

On the computational and statistical complexity of over-parameterized matrix sensing

Jiacheng Zhuo[†] Jeongyeol Kwon^b Nhat Ho[◊] Constantine Caramanis^b

Department of Computer Science, University of Texas at Austin[†],

Department of Electrical and Computer Engineering, University of Texas at Austin^b

Department of Statistics and Data Sciences, University of Texas at Austin[◊]

January 27, 2021

Abstract

We consider solving the low rank matrix sensing problem with Factorized Gradient Descent (FGD) method when the true rank is unknown and over-specified, which we refer to as over-parameterized matrix sensing. If the ground truth signal $\mathbf{X}^* \in \mathbb{R}^{d \times d}$ is of rank r , but we try to recover it using $\mathbf{F}\mathbf{F}^\top$ where $\mathbf{F} \in \mathbb{R}^{d \times k}$ and $k > r$, the existing statistical analysis falls short, due to a flat local curvature of the loss function around the global maxima. By decomposing the factorized matrix \mathbf{F} into separate column spaces to capture the effect of extra ranks, we show that $\|\mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^*\|_F^2$ converges to a statistical error of $\tilde{O}(kd\sigma^2/n)$ after $\tilde{O}(\frac{\sigma_r}{\sigma} \sqrt{\frac{n}{d}})$ number of iterations where \mathbf{F}_t is the output of FGD after t iterations, σ^2 is the variance of the observation noise, σ_r is the r -th largest eigenvalue of \mathbf{X}^* , and n is the number of sample. Our results, therefore, offer a comprehensive picture of the statistical and computational complexity of FGD for the over-parameterized matrix sensing problem.

1 Introduction

We consider the low rank matrix sensing problem: we are given n i.i.d. observations $\{\mathbf{A}_i, y_i\}_{i=1}^n$ from the data generating model $y_i = \langle \mathbf{A}_i, \mathbf{X}^* \rangle + \epsilon_i$, where $\mathbf{A}_i \in \mathbb{R}^{d \times d}$ is a symmetric random sensing matrix, $\mathbf{X}^* \in \mathbb{R}^{d \times d}$ is the target rank r symmetric matrix we want to recover, and ϵ_i is a zero-mean sub-Gaussian noise with variance proxy σ^2 . The low rank matrix sensing problem has found applications in various scenarios, such as multi-task regression, vector auto-regressive process, image processing, metric embedding, quantum tomography, and so on [Candes and Plan, 2011, Negahban and Wainwright, 2011, Recht et al., 2010, Jain et al., 2013, Gross et al., 2010, Candès et al., 2011, Waters et al., 2011, Kalem et al., 2015]. One common approach to recover a low-rank matrix $\mathbf{X} \in \mathbb{R}^{d \times d}$ is to solve the following optimization problem:

$$\arg \min_{\mathbf{X}: \mathbf{X} \succeq 0, \text{rank}(\mathbf{X}) \leq k} \frac{1}{4n} \sum_{i=1}^n (y_i - \langle \mathbf{A}_i, \mathbf{X} \rangle)^2, \quad (1)$$

where k is a chosen rank based on domain knowledge of the data. This problem can be solved by relaxing the rank constraint to nuclear norm constraint [Recht et al., 2010, Candes and Plan, 2011]. However for computational benefits, it is common to reformulate this as a non-convex problem by introducing $\mathbf{F} \in \mathbb{R}^{d \times k}$ such that $\mathbf{X} = \mathbf{F}\mathbf{F}^\top$ and solving the transformed problem [Bhojanapalli et al., 2016a, Chen and Wainwright, 2015, Jain et al., 2013, Hardt, 2014]

$$\arg \min_{\mathbf{F}: \mathbf{F} \in \mathbb{R}^{d \times k}} \mathcal{L}(\mathbf{F}) := \frac{1}{4n} \sum_{i=1}^n \left(y_i - \langle \mathbf{A}_i, \mathbf{F}\mathbf{F}^\top \rangle \right)^2. \quad (2)$$

Solving this formulation directly with gradient descent method on the matrix \mathbf{F} is usually referred to as the Factorized Gradient Descent (FGD) method, which is given by:

$$\mathbf{F}_{t+1} = \mathbf{F}_t - \eta \mathbf{G}_t^n, \quad \text{where} \quad \mathbf{G}_t^n = \nabla \mathcal{L}(\mathbf{F}_t) = \frac{1}{n} \sum_{i=1}^n \left(\langle \mathbf{A}_i, \mathbf{F}_t \mathbf{F}_t^\top - y_i \rangle \right) \mathbf{A}_i \mathbf{F}_t, \quad (3)$$

where η is the step size and \mathbf{G}_t^n denotes the gradient evaluated at iteration t with n i.i.d. samples.

When the specified rank k matches the ground truth rank r , namely, the true rank r is known, FGD converges linearly to a statistical error [Chen and Wainwright, 2015], and the statistical error is minimax optimal up to log factors [Candes and Plan, 2011]. However, in the real world applications, it is often a big challenge to correctly identify the true rank r , and hence the practitioners tend to over-specify the rank. When the rank is over-specified (i.e. $k > r$), we refer to that setting as the *over-parameterized matrix sensing* problem.

The over-parameterized matrix sensing comes with many challenges, and to the best of our knowledge, none of the existing works offer a complete understanding about the computational and statistical performance of FGD under this setting. First and foremost, we are faced a degenerate Hessian around the global maxima caused by the over-specification of the rank. Hence previous works with known rank settings [Bhojanapalli et al., 2016a, Zheng and Lafferty, 2016, Tu et al., 2016] are no longer applicable since they rely on local strong convexity around the global maxima. The analysis of Chen and Wainwright [2015] is also void, because with over-parameterization the ratio of the first and the k -th eigenvalue of \mathbf{X}^* is infinity. Li et al. [2018] focus on the implicit regularization effect with early stopping, and their analysis is limited to the setting where there is no observation noise ($\epsilon_i = 0$), $k = d$ and they can only guarantee recovery within a lower and upper bounded iteration range (as in their Theorem 1). In summary, despite the current progress on the matrix sensing problem, the following questions remain unclear:

If we solve the over-parameterized matrix sensing problem with FGD, (1) what is the achievable statistical error? (2) and how fast we can recover a target matrix \mathbf{X}^ ?*

Contribution. This paper offers a comprehensive analysis of over-parameterized low-rank matrix sensing with the FGD method. We show that $\|\mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^*\|_F^2$ converges to a final statistical error of $\tilde{\mathcal{O}}(kd\sigma^2/n)$ after $\tilde{\mathcal{O}}(\frac{\sigma_r}{\sigma} \sqrt{\frac{n}{d}})$ number of iterations where σ_r and σ respectively the r -th largest eigenvalue of \mathbf{X}^* and the standard deviation of the observation noise. It is different from the computational and statistical behavior of FGD when the true rank is known, namely, the FGD converges to a radius of convergence $\tilde{\mathcal{O}}(rd\sigma^2/n)$ around the true matrix \mathbf{X}^* after $\mathcal{O}(\log(\frac{\sigma_r}{\sigma_1} \cdot \frac{n}{d}))$ iterates [Chen and Wainwright, 2015] where σ_1 is the largest eigenvalue of \mathbf{X}^* . Since we assume no *a priori* knowledge of true rank r , the statistical error $\tilde{\mathcal{O}}(kd\sigma^2/n)$ is also minimax optimal up to logarithmic factors [Candes and Plan, 2011]. Furthermore, the number of iterations $\tilde{\mathcal{O}}(\frac{\sigma_r}{\sigma} \sqrt{\frac{n}{d}})$ is needed in the over-parameterized setting as the local curvature of the loss function (1) around the global maxima is not quadratic and therefore the FGD only converges sub-linearly to the global maxima; see the simulations in Figure 1 for an illustration. Finally, when $\sigma = 0$, i.e., in the noiseless case, we can guarantee the exact recovery similar to when we correctly specify the rank [Chen and Wainwright, 2015].

1.1 Related Work

Works related to Matrix Sensing. Early works on matrix sensing often perform a semidefinite programming (SDP) relaxation, and replace the nonconvex rank constraint with

a convex constraint based on the trace norm or nuclear norm; see for example [Candes and Plan, 2011, Recht et al., 2010, Negahban and Wainwright, 2011, Chen et al., 2013] and the references therein. Candes and Plan [2011] show that for any estimator $\hat{\mathbf{X}}$ based on $\{\mathbf{A}_i, y_i\}_{i=1}^n$ observations, $\|\hat{\mathbf{X}} - \mathbf{X}^*\|_F^2 \geq \frac{dr}{n} \sigma^2$, where \mathbf{X}^* is the ground truth rank r matrix that we want to recover, and σ is the standard deviation of the (sub)-Gaussian observation noise (see Section 1.4 for details). This convex relaxation approach is nearly optimal in this sense. Although we can theoretically solve this convex problem in polynomial time, the computational cost is often prohibitively high for large scale problems. For example, if we solve this SDP problem with the classical interior point method, the computational cost is roughly $\mathcal{O}(d^6)$ [Boyd et al., 2004, Chen and Wainwright, 2015]. Although recently some tailored algorithms [Zheng and Lafferty, 2015, Tu et al., 2016] are developed to solve this convex problem, their computational complexity is at least $\mathcal{O}(d^3)$ since this SDP involves multiplication of two matrices in $\mathbb{R}^{d \times d}$. This computational overhead motivates the study of FGD method. The low rank matrix sensing problem is tightly connected to the low rank matrix completion problem, since they have the same population update when solved by (factorized) gradient method, and they can often be analyzed by very similar techniques [Negahban and Wainwright, 2012, Koltchinskii et al., 2011, Chi et al., 2019].

Works related to FGD. The idea of factorizing the low rank matrix dates back to Burer and Monteiro [2003, 2005]. Bhojanapalli et al. [2016a] characterize the computational convergence behavior of FGD method for general convex and strongly convex function using the restricted strong convexity argument. However, such analysis cannot be converted into statistical analysis. Chen and Wainwright [2015] offer a general theoretical framework for understanding FGD method from both computational and statistical perspective. Specifically, they show that with suitable initialization, FGD converges geometrically up to a statistical precision. However, their analysis only works when we know the ground truth rank ($k = r$).

In this work we focus on local convergence as this is the crux in statistical analysis (see [Chen and Wainwright, 2015]). Initialization condition can be achieved via spectral methods (see [Bhojanapalli et al., 2016a, Tu et al., 2016, Zheng and Lafferty, 2016]). Moreover, the works by Bhojanapalli et al. [2016b], Ge et al. [2016], and Zhang et al. [2019] show that reformulation (2) does not have any spurious local minima from optimization’s perspective, indicating that it is possible to extend our analysis to random initialization.

Recently, Li et al. [2018] look into the implicit regularization effect in the learning of over-parameterized matrix factorization with FGD. They show that if there is no observation noise ($\epsilon_i = 0$) and $k = d$, FGD tends to first recover the majority part of the true signal (that is of rank r) due to the implicit regularization effect of the FGD method. However their analysis can only address the noiseless case, and can not be extended to the more realistic setting when the observation is noisy, i.e., $\epsilon_i \neq 0$. Moreover, they only guarantee recovery within an iteration lower bound and upper bound (e.g., as in the Theorem 1 in Li et al. [2018], the number of iterations to reach the target accuracy has an upper bound and lower bound), which is not in line with the common notion of convergence and statistical rate. We focus on the statistical rate, which means we want to understand the algorithm behavior if run the algorithm for infinitely long. (Further discussion can be found in Section 4).

Localized analysis for degenerate landscape. When the curvature around the local optimum degenerates, first-order methods such as gradient descent slow down due to vanishing gradients as the estimator gets closer to the local optimum. This phenomenon is reported

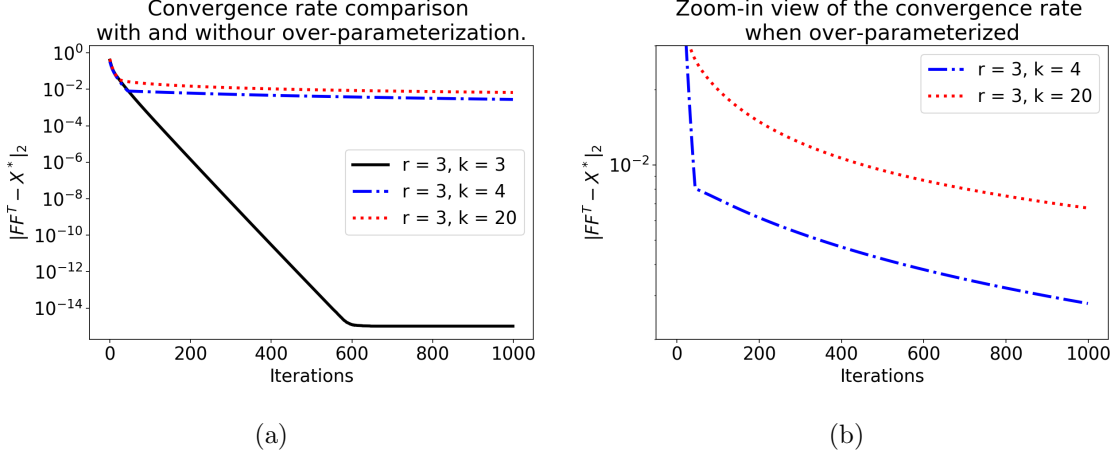


Figure 1. The motivating simulations. (a) When we correctly specify the rank (i.e., $k = r = 3$), the FGD method converges geometrically towards machine precision. But when $k > r$, FGD only converges sub-linearly. (b) A zoom-in view of the convergence rate shows that, FGD might first converge geometrically, and then converge sub-linearly.

in various optimization problems with degenerate landscapes in weakly separated mixture of distributions [Dwivedi et al., 2020a, Kwon et al., 2020]. We can observe the same phenomenon when the rank is over-specified for low-rank matrix factorization problems.

The localization technique is a powerful analysis tool to handle degenerate landscapes with a tight statistical rate. This technique has been used widely in the empirical process theory literature [van der Vaart and Wellner, 2000]. We now see how the localization argument can be applied for a low-rank matrix sensing when we over-specify the rank. In result we obtain a tight statistical rate of FGD which matches the known information-theoretical lower bound for this problem even if we over-specify the rank.

1.2 Motivating Simulations

In the simulations, we consider the dimension $d = 20$, the true rank $r = 3$, and the number of samples $n = 200$. We first generate random orthonormal matrices \mathbf{U} and \mathbf{V} such that the union of their column spaces is \mathbb{R}^d . We set $\mathbf{D}_{\mathbf{S}}^*$ to be a diagonal matrix, with its $(1, 1), (2, 2), (3, 3)$ entries be 1, 0.9, 0.8 respectively, and zero elsewhere. Hence $\mathbf{X}^* = \mathbf{U}\mathbf{D}_{\mathbf{S}}^*\mathbf{U}^\top$. The upper triangle entries of the sensing matrices \mathbf{A}_i are sampled from standard Gaussian distribution, and we fill the lower triangle entries accordingly such that \mathbf{A}_i are symmetric. We further assume that there is no observation noise, so that we have a better understanding of the convergence behavior of the algorithm.

Let $\{\mathbf{F}_t\}_t$ be the sequence generated by the FGD method as in equation (3) with $\eta = 0.1$. The simulation results are shown in Figure 1. When we correctly specify the rank (i.e. $k = r = 3$), the FGD method converge geometrically towards machine precision. However, even if we increase the specified rank by 1, FGD will end up with a much slower convergence rate. A zoom-in view of the convergence rate shows that, FGD might first converge geometrically, and then converge sub-linearly. This phenomenon is not captured by the recent works about FGD [Li et al., 2018, Chen and Wainwright, 2015]. What exactly is the convergence rate? And what about the statistical error? These are the questions that we want to answer in this work.

1.3 Organization

The remainder of the paper is organized as follows. In Section 2, we present the convergence rate of the FGD iterates under the over-parameterized matrix sensing setting. Then, we present the proof sketch of the results in Section 3. The detailed proofs of the main results are deferred to the Appendices while we conclude the paper with a few discussions in Section 4.

1.4 Notations

In the paper, we use bold lower case letters to represent vectors, such as \mathbf{x} , and bold upper case letters to represent matrices, such as \mathbf{X} . When \mathbf{X} is a matrix, we use X_{ij} to represent the element on the i -th row and j -th column of \mathbf{X} , unless otherwise specified. We use $\langle \cdot, \cdot \rangle$ for matrix inner product. For example $\langle \mathbf{A}, \mathbf{X} \rangle = \sum_{ij} A_{ij} X_{ij}$. We denote $\lceil x \rceil$ as the smallest integer greater than or equal to x for any $x \in \mathbb{R}$. We write $\mathbf{A} \succ \mathbf{B}$ (respectively $\mathbf{A} \succeq \mathbf{B}$) if $\mathbf{A} - \mathbf{B}$ is positive definite (respectively positive semidefinite) for square matrices \mathbf{A} and \mathbf{B} . We write $\{\mathbf{A}_i\}_{i=1}^t$ to represent the sequence $\{\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_t\}$. We also use the short hand $\{\mathbf{A}_i\}_i$ to represent $\{\mathbf{A}_1, \mathbf{A}_2, \dots\}$. We use σ_1 and σ_r to denote the first eigenvalue and the r -th eigenvalue of \mathbf{X}^* respectively, which is the ground truth rank r matrix that we want to recover. And we use κ to denote the conditional number: $\kappa := \sigma_1 / \sigma_r$.

We also use the standard asymptotic complexity notation. Specifically, $f(x) = \mathcal{O}(g(x))$ implies $|f(x)| \leq C |g(x)|$ for some constant C and for large enough x , $f(x) = \Omega(g(x))$ implies $|f(x)| \geq C |g(x)|$ for some constant C and for large enough x , and $f(x) = \Theta(g(x))$ implies $C_1 |g(x)| \leq |f(x)| \leq C_2 |g(x)|$ for some constant C_1, C_2 and for large enough x . When log factors are omitted, we use $\tilde{\mathcal{O}}, \tilde{\Omega}, \tilde{\Theta}$ to represent $\mathcal{O}, \Omega, \Theta$ respectively.

Definition 1. (Sub-Gaussian Random Variable). We call a random variable X with mean μ sub-Gaussian with variance proxy $\sigma > 0$ if $\forall \lambda \in \mathbb{R}, \mathbb{E}[\exp(\lambda(X - \mu))] \leq e^{(\sigma^2 \lambda^2 / 2)}$.

Definition 2. (Sub-Gaussian Sensing Matrix). We call a matrix \mathbf{A} a sub-Gaussian sensing matrix if it is sampled as follow: each upper triangle entry ($i < j$) A_{ij} is sampled i.i.d. from a zero-mean sub-Gaussian distribution with variance proxy 1, each lower triangle entry ($i > j$) $A_{ij} = A_{ji}$, and the diagonal entries are sample from i.i.d. from a zero-mean sub-Gaussian distribution with variance proxy 1.

2 Main Result

Before we present our main result, we formally introduce the decomposition notation for \mathbf{X}^* . Let the eigen-decomposition of \mathbf{X}^* (eigenvalues ordered by the absolute values) be given by

$$\mathbf{X}^* = [\mathbf{U} \ \mathbf{V}] \begin{bmatrix} \mathbf{D}_{\mathbf{S}}^* & 0 \\ 0 & \mathbf{D}_{\mathbf{T}}^* \end{bmatrix} [\mathbf{U} \ \mathbf{V}]^\top = \mathbf{U} \mathbf{D}_{\mathbf{S}}^* \mathbf{U}^\top + \mathbf{V} \mathbf{D}_{\mathbf{T}}^* \mathbf{V}^\top,$$

where $\mathbf{U} \in \mathbb{R}^{d \times r}$, $\mathbf{V} \in \mathbb{R}^{d \times (d-r)}$, $\mathbf{D}_{\mathbf{S}}^* \in \mathbb{R}^{r \times r}$, $\mathbf{D}_{\mathbf{T}}^* \in \mathbb{R}^{(d-r) \times (d-r)}$. Without loss of generality we assume that the both \mathbf{U} and \mathbf{V} are orthonormal and $\mathbf{U}^\top \mathbf{V} = 0$ (i.e. \mathbf{U} and \mathbf{V} together span the entire \mathbb{R}^d). Denote σ_1 be the largest value in $\mathbf{D}_{\mathbf{S}}^*$, σ_r be the smallest value in $\mathbf{D}_{\mathbf{S}}^*$, and σ_{r+1} be the largest value in $\mathbf{D}_{\mathbf{T}}^*$. Since we assume \mathbf{X}^* is of approximately rank r , there is a non-trivial gap between σ_r and σ_{r+1} . In this section, we assume that $\sigma_{r+1} \ll \sigma_r$. Since the union of the column space of \mathbf{U} and \mathbf{V} spans the entire \mathbb{R}^d , then for any $\mathbf{F}_t \in \mathbb{R}^{d \times k}$, there exists matrices $\mathbf{S}_t \in \mathbb{R}^{r \times k}$ and $\mathbf{T}_t \in \mathbb{R}^{(d-r) \times k}$ such that

$$\mathbf{F}_t = \mathbf{U} \mathbf{S}_t + \mathbf{V} \mathbf{T}_t.$$

As t goes to infinity, we hope that $\mathbf{S}_t \mathbf{S}_t^\top$ converges to $\mathbf{D}_\mathbf{S}^*$, $\mathbf{T}_t \mathbf{T}_t^\top$ converges to $\mathbf{D}_\mathbf{T}^*$, and $\mathbf{S}_t \mathbf{T}_t^\top$ and $\mathbf{T}_t \mathbf{S}_t^\top$ converges to zero, and hence $\mathbf{F}_t \mathbf{F}_t^\top = \mathbf{U} \mathbf{S}_t \mathbf{S}_t^\top \mathbf{U}^\top + \mathbf{V} \mathbf{T}_t \mathbf{T}_t^\top \mathbf{V}^\top + \mathbf{U} \mathbf{S}_t \mathbf{T}_t^\top \mathbf{V}^\top + \mathbf{V} \mathbf{T}_t \mathbf{S}_t^\top \mathbf{U}^\top$ converges to \mathbf{X}^* .

We introduce the decomposition and study the convergence of $\mathbf{S}_t \mathbf{S}_t^\top$, $\mathbf{T}_t \mathbf{T}_t^\top$, and $\mathbf{S}_t \mathbf{T}_t^\top$ separately. This decomposition technique is essential, since we can then bypass some technical difficulties when we over-specify the rank. For example we do not have to establish the uniqueness (up to rotational ambiguity) of the optimal solution as in the Lemma 1 in [Chen and Wainwright \[2015\]](#). Moreover, this gives more insights about which part is the computational and/or statistical bottleneck. As we will see shortly (both in Theorem 1 and Lemma 3), it is the convergence of $\{\|\mathbf{T}_t \mathbf{T}_t^\top - \mathbf{D}_\mathbf{T}^*\|_2\}_t$ that slows down the entire process of the convergence. Similar decomposition technique is also employed in the work of [Li et al. \[2018\]](#).

Here we focus on the local convergence of FGD method within the following basin of attraction:

Assumption 1. (*Initialization assumption*)

$$\|\mathbf{F}_0 \mathbf{F}_0^\top - \mathbf{X}^*\|_2 \leq \rho \sigma_r, \quad \text{for } \rho \leq 0.07. \quad (4)$$

Note that 0.07 is a universal constant and is chosen for the ease of presentation. Note that one can use spectrum method to achieve this initialization [[Chen and Wainwright, 2015](#), [Bhojanapalli et al., 2016a](#), [Tu et al., 2016](#)]. Connecting the initialization condition to our decomposition strategy, we need to control $\max\{\|\mathbf{D}_\mathbf{T}^* - \mathbf{T}_0 \mathbf{T}_0^\top\|_2, \|\mathbf{D}_\mathbf{S}^* - \mathbf{S}_0 \mathbf{S}_0^\top\|_2, \|\mathbf{S}_0 \mathbf{T}_0^\top\|_2\}$ in our analysis. The following Lemma establish the connection between what we need in the analysis and Assumption 1.

Lemma 1. *If $\|\mathbf{F}_0 \mathbf{F}_0^\top - \mathbf{X}^*\|_2 \leq 0.7 \rho \sigma_r$, then*

$$\max\left\{\|\mathbf{D}_\mathbf{T}^* - \mathbf{T}_0 \mathbf{T}_0^\top\|_2, \|\mathbf{D}_\mathbf{S}^* - \mathbf{S}_0 \mathbf{S}_0^\top\|_2, \|\mathbf{S}_0 \mathbf{T}_0^\top\|_2\right\} \leq \rho \sigma_r.$$

We leave the proof of Lemma 1 to Appendix C.1. Now we are ready to present our main result.

Theorem 1. (*Main result*) *Assume the following settings: (1) $\|\mathbf{D}_\mathbf{T}^*\|_2 < \sqrt{\frac{d \log d}{n}} \sigma$; (2) we have good initialization as in Assumption 1; (3) the sample size $n > C_1 k \kappa^2 d \log^3 d \cdot \max(1, \sigma^2/\sigma_r^2)$ for some universal constant C_1 ; (4) the step size $\eta = \frac{1}{100\sigma_1}$, (5) \mathbf{A}_i s are sub-Gaussian sensing matrices. Let $\{\mathbf{F}_t\}_t$ be the sequence generated by the FGD algorithm as in Equation 3. Then, the following holds:*

- (a) After $t > \left\lceil 2 \log \frac{\sigma_r}{\epsilon_{comp}} \right\rceil$ steps, $\max\{\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2, \|\mathbf{S}_t \mathbf{T}_t^\top\|_2\} < C \epsilon_{comp}$ for some universal constant C , where $\epsilon_{comp} = \sqrt{\frac{k \kappa^2 d \log d}{n}} \sigma_r$.
- (b) After $t \geq \Theta\left(\frac{\sigma_1}{\epsilon_{stat}}\right)$ steps, $\max\{\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2, \|\mathbf{S}_t \mathbf{T}_t^\top\|_2, \|\mathbf{T}_t \mathbf{T}_t^\top - \mathbf{D}_\mathbf{T}^*\|_2\} < C_1 \epsilon_{stat}$, and $\|\mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^*\|_2 \leq C_2 \epsilon_{stat}$ for some constants C_1 and C_2 , where $\epsilon_{stat} := \kappa \sqrt{\frac{d \log d}{n}} \sigma$.

The proof of Theorem 1 is in Appendix B.2. We now have a few remarks with these results.

(1) The sequences $\{\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2\}_t$ and $\{\|\mathbf{S}_t \mathbf{T}_t^\top\|_2\}_t$ first converge linearly and then sub-linearly. Theorem 1 indicates that the sequences $\{\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2\}_t$ and $\{\|\mathbf{S}_t \mathbf{T}_t^\top\|_2\}_t$ first

converge linearly from $0.1\sigma_r$ to ϵ_{comp} , and then converge sub-linearly to $\Omega(\epsilon_{stat})$. Furthermore, the sequence $\{\|\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\|_2\}_t$ always converges sublinearly towards $\Omega(\epsilon_{stat})$. This is consistent with our simulations in Figure 2. As we will see later in Lemma 3, it is the convergence of $\|\mathbf{D}_\mathbf{T}^* - \mathbf{T}_t\mathbf{T}_t^\top\|_2$ that slows down the convergence of $\{\|\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\|_2\}_t$, and incurring the sublinear convergence of $\{\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2\}_t$ and $\{\|\mathbf{S}_t\mathbf{T}_t^\top\|_2\}_t$.

(2) *There is a convergence rate discrepancy between the population and finite-sample versions.* It is often believed that the convergence rate is consistent even if we go from finite n to infinitely large n (i.e., from finite sample scenario to the scenario when we have access to the population gradient). However this is not the case in our setting. As we will show shortly in Lemma 2, if we have access to the population gradient, the convergence rates of the sequences $\{\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2\}_t$ and $\{\|\mathbf{S}_t\mathbf{T}_t^\top\|_2\}_t$ are linear all the way until zero. In our setting, going from population to finite-sample creates an unusual tangling factor, causing the convergence rate discrepancy between the finite-sample and population sequences.

(3) *This achieves nearly minimax-optimal statistical error.* At a glance the statistical error seems too good to be true compared to Yudong’s work, and even better than the minimax rate [Candes and Plan, 2011]. In fact the guarantees we offer are in spectral norm, while the typical rate in the related work is in Frobenius norm. Translating the spectral norm to Frobenius norm will introduce an extra \sqrt{k} factor. That is, the statistical error is $\kappa\sqrt{\frac{kd\log d}{n}}\sigma$ if we evaluate $\|\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\|_F$. This statistical error is similar to the results in Chen and Wainwright [2015] when the rank is known, i.e., $k = r$. Furthermore, we are able to cover both the noisy and noiseless matrix sensing settings. Given that we assume no *a priori* knowledge of true rank r , the statistical error in Theorem 1 is minimax optimal up to log factors [Candes and Plan, 2011].

2.1 Simulation verification of the main result

In this subsection we use the same simulation setup as in Section 1.2. Let $\{\mathbf{F}_t\}_t$ be the sequence generated by the FGD method as in Equation 3, and let \mathbf{S}, \mathbf{T} be defined as in the previous subsection.

The simulation results are shown in Figure 2. In Figure 2a, we plot $\|\mathbf{F}\mathbf{F}^\top - \mathbf{X}^*\|_2$, $\|\mathbf{S}\mathbf{S}^\top - \mathbf{D}_\mathbf{S}^*\|_2$, $\|\mathbf{S}\mathbf{T}^\top\|_2$, and $\|\mathbf{T}\mathbf{T}^\top - \mathbf{D}_\mathbf{T}^*\|_2$ against the algorithm iterations. The simulation results are aligned with our theory. As said in Theorem 1, $\|\mathbf{S}\mathbf{S}^\top - \mathbf{D}_\mathbf{S}^*\|_2$ and $\|\mathbf{S}\mathbf{T}^\top\|_2$ first converge linearly, and then sublinearly. Furthermore, $\|\mathbf{F}\mathbf{F}^\top - \mathbf{X}^*\|_2$ is soon dominated by $\|\mathbf{T}\mathbf{T}^\top - \mathbf{D}_\mathbf{T}^*\|_2$, which converges sub-linearly all the time. Note that these phenomena are when the true rank is 3 and we set $k = 4$. If we correctly specify the rank ($k = r = 3$), the convergences will be linear, as shown in Figure 2b. In Figures 2c and 2d, we re-produce the result as in Figures 2a and 2b with random initialization. This indicates that our assumption of initialization could possibly be waived using recent insights about the global landscape of the matrix sensing problem [Zhang and Zhang, 2020].

3 Proof of the main result

The proof of the main result follows the typical population-sample analysis [Balakrishnan et al., 2017]. We first analyze the convergence behavior of the algorithm when we have access to the population gradient. Then in the finite sample setting, we quantify the difference between the population gradient and the finite sample gradient using concentration arguments,

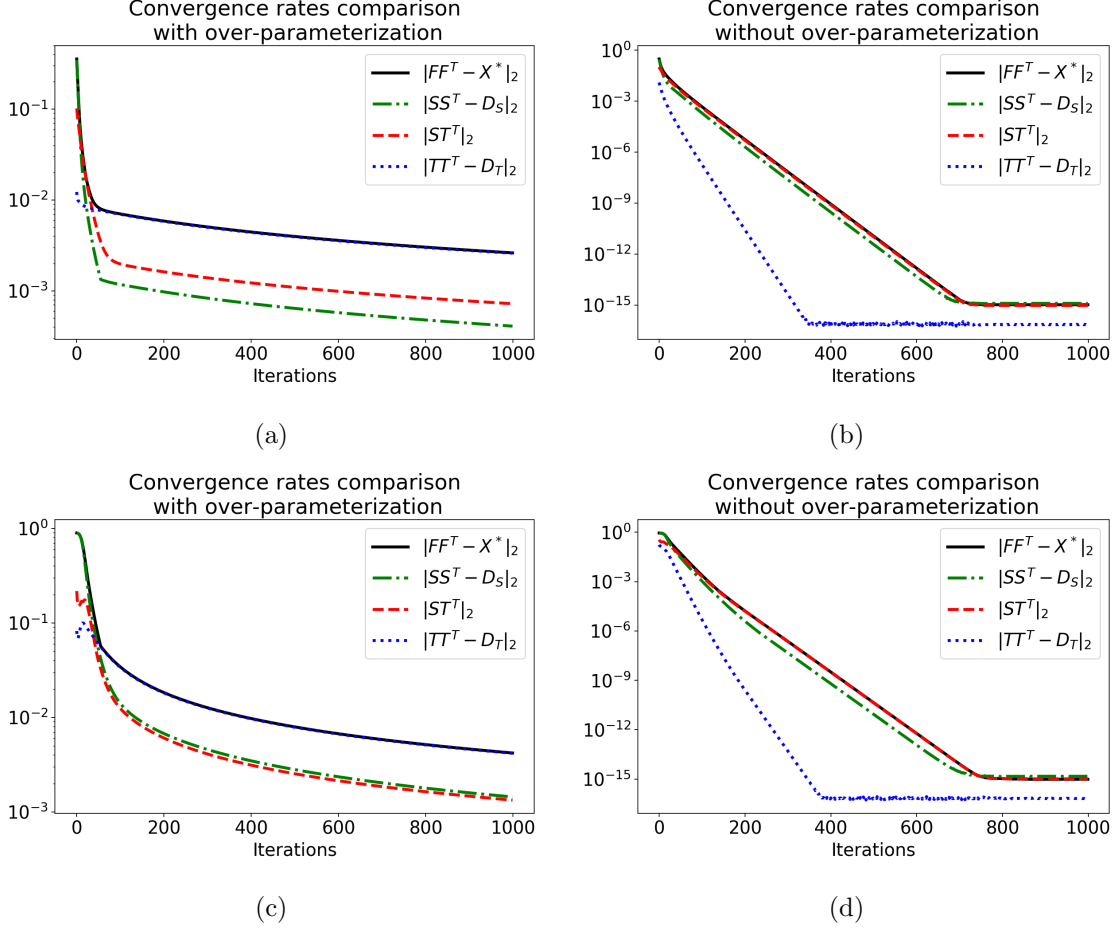


Figure 2. Simulations that verify the main result. (a) Convergence rates of the FGD iterates when we over-specify the rank ($r = 3, k = 4$). (b) Convergence rates of the FGD method when we correctly specify the rank ($r = k = 3$). The Figures in (c) and (d) are executed in the same setting as those in (a) and (b) respectively, except with random initialization around the origin, instead of using Assumption 1.

and use this difference plus the convergence result in population analysis, to characterize the convergence behavior in the finite sample setting.

While it is common to use the Restricted Isometric Property (RIP) as the building block to encapsulate the concentration requirement [Chen and Wainwright, 2015, Chi et al., 2019, Li et al., 2018], we build our results directly based on the concentration of sub-Gaussian sensing matrices for technical convenience. While it is possible to control the Frobenius norm directly, we find it technically easier and more reader friendly to show that the sequence $\|\mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^*\|_2$ converges, and the resulting statistical rate is tight. However, RIP is defined in Frobenius norm since it was first developed for vector and then extended to matrix [Recht et al., 2010, Candes and Plan, 2011]. Translating the Frobenius norm directly to spectral norm will incur a $\Theta(\sqrt{k})$ factor of sub-optimality. That being said, we believe that it is possible to establish similar results for $\|\mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^*\|_F$ directly, and hence we can use the general RIP notion. We leave this for future work.

3.1 Population analysis

The first step of our analysis is to understand the contraction if we have access to the population gradient. One can check that $\mathbb{E}[\langle \mathbf{A}_i, \mathbf{B} \rangle \mathbf{A}_i] = \mathbf{B}$ for any matrix \mathbf{B} with appropriate dimensions. Combined with the fact that $y_i = \langle \mathbf{A}_i, \mathbf{X}^* \rangle + \epsilon$, the population gradient (taking expectation over the observation noise ϵ and the observation matrices \mathbf{A}_i) is

$$\mathbf{G}_t := \mathbb{E}[\mathbf{G}_t^n] = \left(\mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^* \right) \mathbf{F}_t.$$

A closer look at the update in the Factored Gradient Method (Equation 3) with population gradient reveals that at each iteration, the update only changes the coefficient matrices \mathbf{S} and \mathbf{T} . Simple algebra using the last observation yields:

$$\begin{aligned} & \mathbf{F}_t - \eta \mathbf{G}_t \\ &= \mathbf{F}_t - \eta \left(\mathbf{F}_t \mathbf{F}_t^\top \mathbf{F}_t - \mathbf{X}^* \mathbf{F}_t \right) \\ &= \mathbf{U} \mathbf{S}_t + \mathbf{V} \mathbf{T}_t - \eta \left[(\mathbf{U} \mathbf{S}_t + \mathbf{V} \mathbf{T}_t) \left(\mathbf{S}_t^\top \mathbf{S}_t + \mathbf{T}_t^\top \mathbf{T}_t \right) - (\mathbf{U} \mathbf{D}_\mathbf{S}^* \mathbf{S}_t + \mathbf{V} \mathbf{D}_\mathbf{T}^* \mathbf{T}_t) \right] \\ &= \mathbf{U} \mathcal{M}_\mathbf{S}(\mathbf{S}_t) + \mathbf{V} \mathcal{M}_\mathbf{T}(\mathbf{T}_t) \end{aligned}$$

where we define the following operators:

$$\begin{aligned} \mathcal{M}_\mathbf{S}(\mathbf{S}) &= \mathbf{S} - \eta \left(\mathbf{S} \mathbf{S}^\top \mathbf{S} + \mathbf{S} \mathbf{T}^\top \mathbf{T} - \mathbf{D}_\mathbf{S}^* \mathbf{S} \right), \\ \mathcal{M}_\mathbf{T}(\mathbf{T}) &= \mathbf{T} - \eta \left(\mathbf{T} \mathbf{T}^\top \mathbf{T} + \mathbf{T} \mathbf{S}^\top \mathbf{S} - \mathbf{D}_\mathbf{T}^* \mathbf{T} \right). \end{aligned}$$

Lemma 2. (*Contraction per iteration with access to the population gradient.*) Set $\eta = \frac{1}{100\sigma_1}$. We assume good initialization as in Assumption 1. Then we have:

- (a) $\|\mathbf{D}_\mathbf{S}^* - \mathcal{M}_\mathbf{S}(\mathbf{S}) \mathcal{M}_\mathbf{S}(\mathbf{S})^\top\|_2 \leq (1 - \eta\sigma_r) \|\mathbf{D}_\mathbf{S}^* - \mathbf{S} \mathbf{S}^\top\|_2 + 3\eta \|\mathbf{S} \mathbf{T}^\top\|_2^2,$
- (b) $\|\mathcal{M}_\mathbf{S}(\mathbf{S}) \mathcal{M}_\mathbf{T}(\mathbf{T})^\top\|_2 \leq \|\mathbf{S} \mathbf{T}^\top\|_2 (1 - \eta\sigma_r),$
- (c) $\|\mathcal{M}_\mathbf{T}(\mathbf{T}) \mathcal{M}_\mathbf{T}(\mathbf{T})^\top\|_2 \leq \|\mathbf{T} \mathbf{T}^\top\|_2 (1 - \eta \|\mathbf{T} \mathbf{T}^\top\|_2 + 2\eta \|\mathbf{D}_\mathbf{T}^*\|_2),$
- (d) $\|\mathcal{M}_\mathbf{T}(\mathbf{T}) \mathcal{M}_\mathbf{T}(\mathbf{T})^\top - \mathbf{D}_\mathbf{T}^*\|_2 \leq \|\mathbf{T} \mathbf{T}^\top - \mathbf{D}_\mathbf{T}^*\|_2 \|\mathbf{I} - 2\eta \mathbf{T} \mathbf{T}^\top\|_2 + 3\eta \|\mathbf{S} \mathbf{T}^\top\|_2^2.$

The proof of Lemma 2 can be found in Appendix A.1.

According to Lemma 2 above, we have fast convergence in estimating $\mathbf{S} \mathbf{S}^\top$, $\mathbf{S} \mathbf{T}^\top$, but slow convergence in estimating $\mathbf{T} \mathbf{T}^\top$. Intuitively, $\mathbf{T} \mathbf{T}^\top$ is slow because the local curvature of the population version of the loss function (2) is flat, namely, the Hessian matrix around the global maxima $\mathbf{D}_\mathbf{T}^*$ is degenerate. We know that when the curvature of the target matrix is undesirable, we can only guarantee sub-linear convergence rate [Bhojanapalli et al., 2016a].

Note that, we assume that $k > r$ for the above analysis. The case when $k \leq r$ is already covered by various existing works (see Chen and Wainwright [2015], Tu et al. [2016], Bhojanapalli et al. [2016a] and the references therein); therefore, we will not focus on this setting in our analysis.

3.2 Finite sample analysis

On top of our population analysis result, we consider the case when we only have access to the gradient evaluated with finitely many samples. Consider the deviation of the population and sample gradient:

$$\begin{aligned}\mathbf{G}_t^n - \mathbf{G}_t &= \frac{1}{n} \sum_{i=1}^n \left(\langle \mathbf{A}_i, \mathbf{F}_t \mathbf{F}_t^\top \rangle - y_i \right) \mathbf{A}_i \mathbf{F}_t - \left(\mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^* \right) \mathbf{F}_t \\ &= \frac{1}{n} \sum_{i=1}^n \left(\langle \mathbf{A}_i, \mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^* \rangle + \epsilon_i \right) \mathbf{A}_i \mathbf{F}_t - \left(\mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^* \right) \mathbf{F}_t.\end{aligned}$$

We define Δ_t to quantify this deviation:

$$\Delta_t = \frac{1}{n} \sum_i \left(\langle \mathbf{A}_i, \mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^* \rangle + \epsilon_i \right) \mathbf{A}_i - (\mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^*),$$

and hence $\mathbf{G}_t^n - \mathbf{G}_t = \Delta_t \mathbf{F}_t$. If we can control Δ_t , we can have contraction per-iteration, as shown in the lemma below. Note that we make no attempts to optimize the constants.

Lemma 3. (Contraction per iteration.) *Assume that we have the same setting as Theorem 1. Denote $D_t = \max\{\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2, \|\mathbf{T}_t \mathbf{T}_t^\top\|_2, \|\mathbf{S}_t \mathbf{T}_t^\top\|_2\}$, and assume that D_t is still sub-optimal to the statistical error: $D_t > 50\kappa\sqrt{\frac{d \log d}{n}}\sigma$. Suppose*

$$\|\Delta_t\|_2 \leq 10\sqrt{\frac{kd \log d}{n}}D_t + 4\sqrt{\frac{d \log d}{n}}\sigma, \quad (5)$$

Then, we find that

$$\|\mathbf{S}_{t+1} \mathbf{S}_{t+1}^\top - \mathbf{D}_\mathbf{S}^*\|_2 \leq \left(1 - \frac{7}{10}\eta\sigma_r\right) \|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2 + \sqrt{\frac{kd \log d}{n}}D_t + \frac{4}{10}\sqrt{\frac{d \log d}{n}}\sigma, \quad (6)$$

$$\|\mathbf{S}_{t+1} \mathbf{T}_{t+1}^\top\|_2 \leq (1 - \eta\sigma_r) \|\mathbf{S}_t \mathbf{T}_t^\top\|_2 + \sqrt{\frac{kd \log d}{n}}D_t + \frac{4}{10}\sqrt{\frac{d \log d}{n}}\sigma. \quad (7)$$

Moreover, denote $\epsilon_{stat} = \kappa\sqrt{\frac{d \log d}{n}}\sigma$. Then we have

$$(D_{t+1} - 50\epsilon_{stat}) \leq \left[1 - \frac{1}{2}\eta(D_t - 50\epsilon_{stat})\right] (D_t - 50\epsilon_{stat}). \quad (8)$$

The proof of Lemma 3 can be found in Appendix B.1.

Implication of equations (6) and (7): Firstly, when n goes to infinity, the sequence of $\{\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2\}_t$ has constant contraction at each step, and hence achieves a linear convergence after all. This matches our population results in Lemma 2. Secondly, if n is finite, the sequence of $\{\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2\}_t$ still has constant contraction, until roughly the magnitude of $\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2$ reaches $5\sqrt{\frac{kd \log d}{n}}D_t$. This indicates that we will have a linear convergence behavior in the beginning, and then sublinear convergence, as is indicated in Theorem 1.

3.3 Proof sketch for the main theorem

In this subsection we offer a proof sketch for Theorem 1. Detailed proof can be found in Appendix B.2.

Lemma 3 is our key building block towards the main theorem. However there are two missing pieces. (1) Firstly we have to establish equation (5) so that Lemma 3 can be invoked for one iteration. (2) Secondly we have to find a way to correctly invoke Lemma 3 for all iterations and obtain the correct statistical rate.

We resolve the first point by bounding $\|\Delta_t\|_2$ using matrix Bernstein concentration bound [Tropp, 2012] together with the ϵ -net discretization techniques.

Lemma 4. *Let \mathbf{A}_i be symmetric random matrices in $\mathbb{R}^{d \times d}$, with the upper triangle entries ($i \geq j$) being independently sampled from an identical sub-Gaussian distribution whose mean is 0 and variance proxy is 1. Let ϵ_i follows $N(0, \sigma)$. Then*

$$\mathbb{P} \left(\left\| \frac{1}{n} \sum_i^n \mathbf{A}_i \epsilon_i \right\|_2 \geq C \sqrt{\frac{d\sigma^2}{n}} \right) \leq \exp(-C).$$

Lemma 5. *Let \mathbf{A}_i be a symmetric random matrix of dimension d by d . Its upper triangle entries ($i \geq j$) are independently sampled from an identical sub-Gaussian distribution whose mean is 0 and variance proxy is 1. If \mathbf{U} is of rank k and is in a bounded spectral norm ball of radius R (i.e., $\|\mathbf{U}\|_2 \leq R$),*

$$\mathbb{P} \left(\sup_{\mathbf{U}: \|\mathbf{U}\|_2 \leq R} \left\| \frac{1}{n} \sum_i^n (\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U}) \right\|_2 \leq \sqrt{\frac{d \log d}{n}} \sqrt{k} R \right) \geq 1 - \exp(-C_2 \log d).$$

The proof of the above two concentration results can be found in the Appendix D. If we invoke these lemmas for Δ_t , then we can immediately have

$$\begin{aligned} \|\Delta_t\|_2 &= \left\| \frac{1}{n} \sum_i^n \left\langle \mathbf{A}_i, \mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^* \right\rangle \mathbf{A}_i - (\mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^*) + \frac{1}{n} \sum_i^n \epsilon_i \mathbf{A}_i \right\|_2 \\ &\leq 5 \sqrt{\frac{kd \log d}{n}} D_t + \sqrt{\frac{d \log d}{n}} \sigma. \end{aligned}$$

The linear convergence part. Claim (a) in the main theorem is about linear convergence. We mention in the remark that equations (6) and (7) imply constant contractions for $\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2$ and $\|\mathbf{S}_t \mathbf{T}_t^\top\|_2$ respectively. To make the argument more precise, we consider $\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2 > 1000 \sqrt{\frac{k\kappa^2 d \log d}{n}} \sigma_r > 1000 \sqrt{\frac{k\kappa^2 d \log d}{n}} D_t$. Then, we find that $\sqrt{\frac{kd \log d}{n}} D_t \leq 0.1 \eta \sigma_r \|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2$ since $\eta \sigma_r = 0.01/\kappa$. Also, $\frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma < 0.1 \eta \sigma_r \|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2$ by the choice of the constants in the lower bound of n . Hence, when $\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2 > 1000 \sqrt{\frac{k\kappa^2 d \log d}{n}} \sigma_r$, we find that

$$\begin{aligned} \|\mathbf{S}_{t+1} \mathbf{S}_{t+1}^\top - \mathbf{D}_\mathbf{S}^*\|_2 &\leq \left(1 - \frac{7}{10} \eta \sigma_r \right) \|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2 + \sqrt{\frac{kd \log d}{n}} D_t + \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma \\ &\leq \left(1 - \frac{5}{10} \eta \sigma_r \right) \|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2. \end{aligned}$$

The same arguments hold for $\|\mathbf{S}_t \mathbf{T}_t^\top\|_2$. Therefore to obtain the linear convergence result as the part (a) in the main theorem, we can just invoke concentration lemmas for each iteration to obtain constant contraction, and then take union bound over all the iterations.

The sub-linear convergence part. Claim (b) of the main theorem is about sublinear convergence, and is build upon equation (8).

Before we discuss how equation (5) holds in this sub-linear convergence case for all iteration t , we briefly illustrate how equation (8) implies convergence to $\Theta(\epsilon_{stat})$ after $\mathcal{O}(1/\epsilon_{stat})$ iterations. By equation (8), we know that $A_{t+1} \leq (1 - \frac{1}{2}\eta A_t) A_t$ where $A_t = D_t - 50\epsilon_{stat}$. Hence

$$\begin{aligned} A_{t+1} &\leq \left(1 - \frac{1}{2}\eta A_t\right) A_t \stackrel{(1)}{\leq} \left(1 - \frac{2}{t + \frac{4}{\eta A_0}}\right) \frac{4}{\eta t + \frac{4}{A_0}} = \frac{\left(t + \frac{4}{\eta A_0}\right) - 2}{t + \frac{4}{\eta A_0}} \frac{4}{\eta \left(t + \frac{4}{\eta A_0}\right)} \\ &\stackrel{(2)}{\leq} \frac{4}{\eta \left(t + 1 + \frac{4}{\eta A_0}\right)}, \end{aligned}$$

where inequality (1) holds because $(1 - \frac{1}{2}\eta A_t) A_t$ is quadratic with respect to A_t and we plug-in the optimal A_t ; inequality (2) holds because $\frac{\left(t + \frac{4}{\eta A_0}\right) - 2}{\left(t + \frac{4}{\eta A_0}\right)^2} \leq \frac{1}{\left(t + \frac{4}{\eta A_0}\right) + 1}$. Therefore, after $t \geq \Theta\left(\frac{1}{\eta \epsilon_{stat}}\right)$ number of iterations, $A_t = D_t - 50\kappa \sqrt{\frac{d \log d}{n}} \sigma \leq \Theta(\epsilon_{stat})$.

We still need to show that equation (5) holds (with probability at least $1 - d^{-c}$) in this sub-linear convergence case for all iteration t with high probability. To do so, we need to use the localization technique [Kwon et al., 2020, Dwivedi et al., 2020b,a]. Without the localization technique, the statistical error will be proportional to $n^{-1/4}$ which is not tight. With the localization argument, we can improve it to $n^{-1/2}$, making the result nearly minimax optimal. We leave the details of this argument to Appendix B.2.

4 Discussion

In the paper, we provide a comprehensive analysis of the computational and statistical complexity of the Factorized Gradient Descent method under the over-parameterized matrix sensing problem, namely, when the true rank is unknown and over-specified. We show that $\|\mathbf{F}_t \mathbf{F}_t - \mathbf{X}^*\|_F^2$ converges to a minimax optimal radius of convergence $\tilde{\mathcal{O}}(kd/n)$ after $\tilde{\mathcal{O}}(\sqrt{\frac{n}{d}})$ number of iterations where \mathbf{F}_t is the output of FGD after t iterations. We now discuss a few natural questions with this work.

Can the results in Li et al. [2018] imply this work? We would like to explain the difference between our results and those in Li et al. [2018]. If we choose the specified rank k as d , we have the same problem setting, and use the same algorithm. However, the results are different. The key difference here is the sample complexity. As Li et al. [2018] focus on over-parameterization, their analysis requires $\tilde{\mathcal{O}}(dr)$ samples, where r is the rank of the ground truth matrix \mathbf{X}^* , while our analysis requires $\tilde{\mathcal{O}}(dk) = \tilde{\mathcal{O}}(d^2)$ samples when $k = d$. Since they only require $\tilde{\mathcal{O}}(dr)$ samples, they cannot control the error of the over-parameterization part (equivalent to our $\mathbf{T}\mathbf{T}^\top$ part). In fact in their analysis, they only show that in a limited number of step, this error does not blow up. While with $\tilde{\mathcal{O}}(d^2)$ we can show that the over-parameterization part also converges, although with a slower convergence rate. Therefore, their results cannot imply ours.

Extensions to general convex function of low rank problem. One natural question to ask is, if our analysis can be extended to general convex function with respect to a low rank

matrix. In particular, we consider minimizing a convex function $f(\mathbf{X})$, where $\mathbf{X} \in \mathbb{R}^{d \times d}$ is PSD. Let \mathbf{X}^* be the ground truth solution and it is of rank r . We can as well reformulate our problem as minimizing $f(\mathbf{F}\mathbf{F}^\top)$ where $\mathbf{F} \in \mathbb{R}^{d \times k}$. Then as long as the population gradient with respect to \mathbf{X} is $\mathbf{X} - \mathbf{X}^*$, and the sample gradients have good concentration around the population gradient, our analysis techniques should be applicable. For example, matrix completion and principle component analysis fall into this category [Chen and Wainwright, 2015]. However, extension the current results with over-parameterized matrix sensing to general convex function as in Chen and Wainwright [2015], or in Bhojanapalli et al. [2016a] would require more refined analysis. We leave this question for the future work.

5 Acknowledgements

We would like to thank Raaz Dwivedi, Koulik Khamaru, and Martin Wainwright for helpful discussion with this work.

A Proofs for population analysis

In this appendix, we provide all the proofs for population analysis of matrix sensing problem.

A.1 Proof of Lemma 2

We prove the four contraction results separately. To simplify the ensuing presentation, we drop all subscripts t associated with the iteration counter.

Proof of the contraction result (a) in Lemma 2: We would like to prove the following inequality:

$$\left\| \mathbf{D}_S^* - \mathcal{M}_S(\mathbf{S})\mathcal{M}_S(\mathbf{S})^\top \right\|_2 \leq (1 - \eta\sigma_r) \left\| \mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right\|_2 + 3\eta \left\| \mathbf{S}\mathbf{T}^\top \right\|_2^2.$$

Indeed, from the formulation of $\mathcal{M}_S(\mathbf{S})$, we have

$$\begin{aligned} & \mathbf{D}_S^* - \mathcal{M}_S(\mathbf{S})\mathcal{M}_S(\mathbf{S})^\top \\ &= \mathbf{D}_S^* - \left(\mathbf{S} - \eta\mathbf{S}\mathbf{T}^\top\mathbf{T} + \eta \left(\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right) \mathbf{S} \right) \left(\mathbf{S} - \eta\mathbf{S}\mathbf{T}^\top\mathbf{T} + \eta \left(\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right) \mathbf{S} \right)^\top. \end{aligned}$$

We can group the terms in the RHS of the above equation according to whether they contain $\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top$ or not, namely, we find that

$$\begin{aligned} \mathbf{D}_S^* - \mathcal{M}_S(\mathbf{S})\mathcal{M}_S(\mathbf{S})^\top &= \mathbf{I} + \mathbf{II} \\ \text{where, } \mathbf{I} &= \left(\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right) - \eta \left(\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right) \mathbf{S}\mathbf{S}^\top - \eta \mathbf{S}\mathbf{S}^\top \left(\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right) \\ &\quad - \eta^2 \left(\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right) \mathbf{S}\mathbf{S}^\top \left(\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right) \\ &\quad + \eta^2 \left(\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right) \mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top + \eta^2 \mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top \left(\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right), \\ \mathbf{II} &= 2\eta\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top - \eta^2\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top. \end{aligned}$$

We first deal with the I term. A direct application of inequality with operator norm leads to

$$\|I\|_2 \leq \left\| \mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right\|_2 \left\| \mathbf{I} - 2\eta\mathbf{S}\mathbf{S}^\top - \eta^2\mathbf{S}\mathbf{S}^\top (\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top) + 2\eta^2\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top \right\|_2.$$

From the choice of the step size and the initialization condition, the term $\mathbf{I} - 2\eta\mathbf{S}\mathbf{S}^\top - \eta^2\mathbf{S}\mathbf{S}^\top (\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top) + 2\eta^2\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top$ is PSD matrix. Furthermore, for any $\|\mathbf{z}\| = 1$, we have

$$\begin{aligned} & \mathbf{z}^\top \left(\mathbf{I} - 2\eta\mathbf{S}\mathbf{S}^\top - \eta^2\mathbf{S}\mathbf{S}^\top (\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top) + 2\eta^2\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top \right) \mathbf{z} \\ & \leq 1 - 2\eta \|\mathbf{S}\mathbf{z}\|_2^2 + \eta^2 \left\| \mathbf{S}\mathbf{S}^\top \right\|_2 \left\| \mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right\|_2 + 2\eta^2 \left\| \mathbf{T}\mathbf{T}^\top \right\|_2 \|\mathbf{S}\mathbf{z}\|_2^2 \\ & \stackrel{(i)}{\leq} 1 - 2\eta\sigma_r + 3\eta^2\sigma_r\sigma_1 \\ & \stackrel{(ii)}{\leq} 1 - \eta\sigma_r, \end{aligned}$$

where in step (i) we used $0.9\sigma_r\mathbf{I} \preceq \mathbf{S}\mathbf{S}^\top \preceq (\sigma_1 + 0.1\sigma_r)\mathbf{I}$, $\left\| \mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right\|_2 \leq 0.1\sigma_r$ and $\left\| \mathbf{T}\mathbf{T}^\top \right\|_2 \leq 1.1\sigma_r$ by initialization condition and triangular inequality; step (ii) follows from choice of step size $\eta = \frac{1}{100\sigma_1}$, and definition of the conditional number $\kappa = \sigma_1/\sigma_r$. Therefore, we arrive at the following inequality:

$$\left\| \mathbf{I} - 2\eta\mathbf{S}\mathbf{S}^\top - \eta^2\mathbf{S}\mathbf{S}^\top (\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top) + 2\eta^2\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top \right\|_2 \leq 1 - \eta\sigma_r. \quad (9)$$

To deal with the II term, we have to establish the connection between $\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top$ and $\mathbf{S}\mathbf{T}^\top$. Note that, $\left\| \eta^2\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top \right\|_2 \leq \eta^2 \left\| \mathbf{T}\mathbf{T}^\top \right\|_2 \left\| \mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top \right\|_2 \leq \eta \left\| \mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top \right\|_2$ since $\eta \leq 1/\sigma_r$. Hence, we have

$$\|II\|_2 \leq \left\| 2\eta\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top \right\|_2 + \left\| \eta^2\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top \right\|_2 \leq 3\eta \left\| \mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top \right\|_2. \quad (10)$$

Collecting the results from equations (9) and (10), we obtain

$$\left\| \mathbf{D}_S^* - \mathcal{M}_S(\mathbf{S})\mathcal{M}_S(\mathbf{S})^\top \right\|_2 \leq (1 - \eta\sigma_r) \left\| \mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right\|_2 + 3\eta \left\| \mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top \right\|_2.$$

Therefore, we reach the conclusion with claim (a) in Lemma 2.

Proof of the contraction result (b) in Lemma 2: Recall that we want to demonstrate that

$$\left\| \mathcal{M}_S(\mathbf{S})\mathcal{M}_T(\mathbf{T})^\top \right\|_2 \leq \left\| \mathbf{S}\mathbf{T}^\top \right\|_2 (1 - \eta\sigma_r).$$

Firstly, from the formulations of $\mathcal{M}_S(\mathbf{S})$ and $\mathcal{M}_T(\mathbf{T})$, we have the following equations:

$$\begin{aligned} & \mathcal{M}_S(\mathbf{S})\mathcal{M}_T(\mathbf{T})^\top \\ & = \left(\mathbf{S} + \eta \left(\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right) \mathbf{S} - \eta\mathbf{S}\mathbf{T}^\top\mathbf{T} \right) \left(\mathbf{T} + \eta \left(\mathbf{D}_T^* - \mathbf{T}\mathbf{T}^\top \right) \mathbf{T} - \eta\mathbf{T}\mathbf{S}^\top\mathbf{S} \right)^\top \\ & = \frac{1}{2} \left(\mathbf{I} - 2\eta\mathbf{S}\mathbf{S}^\top + 2\eta \left(\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right) - 2\eta^2 \left(\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right) \mathbf{S}\mathbf{S}^\top + 2\eta^2\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top \right) \mathbf{S}\mathbf{T}^\top \\ & \quad + \frac{1}{2} \mathbf{S}\mathbf{T}^\top \left(\mathbf{I} + 2\eta \left(\mathbf{D}_T^* - \mathbf{T}\mathbf{T}^\top \right) - 2\eta\mathbf{T}\mathbf{T}^\top - 2\eta^2\mathbf{T}\mathbf{T}^\top \left(\mathbf{D}_T^* - \mathbf{T}\mathbf{T}^\top \right) \right) \\ & \quad + \eta^2 \left(\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top \right) \mathbf{S}\mathbf{T}^\top \left(\mathbf{D}_T^* - \mathbf{T}\mathbf{T}^\top \right). \end{aligned} \quad (11)$$

Recall that, we have $0.9\sigma_r\mathbf{I} \preceq \mathbf{S}\mathbf{S}^\top \preceq (\sigma_1 + 0.1\sigma_r)\mathbf{I}$, $\|\mathbf{D}_\mathbf{S}^* - \mathbf{S}\mathbf{S}^\top\|_2 \leq 0.1\sigma_r$ and $\|\mathbf{T}\mathbf{T}^\top\|_2 \leq 1.1\sigma_r$ by initialization condition and triangular inequality, and we choose $\eta = \frac{1}{100\sigma_1}$.

For the term in the first line of the RHS of equation (11) we have

$$\begin{aligned} & \left\| \frac{1}{2} \left(\mathbf{I} - 2\eta\mathbf{S}\mathbf{S}^\top + 2\eta \left(\mathbf{D}_\mathbf{S}^* - \mathbf{S}\mathbf{S}^\top \right) - 2\eta^2 \left(\mathbf{D}_\mathbf{S}^* - \mathbf{S}\mathbf{S}^\top \right) \mathbf{S}\mathbf{S}^\top + 2\eta^2 \mathbf{S}\mathbf{T}^\top \mathbf{T}\mathbf{S}^\top \right) \mathbf{S}\mathbf{T}^\top \right\|_2 \\ & \leq \frac{1}{2} \left\| \mathbf{S}\mathbf{T}^\top \right\|_2 \left(\left\| \mathbf{I} - 2\eta\mathbf{S}\mathbf{S}^\top \right\|_2 + 2\eta \left\| \mathbf{D}_\mathbf{S}^* - \mathbf{S}\mathbf{S}^\top \right\|_2 + 2\eta^2 \left\| \mathbf{D}_\mathbf{S}^* - \mathbf{S}\mathbf{S}^\top \right\|_2 \left\| \mathbf{S}\mathbf{S}^\top \right\|_2 + 2\eta^2 \left\| \mathbf{S}\mathbf{T}^\top \mathbf{T}\mathbf{S}^\top \right\|_2 \right) \\ & \leq \frac{1}{2} \left\| \mathbf{S}\mathbf{T}^\top \right\|_2 (1 - 1.8\eta\sigma_r + 0.2\eta\sigma_r + 0.0022\eta\sigma_r + 0.02\eta^2\sigma_r^2) \\ & \leq \frac{1}{2} \left\| \mathbf{S}\mathbf{T}^\top \right\|_2 (1 - 1.5\eta\sigma_r). \end{aligned}$$

For the term in the second line of the RHS of equation (11), direct calculation yields that

$$\begin{aligned} & \left\| \frac{1}{2} \mathbf{S}\mathbf{T}^\top \left(\mathbf{I} + 2\eta \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right) - 2\eta\mathbf{T}\mathbf{T}^\top - 2\eta^2 \mathbf{T}\mathbf{T}^\top \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right) \right) \right\|_2 \\ & \leq \frac{1}{2} \left\| \mathbf{S}\mathbf{T}^\top \right\|_2 \left(\left\| \mathbf{I} - 2\eta\mathbf{T}\mathbf{T}^\top \right\|_2 + 0.2\eta\sigma_r + 2.2\eta^2\sigma_r^2 \right) \\ & \leq \frac{1}{2} \left\| \mathbf{S}\mathbf{T}^\top \right\|_2 (1 + 0.3\eta\rho\sigma_r). \end{aligned}$$

Lastly, for the second order term in the third line of the RHS of equation (11) we have

$$\left\| \eta^2 \left(\mathbf{D}_\mathbf{S}^* - \mathbf{S}\mathbf{S}^\top \right) \mathbf{S}\mathbf{T}^\top \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right) \right\|_2 \leq \eta^2 \rho^2 \sigma_r^2 \left\| \mathbf{S}\mathbf{T}^\top \right\|_2 = \frac{1}{1000} \eta \sigma_r \left\| \mathbf{S}\mathbf{T}^\top \right\|_2.$$

Plugging the above results into equation (11) leads to

$$\left\| \mathcal{M}_\mathbf{S}(\mathbf{S}) \mathcal{M}_\mathbf{T}(\mathbf{T})^\top \right\|_2 \leq \left\| \mathbf{S}\mathbf{T}^\top \right\|_2 (1 - \eta\sigma_r).$$

Hence, we obtain the conclusion of claim (b) in Lemma 2.

Proof of the contraction result (c) in Lemma 2: We would like to establish that

$$\left\| \mathcal{M}_\mathbf{T}(\mathbf{T}) \mathcal{M}_\mathbf{T}(\mathbf{T})^\top \right\|_2 \leq \left\| \mathbf{T}\mathbf{T}^\top \right\|_2 \left(1 - \eta \left\| \mathbf{T}\mathbf{T}^\top \right\|_2 + 2\eta \left\| \mathbf{D}_\mathbf{T}^* \right\|_2 \right).$$

To check the convergence in low SNR, i.e., with small singular values, we assume that $\|\mathbf{D}_\mathbf{T}^*\| \ll \sigma_r$. It suggests that the focus is how fast $\mathbf{T}\mathbf{T}^\top$ converges to 0 when $\|\mathbf{T}\mathbf{T}^\top\| \gg \|\mathbf{D}_\mathbf{T}^*\|$. Indeed, simple algebra indicates that

$$\begin{aligned} & \mathcal{M}_\mathbf{T}(\mathbf{T}) \mathcal{M}_\mathbf{T}(\mathbf{T})^\top \\ & = \left(\mathbf{T} + \eta \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right) \mathbf{T} - \eta \mathbf{T}\mathbf{S}^\top \mathbf{S} \right) \left(\mathbf{T} + \eta \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right) \mathbf{T} - \eta \mathbf{T}\mathbf{S}^\top \mathbf{S} \right)^\top \\ & = \mathbf{T}\mathbf{T}^\top + \eta \mathbf{T}\mathbf{T}^\top \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right) - \eta \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top \\ & \quad + \eta \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right) \mathbf{T}\mathbf{T}^\top + \eta^2 \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right) \mathbf{T}\mathbf{T}^\top \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right) - \eta^2 \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right) \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top \\ & \quad - \eta \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top - \eta^2 \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right)^\top + \eta^2 \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top \\ & = III + IV + V, \end{aligned}$$

where we use the following shorthand notation:

$$\begin{aligned}
III &= \left(\mathbf{T}\mathbf{T}^\top - 2\eta \left(\mathbf{T}\mathbf{T}^\top \right)^2 + \eta^2 \left(\mathbf{T}\mathbf{T}^\top \right)^3 \right), \\
IV &= \eta \left(\mathbf{D}_\mathbf{T}^* \mathbf{T}\mathbf{T}^\top + \mathbf{T}\mathbf{T}^\top \mathbf{D}_\mathbf{T}^* \right) - \left(\eta^2 \mathbf{D}_\mathbf{T}^* \left(\mathbf{T}\mathbf{T}^\top \right)^2 + \eta^2 \left(\mathbf{T}\mathbf{T}^\top \right)^2 \mathbf{D}_\mathbf{T}^* \right) + \eta^2 \mathbf{D}_\mathbf{T}^* \left(\mathbf{T}\mathbf{T}^\top \right) \mathbf{D}_\mathbf{T}^*, \\
V &= -2\eta \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top - \eta^2 \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right)^\top - \eta^2 \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right) \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top \\
&\quad + \eta^2 \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top.
\end{aligned}$$

We first bound the IV term. Inequalities with operator norm show that

$$\begin{aligned}
\left\| \mathbf{D}_\mathbf{T}^* \mathbf{T}\mathbf{T}^\top \right\|_2 &\leq \left\| \mathbf{D}_\mathbf{T}^* \right\|_2 \left\| \mathbf{T}\mathbf{T}^\top \right\|_2, \\
\left\| \mathbf{D}_\mathbf{T}^* \left(\mathbf{T}\mathbf{T}^\top \right)^2 \right\|_2 &\leq \sigma_r \left\| \mathbf{D}_\mathbf{T}^* \right\|_2 \left\| \mathbf{T}\mathbf{T}^\top \right\|_2, \\
\left\| \mathbf{D}_\mathbf{T}^* \left(\mathbf{T}\mathbf{T}^\top \right) \mathbf{D}_\mathbf{T}^* \right\|_2 &\leq \sigma_r \left\| \mathbf{D}_\mathbf{T}^* \right\|_2 \left\| \mathbf{T}\mathbf{T}^\top \right\|_2.
\end{aligned}$$

Given these bounds, we find that

$$\|IV\|_2 \leq (\eta + 3\eta^2 \sigma_r) \left\| \mathbf{D}_\mathbf{T}^* \right\|_2 \left\| \mathbf{T}\mathbf{T}^\top \right\|_2. \quad (12)$$

Now, we move to bound the V term. Indeed, we have

$$\begin{aligned}
&-2\eta \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top - \eta^2 \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right)^\top - \eta^2 \left(\mathbf{D}_\mathbf{T}^* - \mathbf{T}\mathbf{T}^\top \right) \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top + \eta^2 \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top \\
&\preceq (-2\eta + 2\eta^2 \rho \sigma_r + \eta^2 \sigma_1) \mathbf{T}\mathbf{S}^\top \mathbf{S}\mathbf{T}^\top \preceq 0.
\end{aligned} \quad (13)$$

Since $\mathcal{M}_\mathbf{T}(\mathbf{T})\mathcal{M}_\mathbf{T}(\mathbf{T})^\top$ is PSD, we can just relax this term to zero. Finally, we bound the III term. Observe that,

$$\mathbf{T}\mathbf{T}^\top - 2\eta \left(\mathbf{T}\mathbf{T}^\top \right)^2 + \eta^2 \left(\mathbf{T}\mathbf{T}^\top \right)^3 \preceq \mathbf{T}\mathbf{T}^\top - \eta \left(\mathbf{T}\mathbf{T}^\top \right)^2,$$

since $\eta < 1/\sigma_1$ and $\|\mathbf{T}\mathbf{T}^\top\| \leq \rho \sigma_r$. The remaining task is to bound $\mathbf{T}\mathbf{T}^\top - \eta \left(\mathbf{T}\mathbf{T}^\top \right)^2$. Let the singular value decomposition of $\mathbf{T}\mathbf{T}^\top$ as QDQ^\top . Note that D is a diagonal matrix with diagonal entries less than $(1 + \rho)\sigma_r$. We can proceed as

$$\begin{aligned}
\left\| \mathbf{T}\mathbf{T}^\top - \eta \left(\mathbf{T}\mathbf{T}^\top \right)^2 \right\| &= \max_{\|z\|=1} \left(z^\top \mathbf{T}\mathbf{T}^\top z - \eta z^\top \left(\mathbf{T}\mathbf{T}^\top \right)^2 z \right) \\
&= \max_{\|z\|=1} \left(z^\top QDQ^\top z - \eta z^\top QD^2Q^\top z \right) \\
&= \max_{\|z'\|=1} \left(z'^\top D z' - \eta z'^\top D^2 z' \right) \\
&= \max_{\|z'\|=1} \sum_i (d_i - \eta d_i^2) z_i'^2.
\end{aligned}$$

Since $d_i < \sigma_r$ and $1/2\eta \gg \sigma_1$, the above maximum is obtained at the largest singular value of $\mathbf{T}\mathbf{T}^\top$. That is, we have

$$\left\| \mathbf{T}\mathbf{T}^\top - \eta \left(\mathbf{T}\mathbf{T}^\top \right)^2 \right\|_2 \leq \left\| \mathbf{T}\mathbf{T}^\top \right\|_2 \left(1 - \eta \left\| \mathbf{T}\mathbf{T}^\top \right\|_2 \right). \quad (14)$$

Now combining every pieces from equations (12), (13), and (14), we arrive at the following inequality:

$$\left\| \mathcal{M}_{\mathbf{T}}(\mathbf{T}) \mathcal{M}_{\mathbf{T}}(\mathbf{T})^\top - \mathbf{D}_{\mathbf{T}}^* \right\|_2 \leq \left\| \mathbf{T} \mathbf{T}^\top \right\|_2 \left(1 - \eta \left\| \mathbf{T} \mathbf{T}^\top \right\|_2 + 2\eta \left\| \mathbf{D}_{\mathbf{T}}^* \right\|_2 \right).$$

As long as $\left\| \mathbf{D}_{\mathbf{T}}^* \right\|_2 \ll \left\| \mathbf{T} \mathbf{T}^\top \right\|_2$, the contraction rate is roughly $(1 - \eta \left\| \mathbf{T} \mathbf{T}^\top \right\|_2)$. Therefore, we obtain the conclusion of claim (c) in Lemma 2.

Proof of the contraction result (d) in Lemma 2: Direct calculation shows that

$$\begin{aligned} & \mathcal{M}_{\mathbf{T}}(\mathbf{T}) \mathcal{M}_{\mathbf{T}}(\mathbf{T})^\top - \mathbf{D}_{\mathbf{T}}^* \\ &= \left(\mathbf{T} - \eta \left(\mathbf{T} \mathbf{S}^\top \mathbf{S} + (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \mathbf{T} \right) \right) \left(\mathbf{T} - \eta \left(\mathbf{T} \mathbf{S}^\top \mathbf{S} + (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \mathbf{T} \right) \right)^\top - \mathbf{D}_{\mathbf{T}}^* \\ &= VI + VII, \end{aligned}$$

where we denote VI and VII as follows:

$$\begin{aligned} VI &= (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) - \eta ((\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \mathbf{T} \mathbf{T}^\top + \mathbf{T} \mathbf{T}^\top (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*)) \\ &\quad + \eta^2 (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \mathbf{T} \mathbf{T}^\top (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*), \\ VII &= \eta^2 (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \mathbf{T} \mathbf{S}^\top \mathbf{S} \mathbf{T}^\top + \eta^2 \mathbf{T} \mathbf{S}^\top \mathbf{S} \mathbf{T}^\top (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \\ &\quad - 2\eta \mathbf{T} \mathbf{S}^\top \mathbf{S} \mathbf{T}^\top + \eta^2 \mathbf{T} \mathbf{S}^\top \mathbf{S} \mathbf{S}^\top \mathbf{S} \mathbf{T}^\top. \end{aligned}$$

We first show that the $\|VII\|_2 \leq 3\eta \|\mathbf{S} \mathbf{T}^\top\|_2^2$. Firstly, since $\eta \leq \frac{1}{10\sigma_1}$ and the initialization condition $\|\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*\|_2 \leq \rho\sigma_r$, we have

$$\eta^2 \left\| (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \mathbf{T} \mathbf{S}^\top \mathbf{S} \mathbf{T}^\top \right\|_2 \leq \frac{1}{10} \eta \left\| \mathbf{T} \mathbf{S}^\top \mathbf{S} \mathbf{T}^\top \right\|_2.$$

Furthermore, by the choice of η and the fact that $\|\mathbf{S} \mathbf{S}^\top\|_2 \leq \sigma_1$, we find that

$$\eta^2 \left\| \mathbf{T} \mathbf{S}^\top \mathbf{S} \mathbf{S}^\top \mathbf{S} \mathbf{T}^\top \right\|_2 \leq \frac{1}{10} \eta \left\| \mathbf{T} \mathbf{S}^\top \mathbf{S} \mathbf{T}^\top \right\|_2.$$

Putting these results together we have $\|VII\|_2 \leq 3\eta \|\mathbf{S} \mathbf{T}^\top\|_2^2$.

Now for the VI term, direct calculation shows that

$$\begin{aligned} VI &= (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) - \eta (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \mathbf{T} \mathbf{T}^\top \left(I - \frac{\eta}{2} (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \right) \\ &\quad - \eta \left(I - \frac{\eta}{2} (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \right) \mathbf{T} \mathbf{T}^\top (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \\ &= (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \left(\frac{\mathbf{I}}{2} - \eta \mathbf{T} \mathbf{T}^\top \left(I - \frac{\eta}{2} (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \right) \right) \\ &\quad + \left(\frac{\mathbf{I}}{2} - \eta \left(I - \frac{\eta}{2} (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \right) \mathbf{T} \mathbf{T}^\top \right) (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*). \end{aligned}$$

Note that, since $\|\mathbf{T} \mathbf{T}^\top\|_2 \leq \rho\sigma_r$, $\|\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*\|_2 \leq \rho\sigma_r$, and $\eta = \frac{1}{C\sigma_1}$, we obtain that

$$\left\| \frac{\mathbf{I}}{2} - \eta \mathbf{T} \mathbf{T}^\top \left(I - \frac{\eta}{2} (\mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^*) \right) \right\|_2 \leq \left\| \frac{\mathbf{I}}{2} - \eta \mathbf{T} \mathbf{T}^\top \right\|_2.$$

Collecting the above upper bounds with VI and VII, we arrive at

$$\left\| \mathcal{M}_{\mathbf{T}}(\mathbf{T}) \mathcal{M}_{\mathbf{T}}(\mathbf{T})^\top - \mathbf{D}_{\mathbf{T}}^* \right\|_2 \leq \left\| \mathbf{T} \mathbf{T}^\top - \mathbf{D}_{\mathbf{T}}^* \right\|_2 \left\| \mathbf{I} - 2\eta \mathbf{T} \mathbf{T}^\top \right\|_2 + 3\eta \left\| \mathbf{S} \mathbf{T}^\top \right\|_2^2.$$

As a consequence, we reach the conclusion of claim (d) in Lemma 2.

A.2 Additional contraction results for population operators

In this appendix, we offer more population contraction (non-expansion) results, which are useful in showing the contraction results in finite sample setting.

Lemma 6. *Under the same settings as Lemma 2, we have*

- (a) $\|\mathbf{D}_S^* - \mathcal{M}_S(\mathbf{S})\mathbf{S}^\top\|_2 \leq (1 - \eta\sigma_r) \|\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top\|_2 + \eta \|\mathbf{S}\mathbf{T}^\top\|_2^2,$
- (b) $\|\mathcal{M}_S(\mathbf{S})\mathbf{T}^\top\|_2 \leq \|\mathbf{S}\mathbf{T}^\top\|_2,$
- (c) $\|\mathcal{M}_T(\mathbf{T})\mathbf{S}^\top\|_2 \leq \|\mathbf{S}\mathbf{T}^\top\|_2,$
- (d) $\|\mathcal{M}_T(\mathbf{T})\mathbf{T}^\top\|_2 \leq \|\mathbf{T}\mathbf{T}^\top\|_2 + \eta \|\mathbf{S}\mathbf{T}^\top\|_2^2.$

Proof. With $\mathcal{M}_S(\mathbf{S}) = \mathbf{S} - \eta(\mathbf{S}\mathbf{S}^\top\mathbf{S} + \mathbf{S}\mathbf{T}^\top\mathbf{T} - \mathbf{D}_S^*\mathbf{S})$ and simple algebraic manipulations, we obtain

$$\begin{aligned} \mathcal{M}_S(\mathbf{S})\mathbf{S}^\top - \mathbf{D}_S^* &= (\mathbf{S}\mathbf{S}^\top - \mathbf{D}_S^*) - \eta(\mathbf{S}\mathbf{S}^\top\mathbf{S} + \mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top - \mathbf{D}_S^*\mathbf{S}\mathbf{S}^\top) \\ &= (\mathbf{S}\mathbf{S}^\top - \mathbf{D}_S^*)(\mathbf{I} - \eta\mathbf{S}\mathbf{S}^\top) - \eta\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top. \end{aligned}$$

Since $\|\mathbf{S}\mathbf{S}^\top\|_2 \geq 0.9\sigma_r$ by initialization condition and triangular inequality, we know that $\|\mathbf{D}_S^* - \mathcal{M}_S(\mathbf{S})\mathbf{S}^\top\|_2 \leq (1 - 0.9\eta\sigma_r) \|\mathbf{D}_S^* - \mathbf{S}\mathbf{S}^\top\|_2 + \eta \|\mathbf{S}\mathbf{T}^\top\|_2^2$. Therefore, we obtain the conclusion of claim (a).

Move to claim (b), with simple algebraic manipulations, we can show that

$$\begin{aligned} \mathcal{M}_S(\mathbf{S})\mathbf{T}^\top &= \mathbf{S}\mathbf{T}^\top - \eta(\mathbf{S}\mathbf{S}^\top\mathbf{S}\mathbf{T}^\top + \mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{T}^\top - \mathbf{D}_S^*\mathbf{S}\mathbf{T}^\top) \\ &= \left(\frac{1}{2}\mathbf{I} - \eta(\mathbf{S}\mathbf{S}^\top - \mathbf{D}_S^*)\right)\mathbf{S}\mathbf{T}^\top - \mathbf{S}\mathbf{T}^\top\left(\frac{1}{2}\mathbf{I} - \eta\mathbf{T}\mathbf{T}^\top\right). \end{aligned}$$

By initialization condition and triangular inequality, we know that $0 \leq \|(\mathbf{S}\mathbf{S}^\top - \mathbf{D}_S^*)\|_2 \leq \rho\sigma_r$ and $0 \leq \|\mathbf{T}\mathbf{T}^\top\|_2 \leq 1.1\sigma_r$, and hence $\|\mathcal{M}_S(\mathbf{S})\mathbf{T}^\top\|_2 \leq \|\mathbf{S}\mathbf{T}^\top\|_2$. Hence, we reach the conclusion of claim (b).

With $\mathcal{M}_T(\mathbf{T}) = \mathbf{T} - \eta(\mathbf{T}\mathbf{T}^\top\mathbf{T} + \mathbf{T}\mathbf{S}^\top\mathbf{S} - \mathbf{D}_T^*\mathbf{T})$ and direct calculation, we find that

$$\begin{aligned} \mathcal{M}_T(\mathbf{T})\mathbf{S}^\top &= \mathbf{T}\mathbf{S}^\top - \eta(\mathbf{T}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top + \mathbf{T}\mathbf{S}^\top\mathbf{S}\mathbf{S}^\top - \mathbf{D}_T^*\mathbf{T}\mathbf{S}^\top) \\ &= \left(\frac{1}{2}\mathbf{I} - \eta(\mathbf{T}\mathbf{T}^\top - \mathbf{D}_T^*)\right)\mathbf{T}\mathbf{S}^\top - \mathbf{T}\mathbf{S}^\top\left(\frac{1}{2}\mathbf{I} - \eta\mathbf{S}\mathbf{S}^\top\right). \end{aligned}$$

By initialization condition and triangular inequality, we know that $0 \leq \|(\mathbf{T}\mathbf{T}^\top - \mathbf{D}_T^*)\|_2 \leq \rho\sigma_r$ and $0.9\sigma_r \leq \|\mathbf{S}\mathbf{S}^\top\|_2 \leq 0.1\sigma_r + \sigma_1$, and hence $\|\mathcal{M}_T(\mathbf{T})\mathbf{S}^\top\|_2 \leq \|\mathbf{T}\mathbf{S}^\top\|_2$. It leads to the conclusion of claim (c).

Finally, moving to claim (d), simple algebra shows that

$$\begin{aligned} \mathcal{M}_T(\mathbf{T})\mathbf{T}^\top &= \mathbf{T}\mathbf{T}^\top - \eta(\mathbf{T}\mathbf{T}^\top\mathbf{T}\mathbf{T}^\top + \mathbf{T}\mathbf{S}^\top\mathbf{S}\mathbf{T}^\top - \mathbf{D}_T^*\mathbf{T}\mathbf{T}^\top) \\ &= \left(\mathbf{I} - \eta(\mathbf{T}\mathbf{T}^\top - \mathbf{D}_T^*)\right)\mathbf{T}\mathbf{T}^\top - \eta\mathbf{T}\mathbf{S}^\top\mathbf{S}\mathbf{T}^\top. \end{aligned}$$

By initialization condition and triangular inequality, we know that $\|\mathcal{M}_T(\mathbf{T})\mathbf{T}^\top\|_2 \leq \|\mathbf{T}\mathbf{T}^\top\|_2 + \eta \|\mathbf{S}\mathbf{T}^\top\|_2^2$. As a consequence, we obtain the conclusion of claim (d). \square

B Proofs for the finite sample analysis

Recall that, we denote \mathbf{G}_t as the population gradient at iteration t and denote \mathbf{G}_t^n as the corresponding sample gradient with sample size n :

$$\begin{aligned}\mathbf{G}_t &= (\mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^*) \mathbf{F}_t, \\ \mathbf{G}_t^n &= \frac{1}{n} \sum_{i=1}^n \left(\langle \mathbf{A}_i, \mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^* \rangle + \epsilon_i \right) \mathbf{A}_i \mathbf{F}_t.\end{aligned}$$

Then, we can write our update as follows:

$$\mathbf{F}_{t+1} = \mathbf{F}_t - \eta \mathbf{G}_t + \eta \mathbf{G}_t - \eta \mathbf{G}_t^n.$$

We assume the following decomposition by notations: $\mathbf{F} = \mathbf{U}\mathbf{S} + \mathbf{V}\mathbf{T}$. Therefore, we find that

$$\begin{aligned}\mathbf{S}_{t+1} (\mathbf{S}_{t+1})^\top &= \mathbf{U}^\top \mathbf{F}_{t+1} (\mathbf{U}^\top \mathbf{F}_{t+1})^\top \\ &= \mathbf{U}^\top (\mathbf{F}_t - \eta \mathbf{G}_t) (\mathbf{F}_t - \eta \mathbf{G}_t)^\top \mathbf{U} + \eta^2 \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{U} \\ &\quad + \eta \mathbf{U}^\top (\mathbf{F}_t - \eta \mathbf{G}_t) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{U} + \eta \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{F}_t - \eta \mathbf{G}_t)^\top \mathbf{U} \quad (15) \\ &= \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)^\top + \eta^2 \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{U} \\ &\quad + \eta \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{U} + \eta \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)^\top,\end{aligned}$$

where we define $\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)$ as follows:

$$\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) = \mathbf{U}^\top (\mathbf{F}_t - \eta \mathbf{G}_t) = \mathbf{S}_t - \eta (\mathbf{S}_t \mathbf{S}_t^\top \mathbf{S}_t + \mathbf{S}_t \mathbf{T}_t^\top \mathbf{T}_t - \mathbf{D}_{\mathbf{S}}^* \mathbf{S}_t).$$

Furthermore, direct calculation shows that

$$\begin{aligned}\mathbf{S}_{t+1} (\mathbf{T}_{t+1})^\top &= \mathbf{U}^\top \mathbf{F}_{t+1} (\mathbf{V}^\top \mathbf{F}_{t+1})^\top \\ &= \mathbf{U}^\top (\mathbf{F}_t - \eta \mathbf{G}_t) (\mathbf{F}_t - \eta \mathbf{G}_t)^\top \mathbf{V} + \eta^2 \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V} \\ &\quad + \eta \mathbf{U}^\top (\mathbf{F}_t - \eta \mathbf{G}_t) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V} + \eta \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{F}_t - \eta \mathbf{G}_t)^\top \mathbf{V} \quad (16) \\ &= \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top + \eta^2 \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V} \\ &\quad + \eta \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V} + \eta \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top,\end{aligned}$$

where $\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)$ is given by:

$$\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t) = \mathbf{V}^\top (\mathbf{F}_t - \eta \mathbf{G}_t) = \mathbf{T}_t - \eta (\mathbf{T}_t \mathbf{T}_t^\top \mathbf{T}_t + \mathbf{T}_t \mathbf{S}_t^\top \mathbf{S}_t - \mathbf{D}_{\mathbf{T}}^* \mathbf{T}_t).$$

Similarly, we also have

$$\begin{aligned}\mathbf{T}_{t+1} (\mathbf{T}_{t+1})^\top &= \mathbf{V}^\top \mathbf{F}_{t+1} (\mathbf{V}^\top \mathbf{F}_{t+1})^\top \\ &= \mathbf{V}^\top (\mathbf{F}_t - \eta \mathbf{G}_t) (\mathbf{F}_t - \eta \mathbf{G}_t)^\top \mathbf{V} + \eta^2 \mathbf{V}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V} \\ &\quad + \eta \mathbf{V}^\top (\mathbf{F}_t - \eta \mathbf{G}_t) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V} + \eta \mathbf{V}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{F}_t - \eta \mathbf{G}_t)^\top \mathbf{V} \quad (17) \\ &= \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t) \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top + \eta^2 \mathbf{V}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V} \\ &\quad + \eta \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V} + \eta \mathbf{V}^\top (\mathbf{G}_t - \mathbf{G}_t^n) \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top.\end{aligned}$$

Note that, in the above equations, $\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)$ and $\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)$ are the updates of the coefficients when we update S and T using the population gradient. Furthermore, $\eta^2 \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top$, $\eta^2 \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V}$, and $\eta^2 \mathbf{V}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V}$ are second order terms and are relatively small. To facilitate the proof argument, we denote

$$\Delta_t := \frac{1}{n} \sum_{i=1}^n \left(\left\langle \mathbf{A}_i, \mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^* \right\rangle + \epsilon_i \right) \mathbf{A}_i - (\mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^*).$$

We can see that Δ_t is symmetric matrix, and

$$\mathbf{G}_t^m - \mathbf{G}_t = \Delta_t \mathbf{F}_t.$$

B.1 Proof for Lemma 3

Proof. By Lemma 2 and Lemma 6, we have the following contraction results:

$$\begin{aligned} \left\| \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)^\top - \mathbf{D}_{\mathbf{S}}^* \right\|_2 &\leq (1 - \eta \sigma_r) \left\| \mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^* \right\|_2 + 3\eta \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2^2, \\ \left\| \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top \right\|_2 &\leq (1 - \eta \sigma_r) \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2, \\ \left\| \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t) \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top \right\|_2 &\leq \left\| \mathbf{T}_t \mathbf{T}_t^\top \right\|_2 \left(1 - \eta \left\| \mathbf{T}_t \mathbf{T}_t^\top \right\|_2 + 2\eta \left\| \mathbf{D}_{\mathbf{T}}^* \right\|_2 \right), \end{aligned} \quad (18)$$

and the following non-expansion results:

$$\begin{aligned} \left\| \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^* \right\|_2 &\leq (1 - \eta \sigma_r) \left\| \mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^* \right\|_2 + \eta \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2^2, \\ \left\| \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathbf{T}_t^\top \right\|_2 &\leq \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2, \\ \left\| \mathbf{S}_t \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top \right\|_2 &\leq \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2, \\ \left\| \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t) \mathbf{T}_t^\top \right\|_2 &\leq \left\| \mathbf{T}_t \mathbf{T}_t^\top \right\|_2 + \eta \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2^2. \end{aligned} \quad (19)$$

For notation simplicity, let $D_t = \max\{\left\| \mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^* \right\|_2, \left\| \mathbf{T}_t \mathbf{T}_t^\top \right\|_2, \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2\}$, and denote the statistical error $\epsilon_{stat} = \sqrt{\frac{d \log d}{n}} \sigma$. Since Assumption 1 is satisfied, and $\left\| \mathbf{D}_{\mathbf{T}}^* \right\|_2 \leq \epsilon_{stat}$, we have $D_t \leq \sigma_r$ by triangular inequality. Since $\eta \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2 \leq \frac{1}{10} \eta \sigma_r$ by initialization, and $\left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2 \leq D_t$, we have $\eta \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2^2 \leq 0.1 \eta \sigma D_t$. Putting these results together, we have

$$\begin{aligned} \left\| \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)^\top - \mathbf{D}_{\mathbf{S}}^* \right\|_2 &\leq \left(1 - \frac{7}{10} \eta \sigma_r \right) D_t, \\ \left\| \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^* \right\|_2 &\leq \left(1 - \frac{9}{10} \eta \sigma_r \right) D_t, \\ \left\| \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t) \mathbf{T}_t^\top \right\|_2 &\leq \left(1 + \frac{1}{10} \eta \sigma_r \right) D_t. \end{aligned} \quad (20)$$

For the ease of the presentation, we assign a value to the constant C_1 as in the number of sample n . From the requirements of the lemma we know that $\left\| \Delta_t \right\|_2 \leq 10 \sqrt{\frac{kd \log d}{n}} D_t + 4 \sqrt{\frac{d \log d}{n}} \sigma$. We need to connect $\sqrt{\frac{kd \log d}{n}}$ with $\eta \sigma_r$ for the development of the proof. Since $\eta = \frac{1}{100 \sigma_1}$ and $n > C_1 k \kappa^2 d \log^3 d \cdot \max(1, \sigma^2 / \sigma_r^2)$, by choosing $C_1 \geq 1000^2$, we have

$$\sqrt{\frac{kd \log d}{n}} \leq 0.1 \eta \sigma_r. \quad (21)$$

Upper bound for $\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2$: According to equation (15), we have

$$\begin{aligned} \mathbf{S}_{t+1} \mathbf{S}_{t+1}^\top - \mathbf{D}_\mathbf{S}^* &= \underbrace{\mathcal{M}_\mathbf{S}(\mathbf{S}_t) \mathcal{M}_\mathbf{S}(\mathbf{S}_t)^\top - \mathbf{D}_\mathbf{S}^*}_\text{I} + \underbrace{\eta^2 \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{U}}_\text{II} \\ &\quad + \eta \mathcal{M}_\mathbf{S}(\mathbf{S}_t) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{U} + \eta \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) \mathcal{M}_\mathbf{S}(\mathbf{S}_t)^\top, \end{aligned}$$

where we can further expand $\eta \mathcal{M}_\mathbf{S}(\mathbf{S}_t) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{U}$ and $\eta \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) \mathcal{M}_\mathbf{S}(\mathbf{S}_t)^\top$ as follows:

$$\begin{aligned} \eta \mathcal{M}_\mathbf{S}(\mathbf{S}_t) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{U} &= \eta \mathcal{M}_\mathbf{S}(\mathbf{S}_t) \mathbf{F}_t^\top \Delta_t \mathbf{U} = \eta \mathcal{M}_\mathbf{S}(\mathbf{S}_t) (\mathbf{U} \mathbf{S}_t + \mathbf{V} \mathbf{T}_t)^\top \Delta_t \mathbf{U} \\ &= \eta \mathcal{M}_\mathbf{S}(\mathbf{S}_t) \mathbf{S}_t^\top \mathbf{U}^\top \Delta_t \mathbf{U} + \eta \mathcal{M}_\mathbf{S}(\mathbf{S}_t) \mathbf{T}_t^\top \mathbf{V}^\top \Delta_t \mathbf{U} \\ &= \underbrace{\eta \mathcal{M}_\mathbf{S}(\mathbf{S}_t) \mathbf{S}_t^\top \mathbf{U}^\top \Delta_t \mathbf{U} - \eta \mathbf{D}_\mathbf{S}^* \mathbf{U}^\top \Delta_t \mathbf{U}}_\text{III} + \underbrace{\eta \mathcal{M}_\mathbf{S}(\mathbf{S}_t) \mathbf{T}_t^\top \mathbf{V}^\top \Delta_t \mathbf{U}}_\text{IV} + \underbrace{\eta \mathbf{D}_\mathbf{S}^* \mathbf{U}^\top \Delta_t \mathbf{U}}_\text{V}, \end{aligned}$$

and

$$\begin{aligned} \eta \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) \mathcal{M}_\mathbf{S}(\mathbf{S}_t)^\top &= \underbrace{\eta \mathbf{U}^\top \Delta_t \mathbf{U} \mathbf{S}_t \mathcal{M}_\mathbf{S}(\mathbf{S}_t)^\top - \eta \mathbf{U}^\top \Delta_t \mathbf{U} \mathbf{D}_\mathbf{S}^*}_\text{VI} + \underbrace{\eta \mathbf{U}^\top \Delta_t \mathbf{V} \mathbf{T}_t \mathcal{M}_\mathbf{S}(\mathbf{S}_t)^\top}_\text{VII} + \underbrace{\eta \mathbf{U}^\top \Delta_t \mathbf{U} \mathbf{D}_\mathbf{S}^*}_\text{VIII}. \end{aligned}$$

Clearly, our target can be bounded by bounding the eight terms, marked from I to VIII. Note that the spectral norms of the terms (1) III and VI are the same, (2) IV and VII are the same, and (3) V and VIII are the same, which can be upper bounded as follows:

$$\begin{aligned} \text{III \& VI:} \quad & \left\| \eta \mathcal{M}_\mathbf{S}(\mathbf{S}_t) \mathbf{S}_t^\top \mathbf{U}^\top \Delta_t \mathbf{U} - \eta \mathbf{D}_\mathbf{S}^* \mathbf{U}^\top \Delta_t \mathbf{U} \right\|_2 \leq \eta \left\| \mathcal{M}_\mathbf{S}(\mathbf{S}_t) \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^* \right\|_2 \|\Delta_t\|_2, \\ \text{IV \& VII:} \quad & \left\| \eta \mathcal{M}_\mathbf{S}(\mathbf{S}_t) \mathbf{T}_t^\top \mathbf{V}^\top \Delta_t \mathbf{U} \right\|_2 \leq \eta \left\| \mathcal{M}_\mathbf{S}(\mathbf{S}_t) \mathbf{T}_t^\top \right\|_2 \|\Delta_t\|_2, \\ \text{V \& VIII:} \quad & \left\| \eta \mathbf{D}_\mathbf{S}^* \mathbf{U}^\top \Delta_t \mathbf{U} \right\|_2 \leq \eta \|\mathbf{D}_\mathbf{S}^*\|_2 \|\Delta_t\|_2. \end{aligned}$$

Lastly, consider the II term, we have the following bound:

$$\begin{aligned} \left\| \eta^2 \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{U} \right\|_2 &\leq \eta^2 \left\| \Delta_t \mathbf{F}_t \mathbf{F}_t^\top \Delta_t \right\|_2 \\ &\leq \eta^2 \left(\left\| \mathbf{S}_t \mathbf{S}_t^\top \right\|_2 + \left\| \mathbf{T}_t \mathbf{T}_t^\top \right\|_2 + 2 \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2 \right) \|\Delta_t\|_2^2 \quad (22) \\ &\leq \frac{1}{100} \eta \|\Delta_t\|_2^2, \end{aligned}$$

where the last inequality holds by assuming $\rho \leq 0.1$. Putting all the above results together, we obtain that

$$\begin{aligned} \left\| \mathbf{S}_{t+1} \mathbf{S}_{t+1}^\top - \mathbf{D}_\mathbf{S}^* \right\|_2 &\stackrel{(1)}{\leq} \underbrace{\left\| \mathcal{M}_\mathbf{S}(\mathbf{S}_t) \mathcal{M}_\mathbf{S}(\mathbf{S}_t)^\top - \mathbf{D}_\mathbf{S}^* \right\|_2}_\text{I} + \underbrace{2\eta \|\mathbf{D}_\mathbf{S}^*\|_2 \|\Delta_t\|_2}_\text{V + VIII} + \underbrace{\frac{1}{100} \eta \|\Delta_t\|_2^2}_\text{II} \\ &\quad + \underbrace{2\eta \left\| \mathcal{M}_\mathbf{S}(\mathbf{S}_t) \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^* \right\|_2 \|\Delta_t\|_2}_\text{III + VI} + \underbrace{2\eta \left\| \mathcal{M}_\mathbf{S}(\mathbf{S}_t) \mathbf{T}_t^\top \right\|_2 \|\Delta_t\|_2}_\text{IV + VII} \end{aligned}$$

$$\begin{aligned}
&\stackrel{(2)}{\leq} \underbrace{\left(1 - \frac{7}{10}\eta\sigma_r\right) \left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\right\|_2}_{\text{I}} + \underbrace{\frac{1}{50}\|\Delta_t\|_2}_{\text{V} + \text{VIII}} + \underbrace{\frac{1}{100}\eta\|\Delta_t\|_2^2}_{\text{II}} + \underbrace{\frac{4\eta D_t\|\Delta_t\|_2}{\text{III} + \text{VI} + \text{IV} + \text{VII}}} \\
&\stackrel{(3)}{\leq} \left(1 - \frac{7}{10}\eta\sigma_r\right) \left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\right\|_2 + \frac{1}{10}\|\Delta_t\|_2 \\
&\stackrel{(4)}{\leq} \left(1 - \frac{7}{10}\eta\sigma_r\right) \left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\right\|_2 + \sqrt{\frac{kd\log d}{n}}D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma
\end{aligned}$$

where inequality (2) is obtained by the non-expansion property of population update (cf. equations (19) and (20)) ; inequality (3) is obtained by the fact that $\frac{1}{100}\eta\|\Delta_t\|_2 < 0.0001$, and $4\eta D_t < 0.04$; inequality (4) is obtained by plugging in the relaxation of $\|\Delta_t\|_2$ (cf. equation (5)) and organizing according to D_t and σ .

That is, we proved the equation (6) in the Lemma 3, namely, we have

$$\left\|\mathbf{S}_{t+1}\mathbf{S}_{t+1}^\top - \mathbf{D}_\mathbf{S}^*\right\|_2 \leq \left(1 - \frac{7}{10}\eta\sigma_r\right) \left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\right\|_2 + \sqrt{\frac{kd\log d}{n}}D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma. \quad (23)$$

This indicates a contraction with respect to D_t . Since $\sqrt{\frac{kd\log d}{n}} \leq \frac{1}{1000\kappa} = 0.1\eta\sigma_r$ by the choice of n , we find that

$$\left\|\mathbf{S}_{t+1}\mathbf{S}_{t+1}^\top - \mathbf{D}_\mathbf{S}^*\right\|_2 \leq \left(1 - \frac{6}{10}\eta\sigma_r\right) D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma. \quad (24)$$

From the above result, we can verify that

$$\left\|\mathbf{S}_{t+1}\mathbf{S}_{t+1}^\top - \mathbf{D}_\mathbf{S}^*\right\|_2 \leq \left(1 - \frac{6}{10}\eta\sigma_r\right) \left(D_t - 50\kappa\sqrt{\frac{d\log d}{n}}\sigma\right) + 50\kappa\sqrt{\frac{d\log d}{n}}\sigma. \quad (25)$$

Upper bound for $\|\mathbf{S}_t\mathbf{T}_t^\top\|_2$: According to equation (16), we have

$$\begin{aligned}
\mathbf{S}_{t+1}\mathbf{T}_{t+1}^\top &= \underbrace{\mathcal{M}_\mathbf{S}(\mathbf{S}_t)\mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top}_{\text{I}'} + \underbrace{\eta^2\mathbf{U}^\top(\mathbf{G}_t - \mathbf{G}_t^n)(\mathbf{G}_t - \mathbf{G}_t^n)^\top\mathbf{V}}_{\text{II}'} \\
&\quad + \eta\mathcal{M}_\mathbf{S}(\mathbf{S}_t)(\mathbf{G}_t - \mathbf{G}_t^n)^\top\mathbf{V} + \eta\mathbf{U}^\top(\mathbf{G}_t - \mathbf{G}_t^n)\mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top,
\end{aligned}$$

where we can expand $\eta\mathcal{M}_\mathbf{S}(\mathbf{S}_t)(\mathbf{G}_t - \mathbf{G}_t^n)^\top\mathbf{V}$ and $\eta\mathbf{U}^\top(\mathbf{G}_t - \mathbf{G}_t^n)\mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top$ as follows:

$$\begin{aligned}
\eta\mathcal{M}_\mathbf{S}(\mathbf{S}_t)(\mathbf{G}_t - \mathbf{G}_t^n)^\top\mathbf{V} &= \eta\mathcal{M}_\mathbf{S}(\mathbf{S}_t)(\mathbf{U}\mathbf{S}_t + \mathbf{V}\mathbf{T}_t)^\top\Delta_t\mathbf{V} \\
&= \underbrace{\eta\mathcal{M}_\mathbf{S}(\mathbf{S}_t)\mathbf{S}_t^\top\mathbf{U}^\top\Delta_t\mathbf{V} - \eta\mathbf{D}_\mathbf{S}^*\mathbf{U}^\top\Delta_t\mathbf{V}}_{\text{III}'} + \underbrace{\eta\mathcal{M}_\mathbf{S}(\mathbf{S}_t)\mathbf{T}_t^\top\mathbf{V}^\top\Delta_t\mathbf{V}}_{\text{IV}'} \\
&\quad + \underbrace{\eta\mathbf{D}_\mathbf{S}^*\mathbf{U}^\top\Delta_t\mathbf{V}}_{\text{V}'}, \\
\eta\mathbf{U}^\top(\mathbf{G}_t - \mathbf{G}_t^n)\mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top &= \eta\mathbf{U}^\top\Delta_t(\mathbf{U}\mathbf{S}_t + \mathbf{V}\mathbf{T}_t)\mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top \\
&= \underbrace{\eta\mathbf{U}^\top\Delta_t\mathbf{U}\mathbf{S}_t\mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top}_{\text{VI}'} + \underbrace{\eta\mathbf{U}^\top\Delta_t\mathbf{V}\mathbf{T}_t\mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top}_{\text{VII}'}.
\end{aligned}$$

Clearly, our target upper bound for $\|\mathbf{S}_{t+1}\mathbf{T}_{t+1}^\top\|_2$ can be obtained by bounding the seven terms: I' to VII'. Specifically, direct application of inequalities with operator norms leads to

$$\begin{aligned}
\text{III}' : \quad & \left\| \eta \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathbf{S}_t^\top \mathbf{U}^\top \Delta_t \mathbf{V} - \eta \mathbf{D}_{\mathbf{S}}^* \mathbf{U}^\top \Delta_t \mathbf{V} \right\|_2 \leq \eta \left\| \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^* \right\|_2 \|\Delta_t\|_2, \\
\text{IV}' : \quad & \left\| \eta \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathbf{T}_t^\top \mathbf{V}^\top \Delta_t \mathbf{V} \right\|_2 \leq \eta \left\| \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathbf{T}_t^\top \right\|_2 \|\Delta_t\|_2, \\
\text{V}' : \quad & \left\| \eta \mathbf{D}_{\mathbf{S}}^* \mathbf{U}^\top \Delta_t \mathbf{V} \right\|_2 \leq \eta \|\mathbf{D}_{\mathbf{S}}^*\|_2 \|\Delta_t\|_2, \\
\text{VI}' : \quad & \left\| \eta \mathbf{U}^\top \Delta_t \mathbf{U} \mathbf{S}_t \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top \right\|_2 \leq \eta \left\| \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t) \mathbf{S}_t^\top \right\|_2 \|\Delta_t\|_2, \\
\text{VII}' : \quad & \left\| \eta \mathbf{U}^\top \Delta_t \mathbf{V} \mathbf{T}_t \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top \right\|_2 \leq \eta \left\| \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t) \mathbf{T}_t^\top \right\|_2 \|\Delta_t\|_2.
\end{aligned}$$

Lastly, the II term is bounded as in Equation (22), namely, we have

$$\left\| \eta^2 \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V} \right\|_2 \leq \frac{1}{100} \eta \|\Delta_t\|_2^2.$$

Collecting the above results, we find that

$$\begin{aligned}
& \left\| \mathbf{S}_{t+1} \mathbf{T}_{t+1}^\top \right\|_2 \\
& \leq \underbrace{\left\| \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top \right\|_2}_{\text{I}'} + \underbrace{\frac{1}{100} \eta \|\Delta_t\|_2^2}_{\text{II}'} + \underbrace{\eta \left\| \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^* \right\|_2 \|\Delta_t\|_2}_{\text{III}'} + \underbrace{\eta \|\mathbf{D}_{\mathbf{S}}^*\|_2 \|\Delta_t\|_2}_{\text{V}'} \\
& \quad + \underbrace{\eta \left\| \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t) \mathbf{T}_t^\top \right\|_2 \|\Delta_t\|_2}_{\text{IV}'} + \underbrace{\eta \left\| \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t) \mathbf{S}_t^\top \right\|_2 \|\Delta_t\|_2}_{\text{VI}'} + \underbrace{\eta \left\| \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t) \mathbf{T}_t^\top \right\|_2 \|\Delta_t\|_2}_{\text{VII}'} \\
& \stackrel{(1)}{\leq} \underbrace{\left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2}_{\text{I}'} + \underbrace{\frac{1}{100} \eta \|\Delta_t\|_2^2}_{\text{II}'} + \underbrace{\frac{1}{100} \|\Delta_t\|_2}_{\text{V}} + \underbrace{5\eta D_t \|\Delta_t\|_2}_{\text{III}'+\text{IV}'+\text{VI}'+\text{VII}'} \\
& \stackrel{(2)}{\leq} (1 - \eta\sigma_r) \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2 + \frac{1}{10} \|\Delta_t\|_2 \\
& \stackrel{(3)}{\leq} (1 - \eta\sigma_r) \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2 + \sqrt{\frac{kd \log d}{n}} D_t + \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma,
\end{aligned}$$

where inequality (1) is obtained by the non-expansion property of population update (cf. equations (19) and (20)); inequality (2) is obtained by the fact that $\frac{1}{100} \eta \|\Delta_t\|_2 < 0.001$ and $5\eta D_t < 0.05$. In summary, we have

$$\left\| \mathbf{S}_{t+1} \mathbf{T}_{t+1}^\top \right\|_2 \leq (1 - \eta\sigma_r) \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2 + \sqrt{\frac{kd \log d}{n}} D_t + \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma. \quad (26)$$

With similar treatment as in $\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\|_2$, we have the following contraction result with respect to D_t :

$$\left\| \mathbf{S}_{t+1} \mathbf{T}_{t+1}^\top \right\|_2 \leq \left(1 - \frac{9}{10} \eta\sigma_r \right) D_t + \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma. \quad (27)$$

Given the above result, we can verify that

$$\left\| \mathbf{S}_{t+1} \mathbf{T}_{t+1}^\top \right\|_2 \leq \left(1 - \frac{9}{10} \eta\sigma_r \right) \left(D_t - 50\kappa \sqrt{\frac{d \log d}{n}} \sigma \right) + 50\kappa \sqrt{\frac{d \log d}{n}} \sigma. \quad (28)$$

Upper bound for $\|\mathbf{T}_t \mathbf{T}_t^\top\|_2$: According to equation (17) and similar deductions as in previous bounds for $\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2$ and $\|\mathbf{S}_t \mathbf{T}_t^\top\|_2$, we have

$$\begin{aligned} \mathbf{T}_{t+1} \mathbf{T}_{t+1}^\top &= \underbrace{\mathcal{M}_\mathbf{T}(\mathbf{T}_t) \mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top}_{\text{I}''} + \underbrace{\eta^2 \mathbf{V}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V}}_{\text{II}''} \\ &\quad + \eta \mathcal{M}_\mathbf{T}(\mathbf{T}_t) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V} + \eta \mathbf{V}^\top (\mathbf{G}_t - \mathbf{G}_t^n) \mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top, \end{aligned}$$

where the following expansions hold:

$$\begin{aligned} \eta \mathcal{M}_\mathbf{T}(\mathbf{T}_t) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{V} &= \underbrace{\eta \mathcal{M}_\mathbf{T}(\mathbf{T}_t) \mathbf{S}_t^\top \mathbf{U}^\top \Delta_t \mathbf{V}}_{\text{III}''} + \underbrace{\eta \mathcal{M}_\mathbf{T}(\mathbf{T}_t) \mathbf{T}_t^\top \mathbf{V}^\top \Delta_t \mathbf{V}}_{\text{IV}''}, \\ \eta \mathbf{V}^\top (\mathbf{G}_t - \mathbf{G}_t^n) \mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top &= \underbrace{\eta \mathbf{V}^\top \Delta_t \mathbf{U} \mathbf{S}_t \mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top}_{\text{V}''} + \underbrace{\eta \mathbf{V}^\top \Delta_t \mathbf{V} \mathbf{T}_t \mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top}_{\text{VI}''}. \end{aligned}$$

Given the formulations of the terms I''-VI'', we find that

$$\begin{aligned} \text{III}'' \&\text{V}'' : \quad & \left\| \eta \mathcal{M}_\mathbf{T}(\mathbf{T}_t) \mathbf{S}_t^\top \mathbf{U}^\top \Delta_t \mathbf{V} \right\|_2 &\leq \eta \left\| \mathcal{M}_\mathbf{T}(\mathbf{T}_t) \mathbf{S}_t^\top \right\|_2 \|\Delta_t\|_2, \\ \text{IV}'' \&\text{VI}'' : \quad & \left\| \eta \mathcal{M}_\mathbf{T}(\mathbf{T}_t) \mathbf{T}_t^\top \mathbf{V}^\top \Delta_t \mathbf{V} \right\|_2 &\leq \eta \left\| \mathcal{M}_\mathbf{T}(\mathbf{T}_t) \mathbf{T}_t^\top \right\|_2 \|\Delta_t\|_2, \\ \text{II}'' : \quad & \left\| \eta^2 \mathbf{U}^\top (\mathbf{G}_t - \mathbf{G}_t^n) (\mathbf{G}_t - \mathbf{G}_t^n)^\top \mathbf{U} \right\|_2 &\leq \frac{1}{100} \eta \|\Delta_t\|_2^2. \end{aligned}$$

Assume that $\|\mathbf{T}_t \mathbf{T}_t^\top\|_2 = z D_t$ for $0 < z \leq 1$. Note that, z is not necessarily a constant. For notation simplicity we use the short hand that $\epsilon_{stat} = \sqrt{\frac{d \log d}{n}} \sigma$. With the choice of n and equation (5), we have $\|\Delta_t\|_2 \leq \eta \sigma_r D_t + 4\epsilon_{stat}$. Therefore, we obtain that

$$\begin{aligned} & \left\| \mathbf{T}_{t+1} \mathbf{T}_{t+1}^\top \right\|_2 \\ & \stackrel{(1)}{\leq} \underbrace{\left\| \mathcal{M}_\mathbf{T}(\mathbf{T}_t) \mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top \right\|_2}_{\text{I}''} + \underbrace{\frac{1}{100} \eta \|\Delta_t\|_2^2}_{\text{II}''} + \underbrace{2\eta \left\| \mathcal{M}_\mathbf{T}(\mathbf{T}_t) \mathbf{S}_t^\top \right\|_2 \|\Delta_t\|_2}_{\text{III}'' + \text{V}''} + \underbrace{2\eta \left\| \mathcal{M}_\mathbf{T}(\mathbf{T}_t) \mathbf{T}_t^\top \right\|_2 \|\Delta_t\|_2}_{\text{IV}'' + \text{VI}''} \\ & \stackrel{(2)}{\leq} \underbrace{(z - z^2 \eta D_t + 2z\eta \|\mathbf{D}_\mathbf{T}^*\|_2 + 4\eta (\eta \sigma_r D_t + 4\epsilon_{stat})) D_t}_{\text{I}'' + \text{III}'' + \text{V}'' + \text{IV}'' + \text{VI}''} + \underbrace{\frac{1}{100} \eta (\eta \sigma_r D_t + 4\epsilon_{stat})^2}_{\text{II}} \\ & \stackrel{(3)}{\leq} (1 - \eta D_t + 2\eta \|\mathbf{D}_\mathbf{T}^*\|_2 + 4\eta (\eta \sigma_r D_t + 4\epsilon_{stat})) D_t + \frac{1}{100} \eta (\eta \sigma_r D_t + 4\epsilon_{stat})^2, \end{aligned}$$

where inequality (2) is obtained by the non-expansion property of population update (cf. equations (19) and (20)) and the assumption on $\|\Delta_t\|_2$ (cf. equation (5)). For inequality (3), observe that the above quantity is a quadratic formula with respect to z , and the maximum is taken when $z = \frac{1+2\eta \|\mathbf{D}_\mathbf{T}^*\|_2}{2\eta D_t} > 1$. Hence we can just safely plug-in $z = 1$. Now, we arrange by organizing according to D_t and ϵ_{stat} and obtain that

$$\begin{aligned} & \left\| \mathbf{T}_{t+1} \mathbf{T}_{t+1}^\top \right\|_2 \\ & \stackrel{(1)}{\leq} (1 - \eta D_t + 2\eta \epsilon_{stat} + 4\eta^2 \sigma_r D_t + 16\eta \epsilon_{stat}) D_t + \frac{1}{100} \eta (\eta^2 \sigma_r^2 D_t^2 + 16\epsilon_{stat}^2 + 8\epsilon_{stat} \eta \sigma_r D_t) \\ & \stackrel{(2)}{=} (1 - \eta D_t + 4\eta^2 \sigma_r D_t + 0.01\eta^3 \sigma_r^2 D_t) D_t + (0.16\epsilon_{stat} + 0.08\eta \sigma_r D_t + 18D_t) \eta \epsilon_{stat} \\ & \stackrel{(3)}{\leq} (1 - 0.9\eta D_t) D_t + (0.16\epsilon_{stat} + 19D_t) \eta \epsilon_{stat}, \end{aligned}$$

where inequality (1) is obtained by $\|\mathbf{D}_{\mathbf{T}}^*\|_2 \leq \epsilon_{stat}$, and inequality (2) is obtained by organizing according to D_t and σ .

For notation simplicity we introduce $A_t = D_t - 50\epsilon_{stat}$, and hence $D_t = A_t + 50\epsilon_{stat}$ where $\epsilon_{stat} = \kappa\sqrt{\frac{d\log d}{n}}\sigma$. With some algebraic manipulations, we have

$$\left\|\mathbf{T}_{t+1}\mathbf{T}_{t+1}^\top\right\|_2 \leq (1 - 0.9\eta A_t) A_t + 50\kappa\sqrt{\frac{d\log d}{n}}\sigma.$$

Furthermore, from equations (25) and (28), we have

$$D_{t+1} \leq (1 - 0.5\eta A_t) A_t + 50\kappa\sqrt{\frac{d\log d}{n}}\sigma.$$

Putting all these results together yields that

$$A_{t+1} \leq (1 - 0.5\eta A_t) A_t.$$

This completes the proof of the Lemma 3. \square

In the lemma above we use $\sqrt{\frac{kd\log d}{n}}D_t$ to demonstrate how sample iteration converges to population iteration as n increases. However we do not need $\|\Delta_t\|_2 \leq \sqrt{\frac{kd\log d}{n}}D_t$ when n is extremely large such that $\sqrt{\frac{kd\log d}{n}} \gg \eta\sigma_r$. This will lead to a waste of sample and lead to sub-optimal sample complexity result. Hence we introduce the following corollary in order to deal with the scenario .

Corollary 1. *Consider under the same setting and using the same notation as Lemma 3. Suppose $\|\Delta_t\|_2 \leq \eta\sigma_r D_t + 4\sqrt{\frac{d\log d}{n}}\sigma$, Then*

$$\begin{aligned} \left\|\mathbf{S}_{t+1}\mathbf{S}_{t+1}^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 &\leq \left(1 - \frac{7}{10}\eta\sigma_r\right) \left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 + 0.1\eta\sigma_r D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma, \\ \left\|\mathbf{S}_{t+1}\mathbf{T}_{t+1}^\top\right\|_2 &\leq (1 - \eta\sigma_r) \left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2 + 0.1\eta\sigma_r D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma. \end{aligned}$$

Moreover, denote $\epsilon_{stat} = \kappa\sqrt{\frac{d\log d}{n}}\sigma$. Then we have

$$(D_{t+1} - 50\epsilon_{stat}) \leq \left[1 - \frac{1}{2}\eta(D_t - 50\epsilon_{stat})\right] (D_t - 50\epsilon_{stat}). \quad (29)$$

Proof. As in the proof of Lemma 3, if we replace the condition $\|\Delta_t\|_2 \leq 10\sqrt{\frac{kd\log d}{n}}D_t + 4\sqrt{\frac{d\log d}{n}}\sigma$, by $\|\Delta_t\|_2 \leq \eta\sigma_r D_t + 4\sqrt{\frac{d\log d}{n}}\sigma$, nothing changed except the notation, since equation (21) is satisfied directly. Hence this corollary can be proved using the exactly same argument as in the proof in Lemma 3. \square

Note that Lemma 3 is established for $D_t > 50\epsilon_{stat}$. To complete the proof of our main theorem, we want to make sure that D_t do not expand too much after we reaches the statistical accuracy.

Lemma 7. Consider the same setting as Lemma 3, except that $D_t \leq 50\epsilon_{stat}$. We claim that $D_{t+1} \leq 100\epsilon_{stat}$.

Proof. The proof of this Lemma is a simple extension using the proof of Lemma 3. As in the proof of Lemma 3, we know that

$$\left\| \mathbf{S}_{t+1} \mathbf{S}_{t+1}^\top - \mathbf{D}_\mathbf{S}^* \right\|_2 \leq \left(1 - \frac{7}{10} \eta \sigma_r \right) \left\| \mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^* \right\|_2 + \frac{1}{10} \|\Delta_t\|_2.$$

From the hypothesis, $\|\Delta_t\|_2 \leq D_t + \epsilon_{stat} \leq 51\epsilon_{stat}$. Hence, we have

$$\left\| \mathbf{S}_{t+1} \mathbf{S}_{t+1}^\top - \mathbf{D}_\mathbf{S}^* \right\|_2 \leq 100\epsilon_{stat}.$$

Similarly for $\left\| \mathbf{S}_{t+1} \mathbf{T}_{t+1}^\top \right\|_2$, we have

$$\left\| \mathbf{S}_{t+1} \mathbf{T}_{t+1}^\top \right\|_2 \leq (1 - \eta \sigma_r) \left\| \mathbf{S}_t \mathbf{T}_t^\top \right\|_2 + \frac{1}{10} \|\Delta_t\|_2 \leq 100\epsilon_{stat}.$$

Finally, for $\left\| \mathbf{T}_{t+1} \mathbf{T}_{t+1}^\top \right\|_2$, we find that

$$\begin{aligned} & \left\| \mathbf{T}_{t+1} \mathbf{T}_{t+1}^\top \right\|_2 \\ & \stackrel{(1)}{\leq} \underbrace{\left\| \mathcal{M}_\mathbf{T}(\mathbf{T}_t) \mathcal{M}_\mathbf{T}(\mathbf{T}_t)^\top \right\|_2}_\text{I} + \underbrace{\frac{1}{100} \eta \|\Delta_t\|_2^2}_\text{II} + \underbrace{2\eta \left\| \mathcal{M}_\mathbf{T}(\mathbf{T}_t) \mathbf{S}_t^\top \right\|_2 \|\Delta_t\|_2}_\text{III + V} + \underbrace{2\eta \left\| \mathcal{M}_\mathbf{T}(\mathbf{T}_t) \mathbf{T}_t^\top \right\|_2 \|\Delta_t\|_2}_\text{IV + VI} \\ & \stackrel{(2)}{\leq} \left\| \mathbf{T}_t \mathbf{T}_t^\top \right\|_2 \left(1 - \eta \left\| \mathbf{T}_t \mathbf{T}_t^\top \right\|_2 + 2\eta \left\| \mathbf{D}_\mathbf{T}^* \right\|_2 \right) + 5\eta \cdot 50\epsilon_{stat} \cdot 51\epsilon_{stat} \\ & \stackrel{(3)}{\leq} (1 + 300\eta\epsilon_{stat}) 50\epsilon_{stat} \\ & \stackrel{(4)}{\leq} 100\epsilon_{stat} \end{aligned}$$

where inequality (1) is deducted in the proof of Lemma 3; inequality (2) is by relaxing term I using Equation (18), relaxing $\|\Delta_t\|_2 \leq 51\epsilon_{stat}$, and grouping all other terms; inequality (3) is by the assumption that $\left\| \mathbf{D}_\mathbf{T}^* \right\|_2 \leq \epsilon_{stat}$; inequality (4) is by the choice of n such that $\epsilon_{stat} \leq 0.1$.

Putting all the results together, we obtain the conclusion of Lemma 7. \square

B.2 Proof of Theorem 1

Our proof is divided into verifying claim (a) and claim (b).

Proof for claim (a) with the linear convergence: From the result of Lemma 3, we know that

$$\left\| \mathbf{S}_{t+1} \mathbf{S}_{t+1}^\top - \mathbf{D}_\mathbf{S}^* \right\|_2 \leq \left(1 - \frac{7}{10} \eta \sigma_r \right) \left\| \mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^* \right\|_2 + \sqrt{\frac{kd \log d}{n}} D_t + \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma.$$

We consider $\left\| \mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^* \right\|_2 > 1000 \sqrt{\frac{k\kappa^2 d \log d}{n}} \sigma_r > 1000 \sqrt{\frac{k\kappa^2 d \log d}{n}} D_t$. Note that $\eta \sigma_r = 0.01/\kappa$. Then $\sqrt{\frac{kd \log d}{n}} D_t \leq 0.1 \eta \sigma_r \left\| \mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^* \right\|_2$ and $\frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma < 0.1 \eta \sigma_r \left\| \mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^* \right\|_2$

by the choice of the constants in the lower bound of n . Hence, when $\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2 > 1000\sqrt{\frac{k\kappa^2 d \log d}{n}}\sigma_r$, we find that

$$\|\mathbf{S}_{t+1} \mathbf{S}_{t+1}^\top - \mathbf{D}_\mathbf{S}^*\|_2 \leq \left(1 - \frac{5}{10}\eta\sigma_r\right) \|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2. \quad (30)$$

We now have constant contraction for one iteration. We can invoke the Lemma 10 for once, Lemma 12 for t iterations, and take the union bounds, to quantify the probability that equation (5) holds for all iteration t . Shortly we will show that this probability is at least $1 - d^{-c}$ for some universal constant c . But first we need to know how large we need the number of iterations t to be. Suppose equation (5) holds for all iteration t , then

$$\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2 \leq \left(1 - \frac{5}{10}\eta\sigma_r\right)^t \|\mathbf{S}_0 \mathbf{S}_0^\top - \mathbf{D}_\mathbf{S}^*\|_2 \leq \left(1 - \frac{5}{10}\eta\sigma_r\right)^t 0.1\sigma_r,$$

where the final inequality holds by simply plugging in the initialization condition. After at most $t = \frac{1}{\log \frac{1}{1-0.005/\kappa}} \cdot \log \frac{1}{10000\sqrt{\frac{k\kappa^2 d \log d}{n}}}$ iterations, $\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2 < 1000\sqrt{\frac{k\kappa^2 d \log d}{n}}\sigma_r$. Since $\frac{1}{\log \frac{1}{1-0.005/\kappa}} \leq 1.1$, we further simplify this to $t > \log \frac{n}{k\kappa^2 d \log d}$. As a consequence, we claim that after $t = \left\lceil \log \frac{n}{k\kappa^2 d \log d} \right\rceil$ iterations, $\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2 < C\sqrt{\frac{k\kappa^2 d \log d}{n}}\sigma_r$ for some universal constant C .

Now the remaining task is to show that equation (5) holds for all iteration t with probability at least $1 - d^{-c}$. If n is not too large, i.e. $n < d^{c_5}$ for some constant c_5 , then we invoke equation 35 in Lemma 12 for t iterations. This holds with probability at least $1 - td^{-c} > 1 - d^{-c+1}$ for some universal constant c , since $t < \log n < C_5 \log d$. If n is large, i.e., $n^{z_1} > C_2 d \log^3 d k \kappa^2$ for some universal constant $z_1 \in (0, 1)$, then we use Corollary 1 to establish equation (30), and we invoke equation (36) in Lemma 12 for t iterations. This holds with probability at least $1 - t/\exp(n^z) > 1 - d^{-c}$ for some universal constant c , since $t < \log n$.

Therefore equation (5) holds with probability at least $1 - d^{-c}$ for all iteration t , and we complete our proof with $\|\mathbf{S}_t \mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\|_2$.

With the same argument, we also obtain $\|\mathbf{S}_t \mathbf{T}_t^\top\|_2 < C\sqrt{\frac{k\kappa^2 d \log d}{n}}\sigma_r$ after $t = \left\lceil \log \frac{n}{k\kappa^2 d \log d} \right\rceil$ iterations. Therefore, we obtain the conclusion of claim (a) in Theorem 1.

Proof for claim (b) with the sub-linear convergence: For the sublinear convergence part in claim (b), we prove it by induction. We consider the base case. Since $n > C_1 \kappa^2 d \log^3 d \cdot \max(1, \sigma^2/\sigma_r^2)$, we have $50\kappa\sqrt{\frac{d \log d}{n}}\sigma \leq 0.05\sigma_r$ by choosing $\sqrt{C_1} = 1000$. Therefore the base case is correct by the definitions of A_0 and D_0 .

The key induction step is proven in the Lemma 3, as the equation (8). However, as the convergence rate is sub-linear ultimately, it is sub-optimal to directly invoke concentration result (Lemma 12) to establish equation (5) at each iteration and take union bound over all the iterations. Hence, we adapt the standard localization techniques from empirical process theory to sharpen the rates. Note that, these techniques had also been used to study the convergence rates of optimization algorithms in mixture models settings [Dwivedi et al., 2020a, Kwon et al., 2020].

The key idea of the localization technique is that, instead of invoking the concentration result at each iteration, we only do so when D_t is decreased by 2. More precisely, we divide all

the iterations into epochs, where i -th epoch starts at iteration α_i , ends at iteration $\alpha_{i+1} - 1$, and $D_{\alpha_{i+1}} \leq 0.5D_{\alpha_i}$. We invoke Lemma 13 at α_i to establish equation (5) for all the iterations in i -th epoch. Finally, we take a union bound over all the epochs.

By definition, we have

$$\Delta_t = \frac{1}{n} \sum_i^n \langle \mathbf{A}_i, \mathbf{F}_t \mathbf{F}_t^T - \mathbf{X}^* \rangle \mathbf{A}_i - (\mathbf{F}_t \mathbf{F}_t^T - \mathbf{X}^*) + \frac{1}{n} \sum_i^n \epsilon_i \mathbf{A}_i.$$

From Lemma 10, we know that with probability at least $1 - \exp(-C)$,

$$\frac{1}{n} \sum_i^n \epsilon_i \mathbf{A}_i \leq \sqrt{\frac{d \log d}{n}} \sigma.$$

We only have to invoke this concentration result once for the entire algorithm analysis.

At iteration α_i , note that $\|\mathbf{F}_{\alpha_i} \mathbf{F}_{\alpha_i}^T - \mathbf{X}^*\|_2 \leq \|\mathbf{S}_{\alpha_i} \mathbf{S}_{\alpha_i}^T - \mathbf{D}_{\mathbf{S}}^*\|_2 + \|\mathbf{T}_{\alpha_i} \mathbf{T}_{\alpha_i}^T - \mathbf{D}_{\mathbf{T}}^*\|_2 + 2\|\mathbf{S}_{\alpha_i} \mathbf{T}_{\alpha_i}^T\|_2 \leq 4D_{\alpha_i} + \|\mathbf{D}_{\mathbf{T}}^*\|_2 < 5D_{\alpha_i}$. By Lemma 13 we know that, with probability at least $1 - \exp(-C)$, we have

$$\sup_{\|\mathbf{X}\|_2 \leq 5D_{\alpha_i}} \frac{1}{n} \sum_i^n \langle \mathbf{A}_i, \mathbf{X} \rangle \mathbf{A}_i - \mathbf{X} \leq 5\sqrt{\frac{kd \log d}{n}} D_{\alpha_i}.$$

Therefore, we find that

$$\|\Delta_{\alpha_i}\|_2 \leq 5\sqrt{\frac{kd \log d}{n}} D_{\alpha_i} + \sqrt{\frac{d \log d}{n}} \sigma,$$

and equation (5) is satisfied at iteration α_i . For notation simplicity, we define $A_t = D_t - 50\kappa\sqrt{\frac{d \log d}{n}}\sigma$. Invoking Lemma 3, we have

$$D_{\alpha_{i+1}} = A_{\alpha_{i+1}} + 50\kappa\sqrt{\frac{d \log d}{n}}\sigma \leq \left(1 - \frac{1}{2}\eta A_{\alpha_i}\right) A_{\alpha_i} + 50\kappa\sqrt{\frac{d \log d}{n}}\sigma \leq D_{\alpha_i},$$

where the last inequality just comes from $D_{\alpha_i} = A_{\alpha_i} + 50\kappa\sqrt{\frac{d \log d}{n}}\sigma$. At iteration $t \in (\alpha_i, \alpha_{i+1} - 1)$, by induction $D_t = A_t + 50\kappa\sqrt{\frac{d \log d}{n}}\sigma$, and $D_t \leq D_{t-1} \leq D_{\alpha_i}$. Furthermore, we also have $2D_t > D_{\alpha_i}$. Therefore, the following bounds hold:

$$\begin{aligned} \Delta_t &= \frac{1}{n} \sum_i^n \langle \mathbf{A}_i, \mathbf{F}_t \mathbf{F}_t^T - \mathbf{X}^* \rangle \mathbf{A}_i - (\mathbf{F}_t \mathbf{F}_t^T - \mathbf{X}^*) + \frac{1}{n} \sum_i^n \epsilon_i \mathbf{A}_i \\ &\leq 5\sqrt{\frac{d \log d}{n}} D_{\alpha_i} + \sqrt{\frac{d \log d}{n}} \sigma \\ &\leq 10\sqrt{\frac{d \log d}{n}} D_t + \sqrt{\frac{d \log d}{n}} \sigma \\ &\leq \eta \sigma_r D_t + \sqrt{\frac{d \log d}{n}} \sigma. \end{aligned}$$

Hence, equation (5) is satisfied for all iteration $t \in (\alpha_i, \alpha_{i+1} - 1)$. Invoking Lemma 3, we have

$$D_{t+1} = A_{t+1} + 50\kappa\sqrt{\frac{d \log d}{n}}\sigma \leq \left(1 - \frac{1}{2}\eta A_t\right) A_t + 50\kappa\sqrt{\frac{d \log d}{n}}\sigma$$

with probability at least $1 - d^{-c}$ for a universal constant c . This directly implies that

$$A_{t+1} \leq \left(1 - \frac{1}{2}\eta A_t\right) A_t. \quad (31)$$

We first assume that equation (31) holds for all iterations t , and then show that this is true with probability at least $1 - d^{-c}$ for some constant c . With this, we claim that $A_t \leq \frac{4}{\eta t + \frac{4}{A_0}}$.

To see this, we have

$$\begin{aligned} A_{t+1} &\leq \left(1 - \frac{1}{2}\eta A_t\right) A_t \stackrel{(1)}{\leq} \left(1 - \frac{2}{t + \frac{4}{\eta A_0}}\right) \frac{4}{\eta t + \frac{4}{A_0}} \\ &= \frac{\left(t + \frac{4}{\eta A_0}\right) - 2}{t + \frac{4}{\eta A_0}} \frac{4}{\eta \left(t + \frac{4}{\eta A_0}\right)} \\ &\stackrel{(2)}{\leq} \frac{4}{\eta \left(t + 1 + \frac{4}{\eta A_0}\right)} \end{aligned}$$

where inequality (1) holds because $\left(1 - \frac{1}{2}\eta A_t\right) A_t$ is quadratic with respect to A_t and we plug-in the optimal A_t ; inequality (2) holds because $\frac{\left(t + \frac{4}{\eta A_0}\right) - 2}{\left(t + \frac{4}{\eta A_0}\right)^2} \leq \frac{1}{\left(t + \frac{4}{\eta A_0}\right) + 1}$.

Therefore, after $t \geq \Theta\left(\frac{1}{\eta \epsilon_{stat}}\right)$ number of iterations, $A_t = D_t - 50\kappa\sqrt{\frac{d \log d}{n}}\sigma \leq \Theta(\epsilon_{stat})$, which indicates that

$$\max \left\{ \|\mathbf{S}_t \mathbf{S}_t^T - \mathbf{D}_S^*\|_2, \|\mathbf{T}_t \mathbf{T}_t^T\|_2, \|\mathbf{S}_t \mathbf{T}_t^T\|_2 \right\} \leq \Theta(\epsilon_{stat}). \quad (32)$$

Now what is left to be shown is that equation (31) holds for all iterations t with probability at least $1 - d^{-c}$ for some constant c . We first consider $t = \Theta\left(\frac{1}{\eta \epsilon_{stat}}\right)$. If n is not too large, i.e., $n < d^{c_5}$ for some constant c_5 , then we invoke equation (35) in Lemma 12 for t iterations. This holds with probability at least $1 - td^{-c} > 1 - d^{-c+1}$ for some universal constant c , since $t < \log n < C_5 \log d$. If n is large, i.e., $n^{z_1} > C_2 d \log^3 d k \kappa^2$ for some universal constant $z_1 \in (0, 1)$, then we use Corollary 1 to establish equation (31), and we invoke equation (36) in Lemma 12 for t iterations. This holds with probability at least $1 - t/\exp(n^z) > 1 - d^{-c}$ for some universal constant c , since $t < \log n$. If $t > \Theta\left(\frac{1}{\eta \epsilon_{stat}}\right)$, we can show using above argument that after $\Theta\left(\frac{1}{\eta \epsilon_{stat}}\right)$ number of iterations equation (32) holds. After this, by Lemma 7 we know that $D_t = \Theta(\epsilon_{stat})$. Then, we can invoke Lemma 3 or Corollary 1 without further invoking the concentration argument anymore, since the radius in the uniform concentration result does not change.

As a consequence, after $t \geq \Theta\left(\frac{1}{\eta \epsilon_{stat}}\right)$ number of iterations, by triangular inequality, and the assumption that $\|\mathbf{D}_T^*\|_2 \leq \epsilon_{stat}$, we have $\|\mathbf{F}_t \mathbf{F}_t^\top - \mathbf{X}^*\|_2 \leq \Theta(\epsilon_{stat})$. Combined with Lemma 7, we complete the proof of Theorem 1.

C Supporting Lemma

In this appendix, we provide proofs for supporting lemmas in the main text.

C.1 Proof of Lemma 1

Proof. From the definition of operator norm, we have

$$\left\| \mathbf{D}_{\mathbf{T}}^* - \mathbf{T}_0 \mathbf{T}_0^\top \right\|_2 = \max_{\mathbf{x} \in \mathbb{R}^{d-r}: \|\mathbf{x}\|_2 \leq 1} \left| \mathbf{x}^\top \left(\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}_0 \mathbf{T}_0^\top \right) \mathbf{x} \right|.$$

Since $\mathbf{V} \in \mathbb{R}^{d \times (d-r)}$ is an orthonormal matrix, for any $\mathbf{x} \in \mathbb{R}^{d-r}$, we can find a vector $\mathbf{z} \in \mathbb{R}^d$ such that $\mathbf{V}^\top \mathbf{z} = \mathbf{x}$. Hence, we find that

$$\left\| \mathbf{D}_{\mathbf{T}}^* - \mathbf{T}_0 \mathbf{T}_0^\top \right\|_2 = \left\| \mathbf{V} \left(\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}_0 \mathbf{T}_0^\top \right) \mathbf{V}^\top \right\|_2 = \max_{\mathbf{x} \in \mathbb{R}^d: \|\mathbf{x}\|_2 \leq 1} \left| \mathbf{x}^\top \mathbf{V} \left(\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}_0 \mathbf{T}_0^\top \right) \mathbf{V}^\top \mathbf{x} \right|.$$

Without loss of generality we can write any $\mathbf{x} \in \mathbb{R}^d$ as $\mathbf{x} = \mathbf{x}_u + \mathbf{x}_v$, such that $\mathbf{U}\mathbf{z} = \mathbf{x}_u$ for some $\mathbf{z} \in \mathbb{R}^r$, and $\mathbf{V}\mathbf{z}' = \mathbf{x}_v$ for some $\mathbf{z}' \in \mathbb{R}^{d-r}$ since \mathbf{U} and \mathbf{V} are perpendicular to each other and they together span \mathbb{R}^d . If $\mathbf{x}^* = \arg \max_{\mathbf{x} \in \mathbb{R}^d: \|\mathbf{x}\|_2 \leq 1} \left| \mathbf{x}^\top \mathbf{V} \left(\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}_0 \mathbf{T}_0^\top \right) \mathbf{V}^\top \mathbf{x} \right|$ then \mathbf{x}_u^* is zero. It is because if $\mathbf{x}_u^* \neq 0$, one can decrease \mathbf{x}_u^* to zero and increase \mathbf{x}_v^* to $\mathbf{x}_v^* / \|\mathbf{x}_v^*\|_2$, which does make the target quantity smaller. Therefore, we have

$$\begin{aligned} & \left\| \mathbf{D}_{\mathbf{T}}^* - \mathbf{T}_0 \mathbf{T}_0^\top \right\|_2 \\ &= \max_{\mathbf{x} \in \mathbb{R}^d: \|\mathbf{x}\|_2 \leq 1} \left| \mathbf{x}^\top \mathbf{V} \left(\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}_0 \mathbf{T}_0^\top \right) \mathbf{V}^\top \mathbf{x} \right| \\ &= \max_{\substack{\mathbf{x}: \|\mathbf{x}\|_2 \leq 1, \\ \mathbf{U}^\top \mathbf{x} = 0}} \left| \mathbf{x}^\top \mathbf{V} \left(\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}_0 \mathbf{T}_0^\top \right) \mathbf{V}^\top \mathbf{x} + \mathbf{x}^\top \mathbf{U} \left(\mathbf{D}_{\mathbf{S}}^* - \mathbf{S}_0 \mathbf{S}_0^\top \right) \mathbf{U}^\top \mathbf{x} + 2\mathbf{x}^\top \left(\mathbf{U} \mathbf{S}_0 \mathbf{T}_0^\top \mathbf{V}^\top \right) \mathbf{x} \right| \\ &\leq \max_{\mathbf{x}: \|\mathbf{x}\|_2 \leq 1} \left| \mathbf{x}^\top \mathbf{V} \left(\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}_0 \mathbf{T}_0^\top \right) \mathbf{V}^\top \mathbf{x} + \mathbf{x}^\top \mathbf{U} \left(\mathbf{D}_{\mathbf{S}}^* - \mathbf{S}_0 \mathbf{S}_0^\top \right) \mathbf{U}^\top \mathbf{x} + 2\mathbf{x}^\top \left(\mathbf{U} \mathbf{S}_0 \mathbf{T}_0^\top \mathbf{V}^\top \right) \mathbf{x} \right| \\ &= \left\| \mathbf{F}_0 \mathbf{F}_0^\top - \mathbf{X}^* \right\|_2 \leq 0.7\rho\sigma_r, \end{aligned}$$

where the final inequality is due to the Assumption 1. The same techniques can be applied to obtain

$$\left\| \mathbf{D}_{\mathbf{S}}^* - \mathbf{S}_0 \mathbf{S}_0^\top \right\|_2 \leq \left\| \mathbf{F}_0 \mathbf{F}_0^\top - \mathbf{X}^* \right\|_2 \leq 0.7\rho\sigma_r.$$

For $\left\| \mathbf{S}_0 \mathbf{T}_0^\top \right\|_2$, we claim that the following equations hold:

$$\left\| \mathbf{S}_0 \mathbf{T}_0^\top \right\|_2 = \left\| \mathbf{U} \mathbf{S}_0 \mathbf{T}_0^\top \mathbf{V}^\top \right\|_2 = 0.5 \left\| \mathbf{U} \mathbf{S}_0 \mathbf{T}_0^\top \mathbf{V}^\top + \mathbf{V} \mathbf{T}_0 \mathbf{S}_0^\top \mathbf{U}^\top \right\|_2.$$

To see the last equality, let σ_1 be the largest eigen-value (in magnitude) of $\mathbf{U} \mathbf{S}_0 \mathbf{T}_0^\top \mathbf{V}^\top$ and let $\bar{\mathbf{x}}$ be the corresponding eigen-vector. For some $c \in (0, 1)$, let $\bar{\mathbf{x}} = c\bar{\mathbf{x}}_u + \sqrt{1-c^2}\bar{\mathbf{x}}_v$ such that $\mathbf{U}\mathbf{z} = \bar{\mathbf{x}}_u$ for some $\mathbf{z} \in \mathbb{R}^r$, $\mathbf{V}\mathbf{z}' = \bar{\mathbf{x}}_v$ for some $\mathbf{z}' \in \mathbb{R}^{d-r}$, and $\|\bar{\mathbf{x}}_u\|_2 = 1$ and $\|\bar{\mathbf{x}}_v\|_2 = 1$. Then, direct algebra leads to

$$\sigma_1 = (\bar{\mathbf{x}})^\top \mathbf{U} \mathbf{S}_0 \mathbf{T}_0^\top \mathbf{V}^\top \bar{\mathbf{x}} = c\sqrt{1-c^2} (\bar{\mathbf{x}})_u^\top \mathbf{U} \mathbf{S}_0 \mathbf{T}_0^\top \mathbf{V}^\top \bar{\mathbf{x}}_v.$$

For the RHS of the above equation, the optimal choice of c is $1/\sqrt{2}$. We already know that the largest eigen-value (in magnitude) of $\mathbf{V} \mathbf{T}_0 \mathbf{S}_0^\top \mathbf{U}^\top$ is also σ_1 . Therefore, we obtain that

$$(\bar{\mathbf{x}})^\top \mathbf{V} \mathbf{T}_0 \mathbf{S}_0^\top \mathbf{U}^\top \bar{\mathbf{x}} = c\sqrt{1-c^2} (\bar{\mathbf{x}})_u^\top \mathbf{U} \mathbf{S}_0 \mathbf{T}_0^\top \mathbf{V}^\top \bar{\mathbf{x}}_v = \sigma_1.$$

Collecting the above results, we have $\|\mathbf{US}_0\mathbf{T}_0^\top\mathbf{V}^\top\|_2 = 0.5\|\mathbf{US}_0\mathbf{T}_0^\top\mathbf{V}^\top + \mathbf{VT}_0\mathbf{S}_0^\top\mathbf{U}^\top\|_2$. Then, an application of triangular inequality yields that

$$\begin{aligned} 2\|\mathbf{S}_0\mathbf{T}_0^\top\|_2 &= \|\mathbf{US}_0\mathbf{T}_0^\top\mathbf{V}^\top + \mathbf{VT}_0\mathbf{S}_0^\top\mathbf{U}^\top\|_2 \\ &\leq \|\mathbf{F}_0\mathbf{F}_0^\top - \mathbf{X}^*\|_2 + \|\mathbf{D}_\mathbf{T}^* - \mathbf{T}_0\mathbf{T}_0^\top + \mathbf{D}_\mathbf{S}^* - \mathbf{S}_0\mathbf{S}_0^\top\|_2. \end{aligned}$$

We can check that $\|\mathbf{D}_\mathbf{T}^* - \mathbf{T}_0\mathbf{T}_0^\top + \mathbf{D}_\mathbf{S}^* - \mathbf{S}_0\mathbf{S}_0^\top\|_2 \leq 0.7 \cdot \sqrt{2}\rho\sigma_r$ by decomposing the eigenvector $\mathbf{x} = c\mathbf{x}_u + \sqrt{1-c^2}\mathbf{x}_v$ as above. Therefore, $\|\mathbf{S}_0\mathbf{T}_0^\top\|_2 < \rho\sigma_r$.

As a consequence, we obtain the conclusion of the lemma. \square

D Concentration bounds

In this appendix, we want establish the uniform concentration bound for the following term:

$$\frac{1}{n} \sum_{i=1}^n \left(\langle \mathbf{A}_i, \mathbf{FF}^\top - \mathbf{X}^* \rangle + \epsilon_i \right) \mathbf{A}_i - (\mathbf{FF}^\top - \mathbf{X}^*),$$

for any matrix $F \in \mathbb{R}^{d \times k}$ such that $\|\mathbf{FF}^\top - \mathbf{X}^*\|_2 \leq R$ for some radius $R > 0$. To do so, we have to bound the spectral norm of each random observation, and then take Bernstein/Chernoff type bound.

Lemma 8. (*Matrix Bernstein, Theorem 1.4 in [Tropp \[2012\]](#)*) Consider a finite sequence $\{\mathbf{X}_k\}$ of independent, random, self-adjoint matrices with dimension d . Assume that each random matrix satisfies

$$\mathbb{E}[\mathbf{X}_k] = \mathbf{0}, \quad \text{and} \quad \lambda_{\max}(\mathbf{X}_k) \leq R \quad \text{almost surely.}$$

Then, for all $t \geq 0$,

$$\mathbb{P} \left(\lambda_{\max} \left(\sum_k \mathbf{X}_k \right) \geq t \right) \leq d \cdot \exp \left(\frac{-t^2/2}{\sigma^2 + Rt/3} \right) \quad \text{where} \quad \sigma^2 := \left\| \sum_k \mathbb{E}(\mathbf{X}_k^2) \right\|_2. \quad (33)$$

Lemma 9. Let \mathbf{A} be a symmetric random matrix in $\mathbb{R}^{d \times d}$, with the upper triangle entries ($i \geq j$) being independently sampled from an identical sub-Gaussian distribution whose mean is 0 and variance proxy is 1. Let ϵ follows $N(0, \sigma)$. Then

$$\mathbb{P} \left(\|\epsilon\mathbf{A}\|_2 > C_1\sigma\sqrt{d} \right) \leq \exp(-C_2).$$

Proof. As ϵ is sub-Gaussian, we know that for all $t > 0$

$$\mathbb{P}(|\epsilon| > t\sigma) \leq 2 \exp \left(-\frac{t}{2} \right)$$

By standard ϵ -net argument [[Tropp, 2012](#), [Vershynin, 2018](#)], for some universal constant C_1, C_2 , we have

$$\mathbb{P} \left(\|\mathbf{A}\|_2 > C_1\sqrt{d} \right) \leq \exp \left(-\frac{d}{C_2} \right).$$

Applying the union bound to the above concentration results leads to

$$\mathbb{P}\left(|\epsilon| > C_1\sigma \text{ or } \|\mathbf{A}\|_2 > C_2\sqrt{d}\right) \leq 2\exp\left(-\frac{C_1}{2}\right) + \exp\left(-\frac{d}{C_3}\right) \leq \exp(-C_4).$$

Note that, $\|\mathbf{A}\epsilon\|_2 = |\epsilon| \|\mathbf{A}\|_2$. Therefore, we have

$$\mathbb{P}\left(\|\epsilon\mathbf{A}\|_2 > C_1\sigma\sqrt{d}\right) \leq \exp(-C_2).$$

□

Lemma 10. (Lemma 4 re-stated) Let \mathbf{A}_i be symmetric random matrices in $\mathbb{R}^{d \times d}$, with the upper triangle entries ($i \geq j$) being independently sampled from an identical sub-Gaussian distribution whose mean is 0 and variance proxy is 1. Let ϵ_i follows $N(0, \sigma)$. Then

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_i^n \mathbf{A}_i \epsilon_i\right\|_2 \geq C\sqrt{\frac{d\sigma^2}{n}}\right) \leq \exp(-C).$$

Proof. We prove the lemma by applying the matrix Bernstein bound. In fact, direct calculation shows that

$$\mathbb{E}\left((\mathbf{A}_i \epsilon_i)^2\right) = \sigma^2 \mathbb{E}(\mathbf{A}_i^2) = \sigma^2 d \mathbf{I}.$$

Hence, we obtain

$$\left\|\sum_i^n \mathbb{E}\left((\mathbf{A}_i \epsilon_i)^2\right)\right\|_2 \leq n\sigma^2 d.$$

From the matrix Bernstein bound [Wainwright, 2019], we find that

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_i^n \mathbf{A}_i \epsilon_i\right\|_2 \geq t\right) \leq d \cdot \exp\left(\frac{-3t^2 n^2}{6dn\sigma^2 + 2C_1\sigma\sqrt{dtn}}\right) = d \cdot \exp\left(\frac{-3t^2 n}{6d\sigma^2 + 2C_1\sigma\sqrt{dt}}\right).$$

For any $\delta < 1/e$, let $t = \log \frac{1}{\delta} \sqrt{\frac{d\sigma^2}{n}}$. Then, the above bound becomes

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_i^n \mathbf{A}_i \epsilon_i\right\|_2 \geq \log \frac{1}{\delta} \sqrt{\frac{d\sigma^2}{n}}\right) \leq \delta.$$

Or equivalently, for any $C > 1$, let $t = C\sqrt{\frac{d\sigma^2}{n}}$,

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_i^n \mathbf{A}_i \epsilon_i\right\|_2 \geq C\sqrt{\frac{d\sigma^2}{n}}\right) \leq \exp(-C).$$

As a consequence, we reach the conclusion of the lemma. □

Lemma 11. Let \mathbf{A} be a symmetric random matrix in $\mathbb{R}^{d \times d}$, with the upper triangle entries ($i \geq j$) being independently sampled from an identical sub-Gaussian distribution whose mean is 0 and variance proxy is 1. Let \mathbf{U} be a deterministic symmetric matrix of the same dimension. Then, for some universal constant C_1, C_2 , we have

$$\mathbb{P}(\|\langle \mathbf{A}, \mathbf{U} \rangle \mathbf{A} - \mathbf{U}\|_2 \geq C_1 d \|\mathbf{U}\|_F) \leq \exp(-d/C_2).$$

Proof. We show this by standard ϵ -net argument. In particular, we have

$$\begin{aligned}\|\langle \mathbf{A}, \mathbf{U} \rangle \mathbf{A} - \mathbf{U}\|_2 &= \max_{\mathbf{x} \in \mathcal{S}^{d-1}} \mathbf{x}^\top (\langle \mathbf{A}, \mathbf{U} \rangle \mathbf{A} - \mathbf{U}) \mathbf{x} \\ &= \max_{\mathbf{x} \in \mathcal{S}^{d-1}} \langle \mathbf{A}, \mathbf{U} \rangle \langle \mathbf{A}, \mathbf{x} \mathbf{x}^\top \rangle - (\mathbf{x}^\top \mathbf{U} \mathbf{x}).\end{aligned}$$

Note that $\langle \mathbf{A}, \mathbf{U} \rangle = \sum_{i,j} A_{ij} U_{ij}$ is sub-Gaussian with variance proxy $\|\mathbf{U}\|_F^2$, and $\langle \mathbf{A}, \mathbf{x} \mathbf{x}^\top \rangle = \sum_{i,j} A_{ij} x_i x_j$ is sub-Gaussian with variance proxy 1. Therefore $\mathbb{P}(|\langle \mathbf{A}, \mathbf{U} \rangle| > t \|\mathbf{U}\|_F) \leq \exp(-t^2)$ and $\mathbb{P}(|\langle \mathbf{A}, \mathbf{x} \mathbf{x}^\top \rangle| > t) \leq \exp(-t^2)$. By the union bound,

$$\mathbb{P}\left(|\langle \mathbf{A}, \mathbf{U} \rangle \langle \mathbf{A}, \mathbf{x} \mathbf{x}^\top \rangle| > t \|\mathbf{U}\|_F\right) \leq 2 \exp(-t).$$

Since $(\mathbf{x}^\top \mathbf{U} \mathbf{x}) \leq \|\mathbf{U}\|_2 \leq \|\mathbf{U}\|_F$, we have

$$\mathbb{P}\left(\mathbf{x}^\top (\langle \mathbf{A}, \mathbf{U} \rangle \mathbf{A} - \mathbf{U}) \mathbf{x} \geq t \|\mathbf{U}\|_F\right) \leq \exp\left(-\frac{t}{C_1}\right). \quad (34)$$

By the standard ϵ -net argument, let \mathcal{V} be the ϵ covering of \mathcal{S}^{d-1} . Then, we find that

$$\|\langle \mathbf{A}, \mathbf{U} \rangle \mathbf{A} - \mathbf{U}\|_2 \leq \frac{1}{1-2\epsilon} \max_{\mathbf{x} \in \mathcal{V}} \mathbf{x}^\top (\langle \mathbf{A}, \mathbf{U} \rangle \mathbf{A} - \mathbf{U}) \mathbf{x}.$$

Now we fix ϵ to be $1/4$. Then, for equation (34) we take union bound over \mathcal{V} and we have

$$\mathbb{P}\left(\max_{\mathbf{x} \in \mathcal{V}} \mathbf{x}^\top (\langle \mathbf{A}, \mathbf{U} \rangle \mathbf{A} - \mathbf{U}) \mathbf{x} \geq t \|\mathbf{U}\|_F\right) \leq |\mathcal{V}| \exp\left(-\frac{t}{C_1}\right), \quad \text{for } t > C_2$$

and $|\mathcal{V}| = e^{d \log 9}$. By choosing $t = C_1 d$ for reasonably large universal constant C_1 we have

$$\mathbb{P}(\|\langle \mathbf{A}, \mathbf{U} \rangle \mathbf{A} - \mathbf{U}\|_2 \geq C_1 d \|\mathbf{U}\|_F) \leq \exp(-d/C_2).$$

As a consequence, we obtain the conclusion of the lemma. \square

Lemma 12. *Let \mathbf{A}_i be a symmetric random matrix of dimension d by d , with the upper triangle entries ($i \geq j$) being independently sampled from an identical sub-Gaussian distribution whose mean is 0 and variance proxy is 1. Let \mathbf{U} be a deterministic symmetric matrix of the same dimension. Then as long as $n > C_1 d \log^3 d$ for some universal $C_1, C_2 > 10$, we have*

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_i^n (\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U})\right\|_2 \leq \sqrt{\frac{d \log d}{n}} \|\mathbf{U}\|_F\right) \geq 1 - \exp(-C_2 \log d). \quad (35)$$

Moreover when n is larger than the order of d , that is, if there exists a constant $z_1 \in (0, 1)$ such that $n^{z_1} > C_2 d \log^3 d \kappa^2$, for some universal constant $z_2 \in (0, 1)$ we have

$$\mathbb{P}\left(\left\|\frac{1}{n} \sum_i^n (\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U})\right\|_2 \leq \frac{1}{\kappa \log d \sqrt{k} C_2} \|\mathbf{U}\|_F\right) \geq 1 - \exp(-n^{z_2}). \quad (36)$$

Proof. Following Lemma 8, we want to first bound the second order moment of the random matrices. Since $\mathbb{E}\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i = \mathbf{U}$ and \mathbf{U} has no randomness, we have

$$\mathbb{E}(\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U})^2 = \mathbb{E}(\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i)^2 - \mathbf{U}^2.$$

The (m, n) entry of $\mathbb{E}(\langle \mathbf{A}, \mathbf{U} \rangle \mathbf{A})^2$ equals to

$$\sum_{a,b,c,d,j=1}^d \mathbb{E}(A_{ab}A_{cd}U_{ab}U_{cd}A_{mj}A_{jn}).$$

For diagonal entries, i.e., $m = n$, the expectation is not zero if and only if $A_{ab} = A_{cd}$. Hence for diagonal entry (m, m) , its expectation is

$$\sum_{a,b}^d \mathbb{E}(A_{ab}^2 A_{mm}^2) U_{ab}^2 = \sum_{a,b}^d U_{ab}^2 + 2U_{mm}^2 = \|\mathbf{U}\|_F^2 + 2U_{mm}^2.$$

For off diagonal entries, i.e., $m \neq n$, the expectation is not zero for that entry when (1) $A_{ab} = A_{mj}$ and $A_{cd} = A_{jn}$, or when (2) $A_{ab} = A_{jn}$ and $A_{cd} = A_{mj}$. For both cases, the expectation equals the (m, n) entry of \mathbf{U}^2 . Therefore, we obtain that

$$\sum_{j=1}^d \mathbb{E}(A_{mj}^2 A_{jn}^2 U_{mj} U_{jn}) = \sum_{j=1}^d U_{mj} U_{jn}.$$

Hence the (m, n) entry of $\mathbb{E}(\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U})^2$ equals 0 when $m \neq n$, and equals $\|\mathbf{U}\|_F^2 + 2U_{mm}^2 - \sum_j U_{mj}^2$ when $m = n$. Hence $\left\| \mathbb{E}(\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U})^2 \right\|_2 \leq 3\|\mathbf{U}\|_F^2$ and

$$\left\| \sum_i^n \mathbb{E}(\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U})^2 \right\|_2 \leq 3n\|\mathbf{U}\|_F^2.$$

Then, the following inequality holds:

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_i^n (\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U})\right) \geq t\right) \leq d \cdot \exp\left(\frac{-t^2/2}{3n\|\mathbf{U}\|_F^2 + \frac{C_1 d \log d \|\mathbf{U}\|_F^t}{3}}\right)$$

where C_1 is a universal constant inherited from Lemma 11 and

$$\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n} \sum_i^n (\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U})\right) \geq t\right) \leq d \cdot \exp\left(\frac{-3t^2 n}{18\|\mathbf{U}\|_F^2 + 2C_1 d \log d \|\mathbf{U}\|_F t}\right).$$

Let $t = \sqrt{\frac{d \log d}{n}} \|\mathbf{U}\|_F$, and as long as $n > C_2 d \log^3 d$ for some universal constant $C_4 > 1000$, we have

$$\begin{aligned} & \mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n} \sum_i^n (\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U})\right) \geq \sqrt{\frac{d \log d}{n}} \|\mathbf{U}\|_F\right) \\ & \leq d \cdot \exp\left(\frac{-3\left(\sqrt{\frac{d \log d}{n}} \|\mathbf{U}\|_F\right)^2 n}{18\|\mathbf{U}\|_F^2 + 2C_1 d \log d \|\mathbf{U}\|_F \left(\sqrt{\frac{d \log d}{n}} \|\mathbf{U}\|_F\right)}\right) \\ & \leq d \cdot \exp\left(\frac{-d \log d}{C_3 d}\right) \quad \left(\text{for } \sqrt{\frac{d \log d}{n}} \cdot \log d < 1\right) \\ & \leq \exp(-C_4 \log d) \quad \left(\text{for some universal constant } C_4\right). \end{aligned}$$

Hence we finish the proof for equation 35.

For the tightness of our statistical analysis, we need to consider the case when n is larger than the order of polynomial of d . If there exists a constant $z \in (0, 1)$ such that

$$n^z > C_2 d \log^3 dk \kappa^2,$$

then plugging in $t = \sqrt{\frac{d \log d}{C_2 d \log^3 dk \kappa^2}} \|\mathbf{U}\|_F = \frac{1}{\kappa \log d \sqrt{k C_2}} \|\mathbf{U}\|_F$, we have

$$\begin{aligned} & \mathbb{P} \left(\lambda_{\max} \left(\frac{1}{n} \sum_i^n (\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U}) \right) \geq \frac{1}{\kappa \log d \sqrt{k C_2}} \|\mathbf{U}\|_F \right) \\ & \leq d \cdot \exp \left(\frac{-3 \left(\frac{1}{\kappa \log d \sqrt{k C_2}} \|\mathbf{U}\|_F \right)^2 n}{18 \|\mathbf{U}\|_F^2 + 2 C_1 d \log d \|\mathbf{U}\|_F \left(\frac{1}{\kappa \log d \sqrt{k C_2}} \|\mathbf{U}\|_F \right)} \right) \\ & = d \cdot \exp \left(\frac{-3n}{18 (\kappa \log d \sqrt{k C_2})^2 + 2 C_1 d \log d (\kappa \log d \sqrt{k C_2})} \right) \\ & \leq \exp \left(\frac{-n}{C_3 n^{z_1}} \right) \\ & \leq \exp(-n^{z_2}) \quad (\text{for some universal constant } z_2 \in (0, 1)). \end{aligned}$$

In summary, we reach the conclusion of the lemma. \square

Lemma 13. (Lemma 5 re-stated) Let \mathbf{A}_i be a symmetric random matrix of dimension d by d . Its upper triangle entries ($i \geq j$) are independently sampled from an identical sub-Gaussian distribution whose mean is 0 and variance proxy is 1. If \mathbf{U} is of rank k and is in a bounded spectral norm ball of radius R (i.e. $\|\mathbf{U}\|_2 \leq R$), then we have

$$\mathbb{P} \left(\sup_{\mathbf{U}: \|\mathbf{U}\|_2 \leq R} \left\| \frac{1}{n} \sum_i^n (\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U}) \right\|_2 \leq \sqrt{\frac{d \log d}{n}} \sqrt{k} R \right) \geq 1 - \exp(-C_2 \log d). \quad (37)$$

Moreover when n is larger than the order of d , that is, if there exists a constant $z_1 \in (0, 1)$ such that $n^{z_1} > C_2 d \log^3 dk \kappa^2$, for some universal constant $z_2 \in (0, 1)$ we have

$$\mathbb{P} \left(\sup_{\mathbf{U}: \|\mathbf{U}\|_2 \leq R} \left\| \frac{1}{n} \sum_i^n (\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U}) \right\|_2 \leq \frac{1}{\kappa \log d \sqrt{k C_2}} R \right) \geq 1 - \exp(-n^{z_2}). \quad (38)$$

Proof. To show this uniform convergence result, we use the standard discretization techniques (i.e. ϵ -net). In particular, we have

$$\begin{aligned} & \sup_{\mathbf{U}: \|\mathbf{U}\|_2 \leq R} \left\| \frac{1}{n} \sum_i^n (\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U}) \right\|_2 \\ & = \sup_{\mathbf{U}: \|\mathbf{U}\|_2 \leq R} \sup_{\mathbf{x}: \|\mathbf{x}\|_2 \leq 1} \frac{1}{n} \left| \sum_i^n \left(\langle \mathbf{A}_i, \mathbf{U} \rangle \langle \mathbf{A}_i, \mathbf{x} \mathbf{x}^\top \rangle - \langle \mathbf{U}, \mathbf{x} \mathbf{x}^\top \rangle \right) \right|. \end{aligned}$$

Since the above quantity is symmetric, we can take off the absolute value and only look at the one-side deviation. The crux is how to construct the ϵ -net for \mathbf{U} . We decompose \mathbf{U} .

Since \mathbf{U} is of rank k and $\|\mathbf{U}\|_2 \leq R$, we can write $\mathbf{U} = \sum_i^k \mathbf{u}_i \mathbf{u}_i^\top$ where \mathbf{u}_i are vectors, with $\|\mathbf{u}_i\|_2 \leq \sqrt{R}$ and $\mathbf{u}_i^\top \mathbf{u}_j = 0$ for $i \neq j$. Therefore

$$\begin{aligned} & \frac{1}{n} \sum_i^n \left(\langle \mathbf{A}_i, \mathbf{U} \rangle \langle \mathbf{A}_i, \mathbf{x} \mathbf{x}^\top \rangle - \langle \mathbf{U}, \mathbf{x} \mathbf{x}^\top \rangle \right) \\ &= \frac{1}{n} \sum_i^n \left(\langle \mathbf{A}_i, \sum_i^k \mathbf{u}_i \mathbf{u}_i^\top \rangle \langle \mathbf{A}_i, \mathbf{x} \mathbf{x}^\top \rangle - \langle \sum_i^k \mathbf{u}_i \mathbf{u}_i^\top, \mathbf{x} \mathbf{x}^\top \rangle \right). \end{aligned}$$

Now we can construct a standard ϵ -net for each $\mathbf{u}_i \in \mathbb{R}^d$, and in total we construct k such epsilon net for \mathbf{U} . Hence we invoke equation (35) in Lemma 12 for $1/4$ ϵ -net on these k norm balls: $\|\mathbf{u}_i\|_2 \leq R$ and take an union bound, we have

$$\mathbb{P} \left(\sup_{\mathbf{U}: \|\mathbf{U}\|_2 \leq R} \left\| \frac{1}{n} \sum_i^n (\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U}) \right\|_2 \leq \sqrt{\frac{d \log d}{n}} \sqrt{k} R \right) \geq 1 - \exp(-C_2 \log d).$$

Similarly if we invoke the equation (36) for $1/4$ ϵ -net on these k norm balls: $\|\mathbf{u}_i\|_2 \leq R$ and take an union bound, we also have

$$\mathbb{P} \left(\sup_{\mathbf{U}: \|\mathbf{U}\|_2 \leq R} \left\| \frac{1}{n} \sum_i^n (\langle \mathbf{A}_i, \mathbf{U} \rangle \mathbf{A}_i - \mathbf{U}) \right\|_2 \leq \frac{1}{\kappa \log d \sqrt{k} C_2} R \right) \geq 1 - \exp(-n^{z_2}).$$

As a consequence, the conclusion of the lemma follows. \square

References

- S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45:77–120, 2017.
- S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582, 2016a.
- S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016b.
- S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- 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.
- S. Burer and R. D. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.
- E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.
- E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

- 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.
- Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7):4324–4337, 2013.
- Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.
- R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Singularity, misspecification, and the convergence rate of EM. *Annals of Statistics*, 48:3161–3182, 2020a.
- R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Sharp analysis of Expectation-Maximization for weakly identifiable models. In *AISTATS*, 2020b.
- R. Ge, J. D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.
- 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.
- M. Hardt. Understanding alternating minimization for matrix completion. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 651–660. IEEE, 2014.
- P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.
- A. Kalev, R. L. Kosut, and I. H. Deutsch. Quantum tomography protocols with positivity are compressed sensing protocols. *npj Quantum Information*, 1(1):1–6, 2015.
- V. Koltchinskii, K. Lounici, A. B. Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.
- J. Kwon, N. Ho, and C. Caramanis. On the minimax optimality of the EM algorithm for learning two-component mixed linear regression. *arXiv preprint arXiv:2006.02601*, 2020.
- Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.
- S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.
- S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.
- 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.
- J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

- S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.
- A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York, NY, 2000.
- R. Vershynin. *High Dimensional Probability. An Introduction with Applications in Data Science*. Cambridge University Press, 2018.
- M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.
- A. E. Waters, A. C. Sankaranarayanan, and R. Baraniuk. Sparcs: Recovering low-rank and sparse matrices from compressive measurements. In *Advances in neural information processing systems*, pages 1089–1097, 2011.
- J. Zhang and R. Zhang. How many samples is a good initial point worth in low-rank matrix recovery? *Advances in Neural Information Processing Systems*, 33, 2020.
- R. Y. Zhang, S. Sojoudi, and J. Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *Journal of Machine Learning Research*, 20:1–34, 2019.
- Q. Zheng and J. Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. *Advances in Neural Information Processing Systems*, 28:109–117, 2015.
- Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.