# Appendix

## A Properties of Hybrid Regularization Function $\mathcal{R}(\cdot; \mathbf{c})$

Let us briefly study some properties of this hybrid regularization function $\mathcal{R}(\cdot; \mathbf{c})$. In the remainder of this section, we suppress the dependence of the notation on the weights $\mathbf{c}$. From (2), it can be seen that $\mathcal{R}(\cdot)$ satisfies the triangle inequality:

$$\mathcal{R}(\theta + \theta') \leq \mathcal{R}(\theta) + \mathcal{R}(\theta'). \tag{11}$$

Under the additional assumption that each $\mathcal{R}_\alpha(.)$ is 1-sub-homogeneous, so that $\mathcal{R}_\alpha(\beta\theta) \leq \beta\mathcal{R}_\alpha(\theta)$, we get that:

$$\mathcal{R}(\beta\theta + (1-\beta)\theta') \leq \mathcal{R}(\beta\theta) + \mathcal{R}((1-\beta)\theta') \leq \beta\mathcal{R}(\theta) + (1-\beta)\mathcal{R}(\theta'), \tag{12}$$

so that $\mathcal{R}$ is convex. Moreover, provided the individual regularization functions $\mathcal{R}_\alpha(\cdot)$, for $\alpha \in I$, are non-negative, it can be seen that $\mathcal{R}(\cdot)$ is non-negative as well.

From the above properties, it follows that provided the the individual regularization functions $\mathcal{R}_\alpha(\cdot)$, for $\alpha \in I$, are *norms*, $\mathcal{R}(\cdot)$ is a norm as well.

We can also derive the dual of the hybrid regularization function, assuming it is a norm:

$$\begin{aligned} \mathcal{R}^*(u) &= \sup_\theta \frac{\langle \theta, u \rangle}{\mathcal{R}(\theta)} = \sup_{(\theta_\alpha)} \frac{\sum_\alpha \langle u, \theta_\alpha \rangle}{\sum_\alpha c_\alpha \mathcal{R}_\alpha(\theta_\alpha)} = \sup_{(\theta_\alpha)} \frac{\sum_\alpha \langle u/c_\alpha, \theta_\alpha \rangle}{\sum_\alpha \mathcal{R}_\alpha(\theta_\alpha)} \\ &\leq \sup_{(\theta_\alpha)} \frac{\sum_\alpha \mathcal{R}^*_\alpha(u/c_\alpha)\mathcal{R}_\alpha(\theta_\alpha)}{\sum_\alpha \mathcal{R}_\alpha(\theta_\alpha)} \leq \max_{\alpha \in I} \mathcal{R}^*_\alpha(u)/c_\alpha. \end{aligned}$$

Noting that the inequalities become equalities by setting $\theta'_\alpha = 0$, for all $\alpha'$ not attaining the maximum in the last expression, we obtain the required expression for the dual norm:

$$\mathcal{R}^*(u) = \max_{\alpha \in I} \mathcal{R}^*_\alpha(u)/c_\alpha. \tag{13}$$

On the other hand, its Fenchel conjugate (again, assuming it is a norm) is given by:

$$\begin{aligned} \mathcal{R}_f(u) &= \sup_\theta \Big\{ \langle \theta, u \rangle - \mathcal{R}(\theta) \Big\} = \sup_\theta \Big\{ \langle \theta, u \rangle - \inf_{(\theta_\alpha)_{\alpha \in I} : \sum_{\alpha \in I} \theta_\alpha = \theta} \sum_{\alpha \in I} c_\alpha \mathcal{R}_\alpha(\theta_\alpha) \Big\} \\ &= \sum_{\alpha \in I} \sup_{\theta_\alpha} \Big\{ \langle \theta, u/c_\alpha \rangle - \mathcal{R}_\alpha(\theta_\alpha) \Big\} = \sum_{\alpha \in I} \mathcal{R}_{\alpha;f}(u)/c_\alpha. \end{aligned}$$

## B Proof of Theorem 1

Throughout the proof, we make use of the fact that the optimal error vector $\widehat{\theta} - \theta^*$ is guaranteed to belong to a specific set, as a consequence of the decomposability of the respective regularization functions:

**Proposition 4.** *Suppose that conditions (C1) and (C2) are satisfied. Then for any optimal solution* $\widehat{\theta}$ *of* (1)*, with the regularization penalties satisfying* $\lambda_\alpha \geq 2\mathcal{R}^*_\alpha\big(\nabla_{\theta_\alpha} \mathcal{L}(\theta^*; Z_1^n)\big)$*, the error* $\widehat{\Delta}$ *lies in the set*

$$\begin{aligned} \mathbb{C}(\mathcal{M}_1, \overline{\mathcal{M}}_1^\perp, \ldots, \mathcal{M}_{|I|}, \overline{\mathcal{M}}_{|I|}^\perp; \theta^*) := \Big\{ &(\Delta_1, \ldots, \Delta_{|I|}) \in \Omega_1 \times \ldots \times \Omega_{|I|} \Big| \\ &\sum_{\alpha \in I} \lambda_\alpha \mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big) \leq \sum_{\alpha \in I} \lambda_\alpha \Big[ 3\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big) + 4\mathcal{R}_\alpha\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta^*_\alpha)\big) \Big] \Big\} \end{aligned} \tag{14}$$

See Appendix B.1 for the proof of this claim.

To set up the first crucial ingredient of the proof, we define the function $F : \Omega_1 \times \ldots \times \Omega_{|I|} \mapsto \mathbb{R}$:

$$F(\Delta_1, \ldots, \Delta_{|I|}) := \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) + \sum_{\alpha \in I} \Big[ \lambda_\alpha \mathcal{R}_\alpha(\theta^*_\alpha + \Delta_\alpha) - \mathcal{R}_\alpha(\theta^*_\alpha) \Big]. \tag{15}$$

By virtue of its optimality, it can be seen that the error $(\widehat{\Delta}_1, \ldots, \widehat{\Delta}_{|I|})$ of the optimal solution $(\widehat{\theta}_1, \ldots, \widehat{\theta}_{|I|})$ satisfies $F\big(\widehat{\Delta}_1, \ldots, \widehat{\Delta}_{|I|}\big) \leq 0$. Also note that $F(\vec{0}, \ldots, \vec{0}) = 0$.

As the following lemma shows, in order to compute an upper bound on the optimal error, $\sum_{\alpha \in I} \|\widehat{\Delta}_\alpha\|$, it suffices to control the sign of function $F$ over the following set:

$$\mathbb{K}(\delta) := \mathbb{C} \cap \Big\{ \sum_{\alpha \in I} \|\Delta_\alpha\| = \delta \Big\}$$

**Lemma 1.** *Suppose that conditions (C1) and (C2) are satisfied. If* $F(\Delta_1, \ldots, \Delta_{|I|}) > 0$ *for all possible vectors* $(\Delta_1, \ldots, \Delta_{|I|}) \in \mathbb{K}(\delta)$*, then the optimal error* $(\widehat{\Delta}_1, \ldots, \widehat{\Delta}_{|I|})$ *satisfies* $\sum_{\alpha \in I} \|\widehat{\Delta}_\alpha\| \leq \delta$.

See Appendix B.2 for the proof of this claim.

The other crucial ingredient of the proof, is that we show the following “global” restricted strong convexity condition given the “local” restricted strong convexity conditions in (C3) for just individual structures:

**Lemma 2.** *Suppose that conditions (C3) and (C4) are satisfied for the $M$-estimation problem* (1)*. Then, for* $(\theta_1, \ldots, \theta_|I|) \in (\Omega_1 \times \ldots \times \Omega_{|I|})$ *the following “global” restricted strong convexity (RSC) condition holds:*

$$\begin{aligned} \delta\mathcal{L}(\Delta_1, \ldots, \Delta_{|I|}; \theta^*) &:= \mathcal{L}\big(\sum_{\alpha \in I} (\theta^*_\alpha + \Delta_\alpha)\big) - \mathcal{L}\big(\sum_{\alpha \in I} \theta^*_\alpha\big) - \sum_{\alpha \in I} \big\langle \nabla_\theta \mathcal{L}\big(\sum_{\alpha \in I} \theta^*_\alpha\big), \Delta_\alpha \big\rangle \\ &\geq \Big[\frac{\kappa_{\mathcal{L}}}{2} - 32\bar{g}^2 |I| \Phi^2\Big] \sum_{\alpha \in I} \|\Delta_\alpha\|^2 - 32\bar{g}^2 |I| \sum_{\alpha \in I} \Big[\lambda_\alpha^2 \mathcal{R}_\alpha^2\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta^*_\alpha)\big)\Big], \end{aligned} \tag{16}$$

*where $\bar{g}$ is defined as* $\max_\alpha \frac{1}{\lambda_\alpha} \sqrt{g_\alpha + h_\alpha}$.

See Appendix B.3 for the proof of this claim.

We will next prove the following bound on $\sum_{\alpha \in I} \|\widehat{\Delta}_\alpha\|$:

$$\sum_{\alpha \in I} \|\widehat{\Delta}_\alpha\| \leq \frac{|I|}{\bar{\kappa}} \Big( \frac{3}{2} \Phi + \sqrt{\bar{\kappa} \tau_{\mathcal{L}}} \Big), \tag{17}$$

where $\bar{\kappa} := \frac{\kappa_{\mathcal{L}}}{2} - 32\bar{g}^2 |I| \Phi^2$ and $\tau_{\mathcal{L}} := \sum_{\alpha \in I} \Big[ 32\bar{g}^2 \lambda_\alpha^2 \mathcal{R}_\alpha^2\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta^*_\alpha)\big) + \frac{2\lambda_\alpha}{|I|} \mathcal{R}_\alpha\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta^*_\alpha)\big) \Big]$.

Given this bound, we can use the following inequality and complete the proof:

$$\begin{aligned} \|\widehat{\theta} - \theta^*\| &= \| \sum_\alpha \widehat{\theta}_\alpha - \sum_\alpha \theta^*_\alpha \| = \| \sum_\alpha (\widehat{\theta}_\alpha - \theta^*_\alpha) \| \\ &\leq \sum_\alpha \|\widehat{\theta}_\alpha - \theta^*_\alpha\| \leq \sum_{\alpha \in I} \|\widehat{\Delta}_\alpha\| \leq \frac{|I|}{\bar{\kappa}} \Big( \frac{3}{2} \Phi + \sqrt{\bar{\kappa} \tau_{\mathcal{L}}} \Big). \end{aligned}$$

It thus remains to show the bound (17).

As shown in Lemma 2, given conditions (C3) and (C4), the loss function satisfies the following “global” RSC condition with curvature $\bar{\kappa}$ and tolerance $\bar{\tau}$:

$$\delta\mathcal{L}(\Delta_1, \ldots, \Delta_{|I|}; \theta^*) \geq \bar{\kappa} \sum_{\alpha \in I} \|\Delta_\alpha\|^2 - \bar{\tau}(\theta^*).$$

Then, by the construction of $F$, for an arbitrary error vector $(\Delta_1, \ldots, \Delta_{|I|}) \in \mathbb{C}$, we have

$$F(\Delta_1, \ldots, \Delta_{|I|}) \geq \sum_{\alpha \in I} \langle \nabla_\theta \mathcal{L}(\theta^*), \Delta_\alpha \rangle + \bar{\kappa} \sum_{\alpha \in I} \|\Delta_\alpha\|^2 - \bar{\tau}(\theta^*) + \sum_{\alpha \in I} \Big[ \lambda_\alpha \mathcal{R}_\alpha(\theta^*_\alpha + \Delta_\alpha) - \mathcal{R}_\alpha(\theta^*_\alpha) \Big].$$

Using the bounds (18) and (19), we obtain

$$\begin{aligned} F(\Delta_1,\ldots,\Delta_{|I|}) &\geq -\sum_{\alpha\in I}\frac{\lambda_\alpha}{2}\Big[\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big)+\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big)\Big]+\bar{\kappa}\sum_{\alpha\in I}\|\Delta_\alpha\|^2-\bar{\tau}(\theta^*)\\ &\quad+\sum_{\alpha\in I}\lambda_\alpha\Big[\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big)-\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big)-2\mathcal{R}_\alpha\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta_\alpha^*)\big)\Big]\\ &\geq \sum_{\alpha\in I}\Big[\bar{\kappa}\|\Delta_\alpha\|^2-\frac{\lambda_\alpha}{2}\Big[3\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big)+4\mathcal{R}_\alpha\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta_\alpha^*)\big)\Big]\Big]-\bar{\tau}(\theta^*), \end{aligned}$$

where in the last inequality we dropped the term $\frac{\lambda_\alpha}{2}\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big)$, since it is always positive.

The following proposition allows us to reduce the task of bounding a multivariate function to that of bounding a univariate function.

**Proposition 5.** *Consider the $K$-variate quadratic function: $F(x_1,\ldots,x_K)=\sum_{k=1}^{K}(ax_k^2+bx_k)+c$ for some constants a,b and c. Suppose further that: $a>0$, $x_k\geq 0$ for all k and $\sum_k x_k=\delta>0$. Then $F(x_1,\ldots,x_K)$ attains its minimum value at $x_1=\ldots=x_K=\delta/K$.*

Denote $\Phi:=\max_{\alpha\in I}\lambda_\alpha\Psi_\alpha(\overline{\mathcal{M}}_\alpha)$ and $D:=\frac{1}{|I|}\bar{\tau}(\theta^*)+\sum_{\alpha\in I}\frac{2\lambda_\alpha}{|I|}\mathcal{R}_\alpha\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta_\alpha^*)\big)$. Armed with these notations, and using the definition of subspace compatibility constant,

$$\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big)\leq\Psi_\alpha(\overline{\mathcal{M}}_\alpha)\|\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\|\leq\Psi_\alpha(\overline{\mathcal{M}}_\alpha)\|\Delta_\alpha\|,$$

we obtain:

$$F(\Delta_1,\ldots,\Delta_{|I|})\geq\sum_{\alpha\in I}\big(\bar{\kappa}\|\Delta_\alpha\|^2-\frac{3}{2}\Phi\|\Delta_\alpha\|-D\big).$$

Now consider an error vector $(\Delta_1,\ldots,\Delta_{|I|})$ s.t. $\sum_{\alpha\in I}\|\Delta_\alpha\|=\delta$. From Proposition 5,

$$F(\Delta_1,\ldots,\Delta_{|I|})\geq\bar{\kappa}\delta^2-\frac{3}{2}\Phi\delta-D|I|.$$

It follows that $F(\Delta_1,\ldots,\Delta_{|I|})>0$ so long as $\delta>\frac{|I|}{\bar{\kappa}}\big(\frac{3}{2}\Phi+\sqrt{\bar{\kappa}D}\big)$. Therefore, for all error vectors $(\Delta_1,\ldots,\Delta_{|I|})$ such that $\sum_{\alpha\in I}\|\Delta_\alpha\|\geq\frac{|I|}{\bar{\kappa}}\big(\frac{3}{2}\Phi+\sqrt{\bar{\kappa}D}\big)$, and in particular, if $(\Delta_1,\ldots,\Delta_{|I|})\in\mathbb{K}\Big(\frac{|I|}{\bar{\kappa}}\big(\frac{3}{2}\Phi+\sqrt{\bar{\kappa}D}\big)\Big)$, we can guarantee $F(\Delta_1,\ldots,\Delta_{|I|})>0$. This satisfies the conditions for Lemma 1, whose statement then completes the proof.

### B.1 Proof of Proposition 4

In the proof, we use the fact that $F\big(\widehat{\Delta}_1,\ldots,\widehat{\Delta}_{|I|}\big)\leq 0$.

For any decomposable regularizer $\mathcal{R}_\alpha$ on $(\mathcal{M}_\alpha,\overline{\mathcal{M}}_\alpha^\perp)$, it is known that the following inequality holds (See [11] for the proof):

$$\mathcal{R}_\alpha(\theta_\alpha^*+\Delta_\alpha)-\mathcal{R}_\alpha(\theta_\alpha^*)\geq\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big)-\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big)-2\mathcal{R}_\alpha\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta_\alpha^*)\big). \tag{18}$$

At the same time, by the convexity of $\mathcal{L}$, we have

$$\begin{aligned} \mathcal{L}(\theta^*+\Delta)-\mathcal{L}(\theta^*) &= \mathcal{L}\Big(\sum_{\alpha\in I}\theta_\alpha^*+\sum_{\alpha\in I}\Delta_\alpha\Big)-\mathcal{L}\Big(\sum_{\alpha\in I}\theta_\alpha^*\Big)\\ &\geq\sum_{\alpha\in I}\langle\nabla_{\theta_\alpha}\mathcal{L}(\theta^*),\Delta_\alpha\rangle\\ &\overset{(i)}{\geq}-\sum_{\alpha\in I}\mathcal{R}_\alpha^*\big(\nabla_{\theta_\alpha}\mathcal{L}(\theta^*)\big)\mathcal{R}_\alpha(\Delta_\alpha)\\ &\overset{(ii)}{\geq}-\sum_{\alpha\in I}\frac{\lambda_\alpha}{2}\Big[\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big)+\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big)\Big] \qquad (19) \end{aligned}$$

where the inequality (i) comes from the generalized Cauchy-Schwarz inequality, and the inequality (ii) from the triangular inequality and the assumption in the statement.

Combining (18) and (19) yields

$$0 \geq F\big(\widehat{\Delta}_1, \ldots, \widehat{\Delta}_{|I|}\big) \geq \sum_{\alpha \in I} \frac{\lambda_\alpha}{2} \Big[\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\widehat{\Delta}_\alpha)\big) - 3\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha}(\widehat{\Delta}_\alpha)\big) - 4\mathcal{R}_\alpha\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta^*_\alpha)\big)\Big],$$

as claimed.

## B.2 Proof of Lemma 1

We first show that the set $\mathbb{C}$ has some special structure, and then show that that structure guarantees the error upper bounded as in the statement.

Let $(\Delta_1, \ldots, \Delta_{|I|})$ be an arbitrary error vectors in the set $\mathbb{C}$. Then, for any $t \in (0, 1)$, we have

$$\begin{aligned}
&\sum_{\alpha \in I} \lambda_\alpha \mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(t\Delta_\alpha)\big) \\
\overset{(i)}{=}\;& \sum_{\alpha \in I} \lambda_\alpha \mathcal{R}_\alpha\big(t\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big) \\
\overset{(ii)}{=}\;& t\sum_{\alpha \in I} \lambda_\alpha \mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big) \\
\overset{(iii)}{\leq}\;& t\sum_{\alpha \in I} \lambda_\alpha \Big[3\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big) + 4\mathcal{R}_\alpha\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta^*_\alpha)\big)\Big] \\
\overset{(iv)}{=}\;& \sum_{\alpha \in I} \lambda_\alpha \Big[3\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha}(t\Delta_\alpha)\big) + 4t\mathcal{R}_\alpha\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta^*_\alpha)\big)\Big] \\
\overset{(v)}{\leq}\;& \sum_{\alpha \in I} \lambda_\alpha \Big[3\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha}(t\Delta_\alpha)\big) + 4\mathcal{R}_\alpha\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta^*_\alpha)\big)\Big] \qquad (20)
\end{aligned}$$

where step $(i)$ uses the fact that

$$\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(t\Delta_\alpha) = \operatorname*{argmin}_{\gamma \in \bar{\mathcal{M}}^\perp} \|t\Delta_\alpha - \gamma\| = t \operatorname*{argmin}_{\gamma \in \bar{\mathcal{M}}^\perp} \|\Delta_\alpha - \frac{\gamma}{t}\| = t\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha),$$

step $(ii)$ uses the positive homogeneity of norms, and step $(iii)$ holds since $(\Delta_1, \ldots, \Delta_{|I|}) \in \mathbb{C}$. Moreover, step $(iv)$ holds similarly as in equalities $(i)$ and $(ii)$, and finally step $(v)$ trivially holds for any $t \leq 1$. Therefore, if $(\Delta_1, \ldots, \Delta_{|I|}) \in \mathbb{C}$, then the line segment $\{(t\Delta_1, \ldots, t\Delta_{|I|}) \,|\, t \in (0, 1)\}$ between $(\Delta_1, \ldots, \Delta_{|I|})$ and the origin $(\vec{0}, \ldots, \vec{0})$ also lies in $\mathbb{C}$.

Now, we show the statement in the lemma by its contrapositive; suppose $\sum_{\alpha \in I} \|\widehat{\Delta}_\alpha\| > \delta$. Since $\sum_{\alpha \in I} \|t\widehat{\Delta}_\alpha\| = t\sum_{\alpha \in I} \|\widehat{\Delta}_\alpha\|$, there exists some constant $t^* \in (0, 1)$ s.t. $(t^*\widehat{\Delta}_1, \ldots, t^*\widehat{\Delta}_{|I|}) \in \mathbb{K}(\delta)$. At the same time, by the convexity of $\mathcal{L}$ and the regularizers,

$$F(t^*\widehat{\Delta}_1, \ldots, t^*\widehat{\Delta}_{|I|}) \leq t^*F(\widehat{\Delta}_1, \ldots, \widehat{\Delta}_{|I|}) + (1 - t^*)F(\vec{0}, \ldots, \vec{0}) \leq 0.$$

Therefore, $(t^*\widehat{\Delta}_1, \ldots, t^*\widehat{\Delta}_{|I|})$ is in $\mathbb{K}(\delta)$ such that $F(t^*\widehat{\Delta}_1, \ldots, t^*\widehat{\Delta}_{|I|}) \leq 0$ by construction. Hence, by its contrapositive, the claim follows.

## B.3 Proof of Lemma 2

The definition of $\delta\mathcal{L}(\Delta_1, \ldots, \Delta_{|I|}; \theta^*)$ can be rewritten as

$$\begin{aligned}
\delta\mathcal{L}(\Delta_1, \ldots, \Delta_{|I|}; \theta^*) &= \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \sum_{\alpha \in I} \big\langle \nabla_\theta \mathcal{L}(\theta^*), \Delta_\alpha \big\rangle \\
=& \mathcal{L}(\theta^* + \Delta) + (|I| - 1)\mathcal{L}(\theta^*) - \sum_{\alpha \in I} \mathcal{L}\big(\theta^* + \Delta_\alpha\big) \\
&+ \sum_{\alpha \in I} \Big[\mathcal{L}(\theta^* + \Delta_\alpha) - \mathcal{L}(\theta^*) - \langle \nabla_\theta \mathcal{L}(\theta^*), \Delta_\alpha \rangle\Big].
\end{aligned}$$

Then, simply by the inequalities in (C3) and (C4), $\delta\mathcal{L}(\Delta_1, \ldots, \Delta_{|I|}; \theta^*)$ can be lower-bounded by

$$\begin{aligned}&\sum_{\alpha\in I}\Big[\kappa_{\mathcal{L}}\|\Delta_\alpha\|^2 - g_\alpha\mathcal{R}_\alpha^2(\Delta_\alpha)\Big] - \frac{\kappa_{\mathcal{L}}}{2}\sum_{\alpha\in I}\|\Delta_\alpha\|^2 - \sum_{\alpha\in I} h_\alpha\mathcal{R}_\alpha^2(\Delta_\alpha)\\ &\geq \frac{\kappa_{\mathcal{L}}}{2}\sum_{\alpha\in I}\|\Delta_\alpha\|^2 - \sum_{\alpha\in I}\big(g_\alpha + h_\alpha\big)\mathcal{R}_\alpha^2(\Delta_\alpha). \qquad (21)\end{aligned}$$

Now, let us focus on the term $\sum_{\alpha\in I}\big(g_\alpha + h_\alpha\big)\mathcal{R}_\alpha^2(\Delta_\alpha)$ in the RHS of (21). By the basic property of the square, we have the following inequalities:

$$\sum_{\alpha\in I}\big(g_\alpha + h_\alpha\big)\mathcal{R}_\alpha^2(\Delta_\alpha) \leq \Big(\sum_{\alpha\in I}\sqrt{g_\alpha + h_\alpha}\mathcal{R}_\alpha(\Delta_\alpha)\Big)^2 \leq \Big(\bar{g}\sum_{\alpha\in I}\lambda_\alpha\mathcal{R}_\alpha(\Delta_\alpha)\Big)^2 \qquad (22)$$

where in the second inequality we use $\langle x, y\rangle \leq \|x\|_\infty\|y\|_1$ and $\bar{g} := \max_\alpha \frac{1}{\lambda_\alpha}\sqrt{g_\alpha + h_\alpha}$. Since by Lemma 4, for any $(\Delta_1, \ldots, \Delta_{|I|}) \in \mathbb{C}$,

$$\begin{aligned}&\sum_{\alpha\in I}\lambda_\alpha\mathcal{R}_\alpha(\Delta_\alpha) \leq \sum_{\alpha\in I}\lambda_\alpha\Big[\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big) + \mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big)\Big]\\ &\leq \sum_{\alpha\in I}\lambda_\alpha\Big[4\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big) + 4\mathcal{R}_\alpha\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta_\alpha^*)\big)\Big],\end{aligned}$$

the LHS of (22), $\sum_{\alpha\in I}\big(g_\alpha + h_\alpha\big)\mathcal{R}_\alpha^2(\Delta_\alpha)$, can be upper-bounded by

$$\begin{aligned}&32\bar{g}^2|I|\sum_{\alpha\in I}\lambda_\alpha^2\Big[\mathcal{R}_\alpha^2\big(\Pi_{\bar{\mathcal{M}}_\alpha}(\Delta_\alpha)\big) + \mathcal{R}_\alpha^2\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta_\alpha^*)\big)\Big]\\ &\leq 32\bar{g}^2|I|\sum_{\alpha\in I}\lambda_\alpha^2\Big[\Psi_\alpha^2\|\Delta_\alpha\|^2 + \mathcal{R}_\alpha^2\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta_\alpha^*)\big)\Big]. \qquad (23)\end{aligned}$$

By plugging (23) back into (21), we can construct the RSC condition as stated in Lemma 2 and complete the proof:

$$\delta\mathcal{L}(\Delta_1, \ldots, \Delta_{|I|}; \theta^*) \geq \Big[\frac{\kappa_{\mathcal{L}}}{2} - 32\bar{g}^2|I|\Phi^2\Big]\sum_{\alpha\in I}\|\Delta_\alpha\|^2 - 32\bar{g}^2|I|\sum_{\alpha\in I}\lambda_\alpha^2\mathcal{R}_\alpha^2\big(\Pi_{\mathcal{M}_\alpha^\perp}(\theta_\alpha^*)\big).$$

## C Proof of Proposition 2

Given the pair of complementary subspaces $(\overline{\mathcal{M}}, \overline{\mathcal{M}}^\perp)$, the error vector $\Delta_\alpha$ can be written as $\Delta_\alpha = \Pi_{\bar{\mathcal{M}}}(\Delta_\alpha) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta_\alpha)$, we have

$$X\Delta_\alpha = X\big(\Pi_{\bar{\mathcal{M}}}(\Delta_\alpha) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta_\alpha)\big) = (X\mathcal{P}_{\bar{\mathcal{M}}})\Pi_{\bar{\mathcal{M}}}(\Delta_\alpha) + (X\mathcal{P}_{\bar{\mathcal{M}}^\perp})\Pi_{\bar{\mathcal{M}}^\perp}(\Delta_\alpha). \qquad (24)$$

We will now assume that the regularization functions $\mathcal{R}_\alpha$ are atomic norms [4] with respect to the orthonormal basis vectors used to describe the subspaces $(\mathcal{M}_\alpha, \overline{\mathcal{M}}_\alpha)$; though the proof generalizes. Specifically, suppose there is a set of orthonormal vectors or "atoms" $\mathcal{A}_\alpha := \{\mathbf{a}_j\}$, such that the regularization function $\mathcal{R}_\alpha$ can be written as

$$\mathcal{R}_\alpha(\theta_\alpha) = \inf_{\mathbf{c}}\Big\{\sum_j c_j : \theta_\alpha = \sum_j c_j\mathbf{a}_j\Big\}. \qquad (25)$$

We will also assume that the subspace pair $(\mathcal{M}_\alpha, \overline{\mathcal{M}}_\alpha^\perp)$ contain disjoint orthonormal subsets of the basis set; it can be seen that $\mathcal{R}_\alpha$ is decomposable with respect to any such subspace pair.

Let $t$ be the dimensionality of $\overline{\mathcal{M}}_\alpha$. Suppose $\{a_1, \ldots, a_J\} \subseteq \mathcal{A}_\alpha$ are a set of atoms, such that the first $t$ atoms characterize $\overline{\mathcal{M}}_\alpha$, and the remaining $J - t$ atoms characterize $\overline{\mathcal{M}}_\alpha^\perp$. Note that we can then write $\Delta_\alpha$ as

$$\Delta_\alpha = \sum_{j=1}^{J} c_j a_j.$$

We now split the index set $\{1,\dots,J\}$ into subsets $S(1),\dots,S(K)$, such that $S(1)=\{1,\dots,t\}$ contains the atoms characterizing $\overline{\mathcal{M}}_\alpha$, the second set $S(2)$ contains the largest $t$ entries from the remaining coefficients $(c_{t+1},\dots,c_J)$, $S(3)$ contains the largest set of $t$ entries disjoint from $S(1)$ and $S(2)$ and so on. Let $\overline{\mathcal{M}}_i$ denote the subspace characterized by the atoms indexed by set $S(i)$ suppressing the dependency on $\alpha$. With these notations, we can rewrite (24) as

$$X\Delta_\alpha = X\sum_i \sum_{j\in S(i)} c_j\,\mathbf{a}_j = \sum_i \mathcal{P}_{\bar{\mathcal{M}}_i} X \times \big(\sum_{j\in S(i)} c_j\,\mathbf{a}_j\big). \tag{26}$$

Turning to the condition (C4), we have

$$\frac{2}{n}\Big|\sum_{\alpha<\beta}\langle X\Delta_\alpha, X\Delta_\beta\rangle\Big| \le \frac{2}{n}\sum_{\alpha<\beta}\big|\langle X\Delta_\alpha, X\Delta_\beta\rangle\big|. \tag{27}$$

For any pair $\alpha<\beta$ in the index set $I$, we use the equality (26) as follows:

$$\begin{aligned}
&\frac{2}{n}\big|\langle X\Delta_\alpha, X\Delta_\beta\rangle\big| \\
&= \frac{2}{n}\Big|\Big\langle \sum_i \mathcal{P}_{\bar{\mathcal{M}}_i} X\,\big(\sum_{j\in S(i)} c_j\,\mathbf{a}_j\big),\, \sum_k \mathcal{P}_{\bar{\mathcal{M}}_k} X\,\big(\sum_{l\in S(k)} d_l\,\mathbf{b}_l\big)\Big\rangle\Big| \\
&= \frac{2}{n}\Big|\sum_{i,k}\Big\langle \mathcal{P}_{\bar{\mathcal{M}}_i} X\,\big(\sum_{j\in S(i)} c_j\,\mathbf{a}_j\big),\, \mathcal{P}_{\bar{\mathcal{M}}_k} X\,\big(\sum_{l\in S(k)} d_l\,\mathbf{b}_l\big)\Big\rangle\Big| \\
&\le \frac{2}{n}\sum_{i,k}\Big|\Big\langle \mathcal{P}_{\bar{\mathcal{M}}_i} X\,\big(\sum_{j\in S(i)} c_j\,\mathbf{a}_j\big),\, \mathcal{P}_{\bar{\mathcal{M}}_k} X\,\big(\sum_{l\in S(k)} d_l\,\mathbf{b}_l\big)\Big\rangle\Big| \\
&\overset{(i)}{\le} 2\Big[\max_{i,k}\sigma_{\max}\Big(\mathcal{P}_{\bar{\mathcal{M}}_i}\big(\frac{1}{n}X^TX\big)\mathcal{P}_{\bar{\mathcal{M}}_k}\Big)\Big]\sum_{i,k}\big\|\big(\sum_{j\in S(i)} c_j\,\mathbf{a}_j\big)\big\|_2\cdot\big\|\big(\sum_{l\in S(k)} d_l\,\mathbf{b}_l\big)\big\|_2
\end{aligned}$$

where the inequality $(i)$ holds by the multiple applications of Cauchy-Schwarz inequality inequalities and Rayleigh quotient. For now, we define $C_{\max} := 2\Big[\max_{i,k}\sigma_{\max}\Big(\mathcal{P}_{\bar{\mathcal{M}}_i}\,\big(\frac{1}{n}X^TX\big)\,\mathcal{P}_{\bar{\mathcal{M}}_k}\Big)\Big]$. Note that $\mathcal{P}_{\bar{\mathcal{M}}_i}$ denotes the projector matrix corresponding to the structure $\alpha$ while $\mathcal{P}_{\bar{\mathcal{M}}_k}$ does to the structure $\beta$.

Using the fact that the basis vectors are orthonormal, so that $\|\mathbf{a}_j\|_2=1$ and $\Pi_{\mathbf{a}_j}(\mathbf{a}_{j'})=0$ for all $j\neq j'$, we obtain

$$\frac{2}{n}\big|\langle X\Delta_\alpha, X\Delta_\beta\rangle\big| \le C_{\max}\Big(\sum_i\big\|\big(\sum_{j\in S(i)} c_j\big)\big\|_2\Big)\Big(\sum_k\big\|\big(\sum_{l\in S(k)} d_l\big)\big\|_2\Big). \tag{28}$$

To upper-bound $\sum_i\big\|\big(\sum_{j\in S(i)} c_j\big)\big\|_2$, we can directly appeal to the following standard bound in [2].

$$\sum_{i=3}\big\|\sum_{j\in S(i)} c_j\big\|_2 \le t^{-1/2}\sum_{i=2}\sum_{j\in S(i)} c_j.$$

Since $\sum_{i=2}\sum_{j\in S(i)} c_j$ is equal to $\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big)$ by construction, we have

$$\sum_i\big\|\sum_{j\in S(i)} c_j\big\|_2 \le 2\|\Delta_\alpha\|_2 + t^{-1/2}\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big).$$

In addition, since $\Delta_\alpha$ should belong to $\mathbb{C}$ in (14),

$$\lambda_\alpha\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big) \le \sum_{\gamma\in I}\lambda_\gamma\Big[3\mathcal{R}_\gamma\big(\Pi_{\bar{\mathcal{M}}_\gamma}(\Delta_\gamma)\big) + 4\mathcal{R}_\gamma\big(\Pi_{\mathcal{M}_\gamma^\perp}(\theta_\gamma^*)\big)\Big]$$

Under the assumption $\Pi_{\mathcal{M}_\alpha^\perp}(\theta_\alpha^*)=0$ for all $\alpha\in I$ for simplicity,

$$t^{-1/2}\mathcal{R}_\alpha\big(\Pi_{\bar{\mathcal{M}}_\alpha^\perp}(\Delta_\alpha)\big) \le \sum_{\gamma\in I}\frac{3\lambda_\gamma}{\lambda_\alpha\sqrt{t}}\mathcal{R}_\gamma\big(\Pi_{\bar{\mathcal{M}}_\gamma}(\Delta_\gamma)\big) \le \sum_{\gamma\in I}\frac{3\lambda_\gamma\Psi_\gamma(\bar{\mathcal{M}}_\gamma)}{\lambda_\alpha\Psi_\alpha(\bar{\mathcal{M}}_\alpha)}\|\Delta_\gamma\|_2.$$

Let $\Lambda$ be $\max_{\gamma_1,\gamma_2}\Big\{2+\frac{3\lambda_{\gamma_1}\Psi_{\gamma_1}(\bar{\mathcal{M}}_{\gamma_1})}{\lambda_{\gamma_2}\Psi_{\gamma_2}(\bar{\mathcal{M}}_{\gamma_2})}\Big\}$. Then, we obtain

$$\sum_i \Big\| \sum_{j\in S(i)} c_j \Big\|_2 \leq \Lambda \sum_\gamma \|\Delta_\gamma\|_2.$$

Similarly,

$$\sum_k \Big\| \sum_{l\in S(k)} d_l \Big\|_2 \leq \Lambda \sum_\gamma \|\Delta_\gamma\|_2.$$

Combining all the pieces together, we obtain

$$\frac{2}{n}\big|\langle X\Delta_\alpha, X\Delta_\beta\rangle\big| \leq C_{\max}\Big(\Lambda\sum_\gamma \|\Delta_\gamma\|_2\Big)^2,$$

and hence,

$$\frac{2}{n}\Big|\sum_{\alpha<\beta}\langle X\Delta_\alpha, X\Delta_\beta\rangle\Big| \leq C_{\max}\binom{|I|}{2}\Big(\Lambda\sum_\alpha \|\Delta_\alpha\|_2\Big)^2 \leq C_{\max}\binom{|I|}{2}\Lambda^2|I|\sum_\alpha \|\Delta_\alpha\|_2^2. \tag{29}$$

In order to complete the proof, we use the following lemma to control the remaining parameter $C_{\max}$ used in the above inequality.

**Lemma 3.** *Suppose that $X$ has independent sub-Gaussian rows. Then, for any fixed $i,k$ and every $\delta \geq 0$, with probability at least $1-4\exp(-c_2\delta^2)$,*

$$\begin{aligned}&\sigma_{\max}\Big(\mathcal{P}_{\bar{\mathcal{M}}_i}\big(\frac{1}{n}X^TX\big)\mathcal{P}_{\bar{\mathcal{M}}_k}\Big)\\ &\leq \sigma_{\max}\Big(\mathcal{P}_{\bar{\mathcal{M}}_i}\Sigma\,\mathcal{P}_{\bar{\mathcal{M}}_k}\Big) + c_1\max(\eta,\eta^2)\end{aligned} \tag{30}$$

*where $\eta = \sqrt{\frac{t}{n}} + \frac{\delta}{\sqrt{n}}$, and constants $c_1, c_2$ only depend on the distribution of the rows in $X$.*

See Appendix C.1 for the proof of this claim.

In order to build a bound (30) for all $i$, $k$ and additionally for all $\alpha < \beta$, we use the standard union bound with the appropriate choice of $\delta$; let $p'$ be the $\max\{n,p\}$. By choosing $\delta = \sqrt{c_3\log p'}$ and the union bound, (30) holds with $\eta = \sqrt{\frac{t}{n}} + \sqrt{\frac{c_3\log p'}{n}}$ for all $i$ and $k$ with probability at least $1-4\exp(-c_2c_3\log p' + 2\log p + \log\binom{|I|}{2})$ for some large enough constant $c_3$ so that $1-4\exp(-c_2c_3\log p' + 2\log p) \geq 1-4\exp(-\log p') = 1-\frac{4}{p'}$.

Now, for every $i$, $k$, $\alpha$ and $\beta$, since the set $S(i)$ is the subspace of $\overline{\mathcal{M}}_\alpha$ or $\overline{\mathcal{M}}_\alpha^\perp$ and similarly $S(k)$ is that of $\overline{\mathcal{M}}_\beta$ or $\overline{\mathcal{M}}_\beta^\perp$, along with the assumption in the statement, we have

$$\sigma_{\max}\Big(\mathcal{P}_{\bar{\mathcal{M}}_i}\Sigma\,\mathcal{P}_{\bar{\mathcal{M}}_k}\Big) \leq \frac{\kappa_{\mathcal{L}}}{8\binom{|I|}{2}\Lambda^2|I|}.$$

At the same time, provided that $n$ is greater than $128\big(\frac{c_1\binom{|I|}{2}\Lambda^2|I|}{\kappa_{\mathcal{L}}}\big)^2\Big(t + c_3\log p'\Big)$, we have

$$c_1\max(\eta,\eta^2) \leq \frac{\kappa_{\mathcal{L}}}{8\binom{|I|}{2}\Lambda^2|I|}.$$

Combining two inequalities yields

$$\frac{1}{2}C_{\max} = \max_{i,k}\Big(\mathcal{P}_{\bar{\mathcal{M}}_i}\big(\frac{1}{n}X^TX\big)\mathcal{P}_{\bar{\mathcal{M}}_k}\Big) \leq \frac{\kappa_{\mathcal{L}}}{4\binom{|I|}{2}\Lambda^2|I|},$$

hence, from (29)

$$\frac{2}{n}\Big|\sum_{\alpha<\beta}\langle X\Delta_\alpha, X\Delta_\beta\rangle\Big| \leq \frac{\kappa_{\mathcal{L}}}{2}\sum_\alpha \|\Delta_\alpha\|_2^2.$$

as claimed.

### C.1 Proof of Lemma 3

For any fixed $i$ and $k$, in order to bound $\sigma_{\min}\big(X\mathcal{P}_{\bar{\mathcal{M}}_i}\big)$ and $\sigma_{\max}\big(X\mathcal{P}_{\bar{\mathcal{M}}_i}\big)$, we will appeal to the following result in non-asympototic random matrix theory (see, for example, Theorem 5.39 in [13]):

**Theorem 2.** *Let $M$ be a $n \times p$ matrix whose rows $M_i$ are independent sub-Gaussian with its the second moment $\Sigma \in \mathbb{R}^{p\times p}$. Then with probability at least $1 - 2\exp(-c_2\delta^2)$, we have,*

$$\sigma_{\min}(M) \geq \sqrt{\sigma_{\min}(\Sigma)}\Big(\sqrt{n} - c_1\sqrt{p} - \delta\Big)\ ,$$
$$\sigma_{\max}(M) \leq \sqrt{\sigma_{\max}(\Sigma)}\Big(\sqrt{n} + c_1\sqrt{p} + \delta\Big)\ ,$$

*where the constants $c_1$ and $c_2$ only depend on the distribution of the rows $M_i$.*

Therefore, we have

$$\sigma_{\min}\big(X\mathcal{P}_{\bar{\mathcal{M}}_i}\big) \geq \sqrt{\sigma_{\min}(\Sigma\,\mathcal{P}_{\bar{\mathcal{M}}_i})}\Big(\sqrt{n} - c_1\sqrt{t} - \delta\Big)\ , \tag{31}$$
$$\sigma_{\max}\big(X\mathcal{P}_{\bar{\mathcal{M}}_i}\big) \leq \sqrt{\sigma_{\max}(\Sigma\,\mathcal{P}_{\bar{\mathcal{M}}_i})}\Big(\sqrt{n} + c_1\sqrt{t} + \delta\Big)\ , \tag{32}$$

where $t$ is the number of non-zero columns of $X\mathcal{P}_{\bar{\mathcal{M}}_i}$, which is usually same as the number of atoms in the space $\bar{\mathcal{M}}_\alpha$. From (31) and (32), the difference between the mean and the sample mean can be derived by a straight forward way (we can extend Lemma 5.36 in [13]): For every $\delta \geq 0$, with probability at least $1 - 4\exp(-c_2\delta^2)$,

$$\begin{aligned}&\sigma_{\max}\Big[\mathcal{P}_{\bar{\mathcal{M}}_i}\big(\frac{1}{n}X^TX\big)\mathcal{P}_{\bar{\mathcal{M}}_k} - \mathcal{P}_{\bar{\mathcal{M}}_i}\Sigma\,\mathcal{P}_{\bar{\mathcal{M}}_k}\Big]\\ &\leq\ \max(\eta,\eta^2)\sqrt{\sigma_{\max}(\Sigma\,\mathcal{P}_{\bar{\mathcal{M}}_i})}\sqrt{\sigma_{\max}(\Sigma\,\mathcal{P}_{\bar{\mathcal{M}}_k})}\end{aligned}$$

where $\eta = c_1\sqrt{\frac{t}{n}} + \frac{\delta}{\sqrt{n}}$. Since $\sigma_{\max}(\cdot)$ is a norm, by the triangle inequality of a norm: $\|a\| - \|b\| \leq \|a - b\|$, we can finally have

$$\begin{aligned}&\sigma_{\max}\Big(\mathcal{P}_{\bar{\mathcal{M}}_i}\big(\frac{1}{n}X^TX\big)\mathcal{P}_{\bar{\mathcal{M}}_k}\Big)\\ &\leq \sigma_{\max}\Big(\mathcal{P}_{\bar{\mathcal{M}}_i}\Sigma\,\mathcal{P}_{\bar{\mathcal{M}}_k}\Big) + \max(\eta,\eta^2)\sqrt{\sigma_{\max}(\Sigma\,\mathcal{P}_{\bar{\mathcal{M}}_i})}\sqrt{\sigma_{\max}(\Sigma\,\mathcal{P}_{\bar{\mathcal{M}}_k})}.\end{aligned}$$

## D Proof of Proposition 3

The overall proof can be easily extended from the trivial connections between $\|\cdot\|_2$ and $\|\cdot\|_F$, and between $\langle\cdot,\cdot\rangle$ and $\langle\!\langle\cdot,\cdot\rangle\!\rangle$. For the matrix parameter, however, there is a caveat in proof of Lemma 3: even after the projection, the dimension of $X$ will not change unless the left singular vectors are the standard bases. However we can still show the inequalities (31) and (32) from the following reasoning:

$$\begin{aligned}&\sigma_{\max}\big(X\mathcal{P}_{\bar{\mathcal{M}}_i}\big) = \sigma_{\max}(XUU^\top) = \sigma_{\max}(XU)\\ \leq&\sqrt{\sigma_{\max}(\Sigma\,U)}\Big(\sqrt{n} + c_1\sqrt{t} + \delta\Big) = \sqrt{\sigma_{\max}(\Sigma\,\mathcal{P}_{\bar{\mathcal{M}}_i})}\Big(\sqrt{n} + c_1\sqrt{t} + \delta\Big).\end{aligned}$$

## E Proof of Corollary 2

Fist we restate the condition (D3) for individual clean structures, which have been analyzed previously in [11].

$$\begin{aligned}&\frac{1}{n}\|X\Delta_s\|_2^2 \geq \kappa_1\|\Delta_s\|_2^2 - \kappa_2\frac{\log p}{n}\|\Delta_s\|_1^2 \quad \text{for all } \Delta_s \in \mathbb{R}^p\\ &\frac{1}{n}\|X\Delta_g\|_2^2 \geq \kappa_1'\|\Delta_g\|_2^2 - \kappa_2'\mathbb{E}\Big(\Big[\max_{t=1,2,\ldots,q}\frac{\|\epsilon_{G_t}\|_{a^*}}{\sqrt{n}}\Big]\Big)^2\|\Delta_g\|_{1,a}^2 \quad \text{for all } \Delta_g \in \mathbb{R}^p\end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 n)$. Note that $\|\cdot\|_{a^*}$ is dual norm of $\|\cdot\|_a$ and $\epsilon \sim N(0, I_{p\times p})$ is a standard normal vector. As noted in [11], for $a = 2$ as a special case,

$$\frac{1}{n}\|X\Delta_g\|_2^2 \geq \kappa_1' \|\Delta_g\|_2^2 - \kappa_2' \Big(\sqrt{\frac{m}{n}} + \sqrt{\frac{3\log q}{n}}\Big)^2 \|\Delta_g\|_{1,2}^2 \,.$$

The next step is to choose the appropriate $\lambda_\alpha$ in order utilize Theorem 1. Since the dual norm of $\|\cdot\|_1$ is the infinity norm, we need to choose $\lambda_s$ for the sparsity structure such that

$$\lambda_s \geq 2\mathcal{R}_s^*\big(\nabla_{\theta_s}\mathcal{L}(\theta_g^* + \theta_s^*; Z_1^n)\big) = 4\|\frac{1}{n}X^\top w\|_\infty.$$

Under the conditions on the column normalization of $X$ and sub-Gaussian of $w$, [11] showed that $\|\frac{1}{n}X^\top w\|_\infty \leq 2\sigma\sqrt{\frac{\log p}{n}}$ with probability at least $1 - c_1\exp(-c_2 n \lambda_s^2)$.

Moreover, [11] proved that the choice of $\lambda_g$ as in the statement satisfies:,

$$\lambda_g = 8\sigma\Big\{\frac{m^{1-1/a}}{\sqrt{n}} + \sqrt{\frac{\log q}{n}}\Big\} \geq 4 \max_{t=1,2,\ldots,q} \big\|\frac{1}{n}X_{G_t}^\top w\big\|_{a^*} = 2\mathcal{R}_g^*\big(\nabla_{\theta_g}\mathcal{L}(\theta_g^* + \theta_s^*; Z_1^n)\big).$$

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

Finally, when $\theta_s^*$ is exactly $s$-sparse, we use the fact

$$\Psi_s\big(\overline{\mathcal{M}}_s(S)\big) = \sup_{\Delta \in \mathcal{M}\setminus\{0\}} \frac{\|\Delta\|_1}{\|\Delta\|_2} \leq \sqrt{s} \tag{33}$$

where $S$ is equal to the support set of $\theta_s^*$. At the same time, for a exactly group-sparse with non-zero groups $S_G$ with cardinality $s_g$, for any $\Delta \in \overline{\mathcal{M}}_g(S_G)$, we have

$$\sum_{t\in\{1,2,\ldots,q\}} \|\Delta_{G_t}\|_a = \sum_{t\in S_G} \|\Delta_{G_t}\|_a \leq \sum_{t \in S_G} \|\Delta_{G_t}\|_2 \leq \sqrt{s_g}\|\Delta\|_2 \tag{34}$$

where the first inequality holds for $a \geq 2$. Hence, the claim follows from Corollary 1.

## F Proof of Corollary 3

For $n \geq \max(p, m)$, the RSC conditions for each structure are satisfied with $g_\alpha = 0$:

$$\frac{1}{n}|||X\Delta_\alpha|||_F = \frac{1}{n}\sum_{j=1}^{p} \|(X\Delta_\alpha)j\|_2^2 \geq \frac{\sigma_{\min}(X^\top X)}{n}|||\Delta_\alpha|||_F^2.$$

Then, by Theorem 5.39 in [13], we can easily verify that

$$\frac{1}{\sqrt{n}}\sigma_{\min}(X) \geq \sqrt{\sigma_{\min}(\Sigma)}\Big(1 - c\sqrt{\frac{p}{n}} - c'\sqrt{\frac{1}{n}\log\frac{1}{\delta}}\Big) \tag{35}$$

with probability at least $1 - 2\delta$.

At the same time, the subspace compatibility constants for each structure can be easily extended from (33) and (34):

$$\Psi_s(\overline{\mathcal{M}}_s) = \sup_{\Delta\in\mathcal{M}\setminus\{0\}} \frac{\|\Delta\|_1}{|||\Delta|||_F} \leq \sqrt{s}\,,$$

$$\Psi_r(\overline{\mathcal{M}}_r) = \sum_{\Delta\in\mathcal{M}\setminus\{0\}} \frac{\|\Delta\|_{r,a}}{|||\Delta|||_F} \leq \sqrt{s_r}\,,$$

$$\Psi_c(\overline{\mathcal{M}}_c) = \sum_{\Delta\in\mathcal{M}\setminus\{0\}} \frac{\|\Delta\|_{c,a}}{|||\Delta|||_F} \leq \sqrt{s_c}\,.$$

Hence, in order to leverage Corollary 1, it remains to show $\lambda_\alpha \geq 2\mathcal{R}_\alpha^*\big(\nabla_{\theta_\alpha}\mathcal{L}(\theta^*; Z_1^n)\big)$ for each structure. Under the conditions of $\frac{\|X_j\|_2}{\sqrt{n}} \leq 1$, $\sigma_{\max}(X) \leq \sqrt{n}$, and the sub-Gaussian observation noise, we can simply extend the results in [11]:

$$\lambda_s = 8\sigma\sqrt{\frac{\log p + \log m}{n}} \geq 4\|\frac{1}{n}X^\top W\|_\infty = 2\mathcal{R}_s^*\big(\nabla_{\Theta_s}\mathcal{L}(\Theta_r^* + \Theta_c^* + \Theta_s^*; Z_1^n)\big)$$

with probability at least $1 - c_1\exp(-c_2(\log p + \log m))$. Moreover,

$$\lambda_r = 8\sigma\Big\{\frac{m^{1-1/a}}{\sqrt{n}} + \sqrt{\frac{\log p}{n}}\Big\} \geq 4\max_{t=1,2,\ldots,p}\Big\|\frac{1}{n}X^\top W_{t,*}\Big\|_{a^*} = 2\mathcal{R}_r^*\big(\nabla_{\Theta_r}\mathcal{L}(\Theta_r^* + \Theta_c^* + \Theta_s^*; Z_1^n)\big).$$

with probability at least $1 - 2\exp(-2\log p)$. Similarly,

$$\lambda_c = 8\sigma\Big\{\frac{p^{1-1/a}}{\sqrt{n}} + \sqrt{\frac{\log m}{n}}\Big\} \geq 4\max_{t=1,2,\ldots,m}\Big\|\frac{1}{n}X^\top W_{*,t}\Big\|_{a^*} = 2\mathcal{R}_c^*\big(\nabla_{\Theta_c}\mathcal{L}(\Theta_r^* + \Theta_c^* + \Theta_s^*; Z_1^n)\big).$$

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

## G Proof of Corollary 4

Since the PCA model can be understood as the special case of (7) with $X = I_{p\times p}$, the restricted strong convexities for both structures are trivially satisfied with $\kappa_{\mathcal{L}} = 1$ and $g_\alpha = 0$.

As in the previous case, the subspace compatibility can be easily derived as:

$$\Psi_\Theta(\overline{\mathcal{M}}_\Theta) = \sup_{\Delta\in\mathcal{M}\setminus\{0\}} \frac{|||\Delta|||_1}{|||\Delta|||_F} \leq \sqrt{r}\,,$$
$$\Psi_\Gamma(\overline{\mathcal{M}}_\Gamma) = \sup_{\Delta\in\mathcal{M}\setminus\{0\}} \frac{\|\Delta\|_1}{|||\Delta|||_F} \leq \sqrt{s}.$$

Hence, it again remains to show $\lambda_\alpha \geq 2\mathcal{R}_\alpha^*\big(\nabla_{\theta_\alpha}\mathcal{L}(\theta^*; Z_1^n)\big)$ for each structure; we can directly appeal to the results for clean models [1]:

$$\lambda_\Theta = 16\sqrt{|||\Sigma|||_2}\sqrt{\frac{p}{n}} \geq 4|||W|||_2 = 2\mathcal{R}_\Theta^*\big(\nabla_\Theta\mathcal{L}(\Theta + \Gamma; Z_1^n)\big)$$

with probability at least $1 - 2\exp(-c_1 p)$. Also,

$$\lambda_\Gamma = 32\rho(\Sigma)\sqrt{\frac{\log p}{n}} \geq 4\|W\|_\infty = 2\mathcal{R}_\Gamma^*\big(\nabla_\Gamma\mathcal{L}(\Theta + \Gamma; Z_1^n)\big)$$

with probability at least $1 - 2\exp(-c_2\log p)$, which completes the proof.