boldsymbol{\eta}} \leq \max\{\lambda_1\sqrt{s}\,,\, \lambda_2\sqrt{s_{\mathcal{G}}}\} \max\left\{\frac{\eta_1}{\lambda_1}, \frac{\eta_2}{\lambda_2}\right\} \sqrt{2}\Big(\frac{4}{\delta} + 2\Big)\|\mathbf{\Lambda}\|_2\,. \tag{47}$$

From the first inequality of (42), the following inequality can be easily derived in a similar way of constructing (43):

$$\big(\kappa_1 - \mu\big)\|\mathbf{\Lambda}\|_2^2 - \tau_1\|\mathbf{\Lambda}\|_{\boldsymbol{\eta}}^2 \leq \underbrace{\langle \widetilde{\boldsymbol{z}}, \widehat{\boldsymbol{\theta}} \rangle - \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}}}_{\text{(I)}} + \underbrace{\langle \widehat{\boldsymbol{z}}, \mathbf{\Lambda} \rangle}_{\text{(II)}}\,. \tag{48}$$

By the same reasoning in (24),

$$\text{(I)} \leq \|\widehat{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} - \|\widetilde{\boldsymbol{\theta}}\|_{\boldsymbol{\lambda}} \leq \lambda_1\Big(\big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}}\big\|_1 - \big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1\Big) + \lambda_2\Big(\big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}}\big\|_{1,a} - \big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a}\Big)\,. \tag{49}$$

The second term in (48) can be upper bounded as follows: if $\|\widehat{\boldsymbol{z}}_2\|_{\boldsymbol{\lambda}}^* \leq 1 - \delta$ as assumed,

$$\begin{aligned}
\text{(II)} &= \langle \widehat{\boldsymbol{z}}_1, \mathbf{\Lambda}_U \rangle + \langle \widehat{\boldsymbol{z}}_2, \mathbf{\Lambda}_{U^c} \rangle \leq \|\widehat{\boldsymbol{z}}_1\|_{\boldsymbol{\lambda}}^* \|\mathbf{\Lambda}_U\|_{\boldsymbol{\lambda}} + \|\widehat{\boldsymbol{z}}_2\|_{\boldsymbol{\lambda}}^* \|\mathbf{\Lambda}_{U^c}\|_{\boldsymbol{\lambda}} \\
&\leq \|\mathbf{\Lambda}_U\|_{\boldsymbol{\lambda}} + (1-\delta)\|\mathbf{\Lambda}_{U^c}\|_{\boldsymbol{\lambda}} \\
&\leq \lambda_1\Big(\big\|[\mathbf{\Delta}]_U\big\|_1 + (1-\delta)\big\|[\mathbf{\Delta}]_{U^c}\big\|_1\Big) + \lambda_2\Big(\big\|[\mathbf{\Gamma}]_U\big\|_{1,a} + (1-\delta)\big\|[\mathbf{\Gamma}]_{U^c}\big\|_{1,a}\Big) \\
&= \lambda_1\Big(\big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}}\big\|_1 + (1-\delta)\big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1\Big) + \lambda_2\Big(\big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}}\big\|_{1,a} + (1-\delta)\big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a}\Big)
\end{aligned} \tag{50}$$

where the last equality follows the result in Proposition 1.

By (49) and (50),

$$\begin{aligned}
&-\tau_1\|\mathbf{\Lambda}\|_{\boldsymbol{\eta}}^2 \leq \big(\kappa_1 - \mu\big)\|\mathbf{\Lambda}\|_2^2 - \tau_1\|\mathbf{\Lambda}\|_{\boldsymbol{\eta}}^2 \\
&\leq \lambda_1\Big(2\big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}}\big\|_1 - \delta\big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1\Big) + \lambda_2\Big(2\big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}}\big\|_{1,a} - \delta\big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a}\Big)\,.
\end{aligned} \tag{51}$$

Furthermore,

$$\tau_1\|\mathbf{\Lambda}\|_{\boldsymbol{\eta}}^2 \leq 2r\tau_1\|\mathbf{\Lambda}\|_{\boldsymbol{\eta}} \leq 2r\tau_1\Big(\eta_1\|\mathbf{\Delta}\|_1 + \eta_2\|\mathbf{\Gamma}\|_{1,a}\Big)\,. \tag{52}$$

Hence, if $2r\tau_1\eta_1 \leq \frac{\delta}{2}\lambda_1$ and $2r\tau_1\eta_2 \leq \frac{\delta}{2}\lambda_2$ as stated, combining 51 and 52 establishes

$$\begin{aligned}
&-\frac{\delta}{2}\Big(\lambda_1\|\mathbf{\Delta}\|_1 + \lambda_2\|\mathbf{\Gamma}\|_{1,a}\Big) \leq -2r\tau_1\Big(\eta_1\|\mathbf{\Delta}\|_1 + \eta_2\|\mathbf{\Gamma}\|_{1,a}\Big) \\
&\leq \lambda_1\Big(2\big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}}\big\|_1 - \delta\big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1\Big) + \lambda_2\Big(2\big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}}\big\|_{1,a} - \delta\big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a}\Big)
\end{aligned}$$

or equivalently

$$\frac{\delta}{2}\Big(\lambda_1\big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}^c}\big\|_1 + \lambda_2\big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}^c}\big\|_{1,a}\Big) \leq \Big(2 + \frac{\delta}{2}\Big)\Big(\lambda_1\big\|[\mathbf{\Delta}]_{\bar{\boldsymbol{s}}}\big\|_1 + \lambda_2\big\|[\mathbf{\Gamma}]_{\bar{\boldsymbol{b}}}\big\|_{1,a}\Big)\,.$$

Finally, we can have

$$\begin{aligned}
&\|\mathbf{\Lambda}\|_{\boldsymbol{\eta}} \leq \eta_1\|\mathbf{\Delta}\|_1 + \eta_2\|\mathbf{\Gamma}\|_{1,a} \leq \bar{\eta}'\Big(\lambda_1\|\mathbf{\Delta}\|_1 + \lambda_2\|\mathbf{\Gamma}\|_{1,a}\Big) \\
&= \bar{\eta}'\Big(\lambda_1\big\|[\mathbf{\Delta}]_{\bar{\mathbf{s}}}\big\|_1 + \lambda_1\big\|[\mathbf{\Delta}]_{\bar{\mathbf{s}}^c}\big\|_1 + \lambda_2\big\|[\mathbf{\Gamma}]_{\bar{\mathbf{b}}}\big\|_{1,a} + \lambda_2\big\|[\mathbf{\Gamma}]_{\bar{\mathbf{b}}^c}\big\|_{1,a}\Big) \\
&\leq \bar{\eta}'\Big(\frac{4}{\delta}+2\Big)\Big(\lambda_1\big\|[\mathbf{\Delta}]_{\bar{\mathbf{s}}}\big\|_1 + \lambda_2\big\|[\mathbf{\Gamma}]_{\bar{\mathbf{b}}}\big\|_{1,a}\Big) \leq \bar{\eta}'\bar{\lambda}\Big(\frac{4}{\delta}+2\Big)\Big(\|\mathbf{\Delta}\|_2 + \|\mathbf{\Gamma}\|_2\Big) \\
&\leq \sqrt{2}\bar{\eta}'\bar{\lambda}\Big(\frac{4}{\delta}+2\Big)\|\mathbf{\Lambda}\|_2
\end{aligned} \tag{53}$$

where $\bar{\eta}' := \max\{\frac{\eta_1}{\lambda_1}, \frac{\eta_2}{\lambda_2}\}$, $\bar{\lambda} := \max\{\lambda_1\sqrt{s}\,, \lambda_2\sqrt{s_{\mathcal{G}}}\}$, and we have $\|\mathbf{\Delta}\|_2 + \|\mathbf{\Gamma}\|_2 \leq \sqrt{2}\|\mathbf{\Lambda}\|_2$ (since $\frac{1}{2}(\|\mathbf{\Delta}\|_2 + \|\mathbf{\Gamma}\|_2)^2 \leq \|\mathbf{\Delta}\|_2^2 + \|\mathbf{\Gamma}\|_2^2 \leq \|\mathbf{\Lambda}\|_2^2$) by Proposition 1 in the last inequality.

C.2. Proof of Corollary 3

The statement can be shown to hold by combining the result of Theorem 2 and lemmas in (Loh & Wainwright, 2014)

(a) Uniqueness. The proof of Lemma 2 in (Loh & Wainwright, 2014) shows that under (12)

$$v^\top\nabla^2\mathcal{L}(\boldsymbol{\theta})v \geq \kappa_1\|v\|_2^2 - \tau_1\|v\|_{\boldsymbol{\eta}}^2, \quad \forall v \in \{v \in \mathbb{R}^p \,|\, \text{supp}(v) \subseteq U,\, \|v\|_2 = 1\}\,.$$

By definition of $\|\cdot\|_{\boldsymbol{\lambda}}$, for any $v \in \mathbb{R}^U$,

$$\|v\|_{\boldsymbol{\eta}} \leq \eta_1\|v_S\|_1 + \eta_2\|v_B\|_{1,a} \leq \eta_1\sqrt{s}\|v_S\|_2 + \eta_2\sqrt{s_{\mathcal{G}}}\|v_B\|_2 \leq \max\{\eta_1\sqrt{s}, \eta_2\sqrt{s_{\mathcal{G}}}\}\|v\|_2\,.$$

Hence,

$$v^\top\nabla^2\mathcal{L}(\boldsymbol{\theta})v \geq \kappa_1\|v\|_2^2 - \tau_1\big(\max\{\eta_1\sqrt{s}, \eta_2\sqrt{s_{\mathcal{G}}}\}\big)^2\|v\|_2^2, \quad \forall v \in \{v \in \mathbb{R}^p \,|\, \text{supp}(v) \subseteq U,\, \|v\|_2 = 1\}\,.$$

which implies

$$\Big(\nabla^2\mathcal{L}(\boldsymbol{\theta})\Big)_{UU} \succeq \Big(\kappa - \tau_1\big(\max\{\eta_1\sqrt{s}, \eta_2\sqrt{s_{\mathcal{G}}}\}\big)^2\Big)I$$

for all $\boldsymbol{\theta} \in \mathbb{R}^U$. Therefore if $\frac{\kappa_1-\mu}{2} \geq \tau_1\big(\max\{\eta_1\sqrt{s}, \eta_2\sqrt{s_{\mathcal{G}}}\}\big)^2$, it is guaranteed that

$$\Big(\nabla^2\mathcal{L}(\boldsymbol{\theta}) - \frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2\Big)_{UU} \succeq \Big(\kappa - \mu - \tau_1\big(\max\{\eta_1\sqrt{s}, \eta_2\sqrt{s_{\mathcal{G}}}\}\big)^2\Big)I \succeq \tau_1\big(\max\{\eta_1\sqrt{s}, \eta_2\sqrt{s_{\mathcal{G}}}\}\big)^2 I$$

and hence $(\mathcal{L}(\boldsymbol{\theta}) - \frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2) + (\frac{\mu}{2}\|\boldsymbol{\theta}\|_2^2 + \mathcal{R}(\boldsymbol{\theta};\boldsymbol{\lambda})$ is strictly convex over $\mathbb{R}^U$ as the sum of strictly convex and convex function. By Theorem 2, any stationary point $\widetilde{\boldsymbol{\theta}}$ of the program (17) should be in the form of $(\widetilde{\boldsymbol{\theta}}_U, \mathbf{0}_{U^c})$. Moreover, the restricted program is strictly convex as just shown, $\widetilde{\boldsymbol{\theta}}_U$ is unique.

(b) ℓ_∞ error bound. By the definition of $\widehat{Q}$, $\widehat{Q}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \nabla\mathcal{L}(\widehat{\boldsymbol{\theta}}) - \nabla\mathcal{L}(\boldsymbol{\theta}^*)$. As shown in the proof of Theorem 2 of (Loh & Wainwright, 2014),

$$\begin{aligned}
\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty &\leq \Big\|\big(\widehat{Q}_{UU}\big)^{-1}\big(\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U - \nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})_U + \widehat{\boldsymbol{z}}_1\big)\Big\|_\infty \\
&\leq \Big\|\big(\widehat{Q}_{UU}\big)^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U\Big\|_\infty + \min\{\lambda_1, \lambda_2\}\Big\|\big(\widehat{Q}_{UU}\big)^{-1}\Big\|_\infty
\end{aligned} \tag{54}$$

where we use Lemma 6 so that $\|\nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})_U - \widehat{\boldsymbol{z}}_1\|_{\boldsymbol{\lambda}}^* = \max\big\{\|\nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})_U - \widehat{\boldsymbol{z}}_1\|_\infty/\lambda_1, \|\nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})_U - \widehat{\boldsymbol{z}}_1\|_{\infty,a^*}/\lambda_2\big\} \leq 1$ implying $\|\nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})_U - \widehat{\boldsymbol{z}}_1\|_\infty \leq \min\{\lambda_1, \lambda_2\}$.

(c) ℓ_∞ error bound for (μ,γ)-amenable regularizers. By the assumption on $\boldsymbol{\theta}^*_{\min}$, it is guaranteed $|\widehat{\boldsymbol{\theta}}_j| \geq |\boldsymbol{\theta}^*_{\min}| - \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \geq 2\max\{\lambda_1,\lambda_2\}\gamma$ for all $j \in U$. Then, at least either $\widehat{\boldsymbol{\alpha}}_j$ or $\widehat{\boldsymbol{\beta}}_j$ is larger than $\gamma\max\{\lambda_1,\lambda_2\}$, and, as a result,

$$\nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})_U = \widehat{\boldsymbol{z}}_1 \tag{55}$$

and (54) reduces to $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq \left\|\left(\widehat{Q}_{UU}\right)^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U\right\|_\infty$.

In order to show the dual feasibility, we combine the zero-subgradient condition (33) and the definition of$\widehat{Q}$ as described in (Loh & Wainwright, 2014):

$$\begin{bmatrix}\widehat{Q}_{UU} & \widehat{Q}_{UU^c}\\ \widehat{Q}_{U^cU} & \widehat{Q}_{U^cU^c}\end{bmatrix}\begin{bmatrix}\widehat{\boldsymbol{\theta}}_U - \boldsymbol{\theta}^*_U\\ \mathbf{0}\end{bmatrix} + \begin{bmatrix}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U - \nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})_U\\ \nabla\mathcal{L}(\boldsymbol{\theta}^*)_{U^c} - \nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})_{U^c}\end{bmatrix} + \begin{bmatrix}\widehat{\boldsymbol{z}}_1\\ \widehat{\boldsymbol{z}}_2\end{bmatrix} = \mathbf{0}\,,$$

and rearranging it for $\widehat{\boldsymbol{z}}_2$ with (55) and the fact $\nabla q_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}})_{U^c} = 0$, yields

$$\widehat{\boldsymbol{z}}_2 = -\nabla\mathcal{L}(\boldsymbol{\theta}^*)_{U^c} + \widehat{Q}_{U^cU}(\widehat{Q}_{UU})^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U \tag{56}$$

Therefore, if $\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)_{U^c}\|^*_{\boldsymbol{\lambda}} \leq \frac{1-\delta}{2}$ and $\|\widehat{Q}_{U^cU}(\widehat{Q}_{UU})^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U\|^*_{\boldsymbol{\lambda}} \leq \frac{1-\delta}{2}$ additionally hold, the strict dual feasibility holds:

$$\|\widehat{\boldsymbol{z}}_2\|^*_{\boldsymbol{\lambda}} \leq \left\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)_{U^c}\right\|^*_{\boldsymbol{\lambda}} + \left\|\widehat{Q}_{U^cU}(\widehat{Q}_{UU})^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U\right\|^*_{\boldsymbol{\lambda}} \leq \frac{1-\delta}{2} + \frac{1-\delta}{2} \leq 1-\delta\,.$$

C.3. Proof of Corollary 4

For the problem, if $(j,k) \in U^c$, then $[\widehat{\boldsymbol{z}}_2]^{(k)}_j$ in (56), can be written as

$$\frac{1}{n}\left\langle X^{(k)}_j,\, I - \frac{1}{n}X^{(k)}_{U_k}\left(\frac{1}{n}\langle X^{(k)}_{U_k}, X^{(k)}_{U_k}\rangle\right)^{-1}\left(X^{(k)}_{U_k}\right)^\top\right\rangle w^{(k)}\,. \tag{57}$$

In order to show the strict dual feasibility $\|\widehat{\boldsymbol{z}}_2\|^*_{\boldsymbol{\lambda}} \leq 1-\delta$, we need to show both $\max_{(j,k)\in U^c}\left|[\widehat{\boldsymbol{z}}_2]^{(k)}_j\right| \leq (1-\delta)\lambda_1$ and $\max_j\sum_{(j,k)\in U^c}\left|[\widehat{\boldsymbol{z}}_2]^{(k)}_j\right| \leq (1-\delta)\lambda_2$.

Setting $t = 4\sigma\sqrt{\frac{\log(pm)}{n}}$ for (61) of Lemma 9 yields $\max_{(j,k)\in U^c}\left|[\widehat{\boldsymbol{z}}_2]^{(k)}_j\right| \leq 4\sigma\sqrt{\frac{\log(pm)}{n}}$ with probability at least $1 - 2\exp(-3\log(pm))$. Similarly, setting $t = 4\sigma\sqrt{\frac{\log p + m\log 2}{n}}$ for (62) yields $\max_j\sum_{(j,k)\in U^c}\left|[\widehat{\boldsymbol{z}}_2]^{(k)}_j\right| \leq 4\sigma\sqrt{\frac{\log p + m\log 2}{n}}$ with probability at least $1 - 2\exp\big(-3(\log p + m\log 2)\big)$. In addition, by the similar reasoning in the proof of Proposition 2, we can easily show that the RSC condition holds w.h.p. for $\boldsymbol{\eta}$ such that $\max\{\lambda_1/\eta_1, \lambda_2/\eta_2\}$ is some constant depending on σ (the only difference is $a = \infty$ in this example, but $\mathbb{E}(\|\varepsilon\|_\infty)$ for a standard normal vector $\varepsilon \sim N(0, I_{p\times p})$ scales as the same rates of λ_2 by (62) of Lemma 9.).

Therefore, the strict dual feasibility holds for the selection of parameters, and the support set recovery is guaranteed w.h.p.

ℓ_∞ bound is also trivially derived from the combination of (63) of Lemma 9 and Corollary 3 (we can compute the upper bound of $\left\|\left(\widehat{Q}_{UU}\right)^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U\right\|_\infty$ by setting $t = \sigma\sqrt{\frac{100\log(pm)}{nC_{\min}}}$ in (63)).

D. Useful Lemmas for Proofs

Lemma 3. *At any $\boldsymbol{\theta}$, $\|\nabla\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}\|^*_{\boldsymbol{\eta}} \leq \max\left\{\frac{\lambda_1}{\eta_1}, \frac{\lambda_2}{\eta_2}\right\}$.*

Proof. For any fixed $\boldsymbol{\theta}$, let s be the sub-gradient of $\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}$ at $\boldsymbol{\theta}$, $\nabla\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}$, in $\mathbb{R}^p$. By definition of the sub-gradient, s satisfies

$$\|\boldsymbol{\theta} + v\|_{\boldsymbol{\lambda}} \geq \|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}} + \langle s, v\rangle \quad \text{for all } v \in \mathbb{R}^p\,.$$

Taking this inequality the supremum over all v such that $\|v\|_{\boldsymbol{\lambda}} = 1$, we have

$$\sup_{\|v\|_{\boldsymbol{\lambda}}=1} \|\boldsymbol{\theta} + v\|_{\boldsymbol{\lambda}} \geq \|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}} + \sup_{\|v\|_{\boldsymbol{\lambda}}=1} \langle s, v\rangle \,. \tag{58}$$

By rearranging (58),

$$\|s\|_{\boldsymbol{\lambda}}^* \overset{(i)}{:=} \sup_{\|v\|_{\boldsymbol{\lambda}}=1} \langle s, v\rangle \leq \sup_{\|v\|_{\boldsymbol{\lambda}}=1} \|\boldsymbol{\theta} + v\|_{\boldsymbol{\lambda}} - \|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}} \overset{(ii)}{\leq} \sup_{\|v\|_{\boldsymbol{\lambda}}=1} \|v\|_{\boldsymbol{\lambda}}$$

where (i) is the definition of dual norm of $\|\cdot\|_{\boldsymbol{\lambda}}$, and (ii) follows the triangular inequality of the norm $\|\cdot\|_{\boldsymbol{\lambda}}$. Since $\sup_{\|v\|_{\boldsymbol{\lambda}}=1} \|v\|_{\boldsymbol{\lambda}} \leq 1$ by definitions, we obtain $\|s\|_{\boldsymbol{\lambda}}^* = \max\left\{\frac{\|\nabla\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}\|_\infty}{\lambda_1}, \frac{\|\nabla\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}\|_{\infty,a}}{\lambda_2}\right\} \leq 1$, finally implying

$$\|\nabla\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}\|_{\boldsymbol{\eta}}^* \leq \max\left\{\frac{\left\|\nabla\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}\right\|_\infty}{\eta_1}, \frac{\left\|\nabla\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}\right\|_{\infty,a}}{\eta_2}\right\} \leq \max\left\{\frac{\lambda_1}{\eta_1}, \frac{\lambda_2}{\eta_2}\right\}.$$

□

Lemma 4. *At any $\boldsymbol{\theta}$, $\|\nabla\mathcal{R}(\boldsymbol{\theta};\boldsymbol{\lambda})\|_{\boldsymbol{\eta}}^* \leq \max\left\{\frac{\lambda_1}{\eta_1}, \frac{\lambda_2}{\eta_2}\right\}$.*

Proof. At any $\boldsymbol{\theta}$, if we compute the derivative, $\left[\nabla\mathcal{R}(\boldsymbol{\theta};\boldsymbol{\lambda})\right]_j$ is upper bounded by $\max\{\partial\rho_{\lambda_1}([\boldsymbol{\theta}]_j), \partial\phi_{\lambda_2,a}[\boldsymbol{\theta}]_{g_j}\}$ where g_j is the group of indices that j belongs to. As shown in Lemma 8 of (Loh & Wainwright, 2014), $\partial\rho_{\lambda_1}([\boldsymbol{\theta}]_j) \leq \partial\|[\boldsymbol{\theta}]_j\|_1$. Similarly, by definition of ϕ, we have $\partial\phi_{\lambda_2,a}([\boldsymbol{\theta}]_{g_j}) \leq \partial\|[\boldsymbol{\theta}]_{g_j}\|_{1,a}$. Therefore, for every index j, $\left|[\nabla\mathcal{R}(\boldsymbol{\theta};\boldsymbol{\lambda})]_j\right| \leq \left|[\nabla\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}]_j\right|$, and $\left\|\nabla\mathcal{R}(\boldsymbol{\theta};\boldsymbol{\lambda})\right\|_{\boldsymbol{\eta}}^* \leq \left\|\nabla\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}\right\|_{\boldsymbol{\eta}}^*$. The final result comes by Lemma 3. □

Lemma 5. *Consider the non-convex penalty functions in (C1). Then, for any $\boldsymbol{\theta}$,*

$$\left\|\nabla^2 q_{\boldsymbol{\lambda}}(\boldsymbol{\theta})\right\|_2 \leq \mu \,. \tag{59}$$

Proof. For notational simplicity, we define the function $F : \mathbb{R}^p \to \mathbb{R}^p$ as $F(\boldsymbol{\theta};\boldsymbol{\lambda}) := \nabla R(\boldsymbol{\theta};\boldsymbol{\lambda})$, hence the i-th coordinate of F is

$$F_i := \frac{\partial R}{\partial \boldsymbol{\alpha}} \cdot \frac{\partial \boldsymbol{\alpha}}{\partial \theta_i} + \frac{\partial R}{\partial \boldsymbol{\beta}} \cdot \frac{\partial \boldsymbol{\beta}}{\partial \theta_i}$$

where we suppress the dependency on $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ of $F(\boldsymbol{\theta};\boldsymbol{\lambda})$ and $R(\boldsymbol{\theta};\boldsymbol{\lambda})$ for compact presentation. By applying another chain rule to compute ∇F, we immediately obtain

$$\begin{aligned}\frac{\partial F_i}{\partial \theta_j} =& \sum_{k,l} \frac{\partial^2 R}{\partial\alpha_k \partial\alpha_l} \cdot \frac{\partial\alpha_k}{\partial\theta_i}\frac{\partial\alpha_l}{\partial\theta_j} + \sum_k \frac{\partial R}{\partial\alpha_k} \cdot \frac{\partial^2\alpha_k}{\partial\theta_i\partial\theta_j} \\ &+ \sum_{k,l} \frac{\partial^2 R}{\partial\beta_k \partial\beta_l} \cdot \frac{\partial\beta_k}{\partial\theta_i}\frac{\partial\beta_l}{\partial\theta_j} + \sum_k \frac{\partial R}{\partial\beta_k} \cdot \frac{\partial^2\beta_k}{\partial\theta_i\partial\theta_j}\,.\end{aligned}$$

Furthermore, given penalty functions defined in (C1), it can be shown that

1. $\frac{\partial^2 R}{\partial\alpha_k\partial\alpha_l} = \frac{\partial^2 R}{\partial\beta_k\partial\beta_l} = 0$ for all $k \neq l$, and $|\frac{\partial^2 R}{\partial\alpha_k^2}| \leq \mu$ for all k.
2. If $\frac{\partial R}{\partial\alpha_k} = 0$ for some k, then $\frac{\partial\alpha_k}{\partial\theta_i}$ (and hence $\frac{\partial\beta_k}{\partial\theta_i}$) can have any value, but if $\frac{\partial R}{\partial\alpha_k} \neq 0$ then $\frac{\partial\alpha_k}{\partial\theta_i} = 0$ for $i \neq k$, and moreover $\frac{\partial\alpha_k}{\partial\theta_k}$ can be either 1 ($\frac{\partial\beta_k}{\partial\theta_k}$ should be 0 in this case) or 0 ($\frac{\partial\alpha_k}{\partial\theta_k}$ should 1).
3. If $\frac{\partial R}{\partial\alpha_k} \neq 0$, then $\frac{\partial^2\alpha_k}{\partial\theta_i\partial\theta_j} = 0$.

Based on these facts, ∇F can be reduce to the diagonal matrix whose the maximum absolute element is upper bounded by μ, implying (59) (since $\nabla^2\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}$ can be shown to be zero matrix by the same reason above). □

Lemma 6. $\left\|\nabla\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}} - \nabla\mathcal{R}(\boldsymbol{\theta};\boldsymbol{\lambda})\right\|_{\boldsymbol{\eta}}^{*} \leq \max\left\{\frac{\lambda_1}{\eta_1}, \frac{\lambda_2}{\eta_2}\right\}$.

Proof. By definitions of $\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}$ and $R(\boldsymbol{\theta};\boldsymbol{\lambda})$, every pair of coordinates $(\nabla\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}, \nabla\mathcal{R}(\boldsymbol{\theta};\boldsymbol{\lambda}))$ has the same signs, and moreover based on Lemma 3 and 4, the statement holds. □

Lemma 7 (Lemma 8 of (Loh & Wainwright, 2015))**.** *If $\mathcal{L}$ is convex and (12) holds, then (13) holds as well, with $\kappa_2 = \kappa_1$ and $\tau_2 = 2r\tau_1$.*

Proof. As shown in the proof of Lemma 8 of (Loh & Wainwright, 2015), for any $\boldsymbol{\theta}_2 \in \mathbb{R}^p$ such that $\|\boldsymbol{\theta}_2\|_2 \geq 1$,

$$\begin{aligned}\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_1+\boldsymbol{\theta}_2) - \nabla\mathcal{L}(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2\right\rangle &\geq \|\boldsymbol{\theta}_2\|_2\left(\kappa_1 - \tau_1\frac{\|\boldsymbol{\theta}_2\|_{\boldsymbol{\eta}}^2}{\|\boldsymbol{\theta}_2\|_2^2}\right)\\ &\geq \|\boldsymbol{\theta}_2\|_2\left(\kappa_1 - 2r\tau_1\frac{\|\boldsymbol{\theta}_2\|_{\boldsymbol{\eta}}}{\|\boldsymbol{\theta}_2\|_2^2}\right)\\ &\geq \|\boldsymbol{\theta}_2\|_2\left(\kappa_1 - 2r\tau_1\frac{\|\boldsymbol{\theta}_2\|_{\boldsymbol{\eta}}}{\|\boldsymbol{\theta}_2\|_2}\right)\\ &\geq \kappa_1\|\boldsymbol{\theta}_2\|_2 - 2r\tau_1\|\boldsymbol{\theta}_2\|_{\boldsymbol{\eta}}\,.\end{aligned}$$

□

Lemma 8 (Lemma 9 of (Loh & Wainwright, 2015))**.** *If (12) holds globally for all $\boldsymbol{\theta}_2 \in \mathbb{R}^p$, then (13) holds as well, with $\kappa_2 = \kappa_1$ and $\tau_2 = 2r\tau_1$.*

Proof. If $\|\boldsymbol{\theta}_2\|_2 \geq 1$,

$$\begin{aligned}\left\langle \nabla\mathcal{L}(\boldsymbol{\theta}_1+\boldsymbol{\theta}_2) - \nabla\mathcal{L}(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2\right\rangle &\geq \kappa_1\|\boldsymbol{\theta}_2\|_2^2 - \tau_1\|\boldsymbol{\theta}_2\|_{\boldsymbol{\eta}}^2\\ &\geq \kappa_1\|\boldsymbol{\theta}_2\|_2 - 2r\tau_1\|\boldsymbol{\theta}_2\|_{\boldsymbol{\eta}}\,.\end{aligned}$$

□

Lemma 9 (From the proof of Lemma 9 of (Jalali et al., 2010))**.** *Let $W_j^{(k)}$ be*

$$\frac{1}{n}\left\langle X_j^{(k)}, I - \frac{1}{n}X_{U_k}^{(k)}\left(\frac{1}{n}\langle X_{U_k}^{(k)}, X_{U_k}^{(k)}\rangle\right)^{-1}\left(X_{U_k}^{(k)}\right)^{\top}\right\rangle w^{(k)} \tag{60}$$

where U_k^c denotes the support of k-th column of Θ^*. For all $t \geq 0$,

$$\mathbb{P}\left[\max_{(j,k)\in U^c}|W_j^{(k)}| \geq t\right] \leq 2\exp\left(-\frac{t^2 n}{4\sigma^2} + \log(pm)\right), \tag{61}$$

and

$$\mathbb{P}\left[\max_{j}\sum_{(j,k)\in U^c}|W_j^{(k)}| \geq t\right] \leq 2\exp\left(-\frac{t^2 n}{4\sigma^2} + m\log 2 + \log p\right). \tag{62}$$

In addition, for all $t \geq 0$,

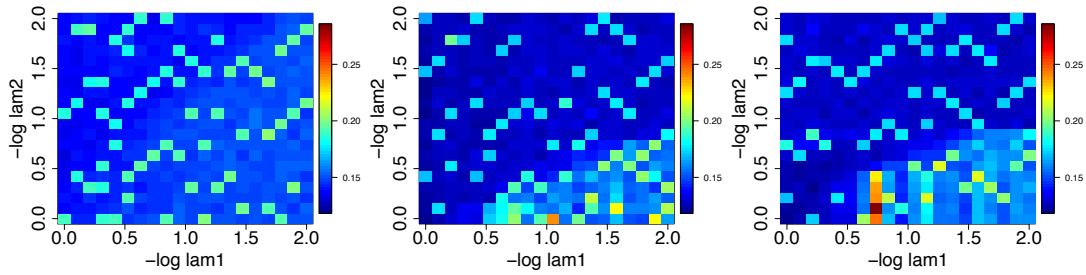
$$\mathbb{P}\left[\max_{k=1,2,\ldots,m}\left\|\left(\frac{1}{n}\langle X_{U_k}^{(k)}, X_{U_k}^{(k)}\rangle\right)^{-1}\frac{1}{n}\left(X_{U_k}^{(k)}\right)^{\top}w^{(k)}\right\|_{\infty} \geq t\right] \leq 2\exp\left(-\frac{t^2 n C_{\min}}{50\sigma^2} + \log(pm)\right) \tag{63}$$

where $C_{\min} := \min_{k=1,2,\ldots,m}\lambda_{\min}\left(\Sigma_{U_k U_k}^{(k)}\right) > 0$.

Note that (Jalali et al., 2010) originally consider only $j \in \cap_{k=1}^m U_k^c$, but (61) and (62) hold by the same reasoning.

E. Additional details on the simulation experiments

Figure 4 provides an example of how the computing time of convex and non-convex dirty models vary with the regularization parameters λ_1, λ_2. The iteration number needed for each method under the same convergence tolerence exhibit a similar profile. In many cases non-convex dirty models converge faster and take less computing time.



Convex DM | Non-convex DM (SCAD) | Non-convex DM (MCP)

Figure 4. Timing for comparison methods for varying penalty parameters.