On the Computational and Statistical Complexity of Over-parameterized Matrix Sensing

Jiacheng Zhuo
Department of Computer Science
University of Texas
Austin, TX 78712, USA

Jeongyeol Kwon
Wisconsin Institute for Discovery
University of Wisconsin-Madison
Madison, WI 53705, USA

Nhat Ho
Department of Statistics and Data Sciences
University of Texas
Austin, TX 78712, USA

Constantine Caramanis
Department of Electrical and Computer Engineering
University of Texas
Austin, TX 78712, USA

Editor: Pradeep Ravikumar

Abstract

We consider solving the low-rank matrix sensing problem with the Factorized Gradient Descent (FGD) method when the specified rank is larger than the true rank. We refer to this as over-parameterized matrix sensing. If the ground truth signal $X^* \in \mathbb{R}^{d \times d}$ is of rank $r$, but we try to recover it using $FF^\top$ where $F \in \mathbb{R}^{d \times k}$ and $k > r$, the existing statistical analysis either no longer holds or produces a vacuous statistical error upper bound (infinity) due to the flat local curvature of the loss function around the global maxima. By decomposing the factorized matrix $F$ into separate column spaces to capture the impact of using $k > r$, we show that $\|F_tF_t - X^*\|_F^2$ converges sub-linearly to a statistical error of $\tilde{O}(kd\sigma^2/n)$ after $\tilde{O}(\frac{\sigma}{\sqrt{n}})$ iterations, where $F_t$ is the output of FGD after $t$ iterations, $\sigma^2$ is the variance of the observation noise, $\sigma_r$ is the $r$-th largest eigenvalue of $X^*$, and $n$ is the number of samples. With a precise characterization of the convergence behavior and the statistical error, our results, therefore, offer a comprehensive picture of the statistical and computational complexity if we solve the over-parameterized matrix sensing problem with FGD.

Keywords: Computational complexity; Statistical complexity; Over-parameterization; Matrix sensing; Matrix regression; Factorized gradient descent.

1. Introduction

We consider the low rank matrix sensing problem: we are given $n$ i.i.d. observations $\{A_i, y_i\}_{i=1}^n$ from the data generating model $y_i = \langle A_i, X^* \rangle + \epsilon_i$, where $A_i \in \mathbb{R}^{d \times d}$ is a
symmetric random sensing matrix, $X^* \in \mathbb{R}^{d \times d}$ is the target rank $r$ symmetric matrix we want to recover, and $\epsilon_i$ is a zero-mean sub-Gaussian noise with variance proxy $\sigma^2$. 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; Kalev et al. 2015). One common approach to recover a low-rank matrix $X \in \mathbb{R}^{d \times d}$ is to solve the following optimization problem:

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

(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) or iterative hard-thresholding (IHT) procedures (Jain et al. (2014a)). However for computational benefits in large-scale problems, i.e., when the dimension $d$ is very large, it is common to reformulate this as a non-convex problem by introducing $F \in \mathbb{R}^{d \times k}$ such that $X = FF^\top$ and solving the transformed problem (Bhojanapalli et al., 2016a; Chen and Wainwright, 2015; Jain et al., 2013; Hardt, 2014; Park et al., 2018):

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

(2)

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

$$F_{t+1} = F_t - \eta G^n_t, \quad \text{where} \quad G^n_t = \nabla \mathcal{L}(F_t) = \frac{1}{n} \sum_{i=1}^{n} \left( \left\langle A_i, F_tF^\top_t \right\rangle - y_i \right) A_i F_t,$$

(3)

where $\eta$ is the step size and $G^n_t$ 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). When the rank is over-specified (i.e., $k > r$), we refer to that setting as the over-parameterized matrix sensing problem. Over-parameterized matrix sensing appears naturally in the real world applications, since the true rank $r$ is often not known and the practitioners have to perform cross-validation, which involves running with $k > r$. Moreover, over-parameterized matrix sensing attracts extra attention lately since it is seen as a sandbox to study the over-parameterization effect in deep learning (Li et al., 2018).

The over-parameterized matrix sensing comes with many challenges. One of the key challenges is that the Hessian is degenerate around the global maxima, because of the over-specification of the rank. Many previous works in the known rank settings (Bhojanapalli et al., 2016a; Zheng and Lafferty, 2016; Tu et al., 2016) have analysis that critically depends on non-degeneracy of the Hessian and hence local strong convexity around the global maxima, and hence do not work in the degenerate Hessian setting we consider.
The analysis of Chen and Wainwright (2015) is also not applicable, because the statistical error upper bound that they obtain implicitly assumes knowledge of the rank, since it scales with the ratio of the first and the \( k \)-th eigenvalue of \( X^* \); when the problem is over-parameterized, this ratio 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).

However, this is not only a problem of analysis. As our simulations and theoretical results below demonstrate, the over-parameterized setting has fundamentally different behavior, with statistical errors that are larger than the setting where the rank is known and used in the algorithm. The computational convergence rate also drops from linear to sublinear. 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 (computational) convergence rate in recovering the solution to the optimization problem, \( \hat{X} \), and (2) what is the statistical error, \( \| \hat{X} - X^* \| ? \)\*

In this work we attempt to answer the above two questions by offering analysis about both the convergence rate and the statistical error of solving over-parameterized low-rank matrix sensing with the FGD method.

**Result overview.** We show that when the number of samples \( n \) is sufficiently large, \( \| F_t F_t^\top - X^* \|_F^2 \) converges to a final statistical error of \( \tilde{O}(kd\sigma^2/n) \) after \( \tilde{O}(\sigma_r^2\sqrt{d}) \) number of iterations, namely sub-linearly, where \( \sigma_r \) and \( \sigma \) are, respectively, the \( r \)-th largest eigenvalue of \( X^* \) 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. In that setting, FGD converges linearly to a radius of convergence \( \tilde{O}(rd\sigma^2/n) \) after \( \tilde{O}(\log(\sigma_r \sigma_1^{-1} \cdot n d)) \) iterates (Chen and Wainwright, 2015), where \( \sigma_1 \) is the largest eigenvalue of \( X^* \). Furthermore, the number of iterations \( \tilde{O}(\sigma_r^2\sqrt{d}) \) 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 \( \hat{X} \) based on \( \{ A_i, y_i \}_{i=1}^n \) observations, \( \| \hat{X} - X^* \|_F^2 \geq \frac{d \sigma^2}{n} \), where \( X^* \) is the ground truth rank \( r \) matrix that we want to recover, and \( \sigma \) is the standard deviation of the (sub)-Gaussian observation noise (see Section 1.4 for details). This convex relaxation approach is nearly
optimal in that it matches this lower bound, up to log factors. Although we can theoretically solve this convex problem in polynomial time, the computational cost is often prohibitively high for large scale problems, which motivates the researchers to study the FGD method (Bhojanapalli et al., 2016a; Park et al., 2018).

It is also worth mentioning that 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; Jain et al., 2014b).

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 applied to the case where \( k > r \). 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 when we know the ground truth rank \( (k = r) \). However it is non-trivial to establish their prerequisite lemma (Lemma 1) when \( k > r \). Even if the analysis still holds when \( k > r \), their results only imply that the statistical error is upper bounded by infinity, since the upper bound scales with the ratio of the first and the \( k \)-th eigenvalue of \( X^* \), and when \( k > r \) this ratio is infinity.

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). This is not in line with the common notion of statistical error where we care about the sub-optimality to the true parameter when iteration counter \( t \) goes to infinity. Such common notion of statistical error is, in contrast, the focus of our work. (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 in various optimization problems with degenerate landscapes in weakly separated
mixture of distributions (Dwivedi et al., 2020a; Kwon et al., 2021). 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 tighter statistical rate (Dwivedi et al., 2020a; Kwon et al., 2021). This technique has been used widely in the empirical process theory literature (van der Vaart and Wellner, 2000). We find that the localization argument can also be applied for a low-rank matrix sensing when we over-specify the rank.

Follow-up work. Since the initial appearance of this work, there has been several studies that offer more precise understanding of the computational and statistical performance of FGD and its variants under this setting. We refer the readers to the above-mentioned papers for the tighter results derived in terms of $k$ Stöger and Soltanolkotabi (2021); Zhang et al. (2021); Soltanolkotabi et al. (2023); Ma and Fattahi (2023); Xu et al. (2023), and the faster linear convergence to true parameters with preconditioning Zhang et al. (2021); Xu et al. (2023); Zhang et al. (2023).

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 $U$ and $V$ such that the union of their column spaces is $\mathbb{R}^d$. We set $D_S^*$ 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 $X^* = UD_S^*U^\top$. The upper triangle entries of the sensing matrices $A_i$ are sampled from standard Gaussian distribution, and we fill the lower triangle entries accordingly such that $A_i$ 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 $\{F_t\}_t$ 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

![Convergence rate comparison with and without over-parameterization.](image1)

(a)

![Zoom-in view of the convergence rate when over-parameterized](image2)

(b)
(i.e., \( k = r = 3 \)), the FGD method converges 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 \( \{ \mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_t \} \). We also use the short hand \( \{ \mathbf{A}_i \}_i \) to represent \( \{ \mathbf{A}_1, \mathbf{A}_2, ... \} \). We use \( \sigma_1 \) and \( \sigma_r \) 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 \( \kappa \) 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 \( \mu \) 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:

\[
\mathbf{A}_i = \frac{1}{2}(\mathbf{R}_i + \mathbf{R}_i^\top),
\]

where \( \mathbf{R}_i \in \mathbb{R}^{d \times d} \), and all elements of \( \mathbf{R}_i \) are independently sampled from an identical sub-Gaussian distribution with zero-mean and variance proxy 1.
2. Main Result

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

\[
X^* = [U \ V] \begin{bmatrix} D^*_S & 0 \\ 0 & D^*_T \end{bmatrix} [U \ V]^\top = UD^*_S U^\top + VD^*_T V^\top,
\]

where \( U \in \mathbb{R}^{d \times r}, V \in \mathbb{R}^{d \times (d-r)}, D^*_S \in \mathbb{R}^{r \times r}, D^*_T \in \mathbb{R}^{(d-r) \times (d-r)} \). Without loss of generality we assume that the both \( U \) and \( V \) are orthonormal and \( U^\top V = 0 \) (i.e., \( U \) and \( V \) together span the entire \( \mathbb{R}^d \)). Denote \( \sigma_1 \) be the largest value in \( D^*_S \), \( \sigma_r \) be the smallest value in \( D^*_S \), and \( \sigma_{r+1} \) be the largest value in \( D^*_T \). If \( X^* \) is exactly rank \( r \), then \( D^*_T \) is a zero matrix. For generality, we allow \( X^* \) to be only approximately rank \( r \), where we assume that \( D^*_T \) is nonzero but \( \sigma_{r+1} \) is very small. Since the union of the column space of \( U \) and \( V \) spans the entire \( \mathbb{R}^d \), then for any \( F_t \in \mathbb{R}^{d \times k} \), there exist matrices \( S_t \in \mathbb{R}^{r \times k} \) and \( T_t \in \mathbb{R}^{(d-r) \times k} \) such that

\[
F_t = US_t + VT_t.
\]

As \( t \) goes to infinity, we hope that \( S_t S_t^\top \) converges to \( D^*_S \), \( T_t T_t^\top \) converges to \( D^*_T \), and \( S_t T_t^\top \) converges to zero, and hence \( F_t F_t^\top = US_t S_t^\top U^\top + VT_t T_t^\top V^\top + US_t T_t^\top V^\top + VT_t S_t^\top U^\top \) converges to \( X^* \).

We introduce the decomposition and study the convergence of \( S_t S_t^\top, T_t T_t^\top, \) and \( S_t T_t^\top \) 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 4 and Lemma 7), it is the convergence of \( \{ \| T_t T_t^\top - D^*_T \|_2 \} \) 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)**

\[
\| F_0 F_0^\top - X^* \|_2 \leq \rho \sigma_r, \quad \text{for} \quad \rho \leq 0.07. \tag{4}
\]

Note that 0.07 is a universal constant and is chosen for the ease of presentation. Note that one can use the spectral 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 \( \max \{ \| D_T^* - T_0 T_0^\top \|_2, \| D_S^* - S_0 S_0^\top \|_2, \| S_0 T_0^\top \|_2 \} \) in our analysis. The following lemma establishes the connection between what we need in the analysis and Assumption 1.

**Lemma 3.** If \( \| F_0 F_0^\top - X^* \|_2 \leq 0.7 \rho \sigma_r \), then

\[
\max \left\{ \| D_T^* - T_0 T_0^\top \|_2, \| D_S^* - S_0 S_0^\top \|_2, \| S_0 T_0^\top \|_2 \right\} \leq \rho \sigma_r.
\]
The statistical error is almost tight compared to the counterpart, with a caveat. At a setting, going from population to finite-sample creates an unusual tangling factor, causing the guarantees we offer are in spectral norm, while the typical rate in the related work is consistent with our simulations in Figure 2. As we will see later in Lemma 7, it is the convergence of \( \| D_T^* - T_i^* \|_2 \) that slows down the convergence of \( \| F_i F_i^T - X^* \|_2 \), and incurring the sublinear convergence of \( \| S_i S_i^T - D_S^* \|_2 \) and \( \| S_i T_i^T \|_2 \).

(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 6, if we have access to the population gradient, the convergence rates of the sequences \( \| S_i S_i^T - D_S^* \|_2 \) and \( \| S_i T_i^T \|_2 \), 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) The statistical error is almost tight compared to the counterpart, with a caveat. A glance the statistical error seems too good to be true compared to the work in (Chen and Wainwright, 2015), 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.

We leave the proof of Lemma 3 to Appendix C.1. Now we are ready to present our main result.

**Theorem 4. (Main result)** Assume the following setting: (1) \( \| D_T \|_2 < \sqrt{\frac{d \log d}{n}} \sigma \); (2) we have good initialization as in Assumption 1; (3) the step size \( \eta = \frac{1}{1000} \); (4) \( A_i \)'s are sub-Gaussian sensing matrices. Let \( \{ F_i \}_i \) be the sequence generated by the FGD algorithm as in equation (3). Then, the following holds:

(a) With sample size \( n > C_1 k \sigma^2 d \log^3 d \cdot \max(1, \sigma^2/\sigma_f^2) \) for some universal constant \( C_1 \), after \( t > \left[ 2 \log \frac{\sigma_f}{\epsilon_{\text{comp}}} \right] \) steps, we have

\[
\max \left\{ \| S_i S_i^T - D_S^* \|_2, \| S_i T_i^T \|_2 \right\} < C \epsilon_{\text{comp}}
\]

for some universal constant \( C \), where \( \epsilon_{\text{comp}} := \sqrt{\frac{k \sigma^2 d \log d}{n}} \sigma_f + \kappa \sqrt{\frac{d \log d}{n}} \sigma_f \).

(b) With sample size \( n > C'_1 k \sigma^2 d \log^3 d \cdot \max(1, \sigma^2/\sigma_f^2) \) for some universal constant \( C'_1 \), after \( t \geq \Theta \left( \frac{\sigma_f}{\epsilon_{\text{stat}}} \right) \) steps, we find that

\[
\max \left\{ \| S_i S_i^T - D_S^* \|_2, \| S_i T_i^T \|_2, \| T_i T_i^T - D_T^* \|_2 \right\} < C_2' \epsilon_{\text{stat}},
\]

and \( \| F_i F_i^T - X^* \|_2 \leq C_3' \epsilon_{\text{stat}} \) for some universal constants \( C_2' \), and \( C_3' \), where \( \epsilon_{\text{stat}} := \kappa \sqrt{\frac{d \log d}{n}} \sigma_f \).

The proof of Theorem 4 is in Appendix B.2. We now have a few remarks with these results:

(1) The sequences \( \{ \| S_i S_i^T - D_S^* \|_2 \}_i \) and \( \{ \| S_i T_i^T \|_2 \}_i \) first converge linearly and then sub-linearly. Theorem 4 indicates that the sequences \( \{ \| S_i S_i^T - D_S^* \|_2 \}_i \) and \( \{ \| S_i T_i^T \|_2 \}_i \) first converge linearly from \( 0.1 \sigma_f \) to \( \epsilon_{\text{comp}} \), and then converge sub-linearly to \( \Omega (\epsilon_{\text{stat}}) \). Furthermore, the sequence \( \{ \| F_i F_i^T - X^* \|_2 \}_i \), always converges sub-linearly towards \( \Omega (\epsilon_{\text{stat}}) \). This is consistent with our simulations in Figure 2. As we will see later in Lemma 7, it is the convergence of \( \| D_T^* - T_i^* \|_2 \) that slows down the convergence of \( \| F_i F_i^T - X^* \|_2 \), and incurring the sublinear convergence of \( \{ \| S_i S_i^T - D_S^* \|_2 \}_i \) and \( \{ \| S_i T_i^T \|_2 \}_i \).

(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 6, if we have access to the population gradient, the convergence rates of the sequences \( \{ \| S_i S_i^T - D_S^* \|_2 \}_i \) and \( \{ \| S_i T_i^T \|_2 \}_i \), 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) The statistical error is almost tight compared to the counterpart, with a caveat. At a glance the statistical error seems too good to be true compared to the work in (Chen and Wainwright, 2015), 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 $\kappa \sqrt{\frac{k d \log d}{n}}$ if we evaluate $\|F_t F_t^\top - X^*\|_F$.

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. While one might read from the statistical error and claim that the sample complexity is $\Omega\left(\frac{k \kappa^2 d \log d \max(1, \sigma^2/\sigma_r^2)}{n} \right)$, a caveat is that, in order to achieve this statistical error, we do require $n > C_1 k^2 \kappa^2 d \log d \max(1, \sigma^2/\sigma_r^2)$ for some universal constant $C_1$. The extra factor of $k$ comes from the uniform concentration bound of random matrices in Lemma 18. We believe that this is an artifact of our proof technique, and this extra factor could potentially be sharpened, though it appears non-trivial, and would require some new developments for the uniform concentration of random matrices. Finally, we would like to remark that the FGD setup with over-specified rank should suffer from dependence on $k$, not the true rank $r$. This is because the weak-signal part suffers from a degenerate landscape, and FGD cannot distinguish whether the weak-signal part has indeed 0 or non-zero energy.

(4) The behaviors of FGD under the noiseless setting. When $\sigma = 0$, namely, the noiseless setting of the matrix sensing problem, with the similar proof argument as that of Theorem 4 we have the following behaviors of the FGD:

**Corollary 5.** (Noiseless Setting) Assume the noiseless setting of the matrix sensing problem with $\|D_t\|_2 = 0$ while the other assumptions on the initialization, the step size, and the sensing matrices are as in Theorem 4. Let $\{F_t\}_t$ be the sequence generated by the FGD algorithm as in equation (3). Then, the following holds:

(a) With $n > C_1 k^2 \kappa^2 d \log d$ for some universal constant $C_1$, after $t > \left\lceil 2 \log \frac{\sigma_r}{\epsilon_{\text{comp\_noiseless}}} \right\rceil$ steps, we have

$$\max \left\{ \|S_t S_t^\top - D_S\|_2, \|S_t T_t^\top\|_2 \right\} < C \epsilon_{\text{comp\_noiseless}}$$

for some universal constant $C$, where $\epsilon_{\text{comp\_noiseless}} = \sqrt{\frac{k \kappa^2 d \log d}{n}} \sigma_r$.

(b) For any $\epsilon > 0$, with $n > C_1' k^2 \kappa^2 d (\log^3 d)$ for some universal constant $C_1'$, after $t \geq \Theta\left( \frac{n}{\epsilon^2} \right)$ steps, we find that

$$\max \left\{ \|S_t S_t^\top - D_S^*\|_2, \|S_t T_t^\top\|_2, \|T_t T_t^\top - D_T^*\|_2 \right\} < C_2' \epsilon.$$

and $\|F_t F_t^\top - X^*\|_2 \leq C_3' \epsilon$ for some universal constants $C_2'$, and $C_3'$.

The results of Corollary 5 indicate that in the noiseless setting of the matrix sensing problem, the sequences $\{\|S_t S_t^\top - D_S\|_2\}_t$ and $\{\|S_t T_t^\top\|_2\}_t$ first converge linearly and then sub-linearly, which share similar behaviors to those in the general noise settings in Theorem 4. The slow convergence is due to the rank over-specification, which significantly flattens the landscape around the global optimum. This is the main distinction between our work and existing literature on rank-specified noiseless cases.
Computational benefits of the FGD over the Iterative Hard Thresholding (IHT) methods.

Now, we would like to compare the FGD to the IHT methods (Jain et al., 2014a), which are useful for solving the over-parameterized matrix sensing problems. The main benefit of the FGD over the IHT methods is its total computational complexity for reaching the final statistical radius around the true matrix $X^\star$. In particular, for the IHT methods, the per iteration cost is $O(nd^2)$ where $n$ is the sample size and $d$ is the dimension while for the FGD, the per iteration cost is $O(ndk)$ where $k$ is the chosen rank. Therefore, based on Theorem 2 of (Jain et al., 2014a), with the computational complexity $O(kd^2)$ from the spectral method for the initialization, the total computational complexity for the IHT methods to reach the statistical radius $O(kd\sigma^2/n)$ around the true matrix $X^\star$ is at the order of $O(kd^2 + nd^2 \log(n/(kd\sigma^2)))$. On the other hand, from Theorem 4, the total computational complexity of the FGD to reach that final statistical radius is at the order of $O(kd^2 + \sigma r n^{3/2}\sqrt{dk})$. As a consequence, as long as $n \ll d^3$ the FGD algorithm has better computational complexity than that of the IHT methods in terms of the dimension $d$, which is crucial in the high dimensional settings of the matrix sensing problems.

2.1 Simulation verification of the main result

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

The simulation results are shown in Figure 2. In Figure 2a, we plot $\|FF^T - X^\star\|_2$, $\|SS^T - D_S\|_2$, $\|ST^T\|_2$, and $\|TT^T - D_T^\star\|_2$ against the algorithm iterations. The simulation results are aligned with our theory. As said in Theorem 4, $\|SS^T - D_S\|_2$ and $\|ST^T\|_2$ first converge linearly, and then sublinearly. Furthermore, $\|FF^T - X^\star\|_2$ is soon dominated by $\|TT^T - D_T^\star\|_2$, 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, 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
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.

\[ \|F_t F_t^T - X^*\|_F \] converges. 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 \(\Theta\left(\sqrt{k}\right)\) factor of sub-optimality. That being said, we believe that it is possible to establish similar results for \(\|F_t F_t^T - X^*\|_F\) 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 \(E[\langle A_i, B \rangle A_i] = B\) for any matrix \(B\) with appropriate dimensions. Combined with the fact that \(y_i = \langle A_i, X^* \rangle + \epsilon\), the population gradient (taking
expectation over the observation noise $\epsilon$ and the observation matrices $A_i$) is

$$G_t := \mathbb{E}[G^n_t] = \left( F_t F_t^\top - X^* \right) F_t.$$  

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 $S$ and $T$. Simple algebra using the last observation yields:

$$F_t - \eta G_t = F_t - \eta \left( F_t F_t^\top F_t - X^* F_t \right) = US_t + VT_t - \eta \left[ (US_t + VT_t) \left( S_t^\top S_t + T_t^\top T_t \right) - (UD_S S_t + VD_T T_t) \right] = UM_S (S_t) + VM_T (T_t)$$

where we define the following operators:

$$M_S(S) = S - \eta \left( SS^\top S + ST^\top T - D_S S \right),$$

$$M_T(T) = T - \eta \left( TT^\top T + TS^\top S - D_T T \right).$$

**Lemma 6. (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) $\|D_S^* - M_S(S) M_S(S)^\top\|_2 \leq (1 - \eta \sigma_r) \|D_S^* - SS^\top\|_2 + 3\eta \|ST^\top\|_2^2,$

(b) $\|M_S(S) M_T(T)^\top\|_2 \leq \|ST^\top\|_2 (1 - \eta \sigma_r),$  

(c) $\|M_T(T) M_T(T)^\top\|_2 \leq \|TT^\top\|_2 (1 - \eta \|TT^\top\|_2 + 2\eta \|D_T\|_2),$  

(d) $\|M_T(T) M_T(T)^\top - D_T^*\|_2 \leq \|TT^\top - D_T\|_2 \|I - 2\eta TT^\top\|_2 + 3\eta \|ST^\top\|_2^2.$

The proof of Lemma 6 can be found in Appendix A.1

According to Lemma 6 above, we have fast convergence in estimating $SS^\top$, $ST^\top$, but slow convergence in estimating $TT^\top$. Intuitively, $TT^\top$ 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 $D_T^*$ 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. We consider the deviation of the
We define \( \Delta \) population and sample gradient as follows:

\[
G^n_t - G_t = \frac{1}{n} \sum_{i=1}^{n} \left( \langle A_i, F_t F_t^\top \rangle - y_i \right) A_i F_t - \left( F_t F_t^\top - X^* \right) F_t
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left( \langle A_i, F_t F_t^\top - X^* \rangle + \epsilon_i \right) A_i F_t - \left( F_t F_t^\top - X^* \right) F_t.
\]

We define \( \Delta_t \) to quantify this deviation:

\[
\Delta_t = \frac{1}{n} \sum_{i=1}^{n} \left( \langle A_i, F_t F_t^\top - X^* \rangle + \epsilon_i \right) A_i - \left( F_t F_t^\top - X^* \right) \text{F}_t.
\]

and hence \( G^n_t - G_t = \Delta_t F_t \). If we can control \( \Delta_t \), we can have contraction per iteration, as shown in the lemma below. Note that we make no attempts to optimize the constants.

**Lemma 7. (Contraction per iteration.)** Assume that we have the same setting as Theorem 4. Denote \( D_t = \max \{ \| S_t S_t^\top - D_t^* \|_2, \| T_t T_t^\top \|_2, \| S_t T_t^\top \|_2 \} \), and assume that \( D_t \) is still sub-optimal to the statistical error: \( D_t > 50 \kappa \sqrt{\frac{d \log d}{n}} \). Suppose

\[
\| \Delta_t \|_2 \leq 5 \sqrt{\frac{kd \log d}{n}} D_t + \sqrt{\frac{d \log d}{n}}. \quad (5)
\]

Then, we find that

\[
\| S_{t+1} S_{t+1}^\top - D_{t+1}^* \|_2 \leq \left( 1 - \frac{7}{10} \eta \sigma_r \right) \| S_t S_t^\top - D_t^* \|_2 + \sqrt{\frac{kd \log d}{n}} D_t + \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma. \quad (6)
\]

\[
\leq \left( 1 - \frac{7}{10} \eta \sigma_r \right) \| S_t S_t^\top - D_t^* \|_2 + \frac{1}{10} \eta \sigma_r D_t + \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma, \quad (7)
\]

\[
\| S_{t+1} T_{t+1}^\top \|_2 \leq (1 - \eta \sigma_r) \| S_t T_t^\top \|_2 + \sqrt{\frac{kd \log d}{n}} D_t + \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma, \quad (8)
\]

\[
\leq (1 - \eta \sigma_r) \| S_t T_t^\top \|_2 + \frac{1}{10} \eta \sigma_r D_t + \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma. \quad (9)
\]

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

\[
(D_{t+1} - 50 \epsilon_{\text{stat}}) \leq \left[ 1 - \frac{1}{2} \eta (D_t - 50 \epsilon_{\text{stat}}) \right] (D_t - 50 \epsilon_{\text{stat}}). \quad (10)
\]

The proof of Lemma 7 can be found in Appendix B.1.

**Implication of equations** (7) and (9): Firstly, when \( n \) goes to infinity, the sequence of \( \{ \| S_t S_t^\top - D_t^* \|_2 \}_t \) has constant contraction at each step, and hence achieves a linear convergence after all. This matches our population results in Lemma 6. Secondly, if \( n \) is finite, the sequence of \( \{ \| S_t S_t^\top - D_t^* \|_2 \}_t \) still has constant contraction, until roughly the magnitude of \( \| S_t S_t^\top - D_t^* \|_2 \) 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 4.
3.3 Proof sketch for the main theorem

In this subsection we offer a proof sketch for Theorem 4. Detailed proof can be found in Appendix B.2.

Lemma 7 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 7 can be invoked for one iteration. (2) Secondly we have to find a way to correctly invoke Lemma 7 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 \( \epsilon \)-net discretization techniques.

**Lemma 8.** Let \( A_i \) be sub-Gaussian sensing matrices. Let \( \epsilon_i \) follows \( N(0, \sigma) \). Then
\[
\mathbb{P} \left( \left\| \frac{1}{n} \sum_i^n A_i \epsilon_i \right\|_2 \geq C \sqrt{\frac{d \sigma^2}{n}} \right) \leq \exp(-C).
\]

**Lemma 9.** Let \( A_i \) be sub-Gaussian sensing matrices. Let \( U \) 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
\[
\mathbb{P} \left( \left\| \frac{1}{n} \sum_i^n (\langle A_i, U \rangle A_i - U) \right\|_2 \leq \sqrt{\frac{d \log d}{n}} \| U \|_F \right) \geq 1 - \exp(-C_2 \log d).
\]

The proof of the above two concentration results can be found in the Appendix D. If we invoke these lemmas for \( \Delta_t \) to ensure that equation (5) holds, then we can immediately have
\[
\| \Delta_t \|_2 = \left\| \frac{1}{n} \sum_i^n \left( \langle A_i, F_t F_t^\top - X^* \rangle A_i - (F_t F_t^\top - X^*) + \frac{1}{n} \sum_i^n \epsilon_i A_i \right) \right\|_2
\leq 5 \sqrt{\frac{kd \log d}{n}} D_t + \sqrt{\frac{d \log d}{n}} \sigma
\leq 0.5 \eta \sigma, D_t + \sqrt{\frac{d \log d}{n}} \sigma
\]
where we use Lemma 9 with \( U = F_t F_t^\top - X^* \) and Assumption 1 with converting it to Frobenius bound (note that \( \| F_t F_t^\top - X^* \|_2 \leq D_t \)), and the last inequality holds because we set \( n > C_1 \kappa^2 d \log^3 d \cdot \max(1, \sigma^2/\sigma^2_r) \).

In order to resolve the second challenge mentioned above, we need a uniform concentration result.

**Lemma 10.** Let \( A_i \) be sub-Gaussian sensing matrices. If \( U \) is of rank \( k \) and is in a bounded spectral norm ball of radius \( R \) (i.e., \( \| U \|_2 \leq R \)),
\[
\mathbb{P} \left( \sup_{\| U \|_2 \leq R} \left\| \frac{1}{n} \sum_i^n (\langle A_i, U \rangle A_i - U) \right\|_2 \leq \sqrt{\frac{d \log d}{n}} k R \right) \geq 1 - \exp(-C_2 \log d).
\]
With \( n > C_1 k^2 \kappa^2 d \log^3 d \cdot \max(1, \sigma^2 / \sigma^2) \), we can apply Lemma 10 with different values of \( R \). For instance, starting from \( R_0 = \sigma \), we can apply Lemma 10 with \( R_0 / 2, R_0 / 4, \ldots \), until reaching \( c_{\text{comp}} \). Then, we can ensure equation (5) by choosing proper \( r \) such that \( r < D_t \leq 2r \) throughout all iterations. In fact, the use of Lemma 10 with multiple levels of \( R \) is the key to obtain the correct statistical rate after linear convergence phase as we describe below (The mathematical details of such argument are in the proof of part (b) of Theorem 1 in Appendix B.2).

Note that, compared to the standard concentration result as in Lemma 9, we have an extra factor of \( k \) in the uniform concentration lemma above. Since the localization analysis for sub-linear convergence requires uniform concentration of random matrices, we naturally introduce an extra factor of \( k \), and hence resulting in the extra factor in the requirement of \( n \) as in the part (b) in Theorem 4.

**The linear convergence part.** Claim (a) in the main theorem is about linear convergence. We mention in the remark that equations (7) and (9) imply constant contractions for \( \| S_t S_t^\top - D_S^* \|_2 \) and \( \| S_t T_t^\top \|_2 \) respectively. To make the argument more precise, we consider\[
\| S_t S_t^\top - D_S^* \|_2 > 1000 \sqrt{\frac{k \kappa^2 d \log d}{n}} \sigma_r > 1000 \sqrt{\frac{k \kappa^2 d \log d}{n}} D_t.
\]
Then, we find that \( \sqrt{\frac{k \kappa d \log d}{n}} D_t \leq 0.1 \eta \sigma_r \| S_t S_t^\top - D_S^* \|_2 \) since \( \eta \sigma_r = 0.01 / \kappa \). Also, \( \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma < 0.1 \eta \sigma_r \| S_t S_t^\top - D_S^* \|_2 \) by the choice of the constants in the lower bound of \( n \). Hence, when \( \| S_t S_t^\top - D_S^* \|_2 > 1000 \sqrt{\frac{k \kappa^2 d \log d}{n}} \sigma_r \), we find that
\[
\| S_{t+1} S_{t+1}^\top - D_S^* \|_2 \leq \left( 1 - \frac{7}{10} \eta \sigma_r \right) \| S_t S_t^\top - D_S^* \|_2 + \sqrt{\frac{k \kappa d \log d}{n}} D_t + \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma
\leq \left( 1 - \frac{5}{10} \eta \sigma_r \right) \| S_t S_t^\top - D_S^* \|_2.
\]
The same arguments hold for \( \| S_t T_t^\top \|_2 \). 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 built upon equation (10).

Before we discuss how equation (5) holds in this sub-linear convergence case for all iteration \( t \), we briefly illustrate how equation (10) implies convergence to \( \Theta(\epsilon_{\text{stat}}) \) after \( O(1 / \epsilon_{\text{stat}}) \) iterations. By equation (10), we know that \( A_{t+1} \leq \left( 1 - \frac{1}{2} \eta A_t \right) A_t \) where \( A_t = D_t - 50 \epsilon_{\text{stat}} \). Hence, we obtain that
\[
A_{t+1} \leq \left( 1 - \frac{1}{2} \eta A_t \right) A_t \stackrel{(1)}{=} \left( 1 - \frac{2}{t + \frac{4}{\eta A_0}} \right) \frac{4}{\eta t + \frac{4}{A_0}} = \frac{t + \frac{4}{\eta A_0}}{t + \frac{4}{\eta A_0}} - 2 \frac{t + \frac{4}{\eta A_0}}{\eta \left( t + \frac{4}{\eta A_0} \right)} \leq \frac{4}{\eta \left( t + 1 + \frac{4}{\eta A_0} \right)}.
\]
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 \(\frac{(t+\frac{1}{\eta A_0}) - 2}{(t+\frac{1}{\eta A_0})^2} \leq \frac{1}{(t+\frac{1}{\eta A_0})^1}\). Therefore, after \(t \geq \Theta\left(\frac{1}{\eta_{\text{stat}}}\right)\) number of iterations, \(A_t = D_t - 50\kappa \sqrt{\frac{d \log d}{n}} \leq \Theta(\epsilon_{\text{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., 2021; 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 \(n^{-1/2}\). We leave the details of this argument to Appendix B.2.

4. Conclusions, discussions, and future works

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 \(\|F_t F^\top_t - X^*\|_F^2\) converges to a radius of \(\tilde{O}(kd/n)\) after \(\tilde{O}(\sqrt{d})\) number of iterations where \(F_t\) is the output of FGD after \(t\) iterations. We now discuss a few natural questions with this work.

**Weaker initialization condition.** In this work we focus on local convergence (see Assumption 1). This is a common practice in the related works (Chen and Wainwright, 2015; Bhojanapalli et al., 2016a; Zheng and Lafferty, 2015), and this initialization can be achieved by the standard spectral method. However, it is natural to ask what would happen if a weaker initialization is performed. When the rank is known, and the number of samples is sufficiently large, global convergence is achievable since factorized matrix sensing problem has no spurious local minima (Zhang et al., 2019; Ge et al., 2017, 2016). However, when the rank is unknown and over-specified, we believe that guaranteeing global convergence is still an open problem and is an interesting future direction. Besides global convergence, another weaker initialization condition is initialization around the origin, namely (randomly) generate a matrix \(F_0\) such that \(\|F_0 F_0^\top\|_2 \leq \delta\) for some small constant \(\delta\). While we show that such initialization works in our simulation (see Figures 2c and 2d), we find that we need at least a constant lower bound on the smallest eigenvalue of \(S_0 S_0^\top\) for the theoretical analysis. Such a lower bound is automatically satisfied when \(k = d\) and we initialize \(F_0\) as \(\delta I\) where \(I\) is the identity matrix, as in Li et al. (2018). However it is fundamentally challenging when \(k < d\) and we cannot guarantee a constant lower bound on the smallest eigenvalue of \(S_0 S_0^\top\).

**Extension to asymmetric matrices.** In this work we focus on the true symmetric matrix \(X^*\) and symmetric sensing matrices, because this captures the essential difficulties of the over-parameterized matrix sensing problem already. For the matrix sensing problem with asymmetric \(X^*\) and asymmetric sensing matrices, one can convert such a problem into a symmetric one without changing the asymptotically statistical and computational complexity. We refer the readers to the Section 5 of Ge et al. (2017) for more details.

**Extensions to other rank-constrained convex optimization problems.** Suppose we want to find a rank \(r\) PSD matrix \(X^* \in \mathbb{R}^{d \times d}\) by minimizing a convex function \(f(X) = \)
\( f(FF^\top) \) where \( F \in \mathbb{R}^{dk} \) and \( k > r \). If the population gradient with respect to \( X \) is \( X - X^* \), and the sample gradients have good concentration around the population gradient (examples can be found in Chen and Wainwright (2015)), our analysis techniques can be directly applied. If the population gradient with respect to \( X \) is explicit and we can still decompose the update process of \( \{F_t\}_t \) into the strong signal part (namely \( \{S_t\}_t \)) and the weak signal part (namely \( \{T_t\}_t \)), similar techniques might also be applicable. However, non-trivial efforts are required to extend our current analysis to general convex functions.

**Extension to projected factorized gradient descent.** So far in our formulation we do not assume structure in the factorized matrix \( \{F_t\}_t \). However for some low-rank problems it is desirable to have constraints on \( \{F_t\}_t \). For example for matrix completion we wish all the iterations \( \{F_t\}_t \) to stay incoherent, and this can be achieved by performing a projection step after each gradient step (Chen and Wainwright, 2015). As long as the projection step is non-expansive (i.e., \( \|\Pi(F)\Pi(F)^\top - X^*\|_2 \leq \|FF^\top - X^*\|_2 \) where \( \Pi \) is the projection operation), our analysis is still applicable. Unfortunately projection required to maintain incoherence is expansive in our analysis (note that it is non-expansive in for example Chen and Wainwright (2015) when we analyze the convergence of Frobenius norm and the corresponding non-expansiveness is defined as \( \|\Pi(F)\Pi(F)^\top - X^*\|_F \leq \|FF^\top - X^*\|_F \)). To get around this issue, one option is to show that all \( \{F_t\}_t \) stay incoherent automatically as in Ma et al. (2018). We leave this direction for future development.

**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 \( \tilde{O}(dr) \) samples, where \( r \) is the rank of the ground truth matrix \( X^* \), while our analysis requires \( \tilde{O}(dk) = \tilde{O}(d^2) \) samples when \( k = d \). Since they only require \( \tilde{O}(dr) \) samples, they cannot control the error of the over-parameterization part (equivalent to our \( TT^\top \) part). In fact in their analysis, they only show that in a limited number of steps, this error does not blow up. While with \( \tilde{O}(dk) \) samples we can show that the over-parameterization part also converges, although with a slower convergence rate. Therefore, their results cannot imply ours.

**Acknowledgements**

We would like to thank Raaz Dwivedi, Koulik Khamaru, and Martin Wainwright for helpful discussion with this work. This work was partially supported by the NSF IFML 2019844 award and research gifts by UT Austin ML grant to NH.

**Appendix 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 6**

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 6: We would like to prove the following inequality:

$$\left\| D_S - M_S(S)M_S(S)^T \right\|_2 \leq (1 - \eta \sigma_r) \left\| D_S^* - SS^T \right\|_2 + 3\eta \left\| ST^T \right\|_2^2.$$ 

Indeed, from the formulation of $M_S(S)$, we have

$$D_S - M_S(S)M_S(S)^T = D_S^* - (S - \eta ST^T S + \eta (D_S^* - SS^T) S) (S - \eta ST^T S + \eta (D_S^* - SS^T) S)^T.$$ 

We can group the terms in the RHS of the above equation according to whether they contain $D_S^* - SS^T$ or not, namely, we find that

$$D_S - M_S(S)M_S(S)^T = I + II$$

where,

$$I = (D_S^* - SS^T) - \eta (D_S^* - SS^T) SS^T - \eta SS^T (D_S^* - SS^T)$$

$$- \eta^2 (D_S^* - SS^T) SS^T (D_S^* - SS^T)$$

$$+ \eta^2 (D_S^* - SS^T) ST^T TS^T + \eta^2 ST^T TS^T (D_S^* - SS^T),$$

$$II = 2\eta ST^T TS^T - \eta^2 ST^T TT^T TS^T.$$ 

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

$$\left\| I \right\|_2 \leq \left\| D_S^* - SS^T \right\|_2 \left\| I - 2\eta SS^T - \eta^2 SS^T (D_S^* - SS^T) + 2\eta^2 ST^T TS^T \right\|_2.$$ 

From the choice of the step size and the initialization condition, the term $I - 2\eta SS^T - \eta^2 SS^T (D_S^* - SS^T) + 2\eta^2 ST^T TS^T$ is PSD matrix. Furthermore, for any $\|z\| = 1$, we have

$$z^T \left( I - 2\eta SS^T - \eta^2 SS^T (D_S^* - SS^T) + 2\eta^2 ST^T TS^T \right) z$$

$$\leq 1 - 2\eta \|Sz\|_2^2 + \eta^2 \left\| SS^T \right\|_2 \left\| D_S^* - SS^T \right\|_2 + 2\eta^2 \left\| TT^T \right\|_2 \|Sz\|_2^2$$

$$\leq 1 - 2\eta \sigma_r + 3\eta^2 \sigma_r \sigma_1$$

$$\leq 1 - \eta \sigma_r,$$

where in step (i) we used $0.9 \sigma_r I \preceq SS^T \preceq (\sigma_1 + 0.1 \sigma_r) I$, $\left\| D_S^* - SS^T \right\|_2 \leq 0.1 \sigma_r$ and $\left\| TT^T \right\|_2 \leq 1.1 \sigma_r$ 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:

$$\left\| I - 2\eta SS^T - \eta^2 SS^T (D_S^* - SS^T) + 2\eta^2 ST^T TS^T \right\|_2 \leq 1 - \eta \sigma_r.$$ (11)
To deal with the $II$ term, we have to establish the connection between $D_S^* - SS^T$ and $ST^T$. Note that, $\|\eta^2 ST^T TT^T TS^T\|_2 \leq \eta^2 \|TT^T\|_2 \|ST^T TS^T\|_2 \leq \eta \|ST^T TS^T\|_2$ since $\eta \leq 1/\sigma_r$. Hence, we have

$$\|II\|_2 \leq \|2\eta ST^T TS^T\|_2 + \|\eta^2 ST^T TT^T TS^T\|_2 \leq 3\eta \|ST^T TS^T\|_2. \quad (12)$$

Collecting the results from equations (11) and (12), we obtain

$$\|D_S^* - M_S(S)M_T(T)^T\|_2 \leq (1 - \eta\sigma_r) \|D_S^* - SS^T\|_2 + 3\eta \|ST^T\|_2^2.$$

Therefore, we reach the conclusion with claim (a) in Lemma 6.

**Proof of the contraction result (b) in Lemma 6:** Recall that we want to demonstrate that

$$\|M_S(S)M_T(T)^T\|_2 \leq \|ST^T\|_2 (1 - \eta\sigma_r).$$

Firstly, from the formulations of $M_S(S)$ and $M_T(T)$, we have the following equations:

$$M_S(S)M_T(T)^T = \left(S + \eta \left(D_S^* - SS^T\right) \right) \left(T + \eta \left(D_T^* - TT^T\right) \right) \left(T - \eta TS^T S\right)^\top$$

$$= \frac{1}{2} \left(I - 2\eta SS^T + 2\eta \left(D_S^* - SS^T\right) - 2\eta^2 \left(D_S^* - SS^T\right) SS^T + 2\eta^2 ST^T TS^T\right) ST^T$$

$$+ \frac{1}{2} ST^T \left(I + 2\eta \left(D_T^* - TT^T\right) - 2\eta TT^* - 2\eta^2 TT^* \left(D_T^* - TT^T\right)\right)$$

$$+ \eta^2 \left(D_S^* - SS^T\right) ST^T \left(D_T^* - TT^T\right) . \quad (13)$$

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

For the term in the first line of the RHS of equation (13) we have

$$\frac{1}{2} \left(I - 2\eta SS^T + 2\eta \left(D_S^* - SS^T\right) - 2\eta^2 \left(D_S^* - SS^T\right) SS^T + 2\eta^2 ST^T TS^T\right) ST^T$$

$$\leq \frac{1}{2} \|ST^T\|_2 \left(\|1 - 2\eta SS^T\|_2 + 2\eta \|D_S^* - SS^T\|_2 + 2\eta^2 \|D_S^* - SS^T\|_2 \|SS^T\|_2 + 2\eta^2 \|ST^T TS^T\|_2\right)$$

$$\leq \frac{1}{2} \|ST^T\|_2 (1 - 1.8\eta\sigma_r + 0.2\eta\sigma_r + 0.0022\eta^2\sigma_r + 0.02\eta^2\sigma_r^2)$$

$$\leq \frac{1}{2} \|ST^T\|_2 (1 - 1.5\eta\sigma_r).$$

For the term in the second line of the RHS of equation (13), direct calculation yields that

$$\frac{1}{2} ST^T \left(I + 2\eta \left(D_T^* - TT^T\right) - 2\eta TT^* - 2\eta^2 TT^* \left(D_T^* - TT^T\right)\right)$$

$$\leq \frac{1}{2} \|ST^T\|_2 \left(\|1 - 2\eta TT^T\|_2 + 0.2\eta\sigma_r + 2.2\eta^2\sigma_r^2\right)$$

$$\leq \frac{1}{2} \|ST^T\|_2 (1 + 0.3\eta\sigma_r).$$
Lastly, for the second order term in the third line of the RHS of equation (13) we have
\[ \left\| \eta^2 \left( D_S^* - SS^T \right) ST^T \left( D_T^* - TT^T \right) \right\|_2 \leq \eta^2 r^2 \sigma_r^2 \left\| ST^T \right\|_2 = \frac{1}{1000} \eta \sigma_r \left\| ST^T \right\|_2. \]

Plugging the above results into equation (13) leads to
\[ \left\| \mathcal{M}_S(S) \mathcal{M}_T(T)^\top \right\|_2 \leq \left\| ST^T \right\|_2 (1 - \eta \sigma_r). \]

Hence, we obtain the conclusion of claim (b) in Lemma 6.

Proof of the contraction result (c) in Lemma 6: We would like to establish that
\[ \left\| \mathcal{M}_T(T) \mathcal{M}_T(T)^\top \right\|_2 \leq \left\| TT^T \right\|_2 \left( 1 - \eta \right) \left\| TT^T \right\|_2 + 2 \eta \| D^*_T \|^2 \right). \]

To check the convergence in low SNR, i.e., with small singular values, we assume that \( \| D_T^* \| \ll \sigma_r \). It suggests that the focus is how fast \( TT^T \) converges to 0 when \( \| TT^T \| \gg \| D_T^* \| \). Indeed, simple algebra indicates that
\[
\mathcal{M}_T(T) \mathcal{M}_T(T)^\top
= \left( T + \eta \left( D_T^* - TT^T \right) T - \eta TS^T S \right) \left( T + \eta \left( D_T^* - TT^T \right) T - \eta TS^T S \right)^\top
= TT^T + \eta TT^T \left( D_T^* - TT^T \right) - \eta TS^T ST^T
+ \eta \left( D_T^* - TT^T \right) TT^T \left( D_T^* - TT^T \right) - \eta^2 \left( D_T^* - TT^T \right) TS^T ST^T
- \eta TS^T ST^T - \eta^2 TS^T ST^T \left( D_T^* - TT^T \right)^\top + \eta^2 TS^T SS^T ST^T
= III + IV + V,
\]
where we use the following shorthand notation:
\[
III = \left( TT^T - 2 \eta \left( TT^T \right)^2 + \eta^2 \left( TT^T \right)^3 \right),
\]
\[
IV = \eta \left( D_T^* TT^T + TT^T D_T^* \right) - \left( \eta^2 D_T^* \left( TT^T \right)^2 + \eta^2 \left( TT^T \right)^2 D_T^* \right) + \eta^2 D_T^* \left( TT^T \right) D_T^*,
\]
\[
V = -2 \eta TS^T ST^T - \eta^2 TS^T ST^T \left( D_T^* - TT^T \right)^\top + \eta^2 \left( D_T^* - TT^T \right) TS^T ST^T
+ \eta^2 TS^T SS^T ST^T.
\]

We first bound the \( IV \) term. Inequalities with operator norm show that
\[
\left\| D_T^* TT^T \right\|_2 \leq \| D_T^* \|_2 \left\| TT^T \right\|_2,
\]
\[
\left\| D_T^* \left( TT^T \right)^2 \right\|_2 \leq \sigma_r \| D_T^* \|_2 \left\| TT^T \right\|_2,
\]
\[
\left\| D_T^* \left( TT^T \right) D_T^* \right\|_2 \leq \sigma_r \| D_T^* \|_2 \left\| TT^T \right\|_2.
\]
Given these bounds, we find that

\[ \| IV \|_2 \leq (\eta + 3\eta^2 \sigma_r) \| D^*_T \|_2 \| TT^\top \|_2. \] (14)

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

\[
- 2\eta TS^\top ST^\top - \eta^2 TS^\top ST^\top (D^*_T - TT^\top)^\top
\leq (\eta + 2\eta^2 \rho \sigma_r + \eta^2 \sigma_1) T \leq 0.
\] (15)

Since \( M_T(T)M_T(T)^\top \) is PSD, we can just relax this term to zero. Finally, we bound the \( III \) term. Observe that,

\[
\| TT^\top - 2\eta (TT^\top)^2 + \eta^2 (TT^\top)^3 \| \leq TT^\top - \eta \left( TT^\top \right)^2,
\]

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

\[
\| TT^\top - \eta \left( TT^\top \right)^2 \| = \max_{\| z \| = 1} \left( z^\top TT^\top z - \eta z^\top \left( TT^\top \right)^2 z \right)
\]

\[
= \max_{\| z \| = 1} \left( z^\top QDQ^\top z - \eta z^\top QD^2 Q^\top z \right)
\]

\[
= \max_{\| z' \| = 1} \left( z'^\top Dz' - \eta z'^\top D^2 z' \right)
\]

\[
= \max_{\| z' \| = 1} \sum_i (d_i - \eta d_i^2) z_i'^2.
\]

Since \( d_i < \sigma_r \) and \( 1/2\eta \gg \sigma_1 \), the above maximum is obtained at the largest singular value of \( TT^\top \). That is, we have

\[
\| TT^\top - \eta \left( TT^\top \right)^2 \| \leq \| TT^\top \|_2 \left( 1 - \eta \| TT^\top \|_2 \right). \] (16)

Now combining every pieces from equations (14), (15), and (16), we arrive at the following inequality:

\[
\| M_T(T)M_T(T)^\top \|_2 \leq \| TT^\top \|_2 \left( 1 - \eta \| TT^\top \|_2 + 2\eta \| D^*_T \|_2 \right).
\]

As long as \( \| D^*_T \|_2 \ll \| TT^\top \|_2 \), the contraction rate is roughly \( 1 - \eta \| TT^\top \|_2 \). Therefore, we obtain the conclusion of claim (c) in Lemma 6.

**Proof of the contraction result (d) in Lemma 6:** Direct calculation shows that

\[
M_T(T)M_T(T)^\top - D^*_T
= \left( T - \eta \left( TS^\top S + (TT^\top - D^*_T) T \right) \right)^\top - D^*_T
= VI + VII,
\]

21
where we denote VI and VII as follows:

\[ VI = (T T^\top - D_T^\top) - \eta ((T T^\top - D_T^\top) T T^\top + T T^\top (T T^\top - D_T^\top)) \]
\[ + \eta^2 (T T^\top - D_T^\top) T T^\top (T T^\top - D_T^\top), \]
\[ VII = \eta^2 (T T^\top - D_T^\top) S S^\top S^\top T^\top + \eta^2 T S^\top S^\top T^\top (T T^\top - D_T^\top) \]
\[ - 2\eta T S^\top S^\top T^\top + \eta^2 T S^\top S S^\top S^\top T^\top. \]

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

\[ \eta^2 \| (T T^\top - D_T^\top) S S^\top S^\top T^\top \|_2 \leq \frac{1}{10} \eta \|S T^\top S^\top T^\top\|_2. \]

Furthermore, by the choice of \( \eta \) and the fact that \( \|S S^\top\|_2 \leq \sigma_1 \), we find that

\[ \eta^2 \|S S^\top S S^\top S^\top T^\top\|_2 \leq \frac{1}{10} \eta \|S T^\top S^\top T^\top\|_2. \]

Putting these results together we have \( \|VII\|_2 \leq 3\eta \|S T^\top\|_2^2 \).

Now for the VI term, direct calculation shows that

\[ VI = (T T^\top - D_T^\top) - \eta (T T^\top - D_T^\top) T T^\top \left( I - \frac{\eta}{2} (T T^\top - D_T^\top) \right) \]
\[ - \eta \left( I - \frac{\eta}{2} (T T^\top - D_T^\top) \right) T T^\top (T T^\top - D_T^\top) \]
\[ = (T T^\top - D_T^\top) \left( \frac{1}{2} - \eta T T^\top \left( I - \frac{\eta}{2} (T T^\top - D_T^\top) \right) \right) \]
\[ + \left( \frac{1}{2} - \eta \right) \left( T T^\top - D_T^\top \right) \left( T T^\top - D_T^\top \right). \]

Note that, since \( \|T T^\top\|_2 \leq \rho \sigma_r \), \( \|T T^\top - D_T^\top\|_2 \leq \rho \sigma_r \), and \( \eta = \frac{1}{10\sigma_1} \), we obtain that

\[ \left\| \frac{1}{2} - \eta T T^\top \left( I - \frac{\eta}{2} (T T^\top - D_T^\top) \right) \right\|_2 \leq \left\| \frac{1}{2} - \eta T T^\top \right\|_2. \]

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

\[ \|M_\tau(T)M_\tau(T)^\top - D_T^\top\|_2 \leq \|T T^\top - D_T^\top\|_2 \|I - 2\eta T T^\top\|_2 + 3\eta \|S T^\top\|_2^2. \]

As a consequence, we reach the conclusion of claim (d) in Lemma 6.

**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 11.** Under the same settings as Lemma 6, we have
\[(a) \| D^*_S - M_S(S)S^T \|_2 \leq (1 - \eta \sigma_r) \| D^*_S - SS^T \|_2 + \eta \| ST^T \|_2^2, \]

\[(b) \| M_S(S)T^T \|_2 \leq \| ST^T \|_2, \]

\[(c) \| M_T(T)S^T \|_2 \leq \| ST^T \|_2, \]

\[(d) \| M_T(T)T^T \|_2 \leq \| TT^T \|_2 + \eta \| ST^T \|_2^2. \]

**Proof** With $M_S(S) = S - \eta (SS^T S + ST^T T - D^*_S S)$ and simple algebraic manipulations, we obtain

\[
M_S(S)S^T - D^*_S = \left( SS^T - D^*_S \right) - \eta \left( SS^T SS^T + ST^T TS^T - D^*_S SS^T \right)
\]

\[
= \left( SS^T - D^*_S \right) \left( I - \eta SS^T \right) - \eta ST^T TS^T.
\]

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

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

\[
M_S(S)T^T = ST^T - \eta \left( SS^T ST^T + ST^T TT^T - D^*_S ST^T \right)
\]

\[
= \left( \frac{1}{2} I - \eta \left( SS^T - D^*_S \right) \right) ST^T - ST^T \left( \frac{1}{2} I - \eta TT^T \right).
\]

By initialization condition and triangular inequality, we know that $0 \leq \| (SS^T - D^*_S) \|_2 \leq \rho \sigma_r$ and $0 \leq \| TT^T \|_2 \leq 1.1 \sigma_r$, and hence $\| M_S(S)T^T \|_2 \leq \| ST^T \|_2$. Hence, we reach the conclusion of claim (b).

With $M_T(T) = T - \eta (TT^T T + TS^T S - D^*_T T)$ and direct calculation, we find that

\[
M_T(T)S^T = TS^T - \eta \left( TT^T TS^T + TS^T SS^T - D^*_T TS^T \right)
\]

\[
= \left( \frac{1}{2} I - \eta \left( TT^T - D^*_T \right) \right) TS^T - TS^T \left( \frac{1}{2} I - \eta SS^T \right).
\]

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

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

\[
M_T(T)T^T = TT^T - \eta \left( TT^T TT^T + TS^T ST^T - D^*_T TT^T \right)
\]

\[
= \left( I - \eta \left( TT^T - D^*_T \right) \right) TT^T - \eta TS^T ST^T.
\]

By initialization condition and triangular inequality, we know that $\| M_T(T)T^T \|_2 \leq \| TT^T \|_2 + \eta \| ST^T \|_2^2$. As a consequence, we obtain the conclusion of claim (d).
Appendix B. Proofs for the finite sample analysis

Recall that, we denote $G_t$ as the population gradient at iteration $t$ and denote $G_t^n$ as the corresponding sample gradient with sample size $n$:

$$G_t = \left(F_t F_t^\top - X^*\right) F_t,$$

$$G_t^n = \frac{1}{n} \sum_{i=1}^{n} \left(\left(A_i, F_i F_i^\top - X^*\right) + \epsilon_i\right) A_i F_t.$$

Then, we can write our update as follows:

$$F_{t+1} = F_t - \eta G_t + \eta G_t^n.$$

We assume the following decomposition by notations: $F = US + VT$. Therefore, we find that

$$S_{t+1} (S_{t+1})^\top = U^T F_{t+1} \left(U^T F_{t+1}\right)^\top = U^T (F_t - \eta G_t) (F_t - \eta G_t)^\top U + \eta^2 U^T (G_t - G_t^n) (G_t - G_t^n)^\top U + \eta U^T (G_t - G_t^n) (F_t - \eta G_t)^\top U,$$

$$= M_S(S_t) M_S(S_t)^\top + \eta^2 U^T (G_t - G_t^n) (G_t - G_t^n)^\top U + \eta U^T (G_t - G_t^n) M_S(S_t)^\top,$$

where we define $M_S(S_t)$ as follows:

$$M_S(S_t) = U^T (F_t - \eta G_t) = S_t - \eta \left(S_t S_t^\top S_t + S_t T_t^\top T_t - D_x S_t\right).$$

Furthermore, direct calculation shows that

$$S_{t+1} (T_{t+1})^\top = U^T F_{t+1} \left(V^T F_{t+1}\right)^\top = U^T (F_t - \eta G_t) (F_t - \eta G_t)^\top V + \eta^2 U^T (G_t - G_t^n) (G_t - G_t^n)^\top V + \eta U^T (G_t - G_t^n) (F_t - \eta G_t)^\top V,$$

$$= M_S(S_t) M_T(T_t)^\top + \eta^2 U^T (G_t - G_t^n) (G_t - G_t^n)^\top V + \eta U^T (G_t - G_t^n) M_T(T_t)^\top,$$

where $M_T(T_t)$ is given by:

$$M_T(T_t) = V^T (F_t - \eta G_t) = T_t - \eta \left(T_t T_t^\top T_t + T_t S_t^\top S_t - D_x T_t\right).$$

Similarly, we also have

$$T_{t+1} (T_{t+1})^\top = V^T F_{t+1} \left(V^T F_{t+1}\right)^\top = V^T (F_t - \eta G_t) (F_t - \eta G_t)^\top V + \eta^2 V^T (G_t - G_t^n) (G_t - G_t^n)^\top V + \eta V^T (G_t - G_t^n) (F_t - \eta G_t)^\top V,$$

$$= M_T(T_t) M_T(T_t)^\top + \eta^2 V^T (G_t - G_t^n) (G_t - G_t^n)^\top V + \eta V^T (G_t - G_t^n) M_T(T_t)^\top.$$

24
Note that, in the above equations, $\mathcal{M}_S(S_t)$ and $\mathcal{M}_T(T_t)$ are the updates of the coefficients when we update $S$ and $T$ using the population gradient. Furthermore, $\eta^2 U^\top (G_t - G_t^\star)(G_t - G_t^\star)^\top \eta^2 V^\top (G_t - G_t^\star)(G_t - G_t^\star)^\top V$, and $\eta^2 V^\top (G_t - G_t^\star)(G_t - G_t^\star)^\top V$ are second order terms and are relatively small. To facilitate the proof argument, we denote

$$\Delta_t := \frac{1}{n} \sum_{i=1}^{n} \left( \langle A_i, F_i F_i^\top - X^\star \rangle + \epsilon_i \right) A_i - (F_i F_i^\top - X^\star).$$

We can see that $\Delta_t$ is symmetric matrix, and

$$G_t^m - G_t = \Delta_t F_t.$$

### B.1 Proof for Lemma 7

#### Proof

By Lemma 6 and Lemma 11, we have the following contraction results:

$$\left\| \mathcal{M}_S(S_t) \mathcal{M}_S(S_t)^\top - D_S^t \right\|_2 \leq (1 - \eta \sigma_r) \left\| S_t S_t^\top - D_S^t \right\|_2 + 3\eta \left\| S_t T_t^\top \right\|_2^2,$$

$$\left\| \mathcal{M}_S(S_t) \mathcal{M}_T(T_t)^\top \right\|_2 \leq (1 - \eta \sigma_r) \left\| S_t T_t^\top \right\|_2,$$  \hspace{1cm} (20)

$$\left\| \mathcal{M}_T(T_t) \mathcal{M}_T(T_t)^\top \right\|_2 \leq \left\| T_t T_t^\top \right\|_2 \left( 1 - \eta \right) \left\| T_t T_t^\top \right\|_2 + 2\eta \left\| D_T \right\|_2,$$

and the following non-expansion results:

$$\left\| \mathcal{M}_S(S_t) S_t^\top - D_S^t \right\|_2 \leq (1 - \eta \sigma_r) \left\| S_t S_t^\top - D_S^t \right\|_2 + \eta \left\| S_t T_t^\top \right\|_2^2,$$

$$\left\| \mathcal{M}_S(S_t) T_t^\top \right\|_2 \leq \left\| S_t T_t^\top \right\|_2,$$  \hspace{1cm} (21)

$$\left\| S_t \mathcal{M}_T(T_t)^\top \right\|_2 \leq \left\| S_t T_t^\top \right\|_2 + \eta \left\| S_t T_t^\top \right\|_2^2.$$

For notation simplicity, let $D_t = \max \{ \left\| S_t S_t^\top - D_S^t \right\|_2, \left\| T_t T_t^\top \right\|_2, \left\| S_t T_t^\top \right\|_2 \}$, and denote the statistical error $\epsilon_{stat} = \sqrt{d \log \frac{d}{\eta \sigma_r}}$. Since Assumption 1 is satisfied, and $\left\| D_T \right\|_2 \leq \epsilon_{stat}$, we have $D_t \leq \sigma_r$ by triangular inequality. Since $\eta \left\| S_t T_t^\top \right\|_2 \leq \frac{1}{10} \eta \sigma_r$ by initialization, and $\left\| S_t T_t^\top \right\|_2 \leq D_t$, we have $\eta \left\| S_t S_t^\top \right\|_2 \leq 0.1 \sigma_r D_t$. Putting these results together, we have

$$\left\| \mathcal{M}_S(S_t) \mathcal{M}_S(S_t)^\top - D_S^t \right\|_2 \leq \left( 1 - \frac{7}{10} \eta \sigma_r \right) D_t,$$

$$\left\| \mathcal{M}_S(S_t) S_t^\top - D_S^t \right\|_2 \leq \left( 1 - \frac{9}{10} \eta \sigma_r \right) D_t,$$  \hspace{1cm} (22)

$$\left\| \mathcal{M}_T(T_t) T_t^\top \right\|_2 \leq \left( 1 + \frac{1}{10} \eta \sigma_r \right) D_t.$$

For the ease of the presentation we need to connect $\sqrt{\frac{k d \log d}{n}}$ with $\eta \sigma_r$ for the development of the proof. Since $n = \frac{1}{100 \log_d}$ and $n > C_1 k \sigma^2 d \log^2 d \cdot \max(1, \sigma^2 / \sigma_r^2)$, by choosing
Lastly, consider the II term, we have the following bound:

\[ \eta \| \Delta_t \|_2 \leq \eta \sigma, D_t + 4\sqrt{\frac{d \log d}{n}}. \]  

Combining equation (23) with equation (5), we obtain that

\[ \| \Delta_t \|_2 \leq \eta \sigma, D_t + 4\sqrt{\frac{d \log d}{n}}. \]  

Upper bound for \( \| S_t S_t^T - D_S^* \|_2 \): According to equation (17), we have

\[ S_{t+1} S_{t+1}^T - D_S^* = \mathcal{M}_S(S_t) \mathcal{M}_S(S_t)^T - D_S^* + \eta^2 U^T (G_t - G_t^n) (G_t - G_t^n)^T U \]

\[ + \eta \mathcal{M}_S(S_t) (G_t - G_t^n)^T U + \eta U^T (G_t - G_t^n) \mathcal{M}_S(S_t)^T, \]

where we can further expand \( \eta \mathcal{M}_S(S_t) (G_t - G_t^n)^T U \) and \( \eta U^T (G_t - G_t^n) \mathcal{M}_S(S_t)^T \) as follows:

\[ \eta \mathcal{M}_S(S_t) (G_t - G_t^n)^T U \]

\[ = \eta \mathcal{M}_S(S_t) F_t^T \Delta_t U = \eta \mathcal{M}_S(S_t) (US_t + VT_t)^T \Delta_t U \]

\[ = \eta \mathcal{M}_S(S_t) S_t^T U^T \Delta_t U + \eta \mathcal{M}_S(S_t) T_t^T V^T \Delta_t U \]

\[ = \eta \mathcal{M}_S(S_t) S_t^T U^T \Delta_t U - \eta D_S^* U^T \Delta_t U + \eta \mathcal{M}_S(S_t) T_t^T V^T \Delta_t U + \eta D_S^* U^T \Delta_t U, \]

and

\[ \eta U^T (G_t - G_t^n) \mathcal{M}_S(S_t)^T \]

\[ = \eta U^T \Delta_t U S_t, \mathcal{M}_S(S_t)^T - \eta U^T \Delta_t U D_S^* + \eta U^T \Delta_t V T_t, \mathcal{M}_S(S_t)^T + \eta U^T \Delta_t U D_S^*. \]

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:

III & VI: \[ \eta \mathcal{M}_S(S_t) S_t^T U^T \Delta_t U - \eta D_S^* U^T \Delta_t U \leq \eta \mathcal{M}_S(S_t) S_t^T - D_S \|_2 \| \Delta_t \|_2, \]

IV & VII: \[ \eta \mathcal{M}_S(S_t) T_t^T V^T \Delta_t U \leq \eta \mathcal{M}_S(S_t) T_t^T \| \Delta_t \|_2, \]

V & VIII: \[ \eta D_S^* U^T \Delta_t U \leq \eta D_S^* \| \Delta_t \|_2. \]

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

\[ \| \eta^2 U^T (G_t - G_t^n) (G_t - G_t^n)^T U \|_2 \leq \eta^2 \| \Delta_t F_t F_t^T \Delta_t \|_2 \]

\[ \leq \eta^2 \left( \| S_t S_t^T \|_2 + \| T_t T_t^T \|_2 + 2 \| S_t T_t \|_2 \right) \| \Delta_t \|_2^2 \]

\[ \leq \frac{1}{100} \eta \| \Delta_t \|_2^2, \]
where the last inequality holds by assuming \( \rho \leq 0.1 \). Putting all the above results together, we obtain that

\[
\| S_{t+1} S_{t+1}^T - D_S^* \|_2 \leq (1 - \frac{7}{10} \eta \sigma_r) \| S_t S_t^T - D_S^* \|_2 + 0.1 \eta \sigma_r D_t + \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma.
\] (26)

This indicates a contraction with respect to \( D_t \):

\[
\| S_{t+1} S_{t+1}^T - D_S^* \|_2 \leq \left(1 - \frac{6}{10} \eta \sigma_r\right) D_t + \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma.
\] (27)

From the above result, we can verify that

\[
\| S_{t+1} S_{t+1}^T - D_S^* \|_2 \leq \left(1 - \frac{6}{10} \eta \sigma_r\right) \left(D_t - 50 \kappa \sqrt{\frac{d \log d}{n}} \sigma\right) + 50 \kappa \sqrt{\frac{d \log d}{n}} \sigma.
\] (28)

**Upper bound for \( \| S_t T_{t}^T \|_2 \):** According to equation (18), we have

\[
S_{t+1} T_{t+1}^T = \mathcal{M}_S(S_t) \mathcal{M}_T(T_t)^T + \eta^2 U^T (G_t - G_t^n) (G_t - G_t^n)^T V + \eta \mathcal{M}_S(S_t) (G_t - G_t^n)^T V + \eta U^T (G_t - G_t^n) \mathcal{M}_T(T_t)^T,
\]

where inequality (2) is obtained by the non-expansion property of population update (cf. equations (21) and (22)); 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 \( \sigma \); inequality (5) is obtained via the bound (24).

That is, we proved the equations (6) and (7) in the Lemma 7, namely, we have

\[
\| S_{t+1} S_{t+1}^T - D_S^* \|_2 \leq \left(1 - \frac{7}{10} \eta \sigma_r\right) \| S_t S_t^T - D_S^* \|_2 + 0.1 \eta \sigma_r D_t + \frac{4}{10} \sqrt{\frac{d \log d}{n}} \sigma.
\]
where we can expand $\eta M_S(S_t)(G_t - G^n_t)^T V$ and $\eta U^T (G_t - G^n_t) M_T(T_t)^T$ as follows:

\[
\eta M_S(S_t)(G_t - G^n_t)^T V = \eta M_S(S_t)(US_t + VT_t)^T \Delta t V \\
= \underbrace{\eta M_S(S_t)S_t^T U^T \Delta t V - \eta D_S^u U^T \Delta t V}_{III'} + \underbrace{\eta M_S(S_t)T_t^T V^T \Delta t V}_{IV'} \\
+ \eta D_S^u U^T \Delta t V, \\
\eta U^T (G_t - G^n_t) M_T(T_t)^T = \eta U^T \Delta t (US_t + VT_t) M_T(T_t)^T \\
= \underbrace{\eta U^T \Delta t US_t M_T(T_t)^T}_{VI'} + \underbrace{\eta U^T \Delta t VT_t M_T(T_t)^T}_{VII'}.
\]

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

\[
III': \quad \left\| \eta M_S(S_t)S_t^T U^T \Delta t V - \eta D_S^u U^T \Delta t V \right\|_2 \leq \eta \left\| M_S(S_t)S_t^T - D_S^u \right\|_2 \| \Delta t \|_2, \\
IV': \quad \left\| \eta M_S(S_t)T_t^T V^T \Delta t V \right\|_2 \leq \eta \left\| M_S(S_t)T_t^T \right\|_2 \| \Delta t \|_2, \\
V': \quad \left\| \eta D_S^u U^T \Delta t V \right\|_2 \leq \eta \| D_S^u \|_2 \| \Delta t \|_2, \\
VI': \quad \left\| \eta U^T \Delta t US_t M_T(T_t)^T \right\|_2 \leq \eta \left\| M_T(T_t)S_t^T \right\|_2 \| \Delta t \|_2, \\
VII': \quad \left\| \eta U^T \Delta t VT_t M_T(T_t)^T \right\|_2 \leq \eta \left\| M_T(T_t)T_t^T \right\|_2 \| \Delta t \|_2.
\]

Lastly, the II term is bounded as in Equation (25), namely, we have

\[
\left\| \eta^2 U^T (G_t - G^n_t)(G_t - G^n_t)^T V \right\|_2 \leq \frac{1}{100} \eta \| \Delta t \|_2^2.
\]
Collecting the above results, we find that

\[ \left\| S_{t+1} T_{t+1}^\top \right\|_2 \]
\[ \leq \left\| M_S(S_t) M_T(T_t) \right\|_2 + \frac{1}{100} \eta \| \Delta_t \|_2^2 + \eta \left\| M_S(S_t) S_t^\top - D_S^t \right\|_2 \| \Delta_t \|_2 + \eta \left\| D_S^t \right\|_2 \| \Delta_t \|_2 \]
\[ + \eta \left\| M_S(S_t) T_t^\top \right\|_2 \| \Delta_t \|_2 + \eta \left\| M_T(T_t) S_t^\top \right\|_2 \| \Delta_t \|_2 + \eta \left\| M_T(T_t) T_t^\top \right\|_2 \| \Delta_t \|_2 \]
\[ \leq (1 - \eta \sigma) \left\| S_t T_t^\top \right\|_2 + \frac{1}{100} \eta \| \Delta_t \|_2^2 + \frac{1}{100} \| \Delta_t \|_2^2 + \frac{5 \eta D_t}{T} \| \Delta_t \|_2 \]
\[ \leq (1 - \eta \sigma) \left\| S_t T_t^\top \right\|_2 + \frac{4}{10} \sqrt{d} \frac{d \log d}{n} + \frac{4}{10} \sqrt{d} \frac{d \log d}{n} \sigma, \]

where inequality (1) is obtained by the non-expansion property of population update (cf. equations (21) and (22)); 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

\[ \left\| S_{t+1} T_{t+1}^\top \right\|_2 \leq (1 - \eta \sigma) \left\| S_t T_t^\top \right\|_2 + 0.1 \eta \sigma D_t + \frac{4}{10} \sqrt{d} \frac{d \log d}{n} \sigma. \]  

(29)

With similar treatment as in \( \left\| S_S^t - D_S^t \right\|_2 \), we have the following contraction result with respect to \( D_t \):

\[ \left\| S_{t+1} T_{t+1}^\top \right\|_2 \leq \left( 1 - \frac{9}{10} \eta \sigma \right) D_t + \frac{4}{10} \sqrt{d} \frac{d \log d}{n} \sigma. \]  

(30)

Given the above result, we can verify that

\[ \left\| S_{t+1} T_{t+1}^\top \right\|_2 \leq \left( 1 - \frac{9}{10} \eta \sigma \right) \left( D_t - 50 \kappa \sqrt{d} \frac{d \log d}{n} \sigma \right) + 50 \kappa \sqrt{d} \frac{d \log d}{n} \sigma. \]  

(31)

**Upper bound for \( \left\| T_t T_t^\top \right\|_2 \):** According to equation (19) and similar deductions as in previous bounds for \( \left\| S_S^t - D_S^t \right\|_2 \) and \( \left\| S_t T_t^\top \right\|_2 \), we have

\[ T_{t+1} T_{t+1}^\top = \left( M_T(T_t) M_T(T_t)^\top \right) + \eta^2 V^\top (G_t - G_t^n) (G_t - G_t^n)^\top V \]
\[ + \eta M_T(T_t) (G_t - G_t^n)^\top V + \eta V^\top (G_t - G_t^n) M_T(T_t)^\top, \]
where the following expansions hold:

\[ \eta M_T(T_t) (G_t - G_t^n) \top V = \eta M_T(T_t) S_t \top U \top \Delta_t V + \eta M_T(T_t) T_t \top V \top \Delta_t V, \]

\[ \eta V \top (G_t - G_t^n) M_T(T_t) \top = \eta V \top \Delta_t U S_t M_T(T_t) \top + \eta V \top \Delta_t V T_t M_T(T_t) \top. \]

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

\[ \text{III}'' \& \text{V}'' : \left\| \eta M_T(T_t) S_t \top U \top \Delta_t V \right\|_2 \leq \eta \left\| M_T(T_t) S_t \top \right\|_2 \left\| \Delta_t \right\|_2, \]
\[ \text{IV}'' \& \text{VI}'' : \left\| \eta M_T(T_t) T_t \top \Delta_t V \right\|_2 \leq \eta \left\| M_T(T_t) T_t \top \right\|_2 \left\| \Delta_t \right\|_2, \]
\[ \text{II}'' : \left\| \eta^2 U \top (G_t - G_t^n) (G_t - G_t^n) \top U \right\|_2 \leq \frac{1}{100} \eta \left\| \Delta_t \right\|_2^2. \]

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

\[ \left\| T_{t+1} T_{t+1} \right\|_2 \leq (z - z^2 \eta D_t + 2z \eta \left\| D_T \right\|_2 + 4 \eta (\eta D_t + 4 \epsilon_{\text{stat}})) D_t + \frac{1}{100} \eta (\eta \sigma_t D_t + 4 \epsilon_{\text{stat}})^2, \]

where inequality (2) is obtained by the non-expansion property of population update (cf. equations (21) and (22)) and the assumption on \( \left\| \Delta_t \right\|_2 \) (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 \( z = \frac{1 + 2 \eta \left\| D_T \right\|_2}{2 \eta D_t} > 1. \) Hence we can just safely plug-in \( z = 1. \) Now, we arrange by organizing according to \( D_t \) and \( \epsilon_{\text{stat}} \) and obtain that

\[ \left\| T_{t+1} T_{t+1} \right\|_2 \leq (1 - \eta D_t + 2 \eta \left\| D_T \right\|_2 + 4 \eta (\eta D_t + 4 \epsilon_{\text{stat}})) D_t + \frac{1}{100} \eta (\eta \sigma_t D_t + 4 \epsilon_{\text{stat}})^2, \]

where inequality (1) is obtained by \( \left\| D_T \right\|_2 \leq \epsilon_{\text{stat}}, \) and inequality (2) is obtained by organizing according to \( D_t \) and \( \sigma. \)
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

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

Furthermore, from equations (28) and (31), 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 7.

Note that Lemma 7 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 reach the statistical accuracy.

**Lemma 12.** Consider the same setting as Lemma 7, 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 7. As in the proof of Lemma 7, we know that

$$\|S_{t+1}S_{t+1}^\top - D_S^*\|_2 \leq \left(1 - \frac{7}{10} \eta \sigma_r\right) \|S_tS_t^\top - D_S^*\|_2 + \frac{1}{10} \|\Delta_t\|_2.$$ 

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

$$\|S_{t+1}S_{t+1}^\top - D_S^*\|_2 \leq 100\epsilon_{stat}.$$ 

Similarly for $\|S_{t+1}T_{t+1}^\top\|_2$, we have

$$\|S_{t+1}T_{t+1}^\top\|_2 \leq (1 - \eta \sigma_r) \|S_tT_t^\top\|_2 + \frac{1}{10} \|\Delta_t\|_2 \leq 100\epsilon_{stat}.$$
Finally, for $\|T_{t+1}T_{t+1}^\top\|_2$, we find that

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

We show later in part (b) that $D_t$ converges sub-linearly and thus $D_t \leq \sigma_r$ throughout the iterations with the initialization condition (Assumption 1). For now, let us assume this is given. Note that an error accumulated each iteration is $\epsilon_0 = \sqrt{\frac{kd\log d}{n}} \sigma_r + \frac{4}{10}\sqrt{\frac{d\log d}{n}} \sigma$.

B.2 Proof of Theorem 4

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

**Proof for claim (a) with the linear convergence:** Recall that

$$
\Delta_t = \frac{1}{n} \sum_i^n \langle A_i, F_tF_t^T - X^* \rangle A_i - \langle F_tF_t^T - X^* \rangle + \frac{1}{n} \sum_i^n \epsilon_i A_i.
$$

With Lemma 15 and Lemma 17, we know that with probability at least $1 - \exp(\log d)$

$$
\Delta_t \leq 5\sqrt{\frac{kd\log d}{n}} D_t + \sqrt{\frac{d\log d}{n}} \sigma.
$$

Therefore, with this inequality, the result from Lemma 7 indicates

$$
\|S_t+1S_{t+1}^\top - D_S\|_2 \leq \left(1 - \frac{7}{10}\eta\sigma_r\right) \|S_tS_t^\top - D_S^*\|_2 + \sqrt{\frac{kd\log d}{n}} D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}} \sigma.
$$

Putting all the results together, we obtain the conclusion of Lemma 12. ■
Thus, we have
\[
\left\| S_{t+1}S_{t+1}^\top - D_S^* \right\|_2 \leq \left( 1 - \frac{7}{10} \eta \sigma_r \right) \left\| S_t S_t^\top - D_S^* \right\|_2 + \epsilon_0 \\
\leq \left( 1 - \frac{7}{10} \eta \sigma_r \right)^2 \left\| S_{t-1} S_{t-1}^\top - D_S^* \right\|_2 + \left( 1 + \left( 1 - \frac{7}{10} \eta \sigma_r \right) \right) \epsilon_0 \\
\leq \ldots \\
\leq \left( 1 - \frac{7}{10} \eta \sigma_r \right)^{t+1} \left\| S_0 S_0^\top - D_S^* \right\|_2 + O(\eta^{-1} \sigma_r^{-1}) \epsilon_0.
\]

We let \( \epsilon_{comp} = \kappa \sqrt{\frac{d \log d}{n}} (\sqrt{k} \sigma_r + \sigma) \). We now have constant contraction for one iteration. We can invoke the Lemma 15 for once, Lemma 17 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. Note that
\[
\left\| S_t S_t^\top - D_S^* \right\|_2 \leq \left( 1 - \frac{7}{10} \eta \sigma_r \right)^t \left\| S_0 S_0^\top - D_S^* \right\|_2 + \epsilon_{comp} \leq \left( 1 - \frac{7}{10} \eta \sigma_r \right)^t 0.1 \sigma_r + \epsilon_{comp},
\]
where the final inequality holds by simply plugging in the initialization condition.

After at most \( t = \frac{1}{\log 1 - 0.005/k} \cdot \log \frac{1}{10000 \sqrt{\frac{k^2 d \log d}{n}}} \) iterations, \( \left\| S_t S_t^\top - D_S^* \right\|_2 < \epsilon_{comp} \).

Since \( \frac{1}{\log 1 - 0.005/k} \leq 1.1 \), we further simplify this to \( t > \log \frac{n}{k \sqrt{2d \log d}} \). As a consequence, we claim that after \( t = \left\lceil \log \frac{n}{k \sqrt{2d \log d}} \right\rceil \) iterations, \( \left\| S_t S_t^\top - D_S^* \right\|_2 < C \epsilon_{comp} \) 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 (36) in Lemma 17 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., \( n^{z_1} > C_2 d \log^3 d \kappa^2 \) for some universal constant \( z_1 \in (0, 1) \), then we should use \( \Delta_t \leq 0.5 \eta \sigma_r D_t + \sqrt{\frac{d \log d}{n}} \sigma \) and the fact that \( \frac{\sqrt{d \log d}}{n} \ll \eta \sigma_r \) to directly establish the above results.

Therefore equation (5) holds with probability at least \( 1 - d^{-c} \) for all iteration \( t \), and we complete our proof with \( \left\| S_t S_t^\top - D_S^* \right\|_2 < C \epsilon_{comp} \).

With the same argument, we also obtain \( \left\| S_t T_t^\top \right\|_2 < C \epsilon_{comp} \) after \( t = \left\lceil \log \frac{n}{k \sqrt{2d \log d}} \right\rceil \) iterations. Therefore, we obtain the conclusion of claim (a) in Theorem 4.

**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 7, as the equation (10). However, as the convergence rate is sub-linear ultimately, it is sub-optimal to directly invoke concentration
result (Lemma 17) 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., 2021).

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 \( \alpha_i \), ends at iteration \( \alpha_{i+1} - 1 \), and \( D_{\alpha_{i+1}} \leq 0.5D_{\alpha_i} \). We invoke Lemma 18 at \( \alpha_i \) 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
\[
\Delta_t = \frac{1}{n} \sum_{i}^n \langle A_i, F_t F_i^T - X^* \rangle A_i - (F_t F_i^T - X^*) + \frac{1}{n} \sum_{i}^n \epsilon_i A_i.
\]

From Lemma 15, we know that with probability at least \( 1 - \exp(-C) \),
\[
\frac{1}{n} \sum_{i}^n \epsilon_i A_i \leq \sqrt{\frac{d \log d}{n}}.
\]

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

At iteration \( \alpha_i \), note that \( \| F_{\alpha_i} F_{\alpha_i}^T - X^* \|_2 \leq \| S_{\alpha_i} S_{\alpha_i}^T - D_{\alpha} \|_2 + \| T_{\alpha_i} T_{\alpha_i}^T - D_{\alpha}^* \|_2 + 2 \| S_{\alpha_i} T_{\alpha_i}^T \|_2 \leq 4D_{\alpha_i} + \| D_T \|_2 < 5D_{\alpha_i} \). By Lemma 18 we know that, with probability at least \( 1 - \exp(-C) \), we have
\[
\sup_{\| X \|_2 \leq 5D_{\alpha_i}} \frac{1}{n} \sum_{i}^n \langle A_i, X \rangle A_i - X \leq 5 \sqrt{\frac{k^2 d \log d}{n}} D_{\alpha_i}.
\]

Therefore, we find that
\[
\| \Delta_{\alpha_i} \|_2 \leq 5 \sqrt{\frac{k^2 d \log d}{n}} D_{\alpha_i} + \sqrt{\frac{d \log d}{n}} \leq 0.5 \eta D_{\alpha_i} + \sqrt{\frac{d \log d}{n}}
\]

where the second inequality is by the choice of \( n \). Hence equation (5) is satisfied at iteration \( \alpha_i \). For notation simplicity, we define \( A_t = D_t - 50\kappa \sqrt{\frac{d \log d}{n}} \). Invoking Lemma 7, we have
\[
D_{\alpha_{i+1}} = A_{\alpha_{i+1}} + 50\kappa \frac{d \log d}{n} \leq \left( 1 - \frac{1}{2} \eta \right) A_{\alpha_i} + 50\kappa \frac{d \log d}{n} \leq D_{\alpha_i},
\]

where the last inequality just comes from \( D_{\alpha_i} = A_{\alpha_i} + 50\kappa \frac{d \log d}{n} \). At iteration \( t \in (\alpha_i, \alpha_{i+1} - 1) \), by induction \( D_t = A_t + 50\kappa \sqrt{\frac{d \log d}{n}} \), 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:
\[
\Delta_t = \frac{1}{n} \sum_{i}^n \langle A_i, F_t F_i^T - X^* \rangle A_i - (F_t F_i^T - X^*) + \frac{1}{n} \sum_{i}^n \epsilon_i A_i
\]

34
\[ \leq 0.5\eta \sigma_r D_{\alpha_i} + \sqrt{\frac{d \log d}{n}} \sigma \]
\[ \leq \eta \sigma_r D_t + \sqrt{\frac{d \log d}{n}} \sigma. \]

Hence, equation (5) is satisfied for all iteration \( t \in (\alpha_i, \alpha_i + 1 - 1) \). Invoking Lemma 7, 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

\[ A_{t+1} \leq \left( 1 - \frac{1}{2} \eta A_t \right) A_t. \] (32)

We first assume that equation (32) 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

\[ A_t \leq 4 \eta t + 4 \kappa \sqrt{\frac{d \log d}{n}} \sigma \]

To see this, we have

\[ A_{t+1} \leq \left( 1 - \frac{1}{2} \eta A_t \right) A_t \overset{(1)}{\leq} \left( 1 - \frac{2}{t + \frac{4}{\eta A_0}} \right) \frac{4}{\eta t + \frac{4}{A_0}} \]
\[ = \frac{t + \frac{4}{\eta A_0}}{t + \frac{4}{\eta A_0}} - 2 \frac{4}{\eta \left( t + \frac{4}{\eta A_0} \right)} \]
\[ \overset{(2)}{\leq} \frac{4}{\eta \left( t + 1 + \frac{4}{\eta A_0} \right)} \]

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 \( \frac{(t + \frac{4}{\eta A_0}) - 2}{(t + \frac{4}{\eta A_0}) + 1} \leq \frac{1}{(t + \frac{4}{\eta A_0}) + 1}. \)

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

\[ \max \{ \| S_t S_t^T - D_* S_* \|_2, \| T_t T_t^T \|_2, \| S_t T_t^T \|_2 \} \leq \Theta (\epsilon_{stat}). \] (33)

Now what is left to be shown is that equation (32) 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 Lemma 18 for each epochs. Let \( T \) be the number of total epochs. For each epoch, the \( D_t \) shrinks by at least a half. To reach \( \epsilon_{stat} \), we need \( T = \Theta (\log (1/\epsilon_{stat})) = \Theta (\log n) \). Equation (32) holds for all epochs with probability at least \( 1 - Td^{-c} > 1 - d^{-c+1} \) for some universal constant \( c \), since \( T = \Theta (\log n) \).

If \( n \) is large, i.e., \( n^{z_1} > C_2 d \log^3 d \kappa \sigma^2 \) for some universal constant \( z_1 \in (0, 1) \), then we establish equation (32), and we invoke equation (37) in Lemma 17 for \( T \) epochs. This
holds with probability at least $1 - T/\exp(n^{\epsilon_1}) > 1 - d^{-c}$ for some universal constant $c$, since $T < \log n$. If $t > \Theta \left( \frac{1}{\eta \sigma_{stat}} \right)$, we can show using above argument that after $\Theta \left( \frac{1}{\eta \sigma_{stat}} \right)$ number of iterations equation (33) holds. After this, by Lemma 12 we know that $D_t = \Theta (\epsilon_{stat})$. Then, we can invoke Lemma 7 or without further invoking the concentration argument anymore, since the radius in the uniform concentration result does not change.

As a consequence, after $t \geq \Theta \left( \frac{1}{\eta \sigma_{stat}} \right)$ number of iterations, by triangular inequality, and the assumption that $\|D^*_T\|_2 \leq \epsilon_{stat}$, we have $\|F_t F^\top_t - X^*\|_2 \leq \Theta (\epsilon_{stat})$. Combined with Lemma 12, we complete the proof of Theorem 4.

Appendix C. Supporting Lemma

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

C.1 Proof of Lemma 3

**Proof** From the definition of operator norm, we have

$$\|D^*_T - T_0 T^\top_0\|_2 = \max_{x \in \mathbb{R}^{d-r}: \|x\|_2 \leq 1} \left| x^\top \left( D^*_T - T_0 T^\top_0 \right) x \right|.$$ 

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

$$\|D^*_T - T_0 T^\top_0\|_2 = \|V \left( D^*_T - T_0 T^\top_0 \right) V^\top\|_2 = \max_{x \in \mathbb{R}^d: \|x\|_2 \leq 1} \left| x^\top V \left( D^*_T - T_0 T^\top_0 \right) V^\top x \right|.$$ 

Without loss of generality we can write any $x \in \mathbb{R}^d$ as $x = x_u + x_v$, such that $Uz = x_u$ for some $z \in \mathbb{R}^{r'}$, and $Vz' = x_v$ for some $z' \in \mathbb{R}^{d-r}$ since $U$ and $V$ are perpendicular to each other and they together span $\mathbb{R}^d$. If $x^* = \arg \max_{x \in \mathbb{R}^d: \|x\|_2 \leq 1} \left| x^\top V \left( D^*_T - T_0 T^\top_0 \right) V^\top x \right|$ then $x_u^*$ is zero. It is because if $x_u^* \neq 0$, one can decrease $x_u^*$ to zero and increase $x_v^*$ to $x_v^*/\|x_v^*\|_2$, which does make the target quantity smaller. Therefore, we have

$$\|D^*_T - T_0 T^\top_0\|_2 = \max_{x \in \mathbb{R}^{d-r}: \|x\|_2 \leq 1} \left| x^\top V \left( D^*_T - T_0 T^\top_0 \right) V^\top x \right|$$

$$= \max_{x: \|x\|_2 \leq 1} \left| x^\top V \left( D^*_T - T_0 T^\top_0 \right) V^\top x + x^\top U \left( D_s^* - S_0 S^\top_0 \right) U^\top x + 2x^\top \left( U S_0 T^\top_0 V^\top \right) x \right|$$

$$\leq \max_{x: \|x\|_2 \leq 1} \left| x^\top V \left( D^*_T - T_0 T^\top_0 \right) V^\top x + x^\top U \left( D_s^* - S_0 S^\top_0 \right) U^\top x + 2x^\top \left( U S_0 T^\top_0 V^\top \right) x \right|$$

$$= \|F^\top_0 F_0 - X^*\|_2 \leq 0.7 \rho \sigma_r,$$

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

$$\|D_s^* - S_0 S^\top_0\|_2 \leq \|F^\top_0 F_0 - X^*\|_2 \leq 0.7 \rho \sigma_r.$$
For $\|S_0T_0^\top\|_2$, we claim that the following equations hold:

$$\left\|S_0T_0^\top\right\|_2 = \left\|US_0T_0^\top V^\top\right\|_2 = 0.5 \left\|US_0T_0^\top V^\top + VT_0S_0^\top U^\top\right\|_2.$$ 

To see the last equality, let $\sigma_1$ be the largest eigen-value (in magnitude) of $US_0T_0^\top V^\top$ and let $\bar{x}$ be the corresponding eigen-vector. For some $c \in (0, 1)$, let $\bar{x} = cx_u + \sqrt{1 - c^2}x_v$ such that $Uz = \bar{x}_u$ for some $z \in \mathbb{R}^r$, $Vz' = x_v$ for some $z' \in \mathbb{R}^{d-r}$, and $\|x_u\|_2 = 1$ and $\|x_v\|_2 = 1$. Then, direct algebra leads to

$$\sigma_1 = (\bar{x})^\top US_0T_0^\top V^\top \bar{x} = c\sqrt{1 - c^2} (\bar{x})^\top_u US_0T_0^\top V^\top x_v = \sigma_1.$$ 

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 $VT_0S_0^\top U^\top$ is also $\sigma_1$. Therefore, we obtain that

$$(\bar{x})^\top VT_0S_0^\top U^\top \bar{x} = c\sqrt{1 - c^2} (\bar{x})^\top_u US_0T_0^\top V^\top x_v = \sigma_1.$$ 

Collecting the above results, we have $\left\|US_0T_0^\top V^\top\right\|_2 = 0.5 \left\|US_0T_0^\top V^\top + VT_0S_0^\top U^\top\right\|_2$. Then, an application of triangular inequality yields that

$$2 \left\|S_0T_0^\top\right\|_2 = \left\|US_0T_0^\top V^\top + VT_0S_0^\top U^\top\right\|_2 \leq \left\|F_0F_0^\top - X^\top\right\|_2 + \left\|D_T^\top - T_0^\top + D_S^\top - S_0^\top\right\|_2.$$ 

We can check that $\left\|D_T^\top - T_0^\top + D_S^\top - S_0^\top\right\|_2 \leq 0.7 \cdot \sqrt{2} \rho \sigma_r$ by decomposing the eigenvector $x = cx_u + \sqrt{1 - c^2}x_v$ as above. Therefore, $\left\|S_0T_0^\top\right\|_2 < \rho \sigma_r$. 

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

Appendix D. Concentration bounds

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

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

for any matrix $F \in \mathbb{R}^{d \times k}$ such that $\|FF^\top - X^\top\|_2 \leq R$ 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 13. (Matrix Bernstein, Theorem 1.4 in Tropp (2012))** Consider a finite sequence $\{X_k\}$ of independent, random, self-adjoint matrices with dimension $d$. Assume that each random matrix satisfies

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

Then, for all $t \geq 0$,

$$\mathbb{P} \left( \lambda_{\max} \left( \sum_k X_k \right) \geq t \right) \leq d \cdot \exp \left( \frac{-t^2/2}{\sigma^2 + Rt/3} \right) \quad \text{where} \quad \sigma^2 := \left\| \sum_k E(X_k^2) \right\|_2.$$  \hspace{1cm} (34)
Lemma 14. Let $A$ 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 $\epsilon$ follow $N(0, \sigma)$. Then

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

Proof

As $\epsilon$ is sub-Gaussian, we know that for all $t > 0$

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

By standard $\epsilon$-net argument (Tropp, 2012; Vershynin, 2018), for some universal constant $C_1, C_2$, we have

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

Applying the union bound to the above concentration results leads to

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

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

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

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

Lemma 15. (Lemma 8 re-stated) Let $A_i$ 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 $\epsilon_i$ follow $N(0, \sigma)$. Then

$$
P \left( \left\| \frac{1}{n} \sum_{i}^{n} A_i \epsilon_i \right\|_2 \geq C \sqrt{\frac{d \sigma^2}{n}} \right) \leq \exp(-C).
$$

Proof

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

$$
E \left( (A_i \epsilon_i)^2 \right) = \sigma^2 E \left( A_i^2 \right) = \sigma^2 d I.
$$

Hence, we obtain

$$
\left\| \sum_{i}^{n} E \left( (A_i \epsilon_i)^2 \right) \right\|_2 \leq n \sigma^2 d.
$$

38
From the matrix Bernstein bound (Wainwright, 2019), we find that
\[
P\left( \left\| \frac{1}{n} \sum_{i=1}^{n} A_i \epsilon_i \right\|_2 \geq t \right) \leq d \cdot \exp\left( \frac{-3t^2n^2}{6dn\sigma^2 + 2C_1\sigma\sqrt{dtn}} \right).
\]

For any \( \delta < 1/e \), let \( t = \log \frac{1}{\delta} \sqrt{\frac{d\sigma^2}{n}} \). Then, the above bound becomes
\[
P\left( \left\| \frac{1}{n} \sum_{i=1}^{n} A_i \epsilon_i \right\|_2 \geq \log \frac{1}{\delta} \sqrt{\frac{d\sigma^2}{n}} \right) \leq \delta.
\]

Or equivalently, for any \( C > 1 \), let \( t = C \sqrt{\frac{d\sigma^2}{n}} \),
\[
P\left( \left\| \frac{1}{n} \sum_{i=1}^{n} A_i \epsilon_i \right\|_2 \geq C \sqrt{\frac{d\sigma^2}{n}} \right) \leq \exp\left( -\frac{C}{C_1} \right).
\]

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

**Lemma 16.** Let \( A \) 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 \( U \) be a deterministic symmetric matrix of the same dimension. Then, for some universal constant \( C_1, C_2 \), we have
\[
P\left( \| \langle A, U \rangle A - U \|_2 \geq C_1 d \| U \|_F \right) \leq \exp\left( -\frac{d}{C_2} \right).
\]

**Proof.** We show this by standard \( \epsilon \)-net argument. In particular, we have
\[
\| \langle A, U \rangle A - U \|_2 = \max_{x \in S^{d-1}} x^\top \left( \langle A, U \rangle A - U \right) x
\]
\[
= \max_{x \in S^{d-1}} \langle A, U \rangle \langle A, xx^\top \rangle - \left( x^\top U x \right).
\]

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

Since \( (x^\top U x) \leq \| U \|_2 \leq \| U \|_F \), we have
\[
P\left( x^\top (\langle A, U \rangle A - U) x \geq t \| U \|_F \right) \leq \exp\left( -\frac{t}{C_1} \right).
\]

By the standard \( \epsilon \)-net argument, let \( V \) be the \( \epsilon \) covering of \( S^{d-1} \). Then, we find that
\[
\| \langle A, U \rangle A - U \|_2 \leq \frac{1}{1 - 2\epsilon} \max_{x \in V} x^\top (\langle A, U \rangle A - U) x.
\]
Now we fix $\epsilon$ to be $1/4$. Then, for equation (35) we take union bound over $V$ and we have

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

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

$$
P \left( \|\langle A, U \rangle A - U\|_2 \geq C_1 d \|U\|_F \right) \leq \exp (-d/C_2).$$

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

Lemma 17. Let $A_i$ 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 $U$ 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

$$
P \left( \left\| \frac{1}{n} \sum_{i=1}^n (\langle A_i, U \rangle A_i - U) \right\|_2^2 \leq \frac{d \log d}{n} \|U\|_F^2 \right) \geq 1 - \exp (-C_2 \log d). \quad (36)$$

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

$$
P \left( \left\| \frac{1}{n} \sum_{i=1}^n (\langle A_i, U \rangle A_i - U) \right\|_2^2 \leq \frac{1}{\kappa \log d \sqrt{k C_2}} \|U\|_F^2 \right) \geq 1 - \exp (-n^{z_2}). \quad (37)$$

Proof Following Lemma 13, we want to first bound the second order moment of the random matrices. Since $E(\langle A_i, U \rangle A_i) = U$ and $U$ has no randomness, we have

$$E(\langle A_i, U \rangle A_i - U)^2 = E(\langle A_i, U \rangle A_i)^2 - U^2.$$ 

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

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

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

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

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 $U^2$. Therefore, we obtain that

$$\sum_{j=1}^d E(A_{mj}^2A_{jn}^2 U_{mj}U_{jn}) = \sum_{j=1}^d U_{mj}U_{jn}. $$
Hence the \((m, n)\) entry of \(E((\langle A_i, U \rangle A_i - U)^2)\) equals 0 when \(m \neq n\), and equals \(\|U\|_F^2 + 2U_m^2 - \sum_j U_j^2\) when \(m = n\). Hence \(\|E((\langle A_i, U \rangle A_i - U)^2)\|_2 \leq 3 \|U\|_F^2\) and

\[
\left\| \sum_i^n E((\langle A_i, U \rangle A_i - U)^2) \right\|_2 \leq 3n \|U\|_F^2.
\]

Then, the following inequality holds:

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

where \(C_1\) is a universal constant inherited from Lemma 16 and

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

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

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

Hence we finish the proof for equation 36.

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

\[
n^z > C_2 d \log^3 d \kappa^{-2},
\]

then the following inequality holds:
then plugging in $t = \sqrt{\frac{d \log d}{C_2 d \log d \cdot dk^2}}\|U\|_F = \frac{1}{\kappa \log d \sqrt{kC_2}}\|U\|_F$, we have

$$\Pr \left( \max_{\lambda} \left( \frac{1}{n} \sum_{t} \langle A_i, U \rangle A_i - U \right) \geq \frac{1}{\kappa \log d \sqrt{kC_2}}\|U\|_F \right)$$

$$\leq d \cdot \exp \left( -3 \left( \frac{1}{\kappa \log d \sqrt{kC_2}}\|U\|_F \right)^2 \frac{n}{18 \|U\|_F^2 + 2C_1 d \log d \|U\|_F \left( \frac{1}{\kappa \log d \sqrt{kC_2}}\|U\|_F \right)} \right)$$

$$= d \cdot \exp \left( -3 \frac{n}{18 (\kappa \log d \sqrt{kC_2})^2 + 2C_1 d \log d (\kappa \log d \sqrt{kC_2})} \right)$$

$$\leq \exp \left( -\frac{n}{C_3 n^2} \right)$$

$$\leq \exp (-n^{z_2}) \quad \text{(for some universal constant } z_2 \in (0, 1))$$

In summary, we reach the conclusion of the lemma.

Lemma 18. (Lemma 10 re-stated) Let $A_i$ 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 $\frac{1}{\kappa \log d \sqrt{kC_2}}\|U\|_F$. If $U$ symmetric is of rank $k$ and is in a bounded spectral norm ball of radius $R$ (i.e. $\|U\|_2 \leq R$), then we have

$$\Pr \left( \frac{1}{n} \sum_{i=1}^{n} (\langle A_i, U \rangle A_i - U) \right) \leq \sqrt{\frac{d \log d}{n \sqrt{kR}}} \geq 1 - \exp (-C_2 d)$$

and

$$\Pr \left( \sup_{U: \|U\|_2 \leq R} \left\| \frac{1}{n} \sum_{i=1}^{n} (\langle A_i, U \rangle A_i - U) \right\|_2 \leq \sqrt{\frac{d \log d}{n \sqrt{kR}}} \right) \geq 1 - \exp (-C_2 kd) .$$

**Proof**

By the standard symmetrization argument,

$$\Pr \left( \sup_{U: \|U\|_2 \leq R} \left\| \frac{1}{n} \sum_{i=1}^{n} (\langle A_i, U \rangle A_i - U) \right\|_2 \geq t \right)$$

$$\leq 2\Pr \left( \sup_{U: \|U\|_2 \leq R} \left\| \frac{1}{n} \sum_{i=1}^{n} \tau_i (\langle A_i, U \rangle A_i) \right\|_2 > t/2 \right)$$

$$= 2\Pr \left( \sup_{U: \|U\|_2 \leq R} \sup_{X: \|X\|_2 \leq 1} \frac{1}{n} \sum_{i=1}^{n} \tau_i \langle A_i, U \rangle \langle A_i, xx^\top \rangle > t/2 \right),$$

where $\tau_i$s are independent Rademacher random variable. Note that $\langle A_i, U \rangle$ is sub-Gaussian with Orlicz norm $O(\|U\|_F)$ and $\langle A_i, xx^\top \rangle$ is sub-Gaussian with Orlicz norm $O(1)$. Hence the product is hence sub-exponential, with Orlicz norm $O(\|U\|_F)$.

Now we need to study the tail behavior by looking at the moment generating function. First we need the following Lemma:
Lemma 19. (Lemma 5.15 in Vershynin (2011)) Let $X$ be a centered sub-exponential random variable. Then, for $t$ such that $t \leq c/\|X\|_{\psi_1}$, one has

$$E \left[ \exp \left( tX \right) \right] \leq \exp \left( C t^2 \|X\|^2_{\psi_1} \right).$$

Hence for $\lambda \leq c/\sqrt{k}$

$$E \left[ \exp \left( \sup_{\|U\|_2 \leq R} \sup_{\|x\|_2 \leq 1} \frac{\lambda}{n} \sum_{i} \tau_i \langle A_i, U \rangle \langle A_i, xx^\top \rangle \right) \right]$$

$$= E \left[ \exp \left( \sup_{\|U\|_2 \leq 1} \sup_{\|x\|_2 \leq 1} \frac{\lambda R}{n} \sum_{i} \tau_i \langle A_i, U \rangle \langle A_i, xx^\top \rangle \right) \right]$$

$$\leq \exp \left( \frac{C_1 \lambda^2 k R^2}{n} + C_2 (k + 1) d \right),$$

where the factor of $(k + 1)d$ comes from the standard $\epsilon$-net argument over $\{U : \|U\|_2 \leq 1\}$ and $\{x : \|x\|_2 \leq 1\}$. By Chernoff inequality,

$$P \left( \sup_{\|U\|_2 \leq R} \sup_{\|x\|_2 \leq 1} \frac{1}{n} \sum_{i} \tau_i \langle A_i, U \rangle \langle A_i, xx^\top \rangle > t/2 \right)$$

$$\leq \exp \left( \frac{C_3 \lambda^2 k R^2}{n} + C_2 (k + 1) d - \lambda t/2 \right).$$

Let $\lambda = \frac{nt}{4C_1 k R^2}$ and choose $t > C_3 \sqrt{kd} \sqrt{kR}$. Then

$$P \left( \sup_{\|U\|_2 \leq R} \left\| \frac{1}{n} \sum_{i} \langle A_i, U \rangle A_i - U \right\|_2 \geq t \right)$$

$$\leq 2 P \left( \sup_{\|U\|_2 \leq R} \sup_{\|x\|_2 \leq 1} \frac{1}{n} \sum_{i} \tau_i \langle A_i, U \rangle \langle A_i, xx^\top \rangle > t/2 \right)$$

$$\leq \exp \left( - \frac{C_4 nt^2}{k R^2} + C_5 kd \right)$$

$$\leq \exp \left( - C_5 kd \right)$$

as long as we choose the constant $C_3$ large enough.

References


Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank mini-
mization and semidefinite programming from random linear measurements. *Advances in*

Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix com-