Submitted to the Annals of Statistics

SUPPLEMENTARY MATERIAL: FAST GLOBAL CONVERGENCE OF GRADIENT METHODS FOR HIGH-DIMENSIONAL STATISTICAL RECOVERY

By Alekh Agarwal[⋆] and Sahand Negahban[‡] and Martin J. Wainwright[†]

UC Berkeley, Department of EECS[⋆,†] and Statistics[†]
MIT, Department of EECS[‡]

In this supplement, we provide the proofs of our main results and their corollaries in Section 7. The more technical arguments are deferred to the appendices.

7. Proofs. Recall that we use $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ to denote the optimization error, and $\Delta^* = \widehat{\theta} - \theta^*$ to denote the statistical error. For future reference, we point out a slight weakening of restricted strong convexity (RSC), useful for obtaining parts of our results. As the proofs to follow reveal, it is only necessary to enforce an RSC condition of the form

$$\mathcal{T}_{\mathcal{L}}(\theta^t; \widehat{\theta}) \geq \frac{\gamma_\ell}{2} \|\theta^t - \widehat{\theta}\|^2 - \tau_\ell(\mathcal{L}_n)\, \mathcal{R}^2(\theta^t - \widehat{\theta}) - \delta^2, \tag{47}$$

which is milder than the original RSC condition (8), in that it applies only to differences of the form $\theta^t - \widehat{\theta}$, and allows for additional slack δ. We make use of this refined notion in the proofs of various results to follow.

With this relaxed RSC condition and the same RSM condition as before, our proof shows that

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t \|\theta^0 - \widehat{\theta}\|^2 + \frac{\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) + 2\delta^2/\gamma_u}{1-\kappa} \tag{48}$$

for all iterations $t = 0, 1, 2, \ldots$. Note that this result reduces to the previous statement when $\delta = 0$. This extension of Theorem 1 is used in the proofs of Corollaries 5 and 6.

We will assume without loss of generality that all the iterates lie in the subset Ω' of Ω. This can be ensured by augmenting the loss with the indicator of Ω' or equivalently performing projections on the set $\Omega' \cap \mathbb{B}_{\mathcal{R}}(\rho)$ as mentioned earlier.

7.1. *Proof of Theorem 1.* Recall that Theorem 1 concerns the constrained problem (1). The proof is based on two technical lemmas. The first lemma guarantees that at each iteration $t = 0, 1, 2, \ldots$, the optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$ belongs to an interesting constraint set defined by the regularizer.

LEMMA 1. *Let $\widehat{\theta}$ be any optimum of the constrained problem* (1) *for which $\mathcal{R}(\widehat{\theta}) = \rho$. Then for any iteration $t = 1, 2, \ldots$ and for any $\mathcal{R}$-decomposable subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ satisfies the inequality*

$$\mathcal{R}(\widehat{\Delta}^t) \leq 2\,\Psi(\overline{\mathcal{M}})\,\|\widehat{\Delta}^t\| + 2\mathcal{R}(\Pi_{\mathcal{M}^{\perp}}(\theta^*)) + 2\mathcal{R}(\Delta^*) + \Psi(\overline{\mathcal{M}})\|\Delta^*\|. \tag{49}$$

For future reference, we use $\mathbb{S}(\mathcal{M}; \overline{\mathcal{M}}; \theta^*)$ to denote the set of all $\Delta \in \Omega$ for which the inequality (49) holds. The proof of this lemma, provided in Appendix A.1, exploits the decomposability of the regularizer in an essential way.

The structure of the set $\mathbb{S}(\mathcal{M}; \overline{\mathcal{M}}; \theta^*)$ takes a simpler form in the special case when $\mathcal{M}$ is chosen to contain θ^* and $\overline{\mathcal{M}} = \mathcal{M}$. In this case, we have $\mathcal{R}(\Pi_{\mathcal{M}^{\perp}}(\theta^*)) = 0$, and hence the optimization error $\widehat{\Delta}^t$ satisfies the inequality

$$\mathcal{R}(\widehat{\Delta}^t) \leq 2\,\Psi(\mathcal{M})\,\{\|\widehat{\Delta}^t\| + \|\Delta^*\|\} + 2\mathcal{R}(\Delta^*). \tag{50}$$

An inequality of this type, when combined with the definitions of RSC/RSM, allows us to establish the curvature conditions required to prove globally geometric rates of convergence.

We now state a second lemma under the more general RSC condition (47):

LEMMA 2. *Under the RSC condition* (47) *and RSM condition* (10)*, for all $t = 0, 1, 2, \ldots$, we have*

$$\begin{aligned} &\gamma_u\,\langle \theta^t - \theta^{t+1},\, \theta^t - \widehat{\theta}\rangle \\ &\geq \Big\{\frac{\gamma_u}{2}\|\theta^t - \theta^{t+1}\|^2 - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t)\Big\} + \Big\{\frac{\gamma_\ell}{2}\|\theta^t - \widehat{\theta}\|^2 - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\theta^t - \widehat{\theta}) - \delta^2\Big\}. \end{aligned} \tag{51}$$

The proof of this lemma, provided in Appendix A.2, follows along the lines of the intermediate result within Theorem 2.2.8 of Nesterov [31], but with some care required to handle the additional terms that arise in our weakened forms of strong convexity and smoothness.

Using these auxiliary results, let us now complete the the proof of Theorem 1. We first note the elementary relation

$$\begin{aligned} \|\theta^{t+1} - \widehat{\theta}\|^2 &= \|\theta^t - \widehat{\theta} - \theta^t + \theta^{t+1}\|^2 \\ &= \|\theta^t - \widehat{\theta}\|^2 + \|\theta^t - \theta^{t+1}\|^2 - 2\langle \theta^t - \widehat{\theta}, \, \theta^t - \theta^{t+1} \rangle. \end{aligned} \tag{52}$$

We now use Lemma 2 and the more general form of RSC (47) to control the cross-term, thereby obtaining that $\|\theta^{t+1} - \widehat{\theta}\|^2$ is upper bounded by

$$\begin{aligned} &\|\theta^t - \widehat{\theta}\|^2 - \frac{\gamma_\ell}{\gamma_u}\|\theta^t - \widehat{\theta}\|^2 + \frac{2\tau_u(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\theta^{t+1} - \theta^t) + \frac{2\tau_\ell(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\theta^t - \widehat{\theta}) + \frac{2\delta^2}{\gamma_u} \\ &= \big(1 - \frac{\gamma_\ell}{\gamma_u}\big)\|\theta^t - \widehat{\theta}\|^2 + \frac{2\tau_u(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\theta^{t+1} - \theta^t) + \frac{2\tau_\ell(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\theta^t - \widehat{\theta}) + \frac{2\delta^2}{\gamma_u}. \end{aligned}$$

We now observe that by triangle inequality and the Cauchy-Schwarz inequality,

$$\begin{aligned} \mathcal{R}^2(\theta^{t+1} - \theta^t) &\leq \big(\mathcal{R}(\theta^{t+1} - \widehat{\theta}) + \mathcal{R}(\widehat{\theta} - \theta^t)\big)^2 \\ &\leq 2\mathcal{R}^2(\theta^{t+1} - \widehat{\theta}) + 2\mathcal{R}^2(\theta^t - \widehat{\theta}). \end{aligned}$$

Recalling the definition of the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$, we find that $\|\widehat{\Delta}^{t+1}\|^2$ is upper bounded by

$$\big(1 - \frac{\gamma_\ell}{\gamma_u}\big)\|\widehat{\Delta}^t\|^2 + \frac{4\tau_u(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\widehat{\Delta}^{t+1}) + \frac{4\tau_u(\mathcal{L}_n) + 2\tau_\ell(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\widehat{\Delta}^t) + \frac{2\delta^2}{\gamma_u}. \tag{53}$$

We now apply Lemma 1 to control the terms involving $\mathcal{R}^2$. In terms of squared quantities, the inequality (49) implies that

$$\mathcal{R}^2(\widehat{\Delta}^t) \leq 4\,\Psi^2(\overline{\mathcal{M}}^\perp)\,\|\widehat{\Delta}^t\|^2 + 2\nu^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \qquad \text{for all } t = 0, 1, 2, \ldots,$$

where we recall that $\Psi^2(\overline{\mathcal{M}}^\perp)$ is the subspace compatibility (12) and $\nu^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ accumulates all the residual terms. Applying this bound twice—once for t and once for $t+1$—and substituting into equation (53) yields that

$$\begin{aligned} \big\{1 - \frac{16\Psi^2(\overline{\mathcal{M}}^\perp)\tau_u(\mathcal{L}_n)}{\gamma_u}\big\}\|\Delta^{t+1}\|^2 &\leq \Big\{1 - \frac{\gamma_\ell}{\gamma_u} + \frac{16\Psi^2(\overline{\mathcal{M}}^\perp)\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)}{\gamma_u}\Big\}\|\Delta^t\|^2 \\ &\quad + \frac{16\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)\nu^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})}{\gamma_u} + \frac{2\delta^2}{\gamma_u}. \end{aligned}$$

Under the assumptions of Theorem 1, we are guaranteed that $\frac{16\Psi^2(\overline{\mathcal{M}}^{\perp})\tau_u(\mathcal{L}_n)}{\gamma_u} < 1/2$, and so we can re-arrange this inequality into the form

$$\|\Delta^{t+1}\|^2 \leq \kappa \|\Delta^t\|^2 + \epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) + \frac{2\delta^2}{\gamma_u} \tag{54}$$

where κ and $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ were previously defined in equations (21) and (22) respectively. Iterating this recursion yields

$$\|\Delta^{t+1}\|^2 \leq \kappa^t \|\Delta^0\|^2 + \left(\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) + \frac{2\delta^2}{\gamma_u}\right)\Big(\sum_{j=0}^{t} \kappa^j\Big).$$

The assumptions of Theorem 1 guarantee that $\kappa \in (0, 1)$, so that summing the geometric series yields the claim (24).

7.2. *Proof of Theorem 2.* The Lagrangian version of the optimization program is based on solving the convex program (2), with the objective function $\phi(\theta) = \mathcal{L}_n(\theta) + \lambda_n \mathcal{R}(\theta)$. Our proof is based on analyzing the error $\phi(\theta^t) - \phi(\widehat{\theta})$ as measured in terms of this objective function. It requires two technical lemmas, both of which are stated in terms of a given tolerance $\bar{\eta} > 0$, and an integer $T > 0$ such that

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \bar{\eta} \qquad \text{for all } t \geq T. \tag{55}$$

Our first technical lemma is analogous to Lemma 1, and restricts the optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$ to a cone-like set.

LEMMA 3 (Iterated Cone Bound (ICB)). *Let $\widehat{\theta}$ be any optimum of the regularized M-estimator* (2). *Under condition* (55) *with parameters $(T, \bar{\eta})$, for any iteration $t \geq T$ and for any $\mathcal{R}$-decomposable subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ satisfies*

$$\mathcal{R}(\widehat{\Delta}^t) \leq 4\Psi(\overline{\mathcal{M}})\|\widehat{\Delta}^t\| + 8\Psi(\overline{\mathcal{M}})\|\Delta^*\| + 8\mathcal{R}(\Pi_{\mathcal{M}^{\perp}}(\theta^*)) + 2\min\left(\frac{\bar{\eta}}{\lambda_n}, \bar{\rho}\right) \tag{56}$$

Our next lemma guarantees sufficient decrease of the objective value difference $\phi(\theta^t) - \phi(\widehat{\theta})$. Lemma 3 plays a crucial role in its proof. Recall the definition (27) of the compound contraction coefficient $\kappa(\mathcal{L}_n; \overline{\mathcal{M}})$, defined in terms of the related quantities $\xi(\overline{\mathcal{M}})$ and $\beta(\overline{\mathcal{M}})$. Throughout the proof, we drop the arguments of κ, ξ and β so as to ease notation.

LEMMA 4. *Under the RSC* (47) *and RSM conditions* (10)*, as well as assumption* (55) *with parameters* $(\bar{\eta}, T)$*, for all* $t \geq T$*, we have*

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \kappa^{t-T}(\phi(\theta^T) - \phi(\widehat{\theta})) + \frac{2}{1-\kappa}\xi(\mathcal{M})\ \beta(\mathcal{M})(\varepsilon^2 + \bar{\epsilon}_{stat}^2),$$

where $\varepsilon := 2\min(\bar{\eta}/\lambda_n, \bar{\rho})$ *and* $\bar{\epsilon}_{stat} := 8\Psi(\overline{\mathcal{M}})\|\Delta^*\| + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))$.

We are now in a position to prove our main theorem, in particular via a recursive application of Lemma 4. At a high level, we divide the iterations $t = 0, 1, 2, \ldots$ into a series of disjoint epochs $[T_k, T_{k+1})$ with

$$0 = T_0 \leq T_1 \leq T_2 \leq \cdots .$$

Moreover, we define an associated sequence of tolerances $\bar{\eta}_0 > \bar{\eta}_1 > \cdots$ such that at the end of epoch $[T_{k-1}, T_k)$, the optimization error has been reduced to $\bar{\eta}_k$. Our analysis guarantees that $\phi(\theta^t) - \phi(\widehat{\theta}) \leq \bar{\eta}_k$ for all $t \geq T_k$, allowing us to apply Lemma 4 with smaller and smaller values of $\bar{\eta}$ until it reduces to the statistical error $\bar{\epsilon}_{\text{stat}}$.

At the first iteration, we have no *a priori* bound on the error $\bar{\eta}_0 = \phi(\theta^0) - \phi(\widehat{\theta})$. However, since Lemma 4 involves the quantity $\varepsilon = \min(\bar{\eta}/\lambda_n, \bar{\rho})$, we may still apply it[1] at the first epoch with $\varepsilon_0 = \bar{\rho}$ and $T_0 = 0$. In this way, we conclude that for all $t \geq 0$,

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \kappa^t(\phi(\theta^0) - \phi(\widehat{\theta})) + \frac{2}{1-\kappa}\xi\beta(\bar{\rho}^2 + \bar{\epsilon}_{\text{stat}}^2).$$

Now since the contraction coefficient $\kappa \in (0, 1)$, for all iterations

$$t \geq T_1 := (\lceil \log(2\,\bar{\eta}_0/\bar{\eta}_1)/\log(1/\kappa) \rceil)_+,$$

we are guaranteed that

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \underbrace{\frac{4\,\xi\beta}{1-\kappa}(\bar{\rho}^2 + \bar{\epsilon}_{\text{stat}}^2)}_{\bar{\eta}_1} \leq \frac{8\xi\beta}{1-\kappa}\max(\bar{\rho}^2, \bar{\epsilon}_{\text{stat}}^2).$$

This same argument can now be applied in a recursive manner. Suppose that for some $k \geq 1$, we are given a pair $(\bar{\eta}_k, T_k)$ such that condition (55) holds. An application of Lemma 4 yields the bound

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \kappa^{t-T_k}(\phi(\theta^{T_k}) - \phi(\widehat{\theta})) + \frac{2\,\xi\beta}{1-\kappa}(\varepsilon_k^2 + \bar{\epsilon}_{\text{stat}}^2) \qquad \text{for all } t \geq T_k.$$

[1]It is for precisely this reason that our regularized M-estimator includes the additional side-constraint defined in terms of $\bar{\rho}$.

We now define $\bar{\eta}_{k+1} := \frac{4\xi\beta}{1-\kappa}(\varepsilon_k^2 + \bar{\epsilon}_{\text{stat}}^2)$. Once again, since $\kappa < 1$ by assumption, we can choose $T_{k+1} := \lceil \log(2\bar{\eta}_k/\bar{\eta}_{k+1})/\log(1/\kappa) \rceil + T_k$, thereby ensuring that for all $t \geq T_{k+1}$, we have

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \frac{8\xi\beta}{1-\kappa} \max(\varepsilon_k^2, \bar{\epsilon}_{\text{stat}}^2).$$

In this way, we arrive at recursive inequalities involving the tolerances $\{\bar{\eta}_k\}_{k=0}^{\infty}$ and time steps $\{T_k\}_{k=0}^{\infty}$—namely

$$\text{(57a)} \quad \bar{\eta}_{k+1} \leq \frac{8\,\xi\beta}{1-\kappa} \max(\varepsilon_k^2, \bar{\epsilon}_{\text{stat}}^2), \qquad \text{where } \varepsilon_k = 2\,\min\{\bar{\eta}_k/\lambda_n, \bar{\rho}\}, \text{ and}$$

$$\text{(57b)} \quad T_k \leq k + \frac{\log(2^k \bar{\eta}_0/\bar{\eta}_k)}{\log(1/\kappa)}.$$

Now we claim that the recursion (57a) can be unwrapped so as to show that

$$\text{(58)} \quad \bar{\eta}_{k+1} \leq \frac{\bar{\eta}_k}{4^{2^{k-1}}} \quad \text{and} \quad \frac{\bar{\eta}_{k+1}}{\lambda_n} \leq \frac{\bar{\rho}}{4^{2^k}} \qquad \text{for all } k = 1, 2, \ldots.$$

Taking these statements as given for the moment, let us now show how they can be used to upper bound the smallest k such that $\bar{\eta}_k \leq \delta^2$. If we are in the first epoch, the claim of the theorem is straightforward from equation (57a). If not, we first use the recursion (58) to upper bound the number of epochs needed and then use the inequality (57b) to obtain the stated result on the total number of iterations needed. Using the second inequality in the recursion (58), we see that it is sufficient to ensure that $\frac{\bar{\rho}\lambda_n}{4^{2^{k-1}}} \leq \delta^2$. Rearranging this inequality, we find that the error drops below δ^2 after at most

$$k_\delta \;\geq\; \log\left(\log\left(\frac{\bar{\rho}\lambda_n}{\delta^2}\right)/\log(4)\right)/\log(2) + 1 = \log_2 \log_2\left(\frac{\bar{\rho}\lambda_n}{\delta^2}\right)$$

epochs. Combining the above bound on k_δ with the recursion 57b, we conclude that the inequality $\phi(\theta^t) - \phi(\widehat{\theta}) \leq \delta^2$ is guaranteed to hold for all iterations

$$t \geq k_\delta\left(1 + \frac{\log 2}{\log(1/\kappa)}\right) + \frac{\log \frac{\bar{\eta}_0}{\delta^2}}{\log(1/\kappa)},$$

which is the desired result.

It remains to prove the recursion (58), which we do via induction on the index k. We begin with base case $k = 1$. Recalling the setting of $\bar{\eta}_1$ and

our assumption on λ_n in the theorem statement (30), we are guaranteed that $\bar{\eta}_1/\lambda_n \leq \bar{\rho}/4$, so that $\varepsilon_1 \leq \varepsilon_0 = \bar{\rho}$. By applying equation (57a) with $\varepsilon_1 = 2\bar{\eta}_1/\lambda_n$ and assuming $\varepsilon_1 \geq \bar{\epsilon}_{\text{stat}}$, we obtain

$$\bar{\eta}_2 \;\leq\; \frac{32\xi\beta\bar{\eta}_1^2}{(1-\kappa)\lambda_n^2} \overset{(i)}{\leq} \frac{32\xi\beta\bar{\rho}\bar{\eta}_1}{(1-\kappa)4\lambda_n} \overset{(ii)}{\leq} \frac{\bar{\eta}_1}{4}, \tag{59}$$

where step (i) uses the fact that $\frac{\bar{\eta}_1}{\lambda_n} \leq \frac{\bar{\rho}}{4}$, and step (ii) uses the condition (30) on λ_n. We have thus verified the first inequality (58) for $k = 1$. Turning to the second inequality in the statement (58), using equation 59, we have

$$\frac{\bar{\eta}_2}{\lambda_n} \;\leq\; \frac{\bar{\eta}_1}{4\lambda_n} \overset{(iii)}{\leq} \frac{\bar{\rho}}{16},$$

where step (iii) follows from the assumption (30) on λ_n. Turning to the inductive step, we again assume that $2\bar{\eta}_k/\lambda_n \geq \bar{\epsilon}_{\text{stat}}$ and obtain from inequality (57a)

$$\bar{\eta}_{k+1} \leq \frac{32\xi\beta\bar{\eta}_k^2}{(1-\kappa)\lambda_n^2} \overset{(iv)}{\leq} \frac{32\xi\beta\bar{\eta}_k\bar{\rho}}{(1-\kappa)\lambda_n 4^{2^{k-1}}} \overset{(v)}{\leq} \frac{\bar{\eta}_k}{4^{2^{k-1}}}.$$

Here step (iv) uses the second inequality of the inductive hypothesis (58) and step (v) is a consequence of the condition on λ_n as before. The second part of the induction is similarly established, completing the proof.

7.3. *Proof of Corollary 1.* In order to prove this claim, we must show that $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, as defined in equation (22), is of order lower than $\mathbb{E}[\|\widehat{\theta} - \theta^*\|^2] = \mathbb{E}[\|\Delta^*\|^2]$. We make use of the following lemma, proved in Appendix C:

LEMMA 5. *If $\rho \leq \mathcal{R}(\theta^*)$, then for any solution $\widehat{\theta}$ of the constrained problem* (1) *and any $\mathcal{R}$-decomposable subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, the statistical error $\Delta^* = \widehat{\theta} - \theta^*$ satisfies the inequality*

$$\mathcal{R}(\Delta^*) \leq 2\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\| + \mathcal{R}(\Pi_{\mathcal{M}^{\perp}}(\theta^*)). \tag{60}$$

Using this lemma, we can complete the proof of Corollary 1. Recalling the form (22), under the condition $\theta^* \in \mathcal{M}$, we have

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) := \frac{32\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)\,\big(2\mathcal{R}(\Delta^*) + \Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|\big)^2}{\gamma_u}.$$

Using the assumption $\frac{(\tau_u(\mathcal{L}_n)+\tau_\ell(\mathcal{L}_n))\Psi^2(\overline{\mathcal{M}}^{\perp})}{\gamma_u} = o(1)$, it suffices to show that $\mathcal{R}(\Delta^*) \leq 2\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|$. Since Corollary 1 assumes that $\theta^* \in \mathcal{M}$ and hence that $\Pi_{\mathcal{M}^{\perp}}(\theta^*) = 0$, Lemma 5 implies that $\mathcal{R}(\Delta^*) \leq 2\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|$, as required.

7.4. *Proofs of Corollaries 2 and 3.* The central challenge in proving this result is verifying that suitable forms of the RSC and RSM conditions hold with sufficiently small parameters $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$.

LEMMA 6. *Define the maximum variance* $\zeta(\Sigma) := \max_{j=1,2,\ldots,d} \Sigma_{jj}$. *Under the conditions of Corollary 2, there are universal positive constants* (c_0, c_1) *such that for all* $\Delta \in \mathbb{R}^d$, *we have*

$$\frac{\|X\Delta\|_2^2}{n} \geq \frac{1}{2}\|\Sigma^{1/2}\Delta\|_2^2 - c_1\zeta(\Sigma)\frac{\log d}{n}\,\|\Delta\|_1^2, \qquad \textit{and} \tag{61a}$$

$$\frac{\|X\Delta\|_2^2}{n} \leq 2\|\Sigma^{1/2}\Delta\|_2^2 + c_1\zeta(\Sigma)\frac{\log d}{n}\,\|\Delta\|_1^2, \tag{61b}$$

with probability at least $1 - \exp(-c_0\, n)$.

Note that this lemma implies that the RSC and RSM conditions both hold with high probability, in particular with parameters

$$\gamma_\ell = \frac{1}{2}\sigma_{\min}(\Sigma), \text{ and} \quad \tau_\ell(\mathcal{L}_n) = c_1\zeta(\Sigma)\frac{\log d}{n}, \qquad \text{for RSC, and}$$

$$\gamma_u = 2\sigma_{\max}(\Sigma) \text{ and} \quad \tau_u(\mathcal{L}_n) = c_1\zeta(\Sigma)\frac{\log d}{n} \qquad \text{for RSM.}$$

This lemma has been proved by Raskutti et al. [34] for obtaining minimax rates in sparse linear regression.

Let us first prove Corollary 2 in the special case of hard sparsity ($q = 0$), in which θ^* is supported on a subset S of cardinality s. Let us define the model subspace $\mathcal{M} := \big\{\theta \in \mathbb{R}^d \mid \theta_j = 0 \quad \text{for all } j \notin S\big\}$, so that $\theta^* \in \mathcal{M}$. Recall from Section 2.4.1 that the ℓ_1-norm is decomposable with respect to $\mathcal{M}$ and $\mathcal{M}^\perp$; as a consequence, we may also set $\overline{\mathcal{M}}^\perp = \mathcal{M}$ in the definitions (21) and (22). By definition (12) of the subspace compatibility between with ℓ_1-norm as the regularizer, and ℓ_2-norm as the error norm, we have $\Psi^2(\mathcal{M}) = s$. Using the settings of $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$ guaranteed by Lemma 6 and substituting into equation (21), we obtain a contraction coefficient

$$\kappa(\Sigma) := \Big\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\Big\}\,\Big\{1 - \chi_n(\Sigma)\Big\}^{-1}, \tag{62}$$

where $\chi_n(\Sigma) := \frac{c_2\zeta(\Sigma)}{\sigma_{\max}(\Sigma)}\,\frac{s\log d}{n}$ for some universal constant c_2. A similar calculation shows that the tolerance term takes the form

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3\,\chi_n(\Sigma)\Big\{\frac{\|\Delta^*\|_1^2}{s} + \|\Delta^*\|_2^2\Big\} \qquad \text{for some constant } c_3.$$

Since $\rho \leq \|\theta^*\|_1$, then Lemma 5 (as exploited in the proof of Corollary 1) shows that $\|\Delta^*\|_1^2 \leq 4s\|\Delta^*\|_2^2$, and hence that $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3\, \chi_n(\Sigma)\, \|\Delta^*\|_2^2$. This completes the proof of the claim (36) for $q = 0$.

We now turn to the case $q \in (0, 1]$, for which we bound the term $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ using a slightly different choice of the subspace pair $\mathcal{M}$ and $\overline{\mathcal{M}}^{\perp}$. For a truncation level $\mu > 0$ to be chosen, define the set

$$S_\mu := \big\{ j \in \{1, 2, \ldots, d\} \mid |\theta^*_j| > \mu \big\},$$

as well as the associated subspaces $\mathcal{M} = \mathcal{M}(S_\mu)$ and $\overline{\mathcal{M}}^{\perp} = \mathcal{M}^{\perp}(S_\mu)$. By combining Lemma 5 and the definition (22) of $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, for any pair $(\mathcal{M}(S_\mu), \mathcal{M}^{\perp}(S_\mu))$, we have

$$\epsilon^2(\Delta^*; \mathcal{M}, \mathcal{M}^{\perp}) \leq \frac{c\,\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{\log d}{n} \big(\|\Pi_{\mathcal{M}^{\perp}}(\theta^*)\|_1 + \sqrt{|S_\mu|}\, \|\Delta^*\|_2 \big)^2,$$

where to simplify notation, we have omitted the dependence of $\mathcal{M}$ and $\mathcal{M}^{\perp}$ on S_μ. We now choose the threshold μ optimally, so as to trade-off the term $\|\Pi_{\mathcal{M}^{\perp}}(\theta^*)\|_1$, which decreases as μ increases, with the term $\sqrt{S_\mu}\|\Delta^*\|_2$, which increases as μ increases.

By definition of $\mathcal{M}^{\perp}(S_\mu)$, we have

$$\|\Pi_{\mathcal{M}^{\perp}}(\theta^*)\|_1 = \sum_{j \notin S_\mu} |\theta^*_j| \;=\; \mu \sum_{j \notin S_\mu} \frac{|\theta^*_j|}{\mu} \;\leq\; \mu \sum_{j \notin S_\mu} \left(\frac{|\theta^*_j|}{\mu} \right)^q,$$

where the inequality holds since $|\theta^*_j| \leq \mu$ for all $j \notin S_\mu$. Now since $\theta^* \in \mathbb{B}_q(R_q)$, we conclude that

$$\|\Pi_{\mathcal{M}^{\perp}}(\theta^*)\|_1 \leq \mu^{1-q} \sum_{j \notin S_\mu} |\theta^*_j|^q \;\leq \mu^{1-q} R_q. \tag{63}$$

On the other hand, again using the inclusion $\theta^* \in \mathbb{B}_q(R_q)$, we have

$$R_q \geq \sum_{j \in S_\mu} |\theta^*_j|^q \;\geq\; |S_\mu|\, \mu^q,$$

which implies that $|S_\mu| \leq \mu^{-q} R_q$. By combining this bound with inequality (63), we obtain the upper bound

$$\begin{aligned} \epsilon^2(\Delta^*; \mathcal{M}, \mathcal{M}^{\perp}) &\leq \frac{c\,\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{\log d}{n} \big(\mu^{2-2q} R_q^2 + \mu^{-q} R_q \|\Delta^*\|_2^2 \big) \\ &= \frac{c\,\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{\log d}{n} \mu^{-q} R_q \big(\mu^{2-q} R_q + \|\Delta^*\|_2^2 \big). \end{aligned}$$

Setting $\mu^2 = \frac{\log d}{n}$ then yields

$$\epsilon^2(\Delta^*; \mathcal{M}, \mathcal{M}^\perp) \leq \chi_n(\Sigma) \left\{ R_q \big(\frac{\log d}{n}\big)^{1-q/2} + \|\Delta^*\|_2^2 \right\},$$

where $\chi_n(\Sigma) := \frac{c\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} R_q \big(\frac{\log d}{n}\big)^{1-q/2}$.

Finally, let us verify the stated form of the contraction coefficient. For the given subspace $\overline{\mathcal{M}}^\perp = \mathcal{M}(S_\mu)$ and choice of μ, we have $\Psi^2(\overline{\mathcal{M}}^\perp) = |S_\mu| \leq \mu^{-q} R_q$. From Lemma 6, we have

$$16\Psi^2(\overline{\mathcal{M}}^\perp) \frac{\tau_\ell(\mathcal{L}_n) + \tau_u(\mathcal{L}_n)}{\gamma_u} \leq \chi_n(\Sigma),$$

and hence, by definition (21) of the contraction coefficient,

$$\kappa \leq \left\{ 1 - \frac{\gamma_\ell}{2\gamma_u} + \chi_n(\Sigma) \right\} \left\{ 1 - \chi_n(\Sigma) \right\}^{-1}.$$

For proving Corollary 3, we observe that the stated settings $\overline{\gamma_\ell}$, $\chi_n(\Sigma)$ and κ follow directly from Lemma 6. The bound for condition 2(a) follows from a standard argument about the suprema of d independent Gaussians with variance ν.

7.5. *Proof of Corollary 4.* This proof is analogous to that of Corollary 2, but appropriately adapted to the matrix setting. We first state a lemma that allows us to establish appropriate forms of the RSC/RSM conditions. Recall that we are studying an instance of matrix regression with random design, where the vectorized form $\text{vec}(X)$ of each matrix is drawn from a $N(0, \Sigma)$ distribution, where $\Sigma \in \mathbb{R}^{d^2 \times d^2}$ is some covariance matrix. In order to state this result, let us define the quantity

$$\zeta_{\text{mat}}(\Sigma) := \sup_{\|u\|_2=1,\, \|v\|_2=1} \text{var}(u^T X v), \quad \text{where } \text{vec}(X) \sim N(0, \Sigma). \tag{64}$$

LEMMA 7. *Under the conditions of Corollary 4, there are universal positive constants* (c_0, c_1) *such that*

$$\frac{\|\mathfrak{X}_n(\Delta)\|_2^2}{n} \geq \frac{1}{2}\, \sigma_{\min}(\Sigma)\, |||\Delta|||_F^2 - c_1 \zeta_{mat}(\Sigma) \frac{d}{n}\, |||\Delta|||_1^2, \qquad \textit{and} \tag{65a}$$

$$\frac{\|\mathfrak{X}_n(\Delta)\|_2^2}{n} \leq 2\, \sigma_{\max}(\Sigma)\, |||\Delta|||_F^2 - c_1\, \zeta_{mat}(\Sigma)\, \frac{d}{n}\, |||\Delta|||_1^2, \qquad \textit{for all } \Delta \in \mathbb{R}^{d \times d}. \tag{65b}$$

with probability at least $1 - \exp(-c_0\, n)$.

Given the quadratic nature of the least-squares loss, the bound (65a) implies that the RSC condition holds with $\gamma_\ell = \frac{1}{2}\sigma_{\min}(\Sigma)$ and $\tau_\ell(\mathcal{L}_n) = c_1\zeta_{\text{mat}}(\Sigma)\frac{d}{n}$, whereas the bound (65b) implies that the RSM condition holds with $\gamma_u = 2\sigma_{\max}(\Sigma)$ and $\tau_u(\mathcal{L}_n) = c_1\zeta_{\text{mat}}(\Sigma)\frac{d}{n}$.

We now prove Corollary 4 in the special case of exactly low rank matrices ($q = 0$), in which Θ^* has some rank $r \leq d$. Given the singular value decomposition $\Theta^* = UDV^T$, let U^r and V^r be the $d \times r$ matrices whose columns correspond to the r non-zero (left and right, respectively) singular vectors of Θ^*. As in Section 2.4.2, define the subspace of matrices

$$\mathcal{M}(U^r, V^r) := \big\{\Theta \in \mathbb{R}^{d\times d} \mid \text{col}(\Theta) \subseteq U^r \text{ and } \text{row}(\Theta) \subseteq V^r\big\}, \tag{66}$$

as well as the associated set $\overline{\mathcal{M}}^\perp(U^r, V^r)$. Note that $\Theta^* \in \mathcal{M}$ by construction, and moreover (as discussed in Section 2.4.2, the nuclear norm is decomposable with respect to the pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$.

By definition (12) of the subspace compatibility with nuclear norm as the regularizer and Frobenius norm as the error norm, we have $\Psi^2(\mathcal{M}) = r$. Using the settings of $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$ guaranteed by Lemma 7 and substituting into equation (21), we obtain a contraction coefficient

$$\kappa(\Sigma) := \Big\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\Big\}\Big\{1 - \chi_n(\Sigma)\Big\}^{-1}, \tag{67}$$

where $\chi_n(\Sigma) := \frac{c_2\zeta_{\text{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)}\frac{rd}{n}$ for some universal constant c_2. A similar calculation shows that the tolerance term takes the form

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3\, \chi_n(\Sigma)\Big\{\frac{|||\Delta^*|||_1^2}{r} + |||\Delta^*|||_F^2\Big\} \qquad \text{for some constant } c_3.$$

Since $\rho \leq |||\Theta^*|||_1$ by assumption, Lemma 5 (as exploited in the proof of Corollary 1) shows that $|||\Delta^*|||_1^2 \leq 4r|||\Delta^*|||_F^2$, and hence that

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3\, \chi_n(\Sigma)\, |||\Delta^*|||_F^2,$$

which show the claim (42) for $q = 0$.

We now turn to the case $q \in (0, 1]$; as in the proof of this case for Corollary 2, we bound $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ using a slightly different choice of the subspace pair. Recall our notation $\sigma_1(\Theta^*) \geq \sigma_2(\Theta^*) \geq \cdots \geq \sigma_d(\Theta^*) \geq 0$ for the ordered singular values of Θ^*. For a threshold μ to be chosen, define $S_\mu = \big\{j \in \{1, 2, \ldots, d\} \mid \sigma_j(\Theta^*) > \mu\big\}$, and $U(S_\mu) \in \mathbb{R}^{d\times|S_\mu|}$ be the matrix of left singular vectors indexed by S_μ, with the matrix $V(S_\mu)$ defined similarly. We then define the subspace $\mathcal{M}(S_\mu) := \mathcal{M}(U(S_\mu), V(S_\mu))$ in an analogous fashion to equation (66), as well as the subspace $\overline{\mathcal{M}}^\perp(S_\mu)$.

Now by a combination of Lemma 5 and the definition (22) of $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, for any pair $(\mathcal{M}(S_\mu), \overline{\mathcal{M}}^\perp(S_\mu))$, we have

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}^\perp) \leq \frac{c\,\zeta_{\mathrm{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{d}{n} \Big(\sum_{j \notin S_\mu} \sigma_j(\Theta^*) + \sqrt{|S_\mu|}\, |||\Delta^*|||_F \Big)^2,$$

where to simplify notation, we have omitted the dependence of $\mathcal{M}$ and $\mathcal{M}^\perp$ on S_μ. As in the proof of Corollary 2, we now choose the threshold μ optimally, so as to trade-off the term $\sum_{j \notin S_\mu} \sigma_j(\Theta^*)$ with its competitor $\sqrt{|S_\mu|}\, |||\Delta^*|||_F$. Exploiting the fact that $\Theta^* \in \mathbb{B}_q(R_q)$ and following the same steps as the proof of Corollary 2 yields the bound

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}^\perp) \leq \frac{c\,\zeta_{\mathrm{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{d}{n} \big(\mu^{2-2q} R_q^2 + \mu^{-q} R_q |||\Delta^*|||_F^2 \big).$$

Setting $\mu^2 = \frac{d}{n}$ then yields

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}^\perp) \leq \chi_n(\Sigma) \left\{ R_q \big(\frac{d}{n}\big)^{1-q/2} + |||\Delta^*|||_F^2 \right\},$$

as claimed. The stated form of the contraction coefficient can be verified by a calculation analogous to the proof of Corollary 2.

7.6. *Proof of Corollary 5.* In this case, we let $\mathfrak{X}_n : \mathbb{R}^{d \times d} \to \mathbb{R}^n$ be the operator defined by the model of random signed matrix sampling [30]. As previously argued, establishing the RSM/RSC property amounts to obtaining a form of uniform control over $\frac{\|\mathfrak{X}_n(\Theta)\|_2^2}{n}$. More specifically, from the proof of Theorem 1, we see that it suffices to have a form of RSC for the difference $\widehat{\Delta}^t = \Theta^t - \widehat{\Theta}$, and a form of RSM for the difference $\Theta^{t+1} - \Theta^t$. The following two lemmas summarize these claims:

LEMMA 8. *There is a constant c such that for all iterations $t = 0, 1, 2, \ldots$ and integers $r = 1, 2, \ldots, d-1$, with probability at least $1 - \exp(-d \log d)$,*

(68)
$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq \frac{1}{2} |||\widehat{\Delta}^t|||_F^2 - \underbrace{c\alpha \sqrt{\frac{r\, d \log d}{n}} \Big\{ \frac{\sum_{j=r+1}^d \sigma_j(\Theta^*)}{\sqrt{r}} + \alpha \sqrt{\frac{rd \log d}{n}} + |||\Delta^*|||_F \Big\}}_{\delta_\ell(r)}.$$

LEMMA 9. *There is a constant c such that for all iterations $t = 0, 1, 2, \ldots$ and integers $r = 1, 2, \ldots, d-1$, with probability at least $1 - \exp(-d \log d)$, the*

difference $\Gamma^t := \Theta^{t+1} - \Theta^t$ *satisfies the inequality* $\frac{\|\mathfrak{X}_n(\Gamma^t)\|_2^2}{n} \leq 2\|\Gamma^t\|_F^2 + \delta_u(r)$, *where*

$$\delta_u(r) := c\alpha\sqrt{\frac{rd\log d}{n}}\Big\{\frac{\sum_{j=r+1}^d \sigma_j(\Theta^*)}{\sqrt{r}} + \alpha\sqrt{\frac{rd\log d}{n}} + \|\Delta^*\|_F + \|\widehat{\Delta}^t\|_F + \|\widehat{\Delta}^{t+1}\|_F\Big\}.$$

We can now complete the proof of Corollary 5 by a minor modification of the proof of Theorem 1. Recalling the elementary relation (52), we have

$$\|\Theta^{t+1} - \widehat{\Theta}\|_F^2 = \|\Theta^t - \widehat{\Theta}\|_F^2 + \|\Theta^t - \Theta^{t+1}\|_F^2 - 2\langle\!\langle \Theta^t - \widehat{\Theta},\ \Theta^t - \Theta^{t+1}\rangle\!\rangle.$$

From the proof of Lemma 2, we see that the combination of Lemma 8 and 9 (with $\gamma_\ell = \frac{1}{2}$ and $\gamma_u = 2$) imply that

$$2\langle\!\langle \Theta^t - \Theta^{t+1},\ \Theta^t - \widehat{\Theta}\rangle\!\rangle \geq \|\Theta^t - \Theta^{t+1}\|_F^2 + \frac{1}{4}\|\Theta^t - \widehat{\Theta}\|_F^2 - \delta_u(r) - \delta_\ell(r)$$

and hence that

$$\|\widehat{\Delta}^{t+1}\|_F^2 \leq \frac{3}{4}\|\widehat{\Delta}^t\|_F^2 + \delta_\ell(r) + \delta_u(r).$$

We substitute the forms of $\delta_\ell(r)$ and $\delta_u(r)$ given in Lemmas 8 and 9 respectively; performing some algebra then yields

$$\Big\{1 - \frac{c\,\alpha\sqrt{\frac{rd\log d}{n}}}{\|\widehat{\Delta}^{t+1}\|_F}\Big\}\|\widehat{\Delta}^{t+1}\|_F^2 \leq \Big\{\frac{3}{4} + \frac{c\alpha\sqrt{\frac{rd\log d}{n}}}{\|\widehat{\Delta}^t\|_F}\Big\}\|\widehat{\Delta}^t\|_F^2 + c'\,\delta_\ell(r).$$

Consequently, as long as $\min\{\|\widehat{\Delta}^t\|_F^2,\ \|\widehat{\Delta}^{t+1}\|_F^2\} \geq c_3\alpha\frac{rd\log d}{n}$ for a sufficiently large constant c_3, we are guaranteed the existence of some $\kappa_t \in (0,1)$ decreasing with t such that

$$\|\widehat{\Delta}^{t+1}\|_F^2 \leq \kappa_t\|\widehat{\Delta}^t\|_F^2 + c'\delta_\ell(r). \tag{69}$$

Since $\delta_\ell(r) = \Omega(\frac{rd\log d}{n})$, this inequality (69) is valid for all $t = 0, 1, 2, \ldots$ as long as c' is sufficiently large. Now iterating this bound, we see that

$$\|\widehat{\Delta}^{t+1}\|_F^2 \leq \Big(\prod_{s=1}^t \kappa_s\Big)\|\widehat{\Delta}^0\|_F^2 + c'\,\delta_\ell(r)\left(\kappa_t + \kappa_t\kappa_{t-1} + \cdots + \prod_{s=2}^t \kappa_s\right).$$

Since κ_t is decreasing in t, we observe that the second term in the above bound is at most

$$c' \, \delta_\ell(r) \left(\kappa_t + \kappa_t \kappa_{t-1} + \cdots + \prod_{s=2}^{t} \kappa_s \right) \leq c' \, \delta_\ell(r) \left(\kappa_1 + \kappa_1^2 + \kappa_1^{t-1} \right) \leq c' \, \frac{\delta_\ell(r)}{1 - \kappa_1}.$$

We also define $\bar{\kappa}_t = (\sum_{s=1}^{t} \kappa_t)/t$. Then the arithmetic mean-geometric mean inequality yields the upper bound $\prod_{s=1}^{t} \kappa_s \leq \bar{\kappa}_t^t$. Combining this with our earlier upper bound further yields the inequality

$$|||\widehat{\Delta}^{t+1}|||_F^2 \leq \bar{\kappa}_t^t |||\widehat{\Delta}^0|||_F^2 + \frac{c'}{1 - \kappa_1} \, \delta_\ell(r). \tag{70}$$

It remains to choose the cut-off $r \in \{1, 2, \ldots, d-1\}$ so as to minimize the term $\delta_\ell(r)$. In particular, when $\Theta^* \in \mathbb{B}_q(R_q)$, then as shown in the paper [29], the optimal choice is $r \asymp \alpha^{-q} R_q \big(\frac{n}{d \log d} \big)^{q/2}$. Substituting into the inequality (70) and performing some algebra yields that there is a universal constant c_4 such that the bound

$$|||\widehat{\Delta}^{t+1}|||_F^2 \leq \bar{\kappa}^t |||\widehat{\Delta}^0|||_F^2 + \frac{c_4}{1 - \kappa_1} \Big\{ R_q \big(\frac{\alpha d \log d}{n} \big)^{1 - q/2} + \sqrt{R_q \big(\frac{\alpha d \log d}{n} \big)^{1 - q/2}} \, |||\Delta^*|||_F \Big\}.$$

holds. Now by the Cauchy-Schwarz inequality we have

$$\sqrt{R_q \big(\frac{\alpha d \log d}{n} \big)^{1 - q/2}} \, |||\Delta^*|||_F \leq \frac{1}{2} R_q \big(\frac{\alpha d \log d}{n} \big)^{1 - q/2} + \frac{1}{2} |||\Delta^*|||_F^2,$$

and the claimed inequality (44) follows.

7.7. *Proof of Corollary 6.* Again the main argument in the proof would be to establish the RSM and RSC properties for the decomposition problem. We define $\widehat{\Delta}_\Theta^t = \Theta^t - \widehat{\Theta}$ and $\widehat{\Delta}_\Gamma^t = \Gamma^t - \widehat{\Gamma}$. We start with giving a lemma that establishes RSC for the differences $(\widehat{\Delta}_\Theta^t, \widehat{\Delta}_\Gamma^t)$. We recall that just like noted in the previous section, it suffices to show RSC only for these differences. Showing RSC/RSM in this example amounts to analyzing $|||\widehat{\Delta}_\Theta^t + \widehat{\Delta}_\Gamma^t|||_F^2$. We recall that this section assumes that Γ^* has only s non-zero columns.

LEMMA 10. *There is a constant c such that for all iterations $t = 0, 1, 2, \ldots,$*

$$|||\widehat{\Delta}_\Theta^t + \widehat{\Delta}_\Gamma^t|||_F^2 \geq \frac{1}{2} \big(|||\widehat{\Delta}_\Theta^t|||_F^2 + |||\widehat{\Delta}_\Gamma^t|||_F^2 \big) - c\alpha \sqrt{\frac{s}{d_2}} \left(|||\widehat{\Gamma} - \Gamma^*|||_F + \alpha \sqrt{\frac{s}{d_2}} \right) \tag{71}$$

This proof of this lemma follows by a straightforward modification of analogous results in the paper [1].

Matrix decomposition has the interesting property that the RSC condition holds in a deterministic sense (as opposed to with high probability). The same deterministic guarantee holds for the RSM condition; indeed, we have

$$\|\widehat{\Delta}^t_\Delta + \widehat{\Delta}^t_\Gamma\|_F^2 \leq 2\big(\|\widehat{\Delta}^t_\Theta\|_F^2 + \|\widehat{\Delta}^t_\Gamma\|_F^2\big), \tag{72}$$

by Cauchy-Schwartz inequality. Now we appeal to the more general form of Theorem 1 as stated in Equation 48, which gives

$$\|\widehat{\Delta}^{t+1}_\Theta\|_F^2 + \|\widehat{\Delta}^{t+1}_\Gamma\|_F^2 \leq \left(\frac{3}{4}\right)^t \big(\|\widehat{\Delta}^0_\Theta\|_F^2 + \|\widehat{\Delta}^0_\Gamma\|_F^2\big) + c\sqrt{\frac{\alpha s}{d_2}}\left(\|\widehat{\Gamma} - \Gamma^*\|_F + \frac{\alpha s}{d_2}\right).$$

The stated form of the corollary follows by an application of Cauchy-Schwarz inequality.

APPENDIX A: AUXILIARY RESULTS FOR THEOREM 1

In this appendix, we provide the proofs of various auxiliary lemmas required in the proof of Theorem 1.

A.1. Proof of Lemma 1. Since θ^t and $\widehat{\theta}$ are both feasible and $\widehat{\theta}$ lies on the constraint boundary, we have $\mathcal{R}(\theta^t) \leq \mathcal{R}(\widehat{\theta})$. Since $\mathcal{R}(\widehat{\theta}) \leq \mathcal{R}(\theta^*) + \mathcal{R}(\widehat{\theta} - \theta^*)$ by triangle inequality, we conclude that

$$\mathcal{R}(\theta^t) \leq \mathcal{R}(\theta^*) + \mathcal{R}(\Delta^*).$$

Since $\theta^* = \Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*)$, a second application of triangle inequality yields

$$\mathcal{R}(\theta^t) \leq \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*). \tag{73}$$

Now define the difference $\Delta^t := \theta^t - \theta^*$. (Note that this is slightly different from $\widehat{\Delta}^t$, which is measured relative to the optimum $\widehat{\theta}$.) With this notation, we have

$$\begin{aligned}
\mathcal{R}(\theta^t) &= \mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*) + \Pi_{\bar{\mathcal{M}}}(\Delta^t) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)\big) \\
&\overset{(i)}{\geq} \mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)\big) - \mathcal{R}\big(\Pi_{\mathcal{M}^\perp}(\theta^*) + \Pi_{\bar{\mathcal{M}}}(\Delta^t)\big) \\
&\overset{(ii)}{\geq} \mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)\big) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)),
\end{aligned}$$

where steps (i) and (ii) each use the triangle inequality. Now by the decomposability condition, we have $\mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)\big) = \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t))$, so that we have shown that

$$\mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) \leq \mathcal{R}(\theta^t).$$

Combining this inequality with the earlier bound (73) yields

$$\mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) \leq \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*).$$

Some algebra then leads to

$$\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)) \leq \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*). \tag{74}$$

The final step is to translate this inequality into one that applies to the optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$. Recalling that $\Delta^* = \widehat{\theta} - \theta^*$, we have $\widehat{\Delta}^t = \Delta^t - \Delta^*$, and hence

$$\mathcal{R}(\widehat{\Delta}^t) \leq \mathcal{R}(\Delta^t) + \mathcal{R}(\Delta^*), \qquad \text{by triangle inequality.} \tag{75}$$

In addition, we have

$$\begin{aligned}\mathcal{R}(\Delta^t) \;\leq\; \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^t)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) &\overset{(i)}{\leq} 2\,\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*) \\ &\overset{(ii)}{\leq} 2\,\Psi(\overline{\mathcal{M}}^\perp)\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*),\end{aligned}$$

where inequality (i) uses the bound (74), and inequality (ii) uses the definition (12) of the subspace compatibility Ψ. Combining with the inequality (75) yields

$$\mathcal{R}(\widehat{\Delta}^t) \leq 2\,\Psi(\overline{\mathcal{M}}^\perp)\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\mathcal{R}(\Delta^*).$$

Since projection onto a subspace is non-expansive, we have $\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\| \leq \|\Delta^t\|$, and hence

$$\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\| \leq \|\widehat{\Delta}^t + \Delta^*\| \;\leq\; \|\widehat{\Delta}^t\| + \|\Delta^*\|.$$

Combining the pieces, we obtain the claim (49).

A.2. Proof of Lemma 2. We start by applying the RSC assumption to the pair $\widehat{\theta}$ and θ^t, thereby obtaining that $\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2$ is lower bounded by

$$\begin{aligned}(76)\quad &\mathcal{L}_n(\theta^t) + \langle \nabla \mathcal{L}_n(\theta^t), \widehat{\theta} - \theta^t \rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}) \\ &= \mathcal{L}_n(\theta^t) + \langle \nabla \mathcal{L}_n(\theta^t), \theta^{t+1} - \theta^t \rangle + \langle \nabla \mathcal{L}_n(\theta^t), \widehat{\theta} - \theta^{t+1} \rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}),\end{aligned}$$

where we have added and subtracted terms. Now for compactness in notation, define

$$\varphi_t(\theta) := \mathcal{L}_n(\theta^t) + \left\langle \nabla \mathcal{L}_n(\theta^t), \theta - \theta^t \right\rangle + \frac{\gamma_u}{2}\|\theta - \theta^t\|^2,$$

and note that by definition of the algorithm, the iterate θ^{t+1} minimizes $\varphi_t(\theta)$ over the ball $\mathbb{B}_{\mathcal{R}}(\rho)$. Moreover, since $\widehat{\theta}$ is feasible, the first-order conditions for optimality imply that $\langle \nabla \varphi_t(\theta^{t+1}), \widehat{\theta} - \theta^{t+1} \rangle \geq 0$, or equivalently that $\langle \nabla \mathcal{L}_n(\theta^t) + \gamma_u(\theta^{t+1} - \theta^t), \widehat{\theta} - \theta^{t+1} \rangle \geq 0$. Applying this inequality to the lower bound (76), we find that $\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2$ is lower bounded by

$$\begin{aligned}&\mathcal{L}_n(\theta^t) + \langle \nabla \mathcal{L}_n(\theta^t), \theta^{t+1} - \theta^t \rangle + \gamma_u \langle \theta^t - \theta^{t+1}, \widehat{\theta} - \theta^{t+1} \rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}) \\ &= \varphi_t(\theta^{t+1}) - \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \gamma_u \langle \theta^t - \theta^{t+1}, \widehat{\theta} - \theta^{t+1} \rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta})\end{aligned}$$

Thus, by adding and subtracting θ^{t+1} in the inner product, we see that

$$\begin{aligned}(77)\quad &\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 \\ &\geq \varphi_t(\theta^{t+1}) + \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \gamma_u \langle \theta^t - \theta^{t+1}, \widehat{\theta} - \theta^t \rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}).\end{aligned}$$

Now by the RSM condition, we have

$$(78)\quad \varphi_t(\theta^{t+1}) \geq \mathcal{L}_n(\theta^{t+1}) - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t) \overset{(a)}{\geq} \mathcal{L}_n(\widehat{\theta}) - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t),$$

where inequality (a) follows by the optimality of $\widehat{\theta}$, and feasibility of θ^{t+1}. Combining this inequality with the previous bound (77) yields that $\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2$ is lower bounded by

$$\begin{aligned}\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \gamma_u \langle \theta^t - \theta^{t+1}, \widehat{\theta} - \theta^t \rangle \\ - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\theta^t - \widehat{\theta}) - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t),\end{aligned}$$

and the claim (51) follows after some simple algebraic manipulations.

APPENDIX B: AUXILIARY RESULTS FOR THEOREM 2

In this appendix, we prove the two auxiliary lemmas required in the proof of Theorem 2.

B.1. Proof of Lemma 3. This result is a generalization of an analogous result in Negahban et al. [28], with some changes required so as to adapt the statement to the optimization setting. Let θ be any vector, feasible for the problem (2), that satisfies the bound

$$\phi(\theta) \leq \phi(\theta^*) + \bar{\eta}, \tag{79}$$

and assume that $\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}_n(\theta^*))$. We then claim that the error vector $\Delta := \theta - \theta^*$ satisfies the inequality

$$\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta)) \;\leq\; 3\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n}, \bar{\rho}\big\}. \tag{80}$$

For the moment, we take this claim as given, returning later to verify its validity.

By applying this intermediate claim (80) in two different ways, we can complete the proof of Lemma 3. First, we observe that when $\theta = \widehat{\theta}$, the optimality of $\widehat{\theta}$ and feasibility of θ^* imply that assumption (79) holds with $\bar{\eta} = 0$, and hence the intermediate claim (80) implies that the statistical error $\Delta^* = \theta^* - \widehat{\theta}$ satisfies the bound

$$\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) \;\leq\; 3\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)). \tag{81}$$

Since $\Delta^* = \Pi_{\bar{\mathcal{M}}}(\Delta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)$, we can write

$$\mathcal{R}(\Delta^*) = \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) \;\leq\; 4\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)), \tag{82}$$

using the triangle inequality in conjunction with our earlier bound (81). Similarly, when $\theta = \theta^t$ for some $t \geq T$, then the given assumptions imply that condition (79) holds with $\bar{\eta} > 0$, so that the intermediate claim (followed by the same argument with triangle inequality) implies that the error $\Delta^t = \theta^t - \theta^*$ satisfies the bound

$$\mathcal{R}(\Delta^t) \;\leq\; 4\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n}, \bar{\rho}\big\}. \tag{83}$$

Now let $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$ be the optimization error at time t, and observe that we have the decomposition $\widehat{\Delta}^t = \Delta^t + \Delta^*$. Consequently, by triangle

inequality

$$\begin{aligned}\mathcal{R}(\widehat{\Delta}^t) &\leq \mathcal{R}(\Delta^t) + \mathcal{R}(\Delta^*) \\ &\overset{(i)}{\leq} 4\Big\{\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^t)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*))\Big\} + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n},\, \bar{\rho}\big\} \\ &\overset{(ii)}{\leq} 4\Psi(\overline{\mathcal{M}})\,\Big\{\|\Pi_{\bar{\mathcal{M}}}(\Delta^t)\| + \|\Pi_{\bar{\mathcal{M}}}(\Delta^*)\|\Big\} + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n},\, \bar{\rho}\big\}\end{aligned} \tag{84}$$

where step (i) follows by applying both equation (82) and (83), step (ii) follows from the definition (12) of the subspace compatibility that relates the regularizer to the norm $\|\cdot\|$. Since projection onto a subspace is non-expansive, we thus obtain

$$\mathcal{R}(\widehat{\Delta}^t) \leq 4\Psi(\overline{\mathcal{M}})\,\Big\{\|\Delta^t\| + \|\Delta^*\|\Big\} + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\big\{\frac{\bar{\eta}}{\lambda_n},\, \bar{\rho}\big\}, \tag{85}$$

Finally, since $\Delta^t = \widehat{\Delta}^t - \Delta^*$, the triangle inequality implies that $\|\Delta^t\| \leq \|\widehat{\Delta}^t\| + \|\Delta^*\|$. Substituting this upper bound into inequality (85) completes the proof of Lemma 3.

It remains to prove the intermediate claim (80). Letting θ be any vector, feasible for the program (2), and satisfying the condition (79), and let $\Delta = \theta - \theta^*$ be the associated error vector. Re-writing the condition (79), we have

$$\mathcal{L}_n(\theta^* + \Delta) + \lambda_n\mathcal{R}(\theta^* + \Delta) \;\leq\; \mathcal{L}_n(\theta^*) + \lambda_n\mathcal{R}(\theta^*) + \bar{\eta}.$$

Subtracting $\big\langle \nabla\mathcal{L}_n(\theta^*),\, \Delta\big\rangle$ from each side and then re-arranging yields the inequality

$$\begin{aligned}\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) - \big\langle \nabla\mathcal{L}_n(\theta^*),\, \Delta\big\rangle + \lambda_n\Big\{\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*)\Big\} \\ \leq\; -\big\langle \nabla\mathcal{L}_n(\theta^*),\, \Delta\big\rangle + \bar{\eta}.\end{aligned}$$

The convexity of $\mathcal{L}_n$ then implies that

$$\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) - \big\langle \nabla\mathcal{L}_n(\theta^*),\, \Delta\big\rangle \geq 0,$$

and hence that

$$\lambda_n\Big\{\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*)\Big\} \;\leq\; -\big\langle \nabla\mathcal{L}_n(\theta^*),\, \Delta\big\rangle + \bar{\eta}.$$

Applying Hölder's inequality to $\langle \nabla \mathcal{L}_n(\theta^*), \Delta \rangle$, as expressed in terms of the dual norms $\mathcal{R}$ and $\mathcal{R}^*$, yields the upper bound

$$\lambda_n \Big\{ \mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \Big\} \leq \mathcal{R}^*(\nabla \mathcal{L}_n(\theta^*)) \, \mathcal{R}(\Delta) + \bar{\eta} \overset{(i)}{\leq} \frac{\lambda_n}{2} \mathcal{R}(\Delta) + \bar{\eta},$$

where step (i) uses the fact that $\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}_n(\theta^*))$ by assumption.

For the remainder of the proof, let us introduce the convenient shorthand $\Delta_{\bar{\mathcal{M}}} := \Pi_{\bar{\mathcal{M}}}(\Delta)$ and $\Delta_{\bar{\mathcal{M}}^\perp} := \Pi_{\bar{\mathcal{M}}^\perp}(\Delta)$, with similar shorthand for projections involving θ^*. Making note of the decomposition $\Delta = \Delta_{\bar{\mathcal{M}}} + \Delta_{\bar{\mathcal{M}}^\perp}$, an application of triangle inequality then yields the upper bound

$$\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \leq \frac{1}{2} \Big\{ \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) \Big\} + \frac{\bar{\eta}}{\lambda_n}, \tag{86}$$

where we have rescaled both sides by $\lambda_n > 0$.

It remains to further lower bound the left-hand side (86). By triangle inequality, we have

$$-\mathcal{R}(\theta^*) \geq -\mathcal{R}(\theta^*_{\mathcal{M}}) - \mathcal{R}(\theta^*_{\mathcal{M}^\perp}). \tag{87}$$

Let us now write $\theta^* + \Delta = \theta^*_{\mathcal{M}} + \theta^*_{\mathcal{M}^\perp} + \Delta_{\bar{\mathcal{M}}} + \Delta_{\bar{\mathcal{M}}^\perp}$. Using this representation and triangle inequality, we have

$$\begin{aligned} \mathcal{R}(\theta^* + \Delta) &\geq \mathcal{R}(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}) - \mathcal{R}(\theta^*_{\mathcal{M}^\perp} + \Delta_{\bar{\mathcal{M}}}) \\ &\geq \mathcal{R}(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}) - \mathcal{R}(\theta^*_{\mathcal{M}^\perp}) - \mathcal{R}(\Delta_{\bar{\mathcal{M}}}). \end{aligned}$$

Finally, since $\theta^*_{\mathcal{M}} \in \mathcal{M}$ and $\Delta_{\bar{\mathcal{M}}^\perp} \in \overline{\mathcal{M}}^\perp$, the decomposability of $\mathcal{R}$ implies that $\mathcal{R}(\theta^*_{\mathcal{M}} + \Delta_{\bar{\mathcal{M}}^\perp}) = \mathcal{R}(\theta^*_{\mathcal{M}}) + \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp})$, and hence that

$$\mathcal{R}(\theta^* + \Delta) \geq \mathcal{R}(\theta^*_{\mathcal{M}}) + \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) - \mathcal{R}(\theta^*_{\mathcal{M}^\perp}) - \mathcal{R}(\Delta_{\bar{\mathcal{M}}}). \tag{88}$$

Adding together equations (87) and (88), we obtain the lower bound

$$\mathcal{R}(\theta^* + \Delta) - \mathcal{R}(\theta^*) \geq \mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) - 2\mathcal{R}(\theta^*_{\mathcal{M}^\perp}) - \mathcal{R}(\Delta_{\bar{\mathcal{M}}}). \tag{89}$$

Combining this lower bound with the earlier inequality (86), some algebra yields the bound

$$\mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) \leq 3\mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp}) + 2\frac{\eta}{\lambda_n},$$

corresponding to the bound (80) when η/λ_n achieves the final minimum. To obtain the final term involving $\bar{\rho}$ in the bound (80), two applications of triangle inequality yields

$$\mathcal{R}(\Delta_{\bar{\mathcal{M}}^\perp}) \leq \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + \mathcal{R}(\Delta) \leq \mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + 2\bar{\rho},$$

where we have used the fact that $\mathcal{R}(\Delta) \leq \mathcal{R}(\theta) + \mathcal{R}(\theta^*) \leq 2\bar{\rho}$, since both θ and θ^* are feasible for the program (2).

B.2. Proof of Lemma 4. The proof of this result follows lines similar to the proof of convergence by Nesterov [32]. Recall our notation $\phi(\theta) = \mathcal{L}_n(\theta) + \lambda_n \mathcal{R}(\theta)$, $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$, and that $\eta_\phi^t = \phi(\theta^t) - \phi(\widehat{\theta})$. We begin by proving that under the stated conditions, a useful version of restricted strong convexity (47) is in force:

LEMMA 11. *Under the assumptions of Lemma 4, we are guaranteed that*

$$\big\{\frac{\gamma_\ell}{2} - 32\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\big\}\|\widehat{\Delta}^t\|^2 \leq 2\tau_\ell(\mathcal{L}_n)\, v^2 \,+\, \phi(\theta^t) - \phi(\widehat{\theta}), \quad \textit{and} \tag{90a}$$

$$\big\{\frac{\gamma_\ell}{2} - 32\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\big\}\|\widehat{\Delta}^t\|^2 \leq 2\,\tau_\ell(\mathcal{L}_n)\, v^2 \,+\, \mathcal{T}_{\mathcal{L}}(\widehat{\theta}; \theta^t), \tag{90b}$$

where $v := \bar{\epsilon}_{stat} + 2\min(\frac{\bar{\eta}}{\lambda_n}, \bar{\rho})$.

See Appendix B.3 for the proof of this claim. So as to ease notation in the remainder of the proof, let us introduce the shorthand

$$\phi_t(\theta) := \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t),\, \theta - \theta^t \big\rangle + \frac{\gamma_u}{2}\|\theta - \theta^t\|^2 + \lambda_n \mathcal{R}(\theta), \tag{91}$$

corresponding to the approximation to the regularized loss function ϕ that is minimized at iteration t of the update (4). Since θ^{t+1} minimizes ϕ_t over the set $\mathbb{B}_{\mathcal{R}}(\bar{\rho})$, we are guaranteed that $\phi_t(\theta^{t+1}) \leq \phi_t(\theta)$ for all $\theta \in \mathbb{B}_{\mathcal{R}}(\bar{\rho})$. In particular, for any $\alpha \in (0,1)$, the vector $\theta_\alpha = \alpha\widehat{\theta} + (1-\alpha)\theta^t$ lies in the convex set $\mathbb{B}_{\mathcal{R}}(\bar{\rho})$, so that

$$\begin{aligned}\phi_t(\theta^{t+1}) \;\leq\; \phi_t(\theta_\alpha) &= \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t),\, \theta_\alpha - \theta^t \big\rangle + \frac{\gamma_u}{2}\|\theta_\alpha - \theta^t\|^2 + \lambda_n\mathcal{R}(\theta_\alpha) \\ &\overset{(i)}{=} \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t),\, \alpha\widehat{\theta} - \alpha\theta^t \big\rangle + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta} - \theta^t\|^2 + \lambda_n\mathcal{R}(\theta_\alpha)\end{aligned}$$

where step (i) follows from substituting the definition of θ_α. Using the convexity of the regularizer $\mathcal{R}$, we then obtain

$$\begin{aligned}\phi_t(\theta^{t+1}) \leq \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t),\, \alpha\widehat{\theta} - \alpha\theta^t \big\rangle + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta} - \theta^t\|^2 \\ + \lambda_n\alpha\mathcal{R}(\widehat{\theta}) + \lambda_n(1-\alpha)\mathcal{R}(\theta^t),\end{aligned}$$

Now the stated conditions of the lemma ensure that

$$\gamma_\ell/2 - 32\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}}) \geq 0,$$

so that by equation (90b), we have

$$\mathcal{L}_n(\widehat{\theta}) + 2\tau_\ell(\mathcal{L}_n)v^2 \geq \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t),\, \widehat{\theta} - \theta^t \big\rangle.$$

Substituting back into our earlier bound yields

$$\phi_t(\theta^{t+1}) \leq (1-\alpha)\mathcal{L}_n(\theta^t) + \alpha\mathcal{L}_n(\widehat{\theta}) + 2\alpha\tau_\ell(\mathcal{L}_n)v^2 + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta}-\theta^t\|^2 + \alpha\lambda_n\mathcal{R}(\widehat{\theta}) + (1-\alpha)\lambda_n\mathcal{R}(\theta^t).$$

By the definition of ϕ and the bound $\alpha \leq 1$, we find that

$$\phi_t(\theta^{t+1}) \leq \phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + 2\tau_\ell(\mathcal{L}_n)v^2 + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta}-\theta^t\|^2. \tag{92}$$

In order to complete the proof, it remains to relate $\phi_t(\theta^{t+1})$ to $\phi(\theta^{t+1})$, which can be performed by exploiting restricted smoothness. In particular, applying the RSM condition at the iterate θ^{t+1} in the direction θ^t yields the upper bound

$$\mathcal{L}_n(\theta^{t+1}) \leq \mathcal{L}_n(\theta^t) + \big\langle \mathcal{L}_n(\theta^t),\, \theta^{t+1}-\theta^t\big\rangle + \frac{\gamma_u}{2}\|\theta^{t+1}-\theta^t\|^2 + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1}-\theta^t),$$

so that

$$\begin{aligned}\phi(\theta^{t+1}) &\leq \mathcal{L}_n(\theta^t) + \big\langle \mathcal{L}_n(\theta^t),\, \theta^{t+1}-\theta^t\big\rangle + \frac{\gamma_u}{2}\|\theta^{t+1}-\theta^t\|^2 + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1}-\theta^t) + \lambda_n\mathcal{R}(\theta^{t+1}) \\ &= \phi_t(\theta^{t+1}) + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1}-\theta^t).\end{aligned}$$

Combining the above bound with the inequality (92) and recalling the notation $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$, we obtain

$$\begin{aligned}\phi(\theta^{t+1}) &\leq \phi(\theta^t) - \alpha(\phi(\theta^t)-\phi(\widehat{\theta})) + \frac{\gamma_u\alpha^2}{2}\|\widehat{\theta}-\theta^t\|^2 + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1}-\theta^t) + 2\tau_\ell(\mathcal{L}_n)v^2 \\ &\overset{(iv)}{\leq} \phi(\theta^t) - \alpha(\phi(\theta^t)-\phi(\widehat{\theta})) + \frac{\gamma_u\alpha^2}{2}\|\widehat{\Delta}^t\|^2 + \tau_u(\mathcal{L}_n)[\mathcal{R}(\widehat{\Delta}^{t+1}) + \mathcal{R}(\widehat{\Delta}^t)]^2 + 2\tau_\ell(\mathcal{L}_n)v^2 \\ &\overset{(v)}{\leq} \phi(\theta^t) - \alpha(\phi(\theta^t)-\phi(\widehat{\theta})) + \frac{\gamma_u\alpha^2}{2}\|\widehat{\Delta}^t\|^2 + 2\tau_u(\mathcal{L}_n)(\mathcal{R}^2(\widehat{\Delta}^{t+1}) + \mathcal{R}^2(\widehat{\Delta}^t)) + 2\tau_\ell(\mathcal{L}_n)v^2.\end{aligned} \tag{93}$$

Here step (iv) uses the fact that $\theta^t - \theta^{t+1} = \widehat{\Delta}^t - \widehat{\Delta}^{t+1}$ and applies triangle inequality to the norm $\mathcal{R}$, whereas step (v) follows from Cauchy-Schwarz inequality.

Next, combining Lemma 3 with the Cauchy-Schwarz inequality inequality yields the upper bound

$$\mathcal{R}^2(\widehat{\Delta}^t) \leq 32\Psi^2(\overline{\mathcal{M}})\|\widehat{\Delta}^t\|^2 + 2v^2 \tag{94}$$

where $v = \bar{\epsilon}_{\text{stat}}(\mathcal{M}, \overline{\mathcal{M}}) + 2\min(\frac{\bar{\eta}}{\lambda_n}, \bar{\rho})$, is a constant independent of θ^t and $\bar{\epsilon}_{\text{stat}}(\mathcal{M}, \overline{\mathcal{M}})$ was previously defined in the lemma statement. Substituting the above bound into inequality (93) yields that $\phi(\theta^{t+1})$ is at most

$$\phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + \frac{\gamma_u \alpha^2}{2}\|\widehat{\Delta}^t\|^2 + 64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\|\widehat{\Delta}^{t+1}\|^2 + 64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\|\widehat{\Delta}^t\|^2 + 8\tau_u(\mathcal{L}_n)v^2 + 2\tau_\ell(\mathcal{L}_n)v^2. \tag{95}$$

The final step is to translate quantities involving $\widehat{\Delta}^t$ to functional values, which may be done using the RSC condition (90a) from Lemma 11. In particular, combining the RSC condition (90a) with the inequality (95) yields

$$\phi(\theta^{t+1}) \leq \phi(\theta^t) - \alpha\eta_\phi^t + \frac{\left(\gamma_u\alpha^2 + 64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\right)}{\overline{\gamma_\ell}}(\eta_\phi^t + 2\tau_\ell(\mathcal{L}_n)v^2) + \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}(\eta_\phi^{t+1} + 2\tau_\ell(\mathcal{L}_n)v^2) + 8\tau_u(\mathcal{L}_n)v^2 + 2\tau_\ell(\mathcal{L}_n)v^2.$$

where we have introduced the shorthand $\overline{\gamma_\ell} := \gamma_\ell - 64\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})$. Recalling the definition of β, adding and subtracting $\phi(\widehat{\theta})$ from both sides, and choosing $\alpha = \frac{\overline{\gamma_\ell}}{2\gamma_u} \in (0,1)$, we obtain

$$\left(1 - \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right)\eta_\phi^{t+1} \leq \left(1 - \frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right)\eta_\phi^t + \beta(\overline{\mathcal{M}})v^2.$$

Recalling the definition of the contraction factor κ from the statement of Theorem 2, the above expression can be rewritten as

$$\eta_\phi^{t+1} \leq \kappa\eta_\phi^t + \beta(\overline{\mathcal{M}})\xi(\overline{\mathcal{M}})v^2, \quad \text{where } \xi(\mathcal{M}) = \left\{1 - \tfrac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right\}^{-1}.$$

Finally, iterating the above expression yields $\eta_\phi^t \leq \kappa^{t-T}\eta_\phi^T + \frac{\xi(\overline{\mathcal{M}})\beta(\overline{\mathcal{M}})v^2}{1-\kappa}$, where we have used the condition $\kappa \in (0,1)$ in order to sum the geometric series, thereby completing the proof.

B.3. Proof of Lemma 11. The key idea to prove the lemma is to use the definition of RSC along with the iterated cone bound of Lemma 3 for simplifying the error terms in RSC.

Let us first show that condition (90a) holds. From the RSC condition assumed in the lemma statement, we have

$$\mathcal{L}_n(\theta^t) - \mathcal{L}_n(\widehat{\theta}) - \langle \nabla\mathcal{L}_n(\widehat{\theta}),\, \theta^t - \widehat{\theta}\rangle \geq \frac{\gamma_\ell}{2}\,\|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\widehat{\theta} - \theta^t). \tag{96}$$

From the convexity of $\mathcal{R}$ and definition of the subdifferential $\partial\mathcal{R}(\theta)$, we obtain

$$\mathcal{R}(\theta^t) - \mathcal{R}(\widehat{\theta}) - \big\langle \partial\mathcal{R}(\widehat{\theta}),\, \theta^t - \widehat{\theta} \big\rangle \geq 0.$$

Adding this lower bound with the inequality (96) yields

$$\phi(\theta^t) - \phi(\widehat{\theta}) - \langle \nabla\phi(\widehat{\theta}),\, \theta^t - \widehat{\theta}\rangle \geq \frac{\gamma_\ell}{2}\,\|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n)\;\mathcal{R}^2(\widehat{\theta} - \theta^t),$$

where we recall that $\phi(\theta) = \mathcal{L}_n(\theta) + \lambda_n\mathcal{R}(\theta)$ is our objective function. By the optimality of $\widehat{\theta}$ and feasibility of θ^t, we are guaranteed that $\langle \nabla\phi(\widehat{\theta}),\, \theta^t - \widehat{\theta}\rangle \geq 0$, and hence

$$\begin{aligned}\phi(\theta^t) - \phi(\widehat{\theta}) &\geq \frac{\gamma_\ell}{2}\,\|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n)\;\mathcal{R}^2(\widehat{\theta} - \theta^t) \\ &\overset{(i)}{\geq} \frac{\gamma_\ell}{2}\,\|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n)\;\; \big\{32\Psi^2(\overline{\mathcal{M}})\|\widehat{\theta} - \theta^t\|^2 + 2v^2\big\}\end{aligned}$$

where step (i) follows by applying Lemma 3. Some algebra then yields the claim (90a).

Finally, let us verify the claim (90b). Using the RSC condition, we have

(97)
$$\mathcal{L}_n(\widehat{\theta}) - \mathcal{L}_n(\theta^t) - \langle \nabla\mathcal{L}_n(\theta^t),\, \widehat{\theta} - \theta^t\rangle \geq \frac{\gamma_\ell}{2}\,\|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n)\;\mathcal{R}^2(\widehat{\theta} - \theta^t).$$

As before, applying Lemma 3 yields

$$\underbrace{\mathcal{L}_n(\widehat{\theta}) - \mathcal{L}_n(\theta^t) - \langle \nabla\mathcal{L}_n(\theta^t),\, \widehat{\theta} - \theta^t\rangle}_{\mathcal{T}_{\mathcal{L}}(\widehat{\theta};\theta^t)} \geq \frac{\gamma_\ell}{2}\,\|\widehat{\theta} - \theta^t\|^2 - \tau_\ell(\mathcal{L}_n)\;\left(32\Psi^2(\overline{\mathcal{M}})\|\widehat{\theta} - \theta^t\|^2 + 2v^2\right),$$

and rearranging the terms and establishes the claim (90b).

APPENDIX C: PROOF OF LEMMA 5

Given the condition $\mathcal{R}(\widehat{\theta}) \leq \rho \leq \mathcal{R}(\theta^*)$, we have

$$\mathcal{R}(\widehat{\theta}) = \mathcal{R}(\theta^* + \Delta^*) \leq \mathcal{R}(\theta^*).$$

Applying triangle inequality then yields

$$\mathcal{R}(\theta^*) = \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*)) \;\leq\; \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)).$$

We then write

$$\begin{aligned}\mathcal{R}(\theta^* + \Delta^*) &= \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*) + \Pi_{\bar{\mathcal{M}}}(\Delta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) \\ &\overset{(i)}{\geq} \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \\ &\overset{(ii)}{=} \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) - \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)),\end{aligned}$$

where the bound (i) follows by triangle inequality, and step (ii) uses the decomposability of $\mathcal{R}$ over the pair $\mathcal{M}$ and $\overline{\mathcal{M}}^\perp$. By combining this lower bound with the previously established upper bound

$$\mathcal{R}(\theta^* + \Delta^*) \leq \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)),$$

we conclude that $\mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*)) \leq \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))$. Finally, by triangle inequality, we have $\mathcal{R}(\Delta^*) \leq \mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + \mathcal{R}(\Pi_{\bar{\mathcal{M}}^\perp}(\Delta^*))$, and hence

$$\begin{aligned}\mathcal{R}(\Delta^*) &\leq 2\mathcal{R}(\Pi_{\bar{\mathcal{M}}}(\Delta^*)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \\ &\overset{(i)}{\leq} 2\,\Psi(\overline{\mathcal{M}}^\perp)\|\Pi_{\bar{\mathcal{M}}}(\Delta^*)\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \\ &\overset{(ii)}{\leq} 2\,\Psi(\overline{\mathcal{M}}^\perp)\|\Delta^*\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)),\end{aligned}$$

where inequality (i) follows from Definition 4 of the subspace compatibility Ψ, and the bound (ii) follows from non-expansivity of projection onto a subspace.

APPENDIX D: A GENERAL RESULT ON GAUSSIAN OBSERVATION OPERATORS

In this appendix, we state a general result about a Gaussian random matrices, and show how it can be adapted to prove Lemmas 6 and 7. Let $X \in \mathbb{R}^{n \times d}$ be a Gaussian random matrix with i.i.d. rows $x_i \sim N(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{d \times d}$ is a covariance matrix. We refer to X as a sample from the Σ-Gaussian ensemble. In order to state the result, we use $\Sigma^{1/2}$ to denote the symmetric matrix square root.

PROPOSITION 1. *Given a random matrix X drawn from the Σ-Gaussian ensemble, there are universal constants c_i, $i = 0, 1$ such that*

$$\frac{\|X\theta\|_2^2}{n} \geq \frac{1}{2}\|\Sigma^{1/2}\theta\|_2^2 - c_1 \frac{(\mathbb{E}[\mathcal{R}^*(x_i)])^2}{n}\mathcal{R}^2(\theta) \qquad \text{and} \tag{98a}$$

$$\frac{\|X\theta\|_2^2}{n} \leq 2\|\Sigma^{1/2}\theta\|_2^2 + c_1 \frac{(\mathbb{E}[\mathcal{R}^*(x_i)])^2}{n}\mathcal{R}^2(\theta) \qquad \text{for all } \theta \in \mathbb{R}^d \tag{98b}$$

with probability greater than $1 - \exp(-c_0\, n)$.

We omit the proof of this result. The two special instances proved in Lemma 6 and 7 have been proved in the papers [35] and [29] respectively. We now show how Proposition 1 can be used to recover various lemmas required in our proofs.

Proof of Lemma 6:. We begin by establishing this auxiliary result required in the proof of Corollary 2. When $\mathcal{R}(\cdot) = \|\cdot\|_1$, we have $\mathcal{R}^*(\cdot) = \|\cdot\|_\infty$. Moreover, the random vector $x_i \sim N(0, \Sigma)$ can be written as $x_i = \Sigma^{1/2} w$, where $w \sim N(0, I_{d\times d})$ is standard normal. Consequently, using properties of Gaussian maxima [22] and defining $\zeta(\Sigma) = \max_{j=1,2,\ldots,d} \Sigma_{jj}$, we have the bound

$$(\mathbb{E}[\|x_i\|_\infty])^2 \;\leq\; \zeta(\Sigma)\,(\mathbb{E}[\|w\|_\infty])^2 \;\leq\; 3\zeta(\Sigma)\,\sqrt{\log d}.$$

Substituting into Proposition 1 yields the claims (61a) and (61b).

Proof of Lemma 7:. In order to prove this claim, we view each random observation matrix $X_i \in \mathbb{R}^{d\times d}$ as a $d = d^2$ vector (namely the quantity $\mathrm{vec}(X_i)$), and apply Proposition 1 in this vectorized setting. Given the standard Gaussian vector $w \in \mathbb{R}^{d^2}$, we let $W \in \mathbb{R}^{d\times d}$ be the random matrix such that $\mathrm{vec}(W) = w$. With this notation, the term $\mathcal{R}^*(\mathrm{vec}(X_i))$ is equivalent to the operator norm $|||X_i|||_{\mathrm{op}}$. As shown in Negahban and Wainwright [29], $\mathbb{E}[|||X_i|||_{\mathrm{op}}] \leq 24\zeta_{\mathrm{mat}}(\Sigma)\,\sqrt{d}$, where ζ_{mat} was previously defined (64).

APPENDIX E: AUXILIARY RESULTS FOR COROLLARY 5

In this section, we provide the proofs of Lemmas 8 and 9 that play a central role in the proof of Corollary 5. In order to do so, we require the following result, which is a re-statement of a theorem due to Negahban and Wainwright [30]:

PROPOSITION 2. *For the matrix completion operator $\mathfrak{X}_n$, there are universal positive constants (c_1, c_2) such that*

$$\left|\frac{\|\mathfrak{X}_n(\Theta)\|_2^2}{n} - |||\Theta|||_F^2\right| \leq c_1\, d\|\Theta\|_\infty\, |||\Theta|||_1 \sqrt{\frac{d\log d}{n}} \;+\; c_2\left(d\|\Theta\|_\infty\sqrt{\frac{d\log d}{n}}\right)^2$$

for all $\Theta \in \mathbb{R}^{d\times d}$ with probability at least $1 - \exp(-d\log d)$.

E.1. Proof of Lemma 8. Applying Proposition 2 to $\widehat{\Delta}^t$ and using the fact that $d\|\widehat{\Delta}^t\|_\infty \leq 2\alpha$ yields

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq |||\widehat{\Delta}^t|||_F^2 - c_1\alpha|||\widehat{\Delta}^t|||_1\sqrt{\frac{d\log d}{n}} - c_2\,\alpha^2\frac{d\log d}{n}, \tag{99}$$

where we recall our convention of allowing the constants to change from line to line. From Lemma 1,

$$\|\widehat{\Delta}^t\|_1 \leq 2\,\Psi(\overline{\mathcal{M}}^{\perp})\,\|\widehat{\Delta}^t\|_F + 2\|\Pi_{\mathcal{M}^{\perp}}(\theta^*)\|_1 + 2\|\Delta^*\|_1 + \Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|_F.$$

Since $\rho \leq \|\Theta^*\|_1$, Lemma 5 implies that $\|\Delta^*\|_1 \leq 2\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|_F + \|\Pi_{\mathcal{M}^{\perp}}(\theta^*)\|_1$, and hence that

$$\|\widehat{\Delta}^t\|_1 \leq 2\,\Psi(\overline{\mathcal{M}}^{\perp})\,\|\widehat{\Delta}^t\|_F + 4\|\Pi_{\mathcal{M}^{\perp}}(\theta^*)\|_1 + 5\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|_F. \tag{100}$$

Combined with the lower bound, we obtain that $\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n}$ is lower bounded by

$$\|\widehat{\Delta}^t\|_F^2\left\{1 - \frac{2c_1\,\alpha\Psi(\overline{\mathcal{M}}^{\perp})\sqrt{\frac{d\log d}{n}}}{\|\widehat{\Delta}^t\|_F}\right\} - 2c_1\,\alpha\sqrt{\frac{d\log d}{n}}\Big\{4\|\Pi_{\mathcal{M}^{\perp}}(\theta^*)\|_1 + 5\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|_F\Big\} - c_2\,\alpha^2\frac{d\log d}{n}.$$

Consequently, for all iterations such that $\|\widehat{\Delta}^t\|_F \geq 4c_1\Psi(\overline{\mathcal{M}}^{\perp})\sqrt{\frac{d\log d}{n}}$, the quantity $\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n}$ is lower bounded by

$$\frac{1}{2}\|\widehat{\Delta}^t\|_F^2 - 2c_1\,\alpha\sqrt{\frac{d\log d}{n}}\Big\{4\|\Pi_{\mathcal{M}^{\perp}}(\theta^*)\|_1 + 5\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|_F\Big\} - c_2\,\alpha^2\frac{d\log d}{n}.$$

By subtracting off an additional term, the bound is valid for all $\widehat{\Delta}^t$—viz.

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq \frac{1}{2}\|\widehat{\Delta}^t\|_F^2 - 2c_1\,\alpha\sqrt{\frac{d\log d}{n}}\Big\{4\|\Pi_{\mathcal{M}^{\perp}}(\theta^*)\|_1 + 5\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|_F\Big\} - c_2\,\alpha^2\frac{d\log d}{n} - 16c_1^2\alpha^2\Psi^2(\overline{\mathcal{M}}^{\perp})\frac{d\log d}{n}.$$

E.2. Proof of Lemma 9. Applying Proposition 2 to Γ^t and using the fact that $d\|\Gamma^t\|_\infty \leq 2\alpha$ yields

$$\frac{\|\mathfrak{X}_n(\Gamma^t)\|_2^2}{n} \leq \|\Gamma^t\|_F^2 + c_1\alpha\|\Gamma^t\|_1\,\sqrt{\frac{d\log d}{n}} + c_2\,\alpha^2\frac{d\log d}{n}, \tag{101}$$

where we recall our convention of allowing the constants to change from line to line. By triangle inequality, we have $\|\Gamma^t\|_1 \leq \|\Theta^t - \widehat{\Theta}\|_1 + \|\Theta^{t+1} - \widehat{\Theta}\|_1 = \|\widehat{\Delta}^t\|_1 + \|\widehat{\Delta}^{t+1}\|_1$. Equation 100 gives us bounds on $\|\widehat{\Delta}^t\|_1$ and $\|\widehat{\Delta}^{t+1}\|_1$. Substituting them into the upper bound (101) yields the claim.

REFERENCES

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Annals of Statistics*, 40(2):1171–1197, 2012.

[2] A. Agarwal, S. Negahban, and M. J. Wainwright. Supplemental file to Fast global convergence of gradient methods for high-dimensional statistical recovery. Technical report, University of California, Berkeley, 2012.

[3] A. A. Amini and M. J. Wainwright. High-dimensional analysis of semdefinite relaxations for sparse principal component analysis. *Annals of Statistics*, 37:2877–2921, 2009.

[4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[5] S. Becker, J. Bobin, and E. J. Candes. Nesta: a fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

[6] D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1995.

[7] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[8] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

[9] K. Bredies and D. A. Lorenz. Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, 14:813–837, 2008.

[10] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? *J. ACM*, 58:11:1–11:37, 2011.

[11] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[12] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. on Optimization*, 21(2):572–596, 2011.

[13] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, 20(1):33–61, 1998.

[14] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *ICML*, 2008.

[15] M. Fazel. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford, 2002. Available online: http://faculty.washington.edu/mfazel/thesis-final.pdf.

[16] R. Garg and R. Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*, 2009.

[17] E. T. Hale, Y. Wotao, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM J. on Optimization*, 19(3):1107–1130, 2008.

[18] D. Hsu, S.M. Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Trans. Info. Theory*, 57(11):7221 –7234, 2011.

[19] J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

[20] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *ICML*, 2009.

[21] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39:2302–2329, 2011.

[22] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

[23] K. Lee and Y. Bresler. Guaranteed minimum rank approximation from linear observations by nuclear norm minimization with an ellipsoidal constraint. Technical report, UIUC, 2009. Available at arXiv:0903.4742.

[24] P. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, 2012. To appear; originally posted as http://arxiv.org/abs/1109.3714.

[25] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. In *COLT*, 2009.

[26] Z. Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46-47:157–178, 1993.

[27] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[28] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *NIPS*, 2009. To appear in Statistical Science.

[29] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011.

[30] S. Negahban and M. J. Wainwright. Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, May 2012.

[31] Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, New York, 2004.

[32] Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 76, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 2007.

[33] H. V. Ngai and J. P. Penot. Paraconvex functions and paraconvex sets. *Studia Mathematica*, 184:1–29, 2008.

[34] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue conditions for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, August 2010.

[35] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Info. Theory*, 57(10):6976—6994, 2011.

[36] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[37] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[38] A. Rohde and A. Tsybakov. Estimation of high-dimensional low-rank matrices. *Annals of Statistics*, 39(2):887–930, 2011.

[39] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. Technical report, University of Michigan, July 2011.

[40] N. Srebro, N. Alon, and T. S. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *NIPS*, 2005.

[41] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[42] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Info. Theory*, 53(12):4655–4666, 2007.

[43] S. van de Geer and P. Buhlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[44] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. *IEEE Trans. Info. Theory*, 58(5):3047 –3064, May 2012.

[45] C. H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

[46] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.

Alekh Agarwal
Department of EECS
University of California Berkeley
Berkeley CA 94720
E-mail: alekh@eecs.berkeley.edu

Sahand Negahban
Department of EECS
Massachusetts Institute of Technology
32 Vassar Street
Cambirdge MA 02139
E-mail: sahandn@mit.edu

Martin J. Wainwright
Department of EECS and Statistics
University of California Berkeley
Berkeley CA 94720
E-mail: wainwrig@stat.berkeley.edu