Global Optimality of Local Search
for Low Rank Matrix Recovery

Srinadh Bhojanapalli,[*] Behnam Neyshabur,[†] Nathan Srebro[‡]

Toyota Technological Institute at Chicago

Abstract

We show that there are no spurious local minima in the non-convex factorized parametrization of low-rank matrix recovery from incoherent linear measurements. With noisy measurements we show all local minima are very close to a global optimum. Together with a curvature bound at saddle points, this yields a polynomial time global convergence guarantee for stochastic gradient descent *from random initialization*.

1 Introduction

Low rank matrix recovery problem is heavily studied and has numerous applications in collaborative filtering, quantum state tomography, clustering, community detection, metric learning and multi-task learning [21, 12, 9, 27].

We consider the "matrix sensing" problem of recovering a low-rank (or approximately low rank) p.s.d. matrix[1] $\boldsymbol{X}^* \in \mathbb{R}^{n \times n}$, given a linear measurement operator $\mathcal{A} : \mathbb{R}^{n \times n} \to \mathbb{R}^m$ and noisy measurements $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}^*) + \boldsymbol{w}$, where $\boldsymbol{w}$ is an i.i.d. noise vector. An estimator for $\boldsymbol{X}^*$ is given by the rank-constrained, non-convex problem

$$\underset{\boldsymbol{X} : \text{rank}(\boldsymbol{X}) \leq r}{\text{minimize}} \quad \|\mathcal{A}(\boldsymbol{X}) - \boldsymbol{y}\|^2. \tag{1}$$

This matrix sensing problem has received considerable attention recently [30, 29, 26]. This and other rank-constrained problems are common in machine learning and related fields, and have been used for applications discussed above. A typical theoretical approach to low-rank problems, including (1) is to relax the low-rank constraint to a convex constraint, such as the trace-norm of $\boldsymbol{X}$. Indeed, for matrix sensing, Recht et al. [20] showed that if the measurements are noiseless and the measurement operator $\mathcal{A}$ satisfies a restricted isometry property, then a low-rank $\boldsymbol{X}^*$ can be recovered as the unique solution to a convex relaxation of (1). Subsequent work established similar guarantees also for the noisy and approximate case [14, 6].

However, convex relaxations to the rank are not the common approach employed in practice. In this and other low-rank problems, the method of choice is typically unconstrained local optimization (via e.g. gradient descent, SGD or alternating minimization) on the factorized parametrization

$$\underset{\boldsymbol{U} \in \mathbb{R}^{n \times r}}{\text{minimize}} \, f(\boldsymbol{U}) = \|\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top) - \boldsymbol{y}\|^2, \tag{2}$$

where the rank constraint is enforced by limiting the dimensionality of $\boldsymbol{U}$. Problem (2) is a non-convex optimization problem that could have many bad local minima (as we show in Section 5), as well as saddle points. Nevertheless, local optimization seems to work very well in practice. Working on (2) is much cheaper computationally and allows scaling to large-sized problems—the number of optimization variables is only $O(nr)$ rather than $O(n^2)$, and the updates are

---

[*]srinadh@ttic.edu

[†]bneyshabur@ttic.edu

[‡]nati@ttic.edu

[1]We study the case where $\boldsymbol{X}^*$ is PSD. We believe the techniques developed here can be used to extend results to the general case.

usually very cheap, especially compared to typical methods for solving the SDP resulting from the convex relaxation. There is therefore a significant disconnect between the theoretically studied and analyzed methods (based on convex relaxations) and the methods actually used in practice.

Recent attempts at bridging this gap showed that, some form of global "initialization", typically relying on singular value decomposition, yields a solution that is already close enough to X^*; that local optimization from this initializer gets to the global optima (or to a good enough solution). Jain et al. [15], Keshavan [17] proved convergence for alternating minimization algorithm provided the starting point is close to the optimum, while Zheng and Lafferty [30], Zhao et al. [29], Tu et al. [26], Chen and Wainwright [8], Bhojanapalli et al. [2] considered gradient descent methods on the factor space and proved local convergence. But all these studies rely on global initialization followed by local convergence, and do not tackle the question of the existence of spurious local minima or deal with optimization starting from random initialization. There is therefore still a disconnect between this theory and the empirical practice of starting from random initialization and relying *only* on the local search to find the global optimum.

In this paper we show that, under a suitable incoherence condition on the measurement operator $\mathcal{A}$ (defined in Section 2), with noiseless measurements and with $\text{rank}(X^*) \leq r$, the problem (2) has no spurious local minima (i.e. all local minima are global and satisfy $X^* = UU^\top$). Furthermore, under the same conditions, all saddle points have a direction with significant negative curvature, and so using a recent result of Ge et al. [10] we can establish that stochastic gradient descent from random initialization converges to X^* in polynomial number of iterations. We extend the results also to the noisy and approximately-low-rank settings, where we can guarantee that every local minima is close to a global minimum. The incoherence condition we require is weaker than conditions used to establish recovery through local search, and so our results also ensures recovery in polynomial time under milder conditions than what was previously known. In particular, with i.i.d. Gaussian measurements, we ensure no spurious local minima and recovery through local search with the optimal number $O(nr)$ of measurements.

Related Work Our work is heavily inspired by Bandeira et al. [1], who recently showed similar behavior for the problem of community detection—this corresponds to a specific rank-1 problem with a linear objective, elliptope constraints and a binary solution. Here we take their ideas, extend them and apply them to matrix sensing with general rank-r matrices. In the past several months, similar type of results were also obtained for other non-convex problems (where the source of non-convexity is *not* a rank constraint), specifically complete dictionary learning [24] and phase recovery [25]. A related recent result of a somewhat different nature pertains to rank unconstrained linear optimization on the elliptope, showing that local minima of the rank-constrained problem approximate well the global optimum of the rank unconstrained convex problem, even though they might *not* be the global minima (in fact, the approximation guarantee for the actual global optimum is better) [18].

Another non-convex low-rank problem long known to not possess spurious local minima is the PCA problem, which can also be phrased as matrix approximation with full observations, namely $\min_{\text{rank}(X) \leq r} \|A - X\|_F$ (e.g. [23]). Indeed, local search methods such as the power-method are routinely used for this problem. Recently local optimization methods for the PCA problem working more directly on the optimized formulation have also been studied, including SGD [22] and Grassmannian optimization [28]. These results are somewhat orthogonal to ours, as they study a setting in which it is well known there are never any spurious local minima, and the challenge is obtaining satisfying convergence rates.

The seminal work of Burer and Monteiro [3] proposed low-rank factorized optimization for SDPs, and showed that for extremely high rank $r > \sqrt{m}$ (number of constraints), an Augmented Lagrangian method converges asymptotically to the optimum. It was also shown that (under mild conditions) any rank deficient local minima is a global minima [4, 16], providing a post-hoc verifiable sufficient condition for global optimality. However, this does not establish any a-priori condition, based on problem structure, implying the lack of spurious local minima.

While preparing this manuscript, we also became aware of parallel work [11] studying the same question for the related but different problem of matrix completion. For this problem they obtain a similar guarantee, though with suboptimal dependence on the incoherence parameters and so suboptimal sample complexity, and requiring adding a specific non-standard regularizer to the objective—this is not needed for our matrix sensing results.

We believe our work, together with the parallel work of [11], are the first to establish the lack of spurious local minima and the global convergence of local search from random initialization for a non-trivial rank-constrained problem (beyond PCA with full observations) with rank $r > 1$.

Notation. For matrices $X, Y \in \mathbb{R}^{n \times n}$, their inner product is $\langle X, Y \rangle = \text{trace}\left(X^\top Y\right)$. We use $\|X\|_F$, $\|X\|_2$ and

$\|X\|_*$ for the Frobenius, spectral and nuclear norms of a matrix respectively. Given a matrix X, we use $\sigma_i(X)$ to denote singular values of X in decreasing order. $X_r = \arg\min_{\mathrm{rank}(Y) \leq r} \|X - Y\|_F$ denotes the rank-r approximation of X, as obtained via its truncated singular value decomposition. We use plain capitals R and Q to denote orthonormal matrices.

2   Formulation and Assumptions

We write the linear measurement operator $\mathcal{A} : \mathbb{R}^{n \times n} \to \mathbb{R}^m$ as $\mathcal{A}(X)_i = \langle A_i, X \rangle$ where $A_i \in \mathbb{R}^{n \times n}$, yielding $y_i = \langle A_i, X^* \rangle + w_i, i = 1, \cdots, m$. We assume $w_i \sim \mathcal{N}(0, \sigma_w^2)$ is i.i.d Gaussian noise. We are generally interested in the high dimensional regime where the number of measurements m is usually much smaller than the dimension n^2.

Even if we know that $\mathrm{rank}(X^*) \leq r$, having many measurements might not be sufficient for recovery if they are not "spread out" enough. E.g., if all measurements only involve the first $n/2$ rows and columns, we would never have any information on the bottom-right block. A sufficient condition for identifiability of a low-rank X^* from linear measurements by Recht et al. [20] is based on restricted isometry property defined below.

Definition 2.1 (Restricted Isometry Property). *Measurement operator $\mathcal{A} : \mathbb{R}^{n \times n} \to \mathbb{R}^m$ (with rows A_i, $i = 1, \cdots, m$) satisfies (r, δ_r) RIP if for any $n \times n$ matrix X with rank $\leq r$,*

$$(1 - \delta_r)\|X\|_F^2 \leq \frac{1}{m} \sum_{i=1}^{m} \langle A_i, X \rangle^2 \leq (1 + \delta_r)\|X\|_F^2. \tag{3}$$

In particular, X^* of rank r is identifiable if $\delta_{2r} < 1$ [see 20, Theorem 3.2]. One situation in which RIP is obtained is for random measurement operators. For example, matrices with i.i.d. $\mathcal{N}(0, 1)$ entries satisfy (r, δ_r)-RIP when $m = O(\frac{nr}{\delta^2})$ [see 6, Theorem 2.3]. This implies identifiability based on i.i.d. Gaussian measurement with $m = O(nr)$ measurements (coincidentally, the number of degrees of freedom in X^*, optimal up to a constant factor).

3   Main Results

We are now ready to present our main result about local minima for the matrix sensing problem (2). We first present the results for noisy sensing of exact low rank matrices, and then generalize the results also to approximately low rank matrices.

Now we will present our result characterizing local minima of $f(U)$, for low-rank X^*. Recall that measurements are $y = \mathcal{A}(X^*) + w$, where entries of w are i.i.d. Gaussian - $w_i \sim \mathcal{N}(0, \sigma_w^2)$.

Theorem 3.1. *Consider the optimization problem* (2) *where $y = \mathcal{A}(X^*) + w$, w is i.i.d. $\mathcal{N}(0, \sigma_w^2)$, $\mathcal{A}$ satisfies $(2r, \delta_{2r})$-RIP with $\delta_{2r} < \frac{1}{10}$, and $\mathrm{rank}(X^*) \leq r$. Then, with probability $\geq 1 - \frac{10}{n^2}$ (over the noise), for any local minimum U of $f(U)$:*

$$\|UU^\top - X^*\|_F \leq 20\sqrt{\frac{\log(n)}{m}} \sigma_w.$$

In particular, in the noiseless case ($\sigma_w = 0$) we have $UU^\top = X^*$ and so $f(U) = 0$ and every local minima is global. In the noiseless case, we can also relax the RIP requirement to $\delta_{2r} < 1/5$ (see Theorem 4.1 in Section 4). In the noisy case we cannot expect to ensure we always get to an exact global minima, since the noise might cause tiny fluctuations very close to the global minima possibly creating multiple very close local minima. But we show that all local minima are indeed very close to some factorization $U^* {U^*}^\top = X^*$ of the true signal, and hence to a global optimum, and this "radius" of local minima decreases as we have more observations.

The proof of the Theorem for the noiseless case is presented in Section 4. The proof for the general setting follows along the same lines and can be found in the Appendix.

So far we have discussed how all local minima are global, or at least very close to a global minimum. Using a recent result by Ge et al. [10] on the convergence of SGD for non-convex functions, we can further obtain a polynomial bound on the number of SGD iterations required to reach the global minima. The main condition that needs to be established

in order to ensure this, is that all saddle points of (2) satisfy the "strict saddle point condition", i.e. have a direction with significant negative curvature:

Theorem 3.2 (Strict saddle). *Consider the optimization problem* (2) *in the noiseless case, where* $y = \mathcal{A}(X^*)$, $\mathcal{A}$ *satisfies* $(2r, \delta_{2r})$*-RIP with* $\delta_{2r} < \frac{1}{10}$, *and* $\operatorname{rank}(X^*) \leq r$. *Let* U *be a first order critical point of* $f(U)$ *with* $UU^\top \neq X^*$. *Then the smallest eigenvalue of the Hessian satisfies*

$$\lambda_{\min}\left[\frac{1}{m}\nabla^2(f(U))\right] \leq \frac{-4}{5}\sigma_r(X^*).$$

Now consider the stochastic gradient descent updates,

$$U^+ = \operatorname{Proj}_b\left(U - \eta\left(\sum_{i=1}^m(\langle A_i, UU^\top\rangle - y_i)A_i U + \psi\right)\right), \tag{4}$$

where ψ is uniformly distributed on the unit sphere and Proj_b is a projection onto $\|U\|_F \leq b$. Using Theorem 3.2 and the result of Ge et al. [10] we can establish:

Theorem 3.3 (Convergence from random initialization). *Consider the optimization problem* (2) *under the same noiseless conditions as in Theorem 3.2. Using* $b \geq \|U^*\|_F$, *for some global optimum* U^* *of* $f(U)$, *for any* $\epsilon, c > 0$, *after* $T = poly\left(\frac{1}{\sigma_r(X^*)}, \sigma_1(X^*), b, \frac{1}{\epsilon}, \log(1/c)\right)$ *iterations of* (4) *with an appropriate stepsize* η, *starting from a random point uniformly distributed on* $\|U\|_F = b$, *with probability at least* $1 - c$, *we reach an iterate* U_T *satisfying*

$$\|U_T - U^*\|_F \leq \epsilon.$$

The above result guarantees convergence of noisy gradient descent to a global optimum. Alternatively, second order methods such as cubic regularization (Nesterov and Polyak [19]) and trust region (Cartis et al. [7]) that have guarantees based on the strict saddle point property can also be used here.

RIP Requirement: Our results require $(2r, 1/10)$-RIP for the noisy case and $(2r, 1/5)$-RIP for the noiseless case. Requiring $(2r, \delta_{2r})$-RIP with $\delta_{2r} < 1$ is sufficient to ensure uniqueness of the global optimum of (1), and thus recovery in the noiseless setting [20], but all known efficient recovery methods require stricter conditions. The best guarantees we are aware of require $(5r, 1/10)$-RIP [20] or $(4r, 0.414)$-RIP [6] using a convex relaxation. Our requirement is not directly comparable to the latter, as we require RIP on a smaller set, but with a lower (stricter) δ. Alternatively, $(6r, 1/10)$-RIP is required for global initialization followed by non-convex optimization [26]—our requirement is strictly better. In terms of requirements on $(2r, \delta_{2r})$-RIP for non-convex methods, the best we are aware of is requiring $\delta_{2r} < \Omega(1/r)$ [15, 29, 30]–this is a much stronger condition than ours, and it yields a suboptimal required number of spherical Gaussian measurements of $\Omega(nr^3)$. So, compared to prior work our requirement is very mild—it ensures efficient recovery even in a regime not previously covered by any guarantee on efficient recovery, and requires the optimal number of spherical Gaussian measurements (up to a constant factor) of $O(nr)$.

Extension to Approximate Low Rank We can also obtain similar results that deteriorate gracefully if X^* is not exactly low rank, but is close to being low-rank (see proof in the Appendix):

Theorem 3.4. *Consider the optimization problem* (2) *where* $y = \mathcal{A}(X^*)$ *and* $\mathcal{A}$ *satisfies* $(2r, \delta_{2r})$*-RIP with* $\delta_{2r} < \frac{1}{100}$, *Then, for any local minima* U *of* $f(U)$:

$$\|UU^\top - X^*\|_F \leq 4(\|X^* - X_r^*\|_F + \delta_{2r}\|X^* - X_r^*\|_*),$$

where X_r^* *is the best rank* r *approximation of* X^*.

This theorem guarantees that any local optimum of $f(U)$ is close to X^* upto an error depending on $\|X^* - X_r^*\|$. For the low-rank noiseless case we have $X^* = X_r^*$ and the right hand side vanishes. When X^* is not exactly low rank, the best recovery error we can hope for is $\|X^* - X_r^*\|_F$, since $UU^\top$ is at most rank k. On the right hand side of Theorem 3.4, we have also a nuclear norm term, which might be higher, but it also gets scaled down by δ_{2r}, and so by the number of measurements.
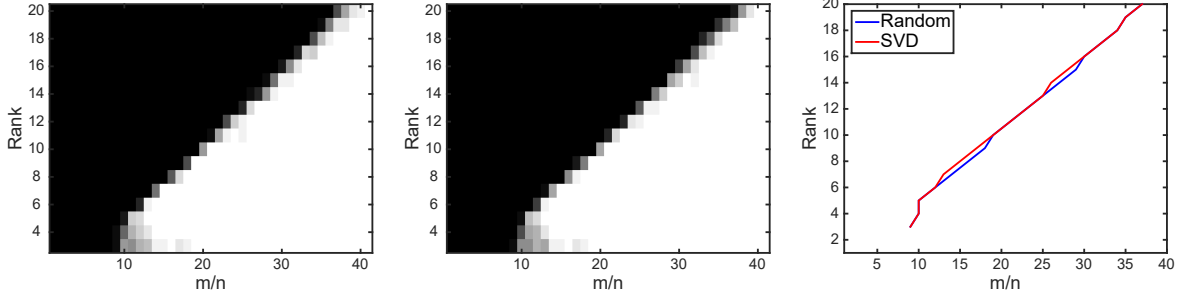
Figure 1: The plots in this figure compare the success probability of gradient descent between (left) random and (center) SVD initialization (suggested in [15]), for problem (2), with increasing number of samples m and various values of rank r. Right most plot is the first m for a given r, where the probability of success reaches the value 0.5. A run is considered success if $\|UU^\top - X^*\|_F/\|X^*\|_F \leq 1e-2$. White cells denote success and black cells denote failure of recovery. We set n to be 100. Measurements y_i are inner product of entrywise i.i.d Gaussian matrix and a rank-r p.s.d matrix with random subspace. We notice no significant difference between the two initialization methods, suggesting absence of local minima as shown. Both methods have phase transition around $m = 2 \cdot n \cdot r$.

4   Proof for the Noiseless Case

In this section we present the proof characterizing the local minima of problem (2). For ease of exposition we first present the results for the noiseless case ($w = 0$). Proof for the general case can be found in the Appendix.

Theorem 4.1. *Consider the optimization problem* (2) *where* $y = \mathcal{A}(X^*)$, $\mathcal{A}$ *satisfies* $(2r, \delta_{2r})$-*RIP with* $\delta_{2r} < \frac{1}{5}$, *and* $\mathrm{rank}(X^*) \leq r$. *Then, for any local minimum* U *of* $f(U)$:

$$UU^\top = X^*.$$

For the proof of this theorem we first discuss the implications of the first and second order optimality conditions and then show how to combine them to yield the result.

Our proof techniques are different from existing results characterizing local minima of dictionary learning [24], phase retrieval [25] and community detection [1, 18]. For most of these problems Hessian is PSD only for points close to optimum. However, Hessian of $f(U)$ can be PSD even for points far from to optima. Hence we need new directions to use the second order conditions.

Invariance of $f(U)$ over $r \times r$ orthonormal matrices introduces additional challenges in comparing a given stationary point to a global optimum. We have to find the best orthonormal matrix R to align a given stationary point U to a global optimum U^*, where $U^*U^{*\top} = X^*$, to combine results from the first and second order conditions, without degrading the isometry constants.

Consider a local optimum U that satisfies first and second order optimality conditions of problem (2). In particular U satisfies $\nabla f(U) = 0$ and $z^\top \nabla^2 f(U) z \geq 0$ for any $z \in \mathbb{R}^{n \cdot r}$. Now we will see how these two conditions constrain the error $UU^\top - U^*U^{*\top}$.

First we present the following consequence of the RIP assumption [see 5, Lemma 2.1].

Lemma 4.1. *Given two* $n \times n$ *rank-r matrices* X *and* Y, *and a* $(2r, \delta)$-*RIP measurement operator* $\mathcal{A}$, *the following holds:*

$$\left| \frac{1}{m} \sum_{i=1}^{m} \langle A_i, X \rangle \langle A_i, Y \rangle - \langle X, Y \rangle \right| \leq \delta \|X\|_F \|Y\|_F. \tag{5}$$

5

4.1 First order optimality

First we will consider the first order condition, $\nabla f(U) = 0$. For any stationary point U this implies

$$\sum_i \left\langle A_i, UU^\top - U^*U^{*\top} \right\rangle A_i U = 0. \tag{6}$$

Now using the isometry property of A_i gives us the following result.

Lemma 4.2. *[First order condition] For any first order stationary point U of $f(U)$, and $\mathcal{A}$ satisfying the $(2r, \delta)$-RIP (3), the following holds:*

$$\|(UU^\top - U^*U^{*\top})QQ^\top\|_F \le \delta \left\|UU^\top - U^*U^{*\top}\right\|_F,$$

where Q is an orthonormal matrix that spans the column space of U.

This lemma states that any stationary point of $f(U)$ is close to a global optimum U^* in the subspace spanned by columns of U. Notice that the error along the orthogonal direction $\|X^*Q_\perp Q_\perp^\top\|_F$ can still be large making the distance between X and X^* arbitrarily far.

Proof of Lemma 4.2. Let $U = QR$, for some orthonormal Q. Consider any matrix of the form $ZQR^{-1^\top}$. The first order optimality condition then implies,

$$\sum_{i=1}^m \left\langle A_i, UU^\top - U^*U^{*\top} \right\rangle \langle A_i, UR^{-1}Q^\top Z^\top \rangle = 0$$

The above equation together with Restricted Isometry Property (equation (5)) gives us the following inequality:

$$\left| \left\langle UU^\top - U^*U^{*\top}, QQ^\top Z^\top \right\rangle \right| \le \delta \left\|UU^\top - U^*U^{*\top}\right\|_F \|QQ^\top Z^\top\|_F.$$

Note that for any matrix A, $\langle A, QQ^\top Z \rangle = \langle QQ^\top A, Z \rangle$. Furthermore, for any matrix A, $\sup_{\{Z : \|Z\|_F \le 1\}} \langle A, Z \rangle = \|A\|_F$. Hence the above inequality implies the lemma statement. $\square$

4.2 Second order optimality

We now consider the second order condition to show that the error along $Q_\perp Q_\perp^\top$ is indeed bounded well. Let $\nabla^2 f(U)$ be the hessian of the objective function. Note that this is an $n \cdot r \times n \cdot r$ matrix. Fortunately for our result we need to only evaluate the Hessian along the direction $\text{vec}(U - U^*R)$ for some orthonormal matrix R. Here $\text{vec}(.)$ denotes writing a matrix in vector form.

Lemma 4.3. *[Hessian computation] Let U be a first order critical point of $f(U)$. Then for any $r \times r$ orthonormal matrix R and $\Delta_j = \Delta e_j e_j^\top$ ($\Delta = U - U^*R$),*

$$\sum_{j=1}^r vec(\Delta_j)^\top \left[\nabla^2 f(U)\right] vec(\Delta_j) = \sum_{i=1}^m (\sum_{j=1}^r 4 \left\langle A_i, U\Delta_j^\top \right\rangle^2 - 2 \left\langle A_i, UU^\top - U^*U^{*\top} \right\rangle^2),$$

Proof of Lemma 4.3. For any matrix Z, taking directional second derivative of the function $f(U)$ with respect to Z we get:

$$\text{vec}(Z)^\top \left[\nabla^2 f(U)\right] \text{vec}(Z) = \text{vec}(Z)^\top \lim_{t \to 0} \left[\frac{\nabla f(U + t(Z)) - \nabla f(U)}{t}\right]$$

$$= 2\sum_{i=1}^m \left[2\left\langle A_i, UZ^\top \right\rangle^2 + \left\langle A_i, UU^\top - U^*U^{*\top} \right\rangle \langle A_i, ZZ^\top \rangle\right]$$

6

Setting $\boldsymbol{Z} = \Delta_j = (\boldsymbol{U} - \boldsymbol{U}^* R)e_j e_j^\top$ and using the first order optimality condition on $\boldsymbol{U}$, we get,

$$
\texttt{vec}\left((\boldsymbol{U} - \boldsymbol{U}^* R)e_j e_j^\top\right)^\top \left[\nabla^2 f(\boldsymbol{U})\right] \texttt{vec}\left((\boldsymbol{U} - \boldsymbol{U}^* R)e_j e_j^\top\right)
$$

$$
= \sum_{i=1}^m 4\left\langle \boldsymbol{A}_i, \boldsymbol{U}\Delta_j^\top\right\rangle^2 + 2\left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\right\rangle\left\langle \boldsymbol{A}_i, \Delta_j\Delta_j^\top\right\rangle
$$

$$
\overset{(i)}{=} \sum_{i=1}^m 4\left\langle \boldsymbol{A}_i, \boldsymbol{U}e_j e_j^\top\Delta_j^\top\right\rangle^2 + 2\left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\right\rangle\left\langle \boldsymbol{A}_i, \boldsymbol{U}^*e_j e_j^\top(\boldsymbol{U}^*e_j e_j^\top)^\top\right\rangle
$$

$$
\overset{(ii)}{=} \sum_{i=1}^m 4\left\langle \boldsymbol{A}_i, \boldsymbol{U}e_j e_j^\top\Delta_j^\top\right\rangle^2 - 2\left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\right\rangle\left\langle \boldsymbol{A}_i, \boldsymbol{U}e_j e_j^\top\boldsymbol{U}^\top - \boldsymbol{U}^*e_j e_j^\top\boldsymbol{U}^{*\top}\right\rangle .
$$

(i) and (ii) follow from the first order optimality condition (6),

$$
\sum_{i=1}^m \left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top\right\rangle \boldsymbol{U}e_j e_j^\top = \sum_{i=1}^m \left\langle \boldsymbol{A}_i, \boldsymbol{U}^*\boldsymbol{U}^{*\top}\right\rangle \boldsymbol{U}e_j e_j^\top ,
$$

for $j = 1 \cdots r$. Finally taking sum over j from 1 to r gives the result. $\qquad \square$

Hence from second order optimality of $\boldsymbol{U}$ we get,

Corollary 4.1. *[Second order optimality] Let $\boldsymbol{U}$ be a local minimum of $f(\boldsymbol{U})$. For any $r \times r$ orthonormal matrix R,*

$$
\sum_{j=1}^r \sum_{i=1}^m 4\left\langle \boldsymbol{A}_i, \boldsymbol{U}\Delta_j^\top\right\rangle^2 \geq \frac{1}{2}\sum_{i=1}^m \left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\right\rangle^2 , \tag{7}
$$

Further for $\mathcal{A}$ satisfying $(2r, \delta)$ -RIP (equation (3)) we have,

$$
\sum_{j=1}^r \|\boldsymbol{U}e_j e_j^\top(\boldsymbol{U} - \boldsymbol{U}^* R)^\top\|_F^2 \geq \frac{1-\delta}{2(1+\delta)}\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\|_F^2 . \tag{8}
$$

The proof of this result follows simply by applying Lemma 4.3. The above Lemma gives a bound on the distance in the factor $(\boldsymbol{U})$ space $\|\boldsymbol{U}(\boldsymbol{U} - \boldsymbol{U}^* R)^\top\|_F^2$. To be able to compare the second order condition to the first order condition we need a relation between $\|\boldsymbol{U}(\boldsymbol{U} - \boldsymbol{U}^* R)^\top\|_F^2$ and $\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2$. Towards this we show the following result.

Lemma 4.4. *Let $\boldsymbol{U}$ and $\boldsymbol{U}^*$ be two $n \times r$ matrices, and Q is an orthonormal matrix that spans the column space of $\boldsymbol{U}$. Then there exists an $r \times r$ orthonormal matrix R such that for any first order stationary point $\boldsymbol{U}$ of $f(\boldsymbol{U})$, the following holds:*

$$
\sum_{j=1}^r \|\boldsymbol{U}e_j e_j^\top(\boldsymbol{U} - \boldsymbol{U}^* R)^\top\|_F^2 \leq \frac{1}{8}\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\|_F^2 + \frac{34}{8}\|(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top})QQ^\top\|_F^2 .
$$

This Lemma bounds the distance in the factor space ($\|(\boldsymbol{U} - \boldsymbol{U}^* R)\boldsymbol{U}^\top\|_F^2$) with $\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\|_F^2$ and $\|(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top})QQ^\top\|_F^2$. Combining this with the result from second order optimality (Corollary 4.1) shows $\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\|_F^2$ is bounded by a constant factor of $\|(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top})QQ^\top\|_F^2$. This implies $\|\boldsymbol{X}^*Q_\perp Q_\perp\|_F$ is bounded, opposite to what the first order condition implied (Lemma 4.2). The proof of the above lemma is in Section E. Hence from the above optimality conditions we get the proof of Theorem 4.1.

Proof of Theorem 4.1. Assuming $\boldsymbol{U}\boldsymbol{U}^\top \neq \boldsymbol{U}^*\boldsymbol{U}^{*\top}$, from Lemmas 4.2, 4.4 and Corollary 4.1 we get,

$$
\left(\frac{1-\delta}{2(1+\delta)} - \frac{1}{8}\right)\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\|_F^2 \leq \frac{34}{8}\delta^2 \left\|(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top})\right\|_F^2 .
$$

If $\delta \leq \frac{1}{5}$ the above inequality holds only if $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{U}^*\boldsymbol{U}^{*\top}$. $\qquad \square$

5   Necessity of RIP

We showed that there are no spurious local minima only under a restricted isometry assumption. A natural question is whether this is necessary, or whether perhaps the problem (2) never has any spurious local minima, perhaps similarly to the non-convex PCA problem $\min_{U} \left\| \boldsymbol{A} - \boldsymbol{U}\boldsymbol{U}^{\top} \right\|$.

A good indication that this is not the case is that (2) is NP-hard, even in the noiseless case when $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}^*)$ for $\mathrm{rank}(\boldsymbol{X}^*) \leq k$ [20] (if we don't require RIP, we can have each $\boldsymbol{A}_i$ be non-zero on a single entry in which case (2) becomes a matrix completion problem, for which hardness has been shown even under fairly favorable conditions [13]). That is, we are unlikely to have a poly-time algorithm that succeeds for any linear measurement operator. Although this doesn't formally preclude the possibility that there are no spurious local minima, but it just takes a very long time to find a local minima, this scenario seems somewhat unlikely.

To resolve the question, we present an explicit example of a measurement operator $\mathcal{A}$ and $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}^*)$ (i.e. $f(\boldsymbol{X}^*) = 0$), with $\mathrm{rank}(\boldsymbol{X}^*) = r$, for which (1), and so also (2), have a non-global local minima.

Example 1: Let $f(\boldsymbol{X}) = (X_{11} + X_{22} - 1)^2 + (X_{11} - 1)^2$ and consider (1) with $r = 1$ (i.e. a rank-1 constraint). For $\boldsymbol{X}^* = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ we have $f(\boldsymbol{X}^*) = 0$ and $\mathrm{rank}(\boldsymbol{X}^*) = 1$. But $\boldsymbol{X} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ is a rank 1 local minimum with $f(\boldsymbol{X}) = 1$.

We can be extended the construction to any rank r by simply adding $\sum_{i=3}^{r+2} (X_{ii} - 1)^2$ to the objective, and padding both the global and local minimum with a diagonal beneath the leading 2×2 block.

In Example 1, we had a rank-r problem, with a rank-r exact solution, and a rank-r local minima. Another question we can ask is what happens if we allow a larger rank than the rank of the optimal solution. That is, if we have $f(\boldsymbol{X}^*) = 0$ with low $\mathrm{rank}(\boldsymbol{X}^*)$, even $\mathrm{rank}(\boldsymbol{X}^*) = 1$, but consider (1) or (2) with a high r. Could we still have non-global local minima? The answer is yes...

Example 2: Let $f(\boldsymbol{X}) = (X_{11} + X_{22} + X_{33} - 1)^2 + (X_{11} - 1)^2 + (X_{22} - X_{33})^2$ and consider the problem (1) with a rank $r = 2$ constraint. We can verify that $\boldsymbol{X}^* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ is a rank=1 global minimum with $f(\boldsymbol{X}^*) = 0$, but $\boldsymbol{X} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \end{bmatrix}$ is a local minimum with $f(\boldsymbol{X}) = 1$. Also for an arbitrary large rank constraint $r > 1$ (taking r to be odd for simplicity), extend the objective to $f(\boldsymbol{X}) = (X_{11} - 1)^2 + \sum_{i=1}^{(r-1)/2} \left[(X_{11} + X_{2i,2i} + X_{(2i+1),(2i+1)} - 1)^2 + (X_{2i,2i} - X_{(2i+1),(2i+1)})^2 \right]$. We still have a rank-1 global minimum $\boldsymbol{X}^*$ with a single non-zero entry $X_{11}^* = 1$, while $\boldsymbol{X} = (I - \boldsymbol{X}^*)/2$ is a local minimum with $f(\boldsymbol{X}) = 1$.

6   Conclusion

We established that under conditions similar to those required for convex relaxation recovery guarantees, the non-convex formulation of matrix sensing (2) does not exhibit any spurious local minima (or, in the noisy and approximate settings, at least not outside some small radius around a global minima), and we can obtain theoretical guarantees on the success of optimizing it using SGD from *random initialization*. This matches the methods frequently used in practice, and can explain their success. This guarantee is very different in nature from other recent work on non-convex optimization for low-rank problems, which relied heavily on initialization to get close to the global optimum, and on local search just for the final local convergence to the global optimum. We believe this is the first result, together with the parallel work of Ge et al. [11], on the global convergence of local search for common rank-constrained problems that are worst-case hard.

Our result suggests that SVD initialization is not necessary for global convergence, and random initialization would succeed under similar conditions (in fact, our conditions are even weaker than in previous work that used SVD initialization). To investigate empirically whether SVD initialization is indeed helpful for ensuring global convergence, in Figure 1 we compare recovery probability of random rank-k matrices for random and SVD initialization—there is no significant difference between the two.

Beyond the implications for matrix sensing, we are hoping these type of results could be a first step and serve as a model for understanding local search in deep networks. Matrix factorization, such as in (2), is a depth-two neural network with linear transfer—an extremely simple network, but already non-convex and arguably the most complicated

network we have a good theoretical understanding of. Deep networks are also hard to optimize in the worst case, but local search seems to do very well in practice. Our ultimate goal is to use the study of matrix recovery as a guide in understating the conditions that enable efficient training of deep networks.

Acknowledgements

References

[1] A. S. Bandeira, N. Boumal, and V. Voroninski. On the low-rank approach for semidefinite programs arising in synchronization and community detection. *arXiv preprint arXiv:1602.04426*, 2016.

[2] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi. Dropping convexity for faster semi-definite optimization. *arXiv preprint arXiv:1509.03917*, 2015.

[3] S. Burer and R. D. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[4] S. Burer and R. D. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

[5] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

[6] E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011.

[7] C. Cartis, N. I. Gould, and P. L. Toint. Complexity bounds for second-order optimality in unconstrained optimization. *Journal of Complexity*, 28(1):93–108, 2012.

[8] Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[9] S. Flammia, D. Gross, Y.-K. Liu, and J. Eisert. Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012.

[10] R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

[11] R. Ge, J. Lee, and T. Ma. Matrix completion has no spurious local minimum. *arXiv preprint arXiv:1605.07272*, 2016.

[12] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.

[13] M. Hardt, R. Meka, P. Raghavendra, and B. Weitz. Computational limits for matrix completion. In *Proceedings of The 27th Conference on Learning Theory*, pages 703–725, 2014.

[14] P. Jain, R. Meka, and I. S. Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010.

[15] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the 45th annual ACM Symposium on theory of computing*, pages 665–674. ACM, 2013.

[16] M. Journée, F. Bach, P.-A. Absil, and R. Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

[17] R. H. Keshavan. *Efficient algorithms for collaborative filtering*. PhD thesis, STANFORD, 2012.

[18] A. Montanari. A grothendieck-type inequality for local maxima. *arXiv preprint arXiv:1603.04064*, 2016.

[19] Y. Nesterov and B. T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

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

[21] J. D. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM, 2005.

[22] C. D. Sa, C. Re, and K. Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2332–2341, 2015.

[23] N. Srebro and T. Jaakkola. Weighted low-rank approximations. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 720–727, 2003.

[24] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery using nonconvex optimization. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 2351–2360, 2015.

[25] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *preprint arXiv:1602.06664*, 2016.

[26] S. Tu, R. Boczar, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via Procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

[27] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon. Large-scale multi-label learning with missing labels. In *Proceedings of The 31st International Conference on Machine Learning*, pages 593–601, 2014.

[28] D. Zhang and L. Balzano. Global convergence of a grassmannian gradient descent algorithm for subspace estimation. *arXiv preprint arXiv:1506.07405*, 2015.

[29] T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.

[30] Q. Zheng and J. Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, pages 109–117, 2015.

A Numerical Simulations

In this section we present simulation results for performance of gradient descent over $f(U)$. We consider measurements $y_i = \langle A_i, X^* \rangle$, where A_i are i.i.d Gaussian with each entry distributed as $\mathcal{N}(0, 1/m)$. X^* is a 100×100 rank r random p.s.d matrix with $\|X^*\|_F = 1$. r is varied from 1 to 20 in the experiments.

We consider both standard gradient descent and noisy gradient descent (4) with step size $\frac{1}{\|U\|_2}$. We add noise of magnitude $1e-4$ for the noisy gradient updates. Each method is run until convergence (max of 200 iterations). Let the output of gradient descent be $\widehat{U}$. A run of this experiment is considered success if the final error $\|\widehat{U}\widehat{U}^\top - X^*\|_F \leq 1e-2$. Each experiment is repeated for 20 times and average probability of success is computed.

We repeat the above procedure starting from both random initialization and SVD initialization. For SVD initialization, the initial point is set to be the rank r approximation of $\sum_{i=1}^m y_i A_i$ as suggested by Jain et al. [15]. In figure 2 we have the plots for the cases discussed above. All of them have phase transition around number of samples $m = 2 \cdot n \cdot r$. This is in agreement with the results in Section 3. $f(U)$ has no local minima once $m \geq 2 \cdot n \cdot r$ and random initialization has same performance as SVD initialization.

In figure 3, the left two plots show error $\|\widehat{U}\widehat{U}^\top - X^*\|_F / \|X^*\|_F$ behaves with varying rank and number of samples for random and SVD initializations. The rightmost plot shows the phase transition for rank 10 case for all the methods. Again we notice no significant difference between these methods.
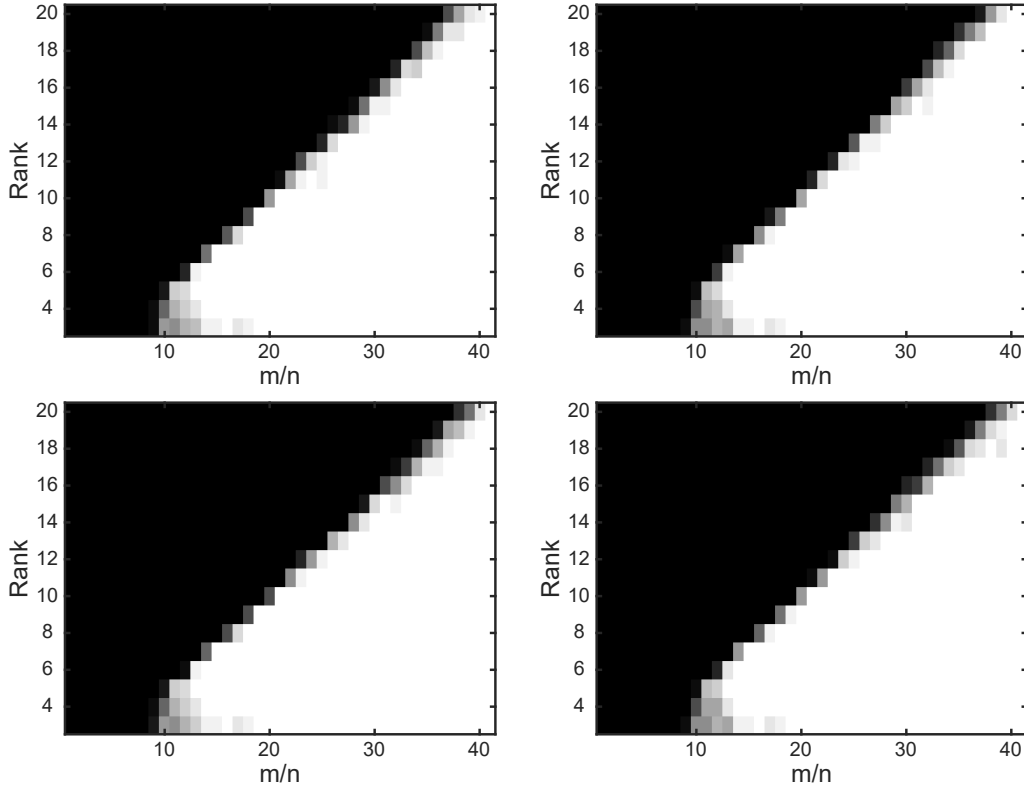

Figure 2: This figure plots the success probability for increasing number of samples m and various values of rank r. The plots on the top are for gradient descent, left for random initialization and the right for SVD initialization. Similarly the bottom plots are for the noisy gradient descent. We notice no significant difference between all these settings. They all have phase transition around $m = 2 \cdot n \cdot r$.
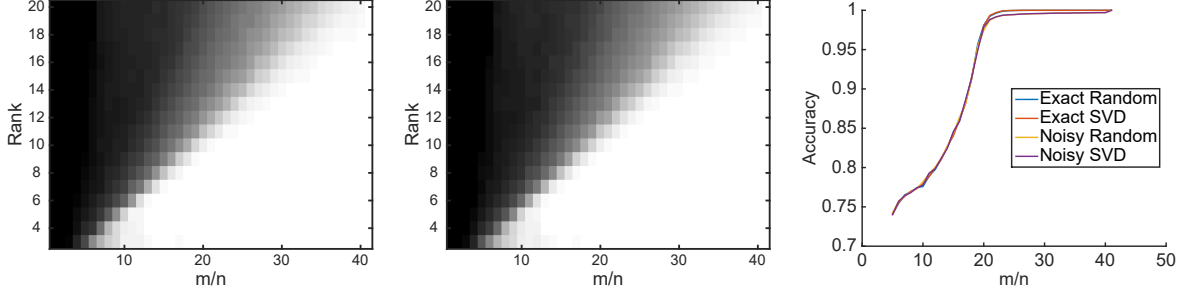
Figure 3: This figure plots the error $\|\widehat{U}\widehat{U}^\top - X^*\|_F / \|X^*\|_F$ for increasing number of samples m. The left plot is for gradient descent with random initialization, center plot corresponds to SVD initialization. Again we notice no difference in error for these two settings. The rightmost figure shows phase transition of low rank recovery for all the different settings when X^* is rank 10.

B   Proof for the Noisy Case

In this section we present the proof characterizing the local minima of problem (2). Recall $y = \mathcal{A}(X^*) + w$, where X^* is a rank-r matrix and w is i.i.d. $\mathcal{N}(0, \sigma_w^2)$.

We consider local optimum that satisfies first and second order optimality conditions of problem (2). In particular U satisfies $\nabla f(U) = 0$ and $z^\top \nabla^2 f(U) z \geq 0$ for any $z \in \mathbb{R}^{n \cdot r}$. Now we will see how these two conditions constrain the error $UU^\top - U^* U^{*\top}$.

B.1   First order optimality

First we will consider the first order condition, $\nabla f(U) = 0$. For any stationary point U this implies

$$\sum_i \left\langle A_i, UU^\top - U^* U^{*\top} \right\rangle A_i U = \sum_{i=1}^m w_i A_i U. \tag{9}$$

Now using the isometry property of A_i gives us the following result.

Lemma B.1. *[First order condition] For any first order stationary point U of $f(U)$, and $\mathcal{A}$ satisfying the $(2r, \delta)$-RIP (3), the following holds:*

$$\|(UU^\top - U^* U^{*\top})QQ^\top\|_F \leq \delta \left\|UU^\top - U^* U^{*\top}\right\|_F + 2\sqrt{\frac{(1+\delta)\log(n)}{m}}\sigma_w,$$

w.p. $\geq 1 - \frac{1}{n^2}$, where Q is an orthonormal matrix that spans the column space of U.

This lemma states that any stationary point of $f(U)$ is close to a global optimum U^* in the subspace spanned by columns of U. Notice that the error along the orthogonal direction $\|X^* Q_\perp Q_\perp^\top\|_F$ can still be large making the distance between X and X^* arbitrarily big.

Proof of Lemma B.1. Let $U = QR$, for some orthonormal Q. Consider any matrix of the form $ZQR^{-1^\top}$. The first order optimality condition then implies,

$$\sum_{i=1}^m \left\langle A_i, UU^\top - U^* U^{*\top} \right\rangle \left\langle A_i, UR^{-1}Q^\top Z^\top \right\rangle = \sum_{i=1}^m w_i A_i UR^{-1}Q^\top Z^\top.$$

The above equation together with Restricted Isometry Property(equation (5)) gives us the following inequality:

$$\left|\left\langle UU^\top - U^* U^{*\top}, QQ^\top Z^\top \right\rangle\right| \leq \delta \left\|UU^\top - U^* U^{*\top}\right\|_F \|QQ^\top Z^\top\|_F + 2\sqrt{\frac{(1+\delta)\log(n)}{m}}\sigma_w \|Z^\top\|_F,$$

12

by Cauchy Schwarz inequality and Lemma E.2. Note that for any matrix A, $\left\langle A, QQ^\top Z\right\rangle = \left\langle AQQ^\top, Z\right\rangle$. Furthermore, for any matrix A, $\sup_{\{Z : \|Z\|_F \leq 1\}} \langle A, Z\rangle = \|A\|_F$. Hence the above inequality implies the lemma statement. $\qquad\square$

B.2 Second order optimality

We will now consider the second order condition to show that the error along $Q_\perp Q_\perp^\top$ is indeed bounded well. Let $\nabla^2 f(U)$ be the hessian of the objective function. Note that this is an $n \cdot r \times n \cdot r$ matrix. Fortunately for our result we need to only evaluate the Hessian along the direction $\text{vec}(U - U^*R)$ for some orthonormal matrix R.

Lemma B.2. *[Hessian computation] Let U be a first order critical point of $f(U)$. Then for any $r \times r$ orthonormal matrix R and $\Delta = U - U^*R$,*

$$
\sum_{j=1}^r vec\left(\Delta_j\right)^\top \left[\nabla^2 f(U)\right] vec\left(\Delta_j\right)
$$

$$
= \sum_{i=1}^m \left(\sum_{j=1}^r 4\left\langle A_i, U\Delta_j^\top\right\rangle^2 - 2\left\langle A_i, UU^\top - U^*U^{*\top}\right\rangle^2 - 2w_i\left\langle A_i, X - X^*\right\rangle\right),
$$

Proof of Lemma B.2. For any matrix Z, taking directional second derivative of the function $f(U)$ with respect to Z we get:

$$
vec\left(Z\right)^\top \left[\nabla^2 f(U)\right] vec\left(Z\right) = vec\left(Z\right)^\top \lim_{t \to 0} \left[\frac{\nabla f\left(U + t(Z)\right) - \nabla f(U)}{t}\right]
$$

$$
= 2\sum_{i=1}^m \left[2\left\langle A_i, UZ^\top\right\rangle^2 + \left(\left\langle A_i, UU^\top - U^*U^{*\top}\right\rangle - w_i\right)\left\langle A_i, ZZ^\top\right\rangle\right]
$$

Setting $Z = \Delta_j = (U - U^*R)e_j e_j^\top$ and using the first order optimality condition on U, we get,

$$
vec\left((U - U^*R)e_j e_j^\top\right)^\top \left[\nabla^2 f(U)\right] vec\left((U - U^*R)e_j e_j^\top\right)
$$

$$
= \sum_{i=1}^m 4\left\langle A_i, U\Delta_j^\top\right\rangle^2 + 2(\left\langle A_i, UU^\top - U^*U^{*\top}\right\rangle - w_i)\left\langle A_i, \Delta_j\Delta_j^\top\right\rangle
$$

$$
= \sum_{i=1}^m 4\left\langle A_i, Ue_j e_j^\top\Delta_j^\top\right\rangle^2 + 2(\left\langle A_i, UU^\top - U^*U^{*\top}\right\rangle - w_i)\left\langle A_i, U^*e_j e_j^\top(U^*e_j e_j^\top)^\top\right\rangle
$$

$$
= \sum_{i=1}^m 4\left\langle A_i, Ue_j e_j^\top\Delta_j^\top\right\rangle^2 - 2\left\langle A_i, UU^\top - U^*U^{*\top}\right\rangle\left\langle A_i, Ue_j e_j^\top U^\top - U^*e_j e_j^\top U^{*\top}\right\rangle
$$

$$
\quad - 2w_i\left\langle A_i, Ue_j e_j^\top U^\top - U^*e_j e_j^\top U^{*\top}\right\rangle. \tag{10}
$$

where the last equality is again by the first order optimality condition (9). $\qquad\square$

Hence from second order optimality of U we get,

Corollary B.1. *[Second order optimality] Let U be a local minimum of $f(U)$. For any $r \times r$ orthonormal matrix R, w.p. $\geq 1 - \frac{1}{n^2}$,*

$$
\frac{1}{2}\sum_{i=1}^m \left\langle A_i, UU^\top - U^*U^{*\top}\right\rangle^2 \leq \sum_{i=1}^m\sum_{j=1}^r \left\langle A_i, U\Delta_j^\top\right\rangle^2 + \sqrt{\log(n)}\sigma_w\|\mathcal{A}(X - X^*)\|_2
$$

$$
\leq \sum_{i=1}^m\sum_{j=1}^r \left\langle A_i, U\Delta_j^\top\right\rangle^2 + 5\log(n)\sigma_w^2 + \frac{1}{20}\sum_{i=1}^m \left\langle A_i, X - X^*\right\rangle^2
$$

13

Further for $\mathcal{A}$ satisfying $(2r, \delta)$ -RIP (equation (3)) we have,

$$\frac{1-\delta}{2(1+\delta)}\|\boldsymbol{UU}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\|_F^2 \leq \sum_{j=1}^r \|\boldsymbol{U}\Delta_j^\top\|_F^2 + \frac{1}{20}\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2 + \frac{5\log(n)}{m(1+\delta)}\sigma_w^2. \tag{11}$$

Hence from the above optimality conditions we get the proof of Theorem 4.1.

Proof of Theorem 3.1. Assuming $\boldsymbol{UU}^\top \neq \boldsymbol{U}^*\boldsymbol{U}^{*\top}$, from Lemma 4.4 and Corollary B.1 we get, with probability $\geq 1 - \frac{2}{n^2}$,

$$\left(\frac{1-\delta}{2(1+\delta)}\right)\|\boldsymbol{UU}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\|_F^2$$

$$\leq \frac{1}{8}\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2 + \frac{34}{8}\|\boldsymbol{X} - \boldsymbol{X}^*QQ^\top\|_F^2 + \frac{1}{20}\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2 + \frac{5\log(n)}{m(1+\delta)}\sigma_w^2$$

$$\overset{(i)}{\leq} \left(\frac{1}{8} + \frac{1}{20}\right)\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2 + \frac{34}{8}\left(2\delta^2\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2 + 8\frac{(1+\delta)\log(n)}{m}\sigma_w^2\right) + \frac{5\log(n)}{m(1+\delta)}\sigma_w^2.$$

(i) follows from Lemma B.1. The above inequality implies,

$$\left(\frac{1-\delta}{2(1+\delta)} - \frac{1}{8} - \frac{1}{20} - \frac{34}{4}\delta^2\right)\|\boldsymbol{UU}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\|_F^2 \leq 34\frac{(1+\delta)\log(n)}{m}\sigma_w^2 + \frac{5\log(n)}{m(1+\delta)}\sigma_w^2.$$

If $\delta \leq \frac{1}{10}$, the above inequality reduces to $\|\boldsymbol{UU}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\|_F \leq c\sqrt{\frac{\log(n)}{m}}\sigma_w$, for some constant $c \leq 17$, w.p $\geq 1 - \frac{2}{n^2}$. $\qquad\square$

C   Proof for the High Rank Case

In this section we will present the proof for the inexact case, where $\mathrm{rank}(\boldsymbol{X}^*) \geq r$. Recall that measurements are $\boldsymbol{y} = \mathcal{A}(\boldsymbol{X}^*)$.

Let SVD of $\boldsymbol{X}^*$ be $Q^*\Sigma^*Q^{*\top}$. With slight abuse of notation we use $\boldsymbol{X}^*_{jr+1:(j+1)r}$ to denote the jth rank r block $Q^*_{jr+1:(j+1)r}\Sigma^*_{jr+1:(j+1)r}Q^*_{jr+1:(j+1)r}{}^\top$, where $Q^*_{jr+1:(j+1)r}$ denotes the restriction of Q to columns $jr + 1$ to $(j+1)r$.

C.1   First order optimality

First we will consider the first order condition, $\nabla f(\boldsymbol{U}) = 0$. For any stationary point $\boldsymbol{U}$ this implies

$$\sum_i \left\langle \boldsymbol{A}_i, \boldsymbol{UU}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\right\rangle \boldsymbol{A}_i\boldsymbol{U} = 0. \tag{12}$$

Now using the isometry property of $\boldsymbol{A}_i$ gives us the following result.

Lemma C.1. *[First order condition] For any first order stationary point $\boldsymbol{U}$ of $f(\boldsymbol{U})$, and $\{\boldsymbol{A}_i\}$ satisfying the $(2r, \delta)$-RIP (3), the following holds:*

$$\|\boldsymbol{X} - QQ^\top\boldsymbol{X}^*_{\boldsymbol{r}}\|_F \leq \delta\,\|\boldsymbol{X} - \boldsymbol{X}^*_{\boldsymbol{r}}\|_F + \|(\boldsymbol{X}^* - \boldsymbol{X}^*_{\boldsymbol{r}})QQ^\top\|_F + \delta\|\boldsymbol{X}^* - \boldsymbol{X}^*_{\boldsymbol{r}}\|_*.$$

where Q is an orthonormal matrix that spans the column space of $\boldsymbol{U}$.

This lemma states that any stationary point of $f(\boldsymbol{U})$ is close to a global optimum $\boldsymbol{U}^*$ in the subspace spanned by columns of $\boldsymbol{U}$. Notice that the error along the orthogonal direction $\|\boldsymbol{X}^*Q_\perp Q_\perp^\top\|_F$ can still be large making the distance between $\boldsymbol{X}$ and $\boldsymbol{X}^*$ arbitrarily big.

14

Proof of Lemma C.1. Let $U = QR$, for some orthonormal Q. Consider any matrix of the form $ZQR^{-\top}$. The first order optimality condition then implies,

$$\sum_{i=1}^{m} \langle A_i, X - X_r^* \rangle \langle A_i, UR^{-1}Q^\top Z^\top \rangle = \sum_{i=1}^{m} \langle A_i, X^* - X_r^* \rangle \langle A_i, UR^{-1}Q^\top Z^\top \rangle.$$

Note that $X - X_r^*$ is atmost rank-$2r$. Hence, the above equation together with Restricted Isometry Property(equation (5)) gives us the following inequality:

$$\left| \langle X - X_r^*, QQ^\top Z^\top \rangle \right| - \delta \|X - X_r^*\|_F \|QQ^\top Z^\top\|_F$$

$$\leq \frac{1}{m} \sum_{i=1}^{m} \left\langle A_i, \sum_j X_{jr+1:(j+1)r}^* \right\rangle \langle A_i, QQ^\top Z^\top \rangle$$

$$\leq \sum_j \left\langle X_{jr+1:(j+1)r}^*, QQ^\top Z^\top \right\rangle + \delta \|X_{jr+1:(j+1)r}^*\|_F$$

$$\leq \|(X^* - X_r^*)QQ^\top\|_F + \delta \|X^* - X_r^*\|_*.$$

The last inequality follows from $\sum_j \|X_{jr+1:(j+1)r}^*\|_F \leq \|X^* - X_r^*\|_*$. The above inequalities are true for any Z.

Further note that for any matrix A, $\langle A, QQ^\top Z \rangle = \langle AQQ^\top, Z \rangle$. Furthermore, for any matrix A, $\sup_{\{Z:\|Z\|_F \leq 1\}} \langle A, Z \rangle = \|A\|_F$. Hence the above inequality implies the Lemma. □

C.2 Second order optimality

We will now consider the second order condition to show that the error along $Q_\perp Q_\perp^\top$ is indeed bounded well. Let $\nabla^2 f(U)$ be the hessian of the objective function. Note that this is an $n \cdot r \times n \cdot r$ matrix. Fortunately for our result we need to only evaluate the Hessian along the direction $\text{vec}(U - U^*R)$ for some orthonormal matrix R.

Lemma C.2. *[Hessian computation] Let U be a first order critical point of $f(U)$. Then for any $n \times r$ matrix Z,*

$$\text{vec}(Z)^\top [\nabla^2 f(U)] \, \text{vec}(Z) = \sum_{i=1}^{m} 4 \langle A_i, UZ^\top \rangle^2 + 2 \langle A_i, UU^\top - U^*U^{*\top} \rangle \langle A_i, ZZ^\top \rangle,$$

Further let U be a local minimum of $f(U)$ and $\mathcal{A}$ satisfying $(2r, \delta)$ -RIP (equation (3)). Then,

$$(1 - 3\delta)\|X - X_r^*\|_F^2 \leq 4(1 + \delta) \sum_{j=1}^{r} \|U\Delta_j^\top\|_F^2 + \|X^* - X_r^*\|_F^2 + \delta \|X^* - X_r^*\|_*^2.$$

Proof of Lemma C.2. For any matrix Z, taking directional second derivative of the function $f(U)$ with respect to Z we get:

$$\text{vec}(Z)^\top [\nabla^2 f(U)] \, \text{vec}(Z) = \text{vec}(Z)^\top \lim_{t \to 0} \left[\frac{\nabla f(U + t(Z)) - \nabla f(U)}{t} \right]$$

$$= 2 \sum_{i=1}^{m} \left[2 \langle A_i, UZ^\top \rangle^2 + \langle A_i, UU^\top - U^*U^{*\top} \rangle \langle A_i, ZZ^\top \rangle \right].$$

Setting $Z = \Delta_j = (U - U^*R)e_j e_j^\top$ we get,

$$\sum_{j=1}^{r} \text{vec}\left((U - U^*R)e_j e_j^\top\right)^\top [\nabla^2 f(U)] \, \text{vec}\left((U - U^*R)e_j e_j^\top\right)$$

$$= \sum_{i=1}^{m} (\sum_{j=1}^{r} 4 \langle A_i, Ue_j e_j^\top (U - U_r^*R)^\top \rangle^2 + 2 \sum_{j=1}^{r} \langle A_i, UU^\top - U^*U^{*\top} \rangle \langle A_i, (U - U_r^*R)e_j e_j^\top (U - U_r^*R)^\top \rangle)$$

$$\overset{(i)}{=} \sum_{i=1}^{m} (\sum_{j=1}^{r} 4 \langle A_i, U\Delta_j^\top \rangle^2 + 2 \langle A_i, UU^\top - U^*U^{*\top} \rangle \langle A_i, U_r^*R(U_r^*R)^\top - X \rangle).$$

15

(i) is by the first order optimality condition (12).

Hence from second order optimality of U we get,

$$\sum_{i=1}^{m} 4 \sum_{j=1}^{r} \left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{\Delta}_j^{\top} \right\rangle^2 \geq \sum_{i=1}^{m} 2 \left\langle \boldsymbol{A}_i, \boldsymbol{X} - \boldsymbol{X}^* \right\rangle \left\langle \boldsymbol{A}_i, \boldsymbol{X} - \boldsymbol{X}_r^* \right\rangle. \tag{13}$$

$$\frac{1}{m} \sum_{i=1}^{m} \left\langle \boldsymbol{A}_i, \boldsymbol{X} - \boldsymbol{X}^* \right\rangle \left\langle \boldsymbol{A}_i, \boldsymbol{X} - \boldsymbol{X}_r^* \right\rangle = \frac{1}{m} \sum_{i=1}^{m} \left\langle \boldsymbol{A}_i, \boldsymbol{X} - \boldsymbol{X}_r^* \right\rangle^2 + \left\langle \boldsymbol{A}_i, \boldsymbol{X}_r^* - \boldsymbol{X}^* \right\rangle \left\langle \boldsymbol{A}_i, \boldsymbol{X} - \boldsymbol{X}_r^* \right\rangle$$

$$\overset{(i)}{\geq} (1 - \delta) \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 - \frac{1}{m} \sum_{i=1}^{m} \left(\sum_{j=1} \left\langle \boldsymbol{A}_i, \boldsymbol{X}_{jr+1:(j+1)r}^* \right\rangle \right) \left\langle \boldsymbol{A}_i, \boldsymbol{X} - \boldsymbol{X}_r^* \right\rangle$$

$$\overset{(ii)}{\geq} (1 - \delta) \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 - \sum_{j=1} \left\langle \boldsymbol{X} - \boldsymbol{X}_r^*, \boldsymbol{X}_{jr+1:(j+1)r}^* \right\rangle - \delta \sum_{j=1} \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F \|\boldsymbol{X}_{jr+1:(j+1)r}^*\|_F$$

$$= (1 - \delta) \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 - \left\langle \boldsymbol{X} - \boldsymbol{X}_r^*, \boldsymbol{X}^* - \boldsymbol{X}_r^* \right\rangle - \delta \sum_{j=1} \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F \|\boldsymbol{X}_{jr+1:(j+1)r}^*\|_F$$

$$\geq (1 - \delta) \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 - \frac{1}{2} \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 - \frac{1}{2} \|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_F^2 - \delta \sum_{j=1} \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F \|\boldsymbol{X}_{jr+1:(j+1)r}^*\|_F$$

$$\overset{(iii)}{\geq} (1 - \delta) \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 - \frac{1}{2} \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 - \frac{1}{2} \|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_F^2 - \delta \frac{1}{2} \left(\|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 + \|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_*^2 \right)$$

$$= \frac{1 - 3\delta}{2} \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 - \frac{1}{2} \|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_F^2 - \frac{\delta}{2} \|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_*^2. \tag{14}$$

(i) is from using RIP and splitting $\boldsymbol{X}^* - \boldsymbol{X}_r^*$ into rank-r components $\boldsymbol{X}^* - \boldsymbol{X}_r^* = \sum_{j=1}^{n/r-1} \boldsymbol{X}_{jr+1:(j+1)r}^*$. (ii) follows from using RIP (5). (iii) follows from $\sum_j \|\boldsymbol{X}_{jr+1:(j+1)r}^*\|_F \leq \|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_*$.

The Lemma now follows by combining equations (13), (14) and using RIP (3). $\qquad\square$

Hence from the above optimality conditions we get the proof of Theorem 3.4.

Proof of Theorem 3.4. Assuming $\boldsymbol{U}\boldsymbol{U}^{\top} \neq \boldsymbol{U}_r^* \boldsymbol{U}_r^{*\top}$, from Lemma 4.4 we know,

$$\sum_{j=1}^{r} \|\boldsymbol{U}\boldsymbol{\Delta}_j^{\top}\|_F^2 \leq \frac{1}{8} \|\boldsymbol{U}\boldsymbol{U}^{\top} - \boldsymbol{U}_r^* \boldsymbol{U}_r^{*\top}\|_F^2 + \frac{34}{8} \|(\boldsymbol{U}\boldsymbol{U}^{\top} - \boldsymbol{U}_r^* \boldsymbol{U}_r^{*\top})QQ^{\top}\|_F^2, \tag{15}$$

for some orthonormal R. Hence combining equations (15), with Lemma C.2 we get,

$$\frac{1 - 3\delta}{2} \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 \leq \frac{1}{2} \|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_F^2 + \frac{\delta}{2} \|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_*^2$$
$$+ 2(1 + \delta) \left(\frac{1}{8} \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 + \frac{34}{8} \|(\boldsymbol{X} - \boldsymbol{X}_r^*)QQ^{\top}\|_F^2 \right).$$

This implies,

$$\frac{1 - 7\delta}{4} \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 \leq \frac{1}{2} \|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_F^2 + \frac{\delta}{2} \|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_*^2 + (1 + \delta) \frac{17}{2} \|(\boldsymbol{X} - \boldsymbol{X}_r^*)QQ^{\top}\|_F^2. \tag{16}$$

Finally from Lemma C.1 we know,

$$\|\boldsymbol{X} - \boldsymbol{X}_r^* QQ^{\top}\|_F^2 \leq \left(\delta \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F + \|(\boldsymbol{X}^* - \boldsymbol{X}_r^*)QQ^{\top}\|_F + \delta \|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_* \right)^2$$
$$\leq \frac{11}{10} \|(\boldsymbol{X}^* - \boldsymbol{X}_r^*)QQ^{\top}\|_F^2 + 22\delta^2 \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 + 22\delta^2 \|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_*^2. \tag{17}$$

The last inequality follows from just using $2ab \leq a^2 + b^2$.

Combining equations (16) and (17) gives,

$$\left(\frac{1 - 7\delta}{4} - \frac{17 * 22(1 + \delta)\delta^2}{2} \right) \|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 \leq \frac{1}{2}\|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_F^2 + \left(\frac{\delta}{2} + \frac{17 * 22\delta^2}{2} \right) \|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_*^2$$
$$+ (1 + \delta)\frac{17 * 11}{20}\|(\boldsymbol{X}^* - \boldsymbol{X}_r^*)QQ^\top\|_F^2$$

Substituting $\delta = \frac{1}{100}$ gives,

$$\|\boldsymbol{X} - \boldsymbol{X}_r^*\|_F^2 \leq \frac{5}{2}\|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_F^2 + 12\delta\|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_*^2 + 10\|(\boldsymbol{X}^* - \boldsymbol{X}_r^*)QQ^\top\|_F^2.$$
$$\leq 13\|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_F^2 + 12\delta\|\boldsymbol{X}^* - \boldsymbol{X}_r^*\|_*^2.$$

$\square$

D Proofs for Section 3

In this section we present the proofs for the strict saddle theorem (Theorem 3.2) and the convergence guarantees (Theorem 3.3). The proofs use the Lemmas developed in Section 4 and the supporting Lemmas from Section E.

Proof of Theorem 3.2. From Lemma 4.3 we know that

$$\sum_{j=1}^r \text{vec}\left(\Delta_j\right)^\top \left[\frac{1}{m}\nabla^2 f(\boldsymbol{U}) \right] \text{vec}\left(\Delta_j\right)$$
$$= \frac{1}{m}\sum_{i=1}^m (\sum_{j=1}^r 4\left\langle \boldsymbol{A}_i, \boldsymbol{U}\Delta_j^\top \right\rangle^2 - 2\left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top} \right\rangle^2$$
$$\leq 4(1 + \delta)\sum_{j=1}^r \|\boldsymbol{U}\Delta_j^\top\|_F^2 - 2(1 - \delta)\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\|_F^2, \tag{18}$$

where the last inequality follows from the RIP (3). Now applying Lemma 4.4 in equation (18) we get,

$$\sum_{j=1}^r \text{vec}\left(\Delta_j\right)^\top \left[\frac{1}{m}\nabla^2 f(\boldsymbol{U}) \right] \text{vec}\left(\Delta_j\right)$$
$$\leq (1 + \delta)\left(\frac{1}{2}\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\|_F^2 + 17\|(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top})QQ^\top\|_F^2 \right) - 2(1 - \delta)\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\|_F^2$$
$$= 17(1 + \delta)\|(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top})QQ^\top\|_F^2 - \frac{(3 - 5\delta)}{2}\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\|_F^2$$
$$\overset{(i)}{\leq} \left[17(1 + \delta)\delta^2 - \frac{(3 - 5\delta)}{2} \right] \left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\right\|_F^2$$
$$\overset{(ii)}{\leq} -1 \cdot \left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\right\|_F^2. \tag{19}$$

(i) follows from Lemma 4.2. (ii) follows from $\delta \leq 1/10$. Now notice that from lemma E.1

$$\|\boldsymbol{X} - \boldsymbol{X}^*\|_F^2 \geq 2(\sqrt{2} - 1)\|(\boldsymbol{U} - \boldsymbol{U}^*R)(\boldsymbol{U}^*R)^\top\|_F^2$$
$$\geq 2(\sqrt{2} - 1)\sigma_r(\boldsymbol{X}^*)\|\boldsymbol{U} - \boldsymbol{U}^*R\|_F^2. \tag{20}$$

17

Finally notice that $\Delta_j = \Delta e_j e_j^\top$ have orthogonal columns. Hence,

$$\lambda_{\min}\left[\frac{1}{m}\nabla^2(f(\boldsymbol{U},\boldsymbol{V}))\right] \leq \frac{1}{\|\boldsymbol{U}-\boldsymbol{U}^*R\|_F^2} \sum_{j=1}^{r} \text{vec}\,(\Delta_j)^\top \left[\frac{1}{m}\nabla^2 f(\boldsymbol{U})\right]\text{vec}\,(\Delta_j)$$

$$\stackrel{(i)}{\leq} \frac{-1}{\|\boldsymbol{U}-\boldsymbol{U}^*R\|_F^2}\left\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\right\|_F^2$$

$$\stackrel{(ii)}{\leq} \frac{-2(\sqrt{2}-1)\sigma_r(\boldsymbol{X}^*)\|\boldsymbol{U}-\boldsymbol{U}^*R\|_F^2}{\|\boldsymbol{U}-\boldsymbol{U}^*R\|_F^2}$$

$$\leq \frac{-4}{5}\sigma_r(\boldsymbol{X}^*).$$

(i) follows from equation (19). (ii) follows from equation (20). □

Proof of Theorem 3.3. To prove this theorem we use Theorem 6 of Ge et al. [10]. We need to show that $f(\boldsymbol{U})$ satisfies, 1) strict saddle property, 2) local strong convexity, 3) f is bounded, smooth and has Lipschitz Hessian.

The boundedness assumption easily follows from assuming we are optimizing over a bounded domain b such that, $\|\boldsymbol{U}^*\|_F \leq b$. Note that we can have any reasonable upper bound on the optimum and we can easily estimate this from $\sum_i y_i^2$ which is $\geq (1-\delta)\|\boldsymbol{X}^*\|_F^2$ for the noiseless case.

Finally all the calculations below are for scaled version of $f(x)$ by $\frac{1}{m}$. Note that this does not change the number of iterations as both smoothness and strong convexity parameters are scaled by the same constant.

Smoothness constant β: Recall that smoothness of f is bounded by maximum eigenvalue of Hessian over the domain. Hence, $\beta = \max_{\boldsymbol{Z}:\|\boldsymbol{Z}\|_F \leq 1} \boldsymbol{Z}^\top \nabla^2 f(\boldsymbol{U})\boldsymbol{Z}$. We have computed this projection of Hessian in Lemma B.2. Hence,

$$\beta = 2\max_{\boldsymbol{Z}:\|\boldsymbol{Z}\|_F^2 \leq 1} \sum_{i=1}^{m}\left[2\left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{Z}^\top\right\rangle^2 + \left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\right\rangle\left\langle \boldsymbol{A}_i, \boldsymbol{Z}\boldsymbol{Z}^\top\right\rangle\right]$$

$$\stackrel{(i)}{\leq} \max_{\boldsymbol{Z}:\|\boldsymbol{Z}\|_F^2 \leq 1} 2\left(2(1+\delta)\|\boldsymbol{U}\|_F^2\|\boldsymbol{Z}\|_F^2 + (1+\delta)\|\boldsymbol{X}-\boldsymbol{X}^*\|_F\|\boldsymbol{Z}\boldsymbol{Z}^\top\|_F\right)$$

$$\leq 4(1+\delta)b^2 + (1+\delta)2b \leq 5b^2 + 3b.$$

(i) follows from the RIP.

ρ- Lipschitz Hessian: Now we will compute the Lipschitz constant of Hessian of $f(\boldsymbol{U})$. We will first bound the spectral norm of difference of Hessian at two points $\boldsymbol{U},\boldsymbol{V}$ in terms of $\|\boldsymbol{U}-\boldsymbol{V}\|_F$ along orthogonal direction $\boldsymbol{Z}_i$ and combine them to get bound on ρ.. Given two $n \times r$ matrices $\boldsymbol{U},\boldsymbol{V}$,

$$\left\langle \nabla^2 f(\boldsymbol{U}) - \nabla^2 f(\boldsymbol{V}), \boldsymbol{Z}\boldsymbol{Z}^\top\right\rangle$$

$$\leq 2\max_{\boldsymbol{Z}:\|\boldsymbol{Z}\|_F^2 \leq 1} \sum_{i=1}^{m}\left[2\left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{Z}^\top\right\rangle^2 + \left\langle \boldsymbol{A}_i, \boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\right\rangle\left\langle \boldsymbol{A}_i, \boldsymbol{Z}\boldsymbol{Z}^\top\right\rangle\right]$$

$$- \sum_{i=1}^{m}\left[2\left\langle \boldsymbol{A}_i, \boldsymbol{V}\boldsymbol{Z}^\top\right\rangle^2 + \left\langle \boldsymbol{A}_i, \boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{U}^*\boldsymbol{U}^{*\top}\right\rangle\left\langle \boldsymbol{A}_i, \boldsymbol{Z}\boldsymbol{Z}^\top\right\rangle\right]$$

$$\leq 4(1+\delta)(\|\boldsymbol{U}\boldsymbol{Z}^\top\|_F^2 - \|\boldsymbol{V}\boldsymbol{Z}^\top\|_F^2) + 2(1+\delta)\|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{V}\boldsymbol{V}^\top\|_F\|\boldsymbol{Z}\boldsymbol{Z}^\top\|_F$$

$$\leq 4(1+\delta)\|\boldsymbol{Z}\|_F^2(\|\boldsymbol{U}-\boldsymbol{V}\|_F^2 + 2\|\boldsymbol{U}\|_F\|\boldsymbol{U}-\boldsymbol{V}\|_F) + 2(1+\delta)\|\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{V}\boldsymbol{V}^\top\|_F$$

$$\leq \|\boldsymbol{Z}\|_F^2\|\boldsymbol{U}-\boldsymbol{V}\|_F\left(8(1+\delta)b + 4(1+\delta)b\right)$$

$$= \|\boldsymbol{Z}\|_F^2\|\boldsymbol{U}-\boldsymbol{V}\|_F\left(12(1+\delta)b\right). \tag{21}$$

Hence, using the variational characterization of the Frobenius norm, the Hessian Lipschitz constant is bounded by $\max\{\boldsymbol{Z}_i\}\sum_i \left\langle \nabla^2 f(\boldsymbol{U}) - \nabla^2 f(\boldsymbol{V}), \boldsymbol{Z}_i\boldsymbol{Z}_i^\top\right\rangle$, where $\boldsymbol{Z}_i$ are orthogonal with $\sum_i \|\boldsymbol{Z}_i\|_F^2 \leq 1$. Hence from equation (21) we get $\rho = O(b)$.

Strict saddle property: So far we have shown regularity properties of $f(\boldsymbol{U})$. Now we will discuss the strict saddle property. Theorem 3.2 shows that $\lambda_{\min}\left[\nabla^2(f(\boldsymbol{U}))\right] \leq \frac{-2}{5}\sigma_r(\boldsymbol{X^*})$. To use results of [10] we need to show this property over an ϵ neighborhood of any saddle point $\boldsymbol{U}$. For this first recall by smoothness, $\|\nabla f(\boldsymbol{U}) - \nabla f(\boldsymbol{V})\|_F \leq \beta\|\boldsymbol{U} - \boldsymbol{V}\|_F$. Therefore $\nabla f(\boldsymbol{V}) \leq \epsilon$, when $\|\boldsymbol{U} - \boldsymbol{V}\|_F \leq \frac{\epsilon}{\beta}$. Further we know the Hessian spectral norm is ρ Lipschitz from equation (21). Hence, for any direction $\boldsymbol{Z}$,

$$\boldsymbol{Z}^\top \left(\nabla^2(f(\boldsymbol{V})) - \nabla^2(f(\boldsymbol{U}))\right) \boldsymbol{Z}^\top \leq \rho\|\boldsymbol{U} - \boldsymbol{V}\|_F \leq \rho\frac{\epsilon}{\beta}.$$

In particular choosing $\boldsymbol{Z}$ to be the projection direction, $\boldsymbol{U} - \boldsymbol{U^*}$ implies from Theorem 3.2,

$$\boldsymbol{Z}^\top \left(\nabla^2(f(\boldsymbol{V}))\right) \boldsymbol{Z}^\top \leq \frac{-2}{5}\sigma_r(\boldsymbol{X^*}) + \rho\frac{\epsilon}{\beta}.$$

Hence for all $\boldsymbol{V}$ in the bowl of radius ϵ around $\boldsymbol{U}$, where $\epsilon \leq \frac{\beta}{5\rho}\sigma_r(\boldsymbol{X^*})$,

$$\lambda_{\min}\left[\nabla^2(f(\boldsymbol{V}))\right] \leq \frac{-1}{5}\sigma_r(\boldsymbol{X^*}). \tag{22}$$

Local strong convexity: Finally we need to show that the function is α strongly convex in a neighborhood θ around the optimum $\boldsymbol{U^*}R$, for any orthonormal R. This easily follows from existing local convergence results for this problem. For example, Lemma 6.1 of Bhojanapalli et al. [2] states that, for $\|\boldsymbol{U} - \boldsymbol{U^*}R\|_F \leq \frac{\sigma_r(\boldsymbol{X^*})}{200\sigma_1(\boldsymbol{X^*})}\sigma_r(\boldsymbol{U^*}R)$,

$$\langle \nabla f(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{U^*}R \rangle \geq \frac{2}{3}\eta\|\nabla f(\boldsymbol{U})\|_F^2 + \frac{27}{200}\sigma_r(\boldsymbol{U^*}R)^2\|\boldsymbol{U} - \boldsymbol{U^*}R\|_F^2. \tag{23}$$

for $\delta = \frac{1}{10}$ and some step size $\eta \propto \frac{1}{\|\boldsymbol{X^*}\|_2}$. Hence $f(\boldsymbol{U})$ is locally strong convex with $\alpha = \frac{27}{200}\sigma_r(\boldsymbol{U^*}R)^2$ in the neighborhood of radius $\theta = \frac{\sigma_r(\boldsymbol{X^*})}{200\sigma_1(\boldsymbol{X^*})}\sigma_r(\boldsymbol{U^*}R)$ around the optimum.

Substituting these parameters in the Theorem 6 of Ge et al. [10] gives the result. $\qquad\square$

E   Supporting Lemmas

In this section we present the supporting results used in the proofs above.

Lemma (4.4)**.** *Let $\boldsymbol{U}$ and $\boldsymbol{U^*}$ be two $n \times r$ matrices, and Q is an orthonormal matrix that spans the column space of $\boldsymbol{U}$. Then there exists an $r \times r$ orthonormal matrix R such that for any first order stationary point $\boldsymbol{U}$ of $f(\boldsymbol{U})$, the following holds:*

$$\sum_{j=1}^r \|\boldsymbol{U}e_j e_j^\top (\boldsymbol{U} - \boldsymbol{U^*}R)^\top\|_F^2 \leq \frac{1}{8}\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U^*}\boldsymbol{U^*}^\top\|_F^2 + \frac{34}{8}\|(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{U^*}\boldsymbol{U^*}^\top)QQ^\top\|_F^2.$$

Proof of Lemma 4.4. To prove this we will expand terms on the both sides in terms of $\boldsymbol{U}$ and $\Delta = \boldsymbol{U} - \boldsymbol{U^*}R$ and then compare. First notice the following properties of R that minimizes $\|\boldsymbol{U^*}R - \boldsymbol{U}\|_F$. Let $LSP^\top$ be the SVD of $\boldsymbol{U^*}^\top\boldsymbol{U}$. Then, $R = LP^\top$. Hence, $R^\top\boldsymbol{U^*}^\top\boldsymbol{U} = PSP^\top = \boldsymbol{U}^\top\boldsymbol{U^*}R$ is a PSD matrix. This implies, $\boldsymbol{U}^\top\Delta = \boldsymbol{U}^\top\boldsymbol{U} - \boldsymbol{U}^\top\boldsymbol{U^*}R = \boldsymbol{U}^\top\boldsymbol{U} - R^\top\boldsymbol{U^*}^\top\boldsymbol{U} = \Delta^\top\boldsymbol{U}$.

Let columns of U be orthogonal, else we can multiply U by an orthonormal matrix and UR will satisfy this. Since UR is also local minimum, and $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{U}RR^\top\boldsymbol{U}^\top$, results for UR will also hold for $\boldsymbol{U}$. Let Q be the orthonormal matrix that spans the column space of $\boldsymbol{U}$ and $Q_\perp Q_\perp^\top = I - QQ^\top$. Similarly let Q_j span $\boldsymbol{U}e_j e_j^\top$. Note that Q_j are orthonormal since columns of U are orthogonal. Hence,

$$\begin{aligned}
\|(\boldsymbol{U} - \boldsymbol{U^*}R)e_j e_j^\top\boldsymbol{U}^\top\|_F^2 &= \|\boldsymbol{U}e_j e_j^\top\boldsymbol{U}^\top - Q_j Q_j^\top\boldsymbol{U^*}Re_j e_j^\top\boldsymbol{U}^\top - Q_{j\perp}Q_{j\perp}^\top\boldsymbol{U^*}Re_j e_j^\top\boldsymbol{U}^\top\|_F^2 \\
&= \|\boldsymbol{U}e_j e_j^\top\boldsymbol{U}^\top - Q_j Q_j^\top\boldsymbol{U^*}Re_j e_j^\top\boldsymbol{U}^\top\|_F^2 + \|Q_{j\perp}Q_{j\perp}^\top\boldsymbol{U^*}Re_j e_j^\top\boldsymbol{U}^\top\|_F^2 \\
&\leq \frac{\|\boldsymbol{U}e_j e_j^\top\boldsymbol{U}^\top - Q_j Q_j^\top\boldsymbol{U^*}Re_j e_j^\top(Q_j Q_j^\top\boldsymbol{U^*}R)^\top\|_F^2}{2(\sqrt{2} - 1)} + \|Q_{j\perp}Q_{j\perp}^\top\boldsymbol{U^*}Re_j e_j^\top\boldsymbol{U}^\top\|_F^2.
\end{aligned}$$
$$\tag{24}$$

The last inequality follows from Lemma E.1 and the fact that $e_j^\top U^\top U^* R e_j \geq 0, \forall j$ as $U^\top U^* R$ is PSD. Now we will bound the second term in the above equation. The main idea here is to split this term into error between the subspaces of X, X^* and then error between their singular values, since both of them are bounded by distance $\|X - X^* Q Q^\top\|_F$. Let Q^* be an orthonormal matrix that spans the column space of X^*. Also let $X = Q \Sigma_U^2 Q^\top$.

$$\|Q_{j\perp} Q_{j\perp}^\top U^* R e_j e_j^\top U^\top\|_F^2 = \operatorname{trace}(e_j^\top R^\top U^{*\top} Q_{j\perp} Q_{j\perp}^\top U^* R e_j e_j^\top U^\top U e_j)$$

$$= \operatorname{trace}\left(e_j^\top R^\top U^{*\top} Q_{j\perp} Q_{j\perp}^\top U^* R e_j \left[e_j^\top U^\top U e_j - e_j^\top R^\top U^{*\top} Q_j Q_j^\top Q_j Q_j^\top U^* R e_j + e_j^\top R^\top U^{*\top} Q_j Q_j^\top U^* R e_j\right]\right)$$

$$\overset{(i)}{\leq} \frac{1}{8} \underbrace{(e_j^\top R^\top U^{*\top} Q_{j\perp} Q_{j\perp}^\top U^* R e_j)^2}_{\text{term1}} + 2 \underbrace{(e_j^\top U^\top U e_j - e_j^\top R^\top U^{*\top} Q_j Q_j^\top U^* R e_j)^2}_{\text{term2}}$$

$$+ \underbrace{(Q_{j\perp} Q_{j\perp}^\top U^* R e_j e_j^\top (Q_j Q_j^\top U^* R)^\top)^2}_{\text{term3}}. \tag{25}$$

where (i) follows from Cauchy-Schwarz inequality.

We will use the following inequality through the rest of the proof. So we state it first for any matrix T.

$$\sum_{j=1}^r (e_j^\top T^\top T e_j)^2 \leq \sum_{j=1}^r \sum_{k=1}^r (e_j^\top T^\top T e_k)^2$$

$$= \sum_{j=1}^r e_j^\top T^\top \left[\sum_{k=1}^r T e_k e_k^\top T^\top\right] T e_j = \sum_{j=1}^r e_j^\top T^\top T T^\top T e_j$$

$$= \|T^\top T\|_F^2 = \|T T^\top\|_F^2. \tag{26}$$

Now we will bound each of the terms in equation .

Term 1: Let, $T = Q_{j\perp} Q_{j\perp}^\top U^* R$. Then applying inequaltiy from equation (26) we get,

$$\sum_{j=1}^r (e_j^\top R^\top U^{*\top} Q_{j\perp} Q_{j\perp}^\top U^* R e_j)^2 = \sum_{j=1}^r (e_j^\top T^\top T e_j)^2$$

$$\leq \|T^\top T\|_F^2 = \|R^\top U^{*\top} Q_\perp Q_\perp^\top U^* R\|_F^2. \tag{27}$$

Further,

$$\|R^\top U^{*\top} Q_\perp Q_\perp^\top U^* R\|_F^2 = \operatorname{trace}(U^{*\top} Q_\perp Q_\perp^\top U^* U^{*\top} Q_\perp Q_\perp^\top U^*)$$

$$= \operatorname{trace}(Q_\perp Q_\perp^\top X^* Q_\perp Q_\perp^\top X^*)$$

$$\leq \|Q_\perp Q_\perp^\top X^*\|_F^2 \leq \|X - X^*\|_F^2. \tag{28}$$

Term 2:

$$(e_j^\top U^\top U e_j - e_j^\top R^\top U^{*\top} Q_j Q_j^\top U^* R e_j)^2$$

$$= (e_j^\top U^\top U e_j)^2 + (e_j^\top R^\top U^{*\top} Q_j Q_j^\top U^* R e_j)^2 - 2 e_j^\top U^\top U e_j e_j^\top R^\top U^{*\top} Q_j Q_j^\top U^* R e_j$$

$$= \|U e_j e_j^\top U^\top\|_F^2 + \|Q_j Q_j^\top U^* R e_j e_j^\top R^\top U^{*\top} Q_j Q_j^\top\|_F^2 - 2 \operatorname{trace}(e_j^\top U^\top U e_j e_j^\top R^\top U^{*\top} Q_j Q_j^\top U^* R e_j)$$

$$\overset{(i)}{=} \|U e_j e_j^\top U^\top\|_F^2 + \|Q_j Q_j^\top U^* R e_j e_j^\top R^\top U^{*\top} Q_j Q_j^\top\|_F^2 - 2 \operatorname{trace}(e_j^\top R^\top U^{*\top} U e_j e_j^\top U^\top U^* R e_j)$$

$$= \|U e_j e_j^\top U^\top - Q_j Q_j^\top U^* R e_j e_j^\top R^\top U^{*\top} Q_j Q_j^\top\|_F^2. \tag{29}$$

(i) follows from $e_j^\top U^\top U e_j = \|U_j\|_F^2$ and $\|U_j\|_F^2 Q_j Q_j^\top = U e_j e_j^\top U^\top$. Now from orthogonality of Q_j we have,

$$\sum_{j=1}^r \|U e_j e_j^\top U^\top - Q_j Q_j^\top U^* R e_j e_j^\top R^\top U^{*\top} Q_j Q_j^\top\|_F^2 \leq \|U U^\top - Q Q^\top U^* U^{*\top} Q Q^\top\|_F^2. \tag{30}$$

Term 3: Finally we bound the last term in equation (25) similar to the first term, which gives,

$$\sum_{j=1}^{r} (Q_{j\perp}Q_{j\perp}^{\top} U^* R e_j e_j^{\top} (Q_j Q_j^{\top} U^* R)^{\top})^2 \leq \|UU^{\top} - U^* U^{*\top} QQ^{\top}\|_F^2.$$

Substituting the above equations (27), (28), (29) and (30) in (24) and (25) gives the result. □

The following lemma relates the error $\|(U - Y)U^{\top}\|_F$ with $\|UU^{\top} - YY^{\top}\|_F$ under some conditions on U and Y. This is a generalization of Lemma 5.4 in [26] and the proof follows similarly.

Lemma E.1. *Let U and Y be two $n \times r$ matrices. Further let $U^{\top}Y = Y^{\top}U$ be a PSD matrix. Then,*

$$\|(U - Y)U^{\top}\|_F^2 \leq \frac{1}{2(\sqrt{2} - 1)}\|UU^{\top} - YY^{\top}\|_F^2.$$

Proof. To prove this we will expand terms on the both sides in terms of U and $\Delta = U - Y$ and then compare.

$$
\begin{aligned}
\|(UU^{\top} - YY^{\top}\|_F^2 &= \|(U\Delta^{\top} + \Delta U^{\top} - \Delta\Delta^{\top})\|_F^2 \\
&= \mathrm{trace}\left(\Delta U^{\top} U\Delta^{\top} + U\Delta^{\top}\Delta U^{\top} + \Delta\Delta^{\top}\Delta\Delta^{\top} + 2\Delta U^{\top}\Delta U^{\top} - 2\Delta\Delta^{\top}\Delta U^{\top} - 2\Delta\Delta^{\top} U\Delta^{\top}\right) \\
&\overset{(i)}{=} \mathrm{trace}\left(2U^{\top} U\Delta^{\top}\Delta + (\Delta^{\top}\Delta)^2 + 2(U^{\top}\Delta)^2 - 4\Delta^{\top}\Delta U^{\top}\Delta\right) \\
&\overset{(ii)}{=} \mathrm{trace}\left(2U^{\top} U\Delta^{\top}\Delta + (\Delta^{\top}\Delta - \sqrt{2}U^{\top}\Delta)^2 - 2(2 - \sqrt{2})\Delta^{\top}\Delta U^{\top}\Delta\right) \\
&\overset{(iii)}{\geq} 2\,\mathrm{trace}\left(\left[U^{\top} U - (2 - \sqrt{2})U^{\top}\Delta\right]\Delta^{\top}\Delta\right) \\
&= 2\,\mathrm{trace}\left(\left[(\sqrt{2} - 1)U^{\top} U + (2 - \sqrt{2})U^{\top} Y\right]\Delta^{\top}\Delta\right) \\
&\overset{(iv)}{\geq} 2\,\mathrm{trace}\left((\sqrt{2} - 1)U^{\top} U\Delta^{\top}\Delta\right).
\end{aligned}
$$

(i) follows from the following properties of trace: $\mathrm{trace}(AB) = \mathrm{trace}(BA)$ and $\mathrm{trace}(A) = \mathrm{trace}(A^{\top})$. (ii) follows from completing the squares. (iii) follows from $\mathrm{trace}(A^2) \geq 0$. (iv) follows from the hypothesis of the lemma ($U^{\top}Y$ is PSD) and $\mathrm{trace}(AB) \geq 0$ for PSD matrices A and B.

Finally notice that $\|(U - Y)U^{\top}\|_F^2 = \mathrm{trace}(U^{\top} U\Delta^{\top}\Delta)$. This completes the proof. □

We recall the standard Gaussian random variable concentration here.

Lemma E.2. *Let $w_i \approx \mathcal{N}(0, \sigma_w)$, then*

$$\sum_{i=1}^{m} w_i x_i \leq 2\sqrt{\log(n)}\sigma_w\|x\|,$$

with probability $\geq 1 - \frac{1}{n^2}$.

Proof. Recall $\mathbb{E}\left[e^{tw_i}\right] = e^{\sigma_w^2 t^2/2}$. Then by Markov's inequality, $P(\sum_{i=1}^{m} w_i x_i \geq c\|x\|) \leq \frac{e^{\sigma_w^2\|x\|^2 t^2/2}}{e^{tc\|x\|}} \leq e^{-c^2/2\sigma_w^2}$, by setting $t = \frac{c}{\sigma_w^2\|x\|}$. Choosing $c = 2\sqrt{\log(n)}\sigma_w$ completes the proof.

□