Journal of Machine Learning Research 25 (2024) 1-47 Submitted 12/21; Revised 10/23; Published 2/24

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

Jiacheng Zhuo* JZHUO@UTEXAS.EDU
Department of Computer Science
University of Texas
Austin, TX 78712, USA

Jeongyeol Kwon JEONGYEOL.KWON@WISC.EDU
Wisconsin Institute for Discovery
University of Wisconsin-Madison
Madison, WI 53705, USA

Nhat Ho MINHNHAT@UTEXAS.EDU
Department of Statistics and Data Sciences
University of Texas
Austin, TX 78712, USA

Constantine Caramanis CONSTANTINE@UTEXAS.EDU
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 $\mathbf{X}^* \in \mathbb{R}^{d\times d}$ is of rank r, but we try to recover it using $\mathbf{F}\mathbf{F}^\top$ where $\mathbf{F} \in \mathbb{R}^{d\times k}$ and $k > r$, the existing statistical analysis 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 $\mathbf{F}$ into separate column spaces to capture the impact of using $k > r$, we show that $\|\mathbf{F}_t\mathbf{F}_t - \mathbf{X}^*\|_F^2$ converges sub-linearly to a statistical error of $\tilde{\mathcal{O}}\left(kd\sigma^2/n\right)$ after $\tilde{\mathcal{O}}(\frac{\sigma_r}{\sigma}\sqrt{\frac{n}{d}})$ iterations, where $\mathbf{F}_t$ is the output of FGD after t iterations, σ^2 is the variance of the observation noise, σ_r is the r-th largest eigenvalue of $\mathbf{X}^*$, and n is the number of 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 $\{\mathbf{A}_i, y_i\}_{i=1}^n$ from the data generating model $y_i = \langle \mathbf{A}_i, \mathbf{X}^* \rangle + \epsilon_i$, where $\mathbf{A}_i \in \mathbb{R}^{d\times d}$ is a

symmetric random sensing matrix, $\mathbf{X}^* \in \mathbb{R}^{d\times d}$ is the target rank r symmetric matrix we want to recover, and ϵ_i is a zero-mean sub-Gaussian noise with variance proxy σ^2. The low rank matrix sensing problem has found applications in various scenarios, such as multi-task regression, vector auto-regressive process, image processing, metric embedding, quantum tomography, and so on (Candes and Plan, 2011; Negahban and Wainwright, 2011; Recht et al., 2010; Jain et al., 2013; Gross et al., 2010; Candès et al., 2011; Waters et al., 2011; Kalev et al., 2015). One common approach to recover a low-rank matrix $\mathbf{X} \in \mathbb{R}^{d\times d}$ is to solve the following optimization problem:

$$\underset{\mathbf{X}:\mathbf{X}\succeq 0,\mathrm{rank}(\mathbf{X})\leq k}{\arg\min} \frac{1}{4n}\sum_{i=1}^{n}\left(y_i - \langle \mathbf{A}_i, \mathbf{X}\rangle\right)^2, \tag{1}$$

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 $\mathbf{F} \in \mathbb{R}^{d\times k}$ such that $\mathbf{X} = \mathbf{F}\mathbf{F}^\top$ and solving the transformed problem (Bhojanapalli et al., 2016a; Chen and Wainwright, 2015; Jain et al., 2013; Hardt, 2014; Park et al., 2018):

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

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

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

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

When the specified rank k matches the ground truth rank r, namely, the true rank r is known, FGD converges linearly to a statistical error (Chen and Wainwright, 2015), and the statistical error is minimax optimal up to log factors (Candes and Plan, 2011). 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 $\mathbf{X}^*$; 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, $\left\|\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\right\|_F^2$ converges to a final statistical error of $\tilde{\mathcal{O}}\left(kd\sigma^2/n\right)$ after $\tilde{\mathcal{O}}(\frac{\sigma_r}{\sigma}\sqrt{\frac{n}{d}})$ number of iterations, namely sub-linearly, where σ_r and σ are, respectively, the r-th largest eigenvalue of $\mathbf{X}^*$ and the standard deviation of the observation noise. It is different from the computational and statistical behavior of FGD when the true rank is known. In that setting, FGD converges linearly to a radius of convergence $\tilde{\mathcal{O}}(rd\sigma^2/n)$ around the true matrix $\mathbf{X}^*$ after $\mathcal{O}(\log(\frac{\sigma_r}{\sigma_1}\cdot\frac{n}{d}))$ iterates (Chen and Wainwright, 2015), where σ_1 is the largest eigenvalue of $\mathbf{X}^*$. Furthermore, the number of iterations $\tilde{\mathcal{O}}(\frac{\sigma_r}{\sigma}\sqrt{\frac{n}{d}})$ is needed in the over-parameterized setting as the local curvature of the loss function (1) around the global maxima is not quadratic and therefore the FGD only converges sub-linearly to the global maxima; see the simulations in Figure 1 for an illustration. Finally, when $\sigma = 0$, i.e., in the noiseless case, we can guarantee the exact recovery similar to when we correctly specify the rank (Chen and Wainwright, 2015).

1.1 Related Work

Works related to Matrix Sensing. Early works on matrix sensing often perform a semidefinite programming (SDP) relaxation, and replace the nonconvex rank constraint with a convex constraint based on the trace norm or nuclear norm; see for example (Candes and Plan, 2011; Recht et al., 2010; Negahban and Wainwright, 2011; Chen et al., 2013) and the references therein. Candes and Plan (2011) show that for any estimator $\hat{\mathbf{X}}$ based on $\{\mathbf{A}_i, y_i\}_{i=1}^n$ observations, $\left\|\hat{\mathbf{X}} - \mathbf{X}^*\right\|_F^2 \geq \frac{dr}{n}\sigma^2$, where $\mathbf{X}^*$ is the ground truth rank r matrix that we want to recover, and σ is the standard deviation of the (sub)-Gaussian observation noise (see Section 1.4 for details). This convex relaxation approach is nearly

optimal in 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).

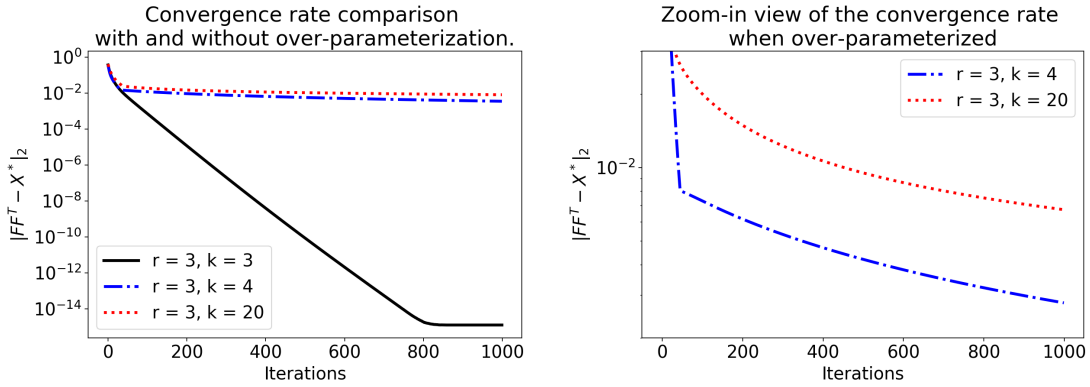
It is also worth mentioning that the low rank matrix sensing problem is tightly connected to the low rank matrix completion problem, since they have the same population update when solved by (factorized) gradient method, and they can often be analyzed by very similar techniques (Negahban and Wainwright, 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 $\mathbf{X}^*$, 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



(a) (b)

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

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

Let $\{\mathbf{F}_t\}_t$ be the sequence generated by the FGD method as in equation (3) with $\eta = 0.1$. The simulation results are shown in Figure 1. When we correctly specify the rank

(i.e., $k = r = 3$), the FGD method 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 σ_1 and σ_r to denote the first eigenvalue and the r-th eigenvalue of $\mathbf{X}^*$ respectively, which is the ground truth rank r matrix that we want to recover. And we use κ to denote the conditional number: $\kappa := \sigma_1 / \sigma_r$.

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

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

Definition 2. ***(Sub-Gaussian Sensing Matrix).*** *We call a matrix $\mathbf{A}$ a sub-Gaussian sensing matrix if it is sampled as follow:*

$$\mathbf{A}_i = \frac{1}{2}(R_i + R_i^\top),$$

where $R_i \in \mathbb{R}^{d \times d}$, and all elements of 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 $\mathbf{X}^*$. Let the eigen-decomposition of $\mathbf{X}^*$ (eigenvalues ordered by the absolute values) be given by

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

where $\mathbf{U} \in \mathbb{R}^{d*r}$, $\mathbf{V} \in \mathbb{R}^{d*(d-r)}$, $\mathbf{D}_{\mathbf{S}}^* \in \mathbb{R}^{r*r}$, $\mathbf{D}_{\mathbf{T}}^* \in \mathbb{R}^{(d-r)*(d-r)}$. Without loss of generality we assume that the both $\mathbf{U}$ and $\mathbf{V}$ are orthonormal and $\mathbf{U}^\top\mathbf{V} = 0$ (i.e., $\mathbf{U}$ and $\mathbf{V}$ together span the entire $\mathbb{R}^d$). Denote σ_1 be the largest value in $\mathbf{D}_{\mathbf{S}}^*$, σ_r be the smallest value in $\mathbf{D}_{\mathbf{S}}^*$, and σ_{r+1} be the largest value in $\mathbf{D}_{\mathbf{T}}^*$. If $\mathbf{X}^*$ is exactly rank r, then $\mathbf{D}_{\mathbf{T}}^*$ is a zero matrix. For generality, we allow $\mathbf{X}^*$ to be only approximately rank r, where we assume that $\mathbf{D}_{\mathbf{T}}^*$ is nonzero but σ_{r+1} is very small. Since the union of the column space of $\mathbf{U}$ and $\mathbf{V}$ spans the entire $\mathbb{R}^d$, then for any $\mathbf{F}_t \in \mathbb{R}^{d*k}$, there exist matrices $\mathbf{S}_t \in \mathbb{R}^{r*k}$ and $\mathbf{T}_t \in \mathbb{R}^{(d-r)*k}$ such that

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

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

We introduce the decomposition and study the convergence of $\mathbf{S}_t\mathbf{S}_t^\top$, $\mathbf{T}_t\mathbf{T}_t^\top$, and $\mathbf{S}_t\mathbf{T}_t^\top$ separately. This decomposition technique is essential, since we can then bypass some technical difficulties when we over-specify the rank. For example we do not have to establish the uniqueness (up to rotational ambiguity) of the optimal solution as in the Lemma 1 in Chen and Wainwright (2015). Moreover, this gives more insights about which part is the computational and/or statistical bottleneck. As we will see shortly (both in Theorem 4 and Lemma 7), it is the convergence of $\left\{\left\|\mathbf{T}_t\mathbf{T}_t^\top - \mathbf{D}_{\mathbf{T}}^*\right\|_2\right\}_t$ that slows down the entire process of the convergence. Similar decomposition technique is also employed in the work of Li et al. (2018).

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

Assumption 1. ***(Initialization assumption)***

$$\left\|\mathbf{F}_0\mathbf{F}_0^\top - \mathbf{X}^*\right\|_2 \leq \rho\sigma_r, \quad \textit{for} \quad \rho \leq 0.07. \tag{4}$$

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 $\max\left\{\left\|\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}_0\mathbf{T}_0^\top\right\|_2, \left\|\mathbf{D}_{\mathbf{S}}^* - \mathbf{S}_0\mathbf{S}_0^\top\right\|_2, \left\|\mathbf{S}_0\mathbf{T}_0^\top\right\|_2\right\}$ in our analysis. The following lemma establishes the connection between what we need in the analysis and Assumption 1.

Lemma 3. *If* $\left\|\mathbf{F}_0\mathbf{F}_0^\top - \mathbf{X}^*\right\|_2 \leq 0.7\rho\sigma_r$, *then*

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

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)* $\|\mathbf{D}_{\mathbf{T}}^*\|_2 < \sqrt{\frac{d\log d}{n}}\sigma$*; (2) we have good initialization as in Assumption 1; (3) the step size* $\eta = \frac{1}{100\sigma_1}$*, (4)* $\mathbf{A}_i$*'s are sub-Gaussian sensing matrices. Let* $\{\mathbf{F}_t\}_t$ *be the sequence generated by the FGD algorithm as in equation* (3)*. Then, the following holds:*

(a) With sample size $n > C_1 k\kappa^2 d\log^3 d \cdot \max(1, \sigma^2/\sigma_r^2)$ *for some universal constant* C_1*, after* $t > \left\lceil 2\log\frac{\sigma_r}{\epsilon_{comp}} \right\rceil$ *steps, we have*

$$\max\left\{\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2, \left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2\right\} < C\epsilon_{comp}$$

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

(b) With sample size $n > C_1' k^2\kappa^2 d(\log^3 d)\cdot\max(1, \sigma^2/\sigma_r^2)$ *for some universal constant* C_1'*, after* $t \geq \Theta\left(\frac{\sigma_1}{\epsilon_{stat}}\right)$ *steps, we find that*

$$\max\left\{\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2, \left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2, \left\|\mathbf{T}_t\mathbf{T}_t^\top - \mathbf{D}_{\mathbf{T}}^*\right\|_2\right\} < C_2'\epsilon_{stat},$$

and $\left\|\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\right\|_2 \leq C_3'\epsilon_{stat}$ *for some universal constants* C_2'*, and* C_3'*, where* $\epsilon_{stat} := \kappa\sqrt{\frac{d\log d}{n}}\sigma$*.*

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

(1) The sequences $\left\{\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2\right\}_t$ *and* $\left\{\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2\right\}_t$ *first converge linearly and then sub-linearly.* Theorem 4 indicates that the sequences $\left\{\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2\right\}_t$ and $\left\{\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2\right\}_t$ first converge linearly from $0.1\sigma_r$ to ϵ_{comp}, and then converge sub-linearly to $\Omega(\epsilon_{stat})$. Furthermore, the sequence $\left\{\left\|\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\right\|_2\right\}_t$ always converges sublinearly towards $\Omega(\epsilon_{stat})$. This is consistent with our simulations in Figure 2. As we will see later in Lemma 7, it is the convergence of $\left\|\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}_t\mathbf{T}_t^\top\right\|_2$ that slows down the convergence of $\left\{\left\|\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\right\|_2\right\}_t$, and incurring the sublinear convergence of $\left\{\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2\right\}_t$ and $\left\{\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2\right\}$.

(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 $\left\{\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2\right\}_t$ and $\left\{\left\|\mathbf{S}_t\mathbf{T}_t^T\right\|_2\right\}_t$ are linear all the way until zero. In our setting, going from population to finite-sample creates an unusual tangling factor, causing the convergence rate discrepancy between the finite-sample and population sequences.

(3) 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 $\kappa\sqrt{\frac{kd\log d}{n}}\sigma$ if we evaluate $\left\|\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\right\|_F$. 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 $\Omega\left(k\kappa^2 d\log^3 d\cdot\max(1,\sigma^2/\sigma_r^2)\right)$, a caveat is that, in order to achieve this statistical error, we do require $n > C_1k^2\kappa^2 d\log^3 d\cdot\max(1,\sigma^2/\sigma_r^2)$ 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 $\|\mathbf{D}_\mathbf{T}^*\|_2 = 0$ while the other assumptions on the initialization, the step size, and the sensing matrices are as in Theorem 4. Let $\{\mathbf{F}_t\}_t$ be the sequence generated by the FGD algorithm as in equation* (3)*. Then, the following holds:*

(a) With $n > C_1k\kappa^2 d\log^3 d$ for some universal constant C_1, after $t > \left\lceil 2\log\frac{\sigma_r}{\epsilon_{comp\_noiseless}}\right\rceil$ steps, we have

$$\max\left\{\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\right\|_2, \left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2\right\} < C\epsilon_{comp\_noiseless}$$

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

(b) For any $\epsilon > 0$, with $n > C_1'k^2\kappa^2 d(\log^3 d)$ for some universal constant C_1', after $t \geq \Theta\left(\frac{\sigma_1}{\epsilon}\right)$ steps, we find that

$$\max\left\{\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\right\|_2, \left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2, \left\|\mathbf{T}_t\mathbf{T}_t^\top - \mathbf{D}_\mathbf{T}^*\right\|_2\right\} < C_2'\epsilon.$$

and $\left\|\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\right\|_2 \leq C_3'\epsilon$ 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 $\left\{\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_\mathbf{S}^*\right\|_2\right\}_t$ and $\left\{\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2\right\}_t$ 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.

(5) 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 $\mathbf{X}^*$. In particular, for the IHT methods, the per iteration cost is $\mathcal{O}(nd^2)$ where n is the sample size and d is the dimension while for the FGD, the per iteration cost is $\mathcal{O}(ndk)$ where k is the chosen rank. Therefore, based on Theorem 2 of (Jain et al., 2014a), with the computational complexity $\mathcal{O}(kd^2)$ from the spectral method for the initialization, the total computational complexity for the IHT methods to reach the statistical radius $\mathcal{O}(kd\sigma^2/n)$ around the true matrix X^* is at the order of $\mathcal{O}(kd^2 + nd^2 \log(n/(kd\sigma^2)))$. 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 $\mathcal{O}(kd^2 + \frac{\sigma_r}{\sigma} n^{3/2}\sqrt{d}k)$. 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 $\{\mathbf{F}_t\}_t$ be the sequence generated by the FGD method as in Equation (3), and let $\mathbf{S}, \mathbf{T}$ be defined as in the previous subsection.

The simulation results are shown in Figure 2. In Figure 2a, we plot $\left\|\mathbf{F}\mathbf{F}^\top - \mathbf{X}^*\right\|_2$, $\left\|\mathbf{S}\mathbf{S}^\top - \mathbf{D}^*_{\mathbf{S}}\right\|_2$, $\left\|\mathbf{S}\mathbf{T}^\top\right\|_2$, and $\left\|\mathbf{T}\mathbf{T}^\top - \mathbf{D}^*_{\mathbf{T}}\right\|_2$ against the algorithm iterations. The simulation results are aligned with our theory. As said in Theorem 4, $\left\|\mathbf{S}\mathbf{S}^\top - \mathbf{D}^*_{\mathbf{S}}\right\|_2$ and $\left\|\mathbf{S}\mathbf{T}^\top\right\|_2$ first converge linearly, and then sublinearly. Furthermore, $\left\|\mathbf{F}\mathbf{F}^\top - \mathbf{X}^*\right\|_2$ is soon dominated by $\left\|\mathbf{T}\mathbf{T}^\top - \mathbf{D}^*_{\mathbf{T}}\right\|_2$, 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

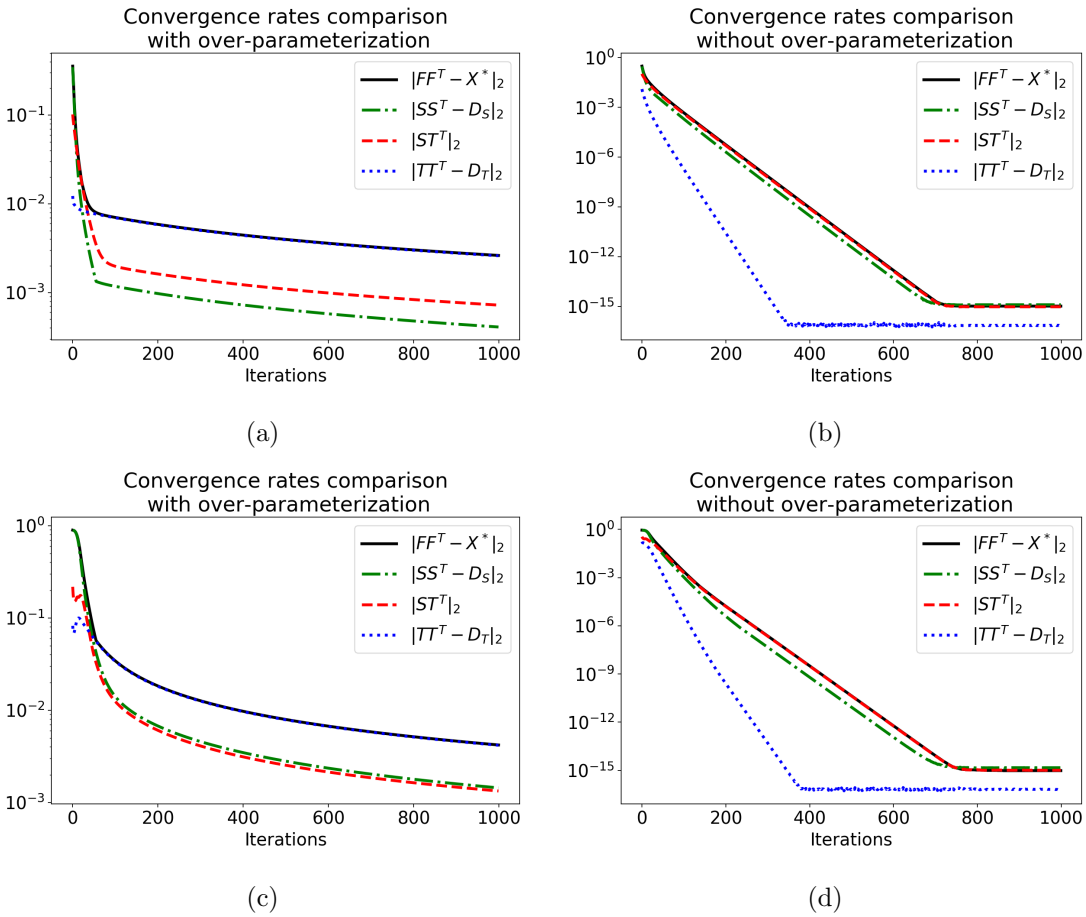


Figure 2. Simulations that verify the main result. (a) Convergence rates of the FGD iterates when we over-specify the rank ($r = 3, k = 4$). (b) Convergence rates of the FGD method when we correctly specify the rank ($r = k = 3$). The Figures in (c) and (d) are executed in the same setting as those in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

$\left\|\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\right\|_2$ 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 $\left\|\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\right\|_F$ directly, and hence we can use the general RIP notion. We leave this for future work.

3.1 Population analysis

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

expectation over the observation noise ϵ and the observation matrices $\mathbf{A}_i$) is

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

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

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

where we define the following operators:

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

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

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

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

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

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

According to Lemma 6 above, we have fast convergence in estimating $\mathbf{S}\mathbf{S}^\top$, $\mathbf{S}\mathbf{T}^\top$, but slow convergence in estimating $\mathbf{T}\mathbf{T}^\top$. Intuitively, $\mathbf{T}\mathbf{T}^\top$ is slow because the local curvature of the population version of the loss function (2) is flat, namely, the Hessian matrix around the global maxima $\mathbf{D}_{\mathbf{T}}^*$ is degenerate. We know that when the curvature of the target matrix is undesirable, we can only guarantee sub-linear convergence rate (Bhojanapalli et al., 2016a).

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

3.2 Finite sample analysis

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

population and sample gradient as follows:

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

We define Δ_t to quantify this deviation:

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

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

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

$$\|\Delta_t\|_2 \le 5\sqrt{\frac{kd\log d}{n}}D_t + \sqrt{\frac{d\log d}{n}}\sigma. \tag{5}$$

Then, we find that

$$\begin{aligned}\left\|\mathbf{S}_{t+1}\mathbf{S}_{t+1}^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 \le& \left(1 - \frac{7}{10}\eta\sigma_r\right)\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 + \sqrt{\frac{kd\log d}{n}}D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma & (6)\\ \le& \left(1 - \frac{7}{10}\eta\sigma_r\right)\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 + \frac{1}{10}\eta\sigma_r D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma, & (7)\\ \left\|\mathbf{S}_{t+1}\mathbf{T}_{t+1}^\top\right\|_2 \le& (1 - \eta\sigma_r)\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2 + \sqrt{\frac{kd\log d}{n}}D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma & (8)\\ \le& (1 - \eta\sigma_r)\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2 + \frac{1}{10}\eta\sigma_r D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma. & (9)\end{aligned}$$

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

$$(D_{t+1} - 50\epsilon_{stat}) \le \left[1 - \frac{1}{2}\eta\left(D_t - 50\epsilon_{stat}\right)\right]\left(D_t - 50\epsilon_{stat}\right). \tag{10}$$

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 $\left\{\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2\right\}_t$ 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 $\left\{\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2\right\}_t$ still has constant contraction, until roughly the magnitude of $\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2$ 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 ϵ-net discretization techniques.

Lemma 8. *Let $\mathbf{A}_i$ be sub-Gaussian sensing matrices. Let ϵ_i follows $N(0,\sigma)$. Then*

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

Lemma 9. *Let $\mathbf{A}_i$ be sub-Gaussian sensing matrices. Let $\mathbf{U}$ be a deterministic symmetric matrix of the same dimension. Then as long as $n > C_1 d\log^3 d$ for some universal $C_1, C_2 > 10$, we have*

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

The proof of the above two concentration results can be found in the Appendix D. If we invoke these lemmas for Δ_t to ensure that equation (5) holds, then we can immediately have

$$\begin{aligned}
\|\Delta_t\|_2 &= \left\|\frac{1}{n}\sum_i^n \left\langle \mathbf{A}_i, \mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\right\rangle \mathbf{A}_i - (\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*) + \frac{1}{n}\sum_i^n \epsilon_i\mathbf{A}_i\right\|_2 \\
&\leq 5\sqrt{\frac{kd\log d}{n}}D_t + \sqrt{\frac{d\log d}{n}}\sigma \\
&\leq 0.5\eta\sigma_r D_t + \sqrt{\frac{d\log d}{n}}\sigma
\end{aligned}$$

where we use Lemma 9 with $\mathbf{U} = \mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*$ and Assumption 1 with converting it to Frobenius bound (note that $\|\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\|_2 \leq D_t$), and the last inequality holds because we set $n > C_1 k\kappa^2 d\log^3 d \cdot \max(1, \sigma^2/\sigma_r^2)$.

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

Lemma 10. *Let $\mathbf{A}_i$ be sub-Gaussian sensing matrices. If $\mathbf{U}$ is of rank k and is in a bounded spectral norm ball of radius R (i.e., $\|\mathbf{U}\|_2 \leq R$),*

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

With $n > C_1 k^2 \kappa^2 d \log^3 d \cdot \max(1, \sigma^2/\sigma_r^2)$, we can apply Lemma 10 with different values of R. For instance, starting from $R_0 = \sigma_r$, we can apply Lemma 10 with $R_0/2$, $R_0/4$, ..., until reaching ϵ_{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 $\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2$ and $\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2$ respectively. To make the argument more precise, we consider $\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 > 1000\sqrt{\frac{k\kappa^2 d\log d}{n}}\sigma_r > 1000\sqrt{\frac{k\kappa^2 d\log d}{n}}D_t$. Then, we find that $\sqrt{\frac{kd\log d}{n}}D_t \leq 0.1\eta\sigma_r \left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2$ since $\eta\sigma_r = 0.01/\kappa$. Also, $\frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma < 0.1\eta\sigma_r \left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2$ by the choice of the constants in the lower bound of n. Hence, when $\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 > 1000\sqrt{\frac{k\kappa^2 d\log d}{n}}\sigma_r$, we find that

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

The same arguments hold for $\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2$. 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 $\Theta(\epsilon_{stat})$ after $\mathcal{O}(1/\epsilon_{stat})$ iterations. By equation (10), we know that $A_{t+1} \leq \left(1 - \frac{1}{2}\eta A_t\right) A_t$ where $A_t = D_t - 50\epsilon_{stat}$. Hence, we obtain that

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

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

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

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 $\mathbf{F}_0$ such that $\left\|\mathbf{F}_0\mathbf{F}_0^\top\right\|_2 \leq \delta$ for some small constant δ. 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 $\mathbf{S}_0\mathbf{S}_0^\top$ for the theoretical analysis. Such a lower bound is automatically satisfied when $k = d$ and we initialize $\mathbf{F}_0$ as $\delta\mathbf{I}$ where $\mathbf{I}$ 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 $\mathbf{S}_0\mathbf{S}_0^\top$.

Extension to asymmetric matrices. In this work we focus on the true symmetric matrix $\mathbf{X}^*$ 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 $\mathbf{X}^*$ 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 $\mathbf{X}^* \in \mathbb{R}^{d*d}$ by minimizing a convex function $f\left(\mathbf{X}\right) =$

$f\left(\mathbf{F}\mathbf{F}^\top\right)$ where $\mathbf{F} \in \mathbb{R}^{d*k}$ and $k > r$. If the population gradient with respect to $\mathbf{X}$ is $\mathbf{X} - \mathbf{X}^*$, 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 $\mathbf{X}$ is explicit and we can still decompose the update process of $\{\mathbf{F}_t\}_t$ into the strong signal part (namely $\{\mathbf{S}_t\}_t$) and the weak signal part (namely $\{\mathbf{T}_t\}_t$), 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 $\{\mathbf{F}_t\}_t$. However for some low-rank problems it is desirable to have constraints on $\{\mathbf{F}_t\}_t$. For example for matrix completion we wish all the iterations $\{\mathbf{F}_t\}_t$ 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., $\left\|\mathbf{\Pi}(\mathbf{F})\mathbf{\Pi}(\mathbf{F})^\top - \mathbf{X}^*\right\|_2 \leq \left\|\mathbf{F}\mathbf{F}^\top - \mathbf{X}^*\right\|_2$ where $\mathbf{\Pi}$ 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 $\left\|\mathbf{\Pi}(\mathbf{F})\mathbf{\Pi}(\mathbf{F})^\top - \mathbf{X}^*\right\|_F \leq \left\|\mathbf{F}\mathbf{F}^\top - \mathbf{X}^*\right\|_F$). To get around this issue, one option is to show that all $\{\mathbf{F}_t\}_t$ 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 $\tilde{\mathcal{O}}(dr)$ samples, where r is the rank of the ground truth matrix $\mathbf{X}^*$, while our analysis requires $\tilde{\mathcal{O}}(dk) = \tilde{\mathcal{O}}(d^2)$ samples when $k = d$. Since they only require $\tilde{\mathcal{O}}(dr)$ samples, they cannot control the error of the over-parameterization part (equivalent to our $\mathbf{T}\mathbf{T}^\top$ part). In fact in their analysis, they only show that in a limited number of steps, this error does not blow up. While with $\tilde{\mathcal{O}}(dk)$ 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:

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

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

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

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

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

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

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

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

$$\begin{aligned}&\mathbf{z}^{\top}\left(\mathbf{I}-2\eta\mathbf{S}\mathbf{S}^{\top}-\eta^{2}\mathbf{S}\mathbf{S}^{\top}\left(\mathbf{D}_{\mathbf{S}}^{*}-\mathbf{S}\mathbf{S}^{\top}\right)+2\eta^{2}\mathbf{S}\mathbf{T}^{\top}\mathbf{T}\mathbf{S}^{\top}\right)\mathbf{z}\\&\qquad\leq1-2\eta\left\|\mathbf{S}\mathbf{z}\right\|_{2}^{2}+\eta^{2}\left\|\mathbf{S}\mathbf{S}^{\top}\right\|_{2}\left\|\mathbf{D}_{\mathbf{S}}^{*}-\mathbf{S}\mathbf{S}^{\top}\right\|_{2}+2\eta^{2}\left\|\mathbf{T}\mathbf{T}^{\top}\right\|_{2}\left\|\mathbf{S}\mathbf{z}\right\|_{2}^{2}\\&\qquad\overset{(i)}{\leq}1-2\eta\sigma_{r}+3\eta^{2}\sigma_{r}\sigma_{1}\\&\qquad\overset{(ii)}{\leq}1-\eta\sigma_{r},\end{aligned}$$

where in step (i) we used $0.9\sigma_{r}\mathbf{I}\preceq\mathbf{S}\mathbf{S}^{\top}\preceq(\sigma_{1}+0.1\sigma_{r})\mathbf{I}$, $\left\|\mathbf{D}_{\mathbf{S}}^{*}-\mathbf{S}\mathbf{S}^{\top}\right\|_{2}\leq0.1\sigma_{r}$ and $\left\|\mathbf{T}\mathbf{T}^{\top}\right\|_{2}\leq1.1\sigma_{r}$ by initialization condition and triangular inequality; step (ii) follows from choice of step size $\eta=\frac{1}{100\sigma_{1}}$, and definition of the conditional number $\kappa=\sigma_{1}/\sigma_{r}$. Therefore, we arrive at the following inequality:

$$\left\|\mathbf{I}-2\eta\mathbf{S}\mathbf{S}^{\top}-\eta^{2}\mathbf{S}\mathbf{S}^{\top}\left(\mathbf{D}_{\mathbf{S}}^{*}-\mathbf{S}\mathbf{S}^{\top}\right)+2\eta^{2}\mathbf{S}\mathbf{T}^{\top}\mathbf{T}\mathbf{S}^{\top}\right\|_{2}\leq1-\eta\sigma_{r}.\tag{11}$$

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

$$\|II\|_2 \leq \left\|2\eta\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top\right\|_2 + \left\|\eta^2\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top\right\|_2 \leq 3\eta\left\|\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top\right\|_2. \tag{12}$$

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

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

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

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

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

$$\begin{aligned}
&\mathcal{M}_{\mathbf{S}}(\mathbf{S})\mathcal{M}_{\mathbf{T}}(\mathbf{T})^\top \\
&\quad= \left(\mathbf{S} + \eta\left(\mathbf{D}_{\mathbf{S}}^* - \mathbf{S}\mathbf{S}^\top\right)\mathbf{S} - \eta\mathbf{S}\mathbf{T}^\top\mathbf{T}\right)\left(\mathbf{T} + \eta\left(\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}\mathbf{T}^\top\right)\mathbf{T} - \eta\mathbf{T}\mathbf{S}^\top\mathbf{S}\right)^\top \\
&\quad= \frac{1}{2}\left(\mathbf{I} - 2\eta\mathbf{S}\mathbf{S}^\top + 2\eta\left(\mathbf{D}_{\mathbf{S}}^* - \mathbf{S}\mathbf{S}^\top\right) - 2\eta^2\left(\mathbf{D}_{\mathbf{S}}^* - \mathbf{S}\mathbf{S}^\top\right)\mathbf{S}\mathbf{S}^\top + 2\eta^2\mathbf{S}\mathbf{T}^\top\mathbf{T}\mathbf{S}^\top\right)\mathbf{S}\mathbf{T}^\top \\
&\qquad+ \frac{1}{2}\mathbf{S}\mathbf{T}^\top\left(\mathbf{I} + 2\eta\left(\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}\mathbf{T}^\top\right) - 2\eta\mathbf{T}\mathbf{T}^\top - 2\eta^2\mathbf{T}\mathbf{T}^\top\left(\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}\mathbf{T}^\top\right)\right) \\
&\qquad+ \eta^2\left(\mathbf{D}_{\mathbf{S}}^* - \mathbf{S}\mathbf{S}^\top\right)\mathbf{S}\mathbf{T}^\top\left(\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}\mathbf{T}^\top\right).
\end{aligned} \tag{13}$$

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

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

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

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

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

Lastly, for the second order term in the third line of the RHS of equation (13) we have

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

Plugging the above results into equation (13) leads to

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

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

Proof of the contraction result (c) in Lemma 6: We would like to establish that

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

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

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

where we use the following shorthand notation:

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

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

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

Given these bounds, we find that

$$\|IV\|_2 \leq \left(\eta + 3\eta^2\sigma_r\right) \|\mathbf{D}_{\mathbf{T}}^*\|_2 \left\|\mathbf{T}\mathbf{T}^\top\right\|_2. \tag{14}$$

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

$$\begin{aligned}
&-2\eta\mathbf{T}\mathbf{S}^\top\mathbf{S}\mathbf{T}^\top - \eta^2\mathbf{T}\mathbf{S}^\top\mathbf{S}\mathbf{T}^\top\left(\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}\mathbf{T}^\top\right)^\top - \eta^2\left(\mathbf{D}_{\mathbf{T}}^* - \mathbf{T}\mathbf{T}^\top\right)\mathbf{T}\mathbf{S}^\top\mathbf{S}\mathbf{T}^\top + \eta^2\mathbf{T}\mathbf{S}^\top\mathbf{S}\mathbf{S}^\top\mathbf{S}\mathbf{T}^\top \\
&\preceq \left(-2\eta + 2\eta^2\rho\sigma_r + \eta^2\sigma_1\right)\mathbf{T}\mathbf{S}^\top\mathbf{S}\mathbf{T}^\top \preceq 0.
\end{aligned} \tag{15}$$

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

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

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

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

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

$$\left\|\mathbf{T}\mathbf{T}^\top - \eta\left(\mathbf{T}\mathbf{T}^\top\right)^2\right\|_2 \leq \left\|\mathbf{T}\mathbf{T}^\top\right\|_2\left(1 - \eta\left\|\mathbf{T}\mathbf{T}^\top\right\|_2\right). \tag{16}$$

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

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

As long as $\|\mathbf{D}_{\mathbf{T}}^*\|_2 \ll \left\|\mathbf{T}\mathbf{T}^\top\right\|_2$, the contraction rate is roughly $(1 - \eta\left\|\mathbf{T}\mathbf{T}^\top\right\|_2)$. Therefore, we obtain the conclusion of claim (c) in Lemma 6.

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

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

where we denote VI and VII as follows:

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

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

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

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

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

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

Now for the VI term, direct calculation shows that

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

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

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

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

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

As a consequence, we reach the conclusion of claim (d) in Lemma 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) $\left\|\mathbf{D}_{\mathbf{S}}^{*}-\mathcal{M}_{\mathbf{S}}(\mathbf{S})\mathbf{S}^{\top}\right\|_{2} \leq (1-\eta\sigma_r)\left\|\mathbf{D}_{\mathbf{S}}^{*}-\mathbf{S}\mathbf{S}^{\top}\right\|_{2}+\eta\left\|\mathbf{S}\mathbf{T}^{\top}\right\|_{2}^{2}$,

(b) $\left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S})\mathbf{T}^{\top}\right\|_{2} \leq \left\|\mathbf{S}\mathbf{T}^{\top}\right\|_{2}$,

(c) $\left\|\mathcal{M}_{\mathbf{T}}(\mathbf{T})\mathbf{S}^{\top}\right\|_{2} \leq \left\|\mathbf{S}\mathbf{T}^{\top}\right\|_{2}$,

(d) $\left\|\mathcal{M}_{\mathbf{T}}(\mathbf{T})\mathbf{T}^{\top}\right\|_{2} \leq \left\|\mathbf{T}\mathbf{T}^{\top}\right\|_{2}+\eta\left\|\mathbf{S}\mathbf{T}^{\top}\right\|_{2}^{2}$.

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

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

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

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

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

By initialization condition and triangular inequality, we know that $0 \leq \left\|\left(\mathbf{S}\mathbf{S}^{\top}-\mathbf{D}_{\mathbf{S}}^{*}\right)\right\|_{2} \leq \rho\sigma_r$ and $0 \leq \left\|\mathbf{T}\mathbf{T}^{\top}\right\|_{2} \leq 1.1\sigma_r$, and hence $\left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S})\mathbf{T}^{\top}\right\|_{2} \leq \left\|\mathbf{S}\mathbf{T}^{\top}\right\|_{2}$. Hence, we reach the conclusion of claim (b).

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

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

By initialization condition and triangular inequality, we know that $0 \leq \left\|\left(\mathbf{T}\mathbf{T}^{\top}-\mathbf{D}_{\mathbf{T}}^{*}\right)\right\|_{2} \leq \rho\sigma_r$ and $0.9\sigma_r \leq \left\|\mathbf{S}\mathbf{S}^{\top}\right\|_{2} \leq 0.1\sigma_r+\sigma_1$, and hence $\left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S})\mathbf{T}^{\top}\right\|_{2} \leq \left\|\mathbf{S}\mathbf{T}^{\top}\right\|_{2}$. It leads to the conclusion of claim (c).

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

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

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

■

Appendix B. Proofs for the finite sample analysis

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

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

Then, we can write our update as follows:

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

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

$$\begin{aligned}
\mathbf{S}_{t+1}\left(\mathbf{S}_{t+1}\right)^\top =& \mathbf{U}^\top\mathbf{F}_{t+1}\left(\mathbf{U}^\top\mathbf{F}_{t+1}\right)^\top \\
=& \mathbf{U}^\top\left(\mathbf{F}_t - \eta\mathbf{G}_t\right)\left(\mathbf{F}_t - \eta\mathbf{G}_t\right)^\top\mathbf{U} + \eta^2\mathbf{U}^\top\left(\mathbf{G}_t - \mathbf{G}_t^n\right)\left(\mathbf{G}_t - \mathbf{G}_t^n\right)^\top\mathbf{U} \\
&+ \eta\mathbf{U}^\top\left(\mathbf{F}_t - \eta\mathbf{G}_t\right)\left(\mathbf{G}_t - \mathbf{G}_t^n\right)^\top\mathbf{U} + \eta\mathbf{U}^\top\left(\mathbf{G}_t - \mathbf{G}_t^n\right)\left(\mathbf{F}_t - \eta\mathbf{G}_t\right)^\top\mathbf{U} \\
=& \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)^\top + \eta^2\mathbf{U}^\top\left(\mathbf{G}_t - \mathbf{G}_t^n\right)\left(\mathbf{G}_t - \mathbf{G}_t^n\right)^\top\mathbf{U} \\
&+ \eta\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\left(\mathbf{G}_t - \mathbf{G}_t^n\right)^\top\mathbf{U} + \eta\mathbf{U}^\top\left(\mathbf{G}_t - \mathbf{G}_t^n\right)\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)^\top,
\end{aligned} \tag{17}$$

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

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

Furthermore, direct calculation shows that

$$\begin{aligned}
\mathbf{S}_{t+1}\left(\mathbf{T}_{t+1}\right)^\top =& \mathbf{U}^\top\mathbf{F}_{t+1}\left(\mathbf{V}^T\mathbf{F}_{t+1}\right)^\top \\
=& \mathbf{U}^\top\left(\mathbf{F}_t - \eta\mathbf{G}_t\right)\left(\mathbf{F}_t - \eta\mathbf{G}_t\right)^\top\mathbf{V} + \eta^2\mathbf{U}^\top\left(\mathbf{G}_t - \mathbf{G}_t^n\right)\left(\mathbf{G}_t - \mathbf{G}_t^n\right)^\top\mathbf{V} \\
&+ \eta\mathbf{U}^\top\left(\mathbf{F}_t - \eta\mathbf{G}_t\right)\left(\mathbf{G}_t - \mathbf{G}_t^n\right)^\top\mathbf{V} + \eta\mathbf{U}^\top\left(\mathbf{G}_t - \mathbf{G}_t^n\right)\left(\mathbf{F}_t - \eta\mathbf{G}_t\right)^\top\mathbf{V} \\
=& \mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top + \eta^2\mathbf{U}^\top\left(\mathbf{G}_t - \mathbf{G}_t^n\right)\left(\mathbf{G}_t - \mathbf{G}_t^n\right)^\top\mathbf{V} \\
&+ \eta\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\left(\mathbf{G}_t - \mathbf{G}_t^n\right)^\top\mathbf{V} + \eta\mathbf{U}^\top\left(\mathbf{G}_t - \mathbf{G}_t^n\right)\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top,
\end{aligned} \tag{18}$$

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

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

Similarly, we also have

$$\begin{aligned}
\mathbf{T}_{t+1}\left(\mathbf{T}_{t+1}\right)^\top =& \mathbf{V}^\top\mathbf{F}_{t+1}\left(\mathbf{V}^\top\mathbf{F}_{t+1}\right)^\top \\
=& \mathbf{V}^\top\left(\mathbf{F}_t - \eta\mathbf{G}_t\right)\left(\mathbf{F}_t - \eta\mathbf{G}_t\right)^\top\mathbf{V} + \eta^2\mathbf{V}^\top\left(\mathbf{G}_t - \mathbf{G}_t^n\right)\left(\mathbf{G}_t - \mathbf{G}_t^n\right)^\top\mathbf{V} \\
&+ \eta\mathbf{V}^\top\left(\mathbf{F}_t - \eta\mathbf{G}_t\right)\left(\mathbf{G}_t - \mathbf{G}_t^n\right)^\top\mathbf{V} + \eta\mathbf{V}^T\left(\mathbf{G}_t - \mathbf{G}_t^n\right)\left(\mathbf{F}_t - \eta\mathbf{G}_t\right)^\top\mathbf{V} \\
=& \mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top + \eta^2\mathbf{V}^\top\left(\mathbf{G}_t - \mathbf{G}_t^n\right)\left(\mathbf{G}_t - \mathbf{G}_t^n\right)^\top\mathbf{V} \\
&+ \eta\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)\left(\mathbf{G}_t - \mathbf{G}_t^n\right)^\top\mathbf{V} + \eta\mathbf{V}^\top\left(\mathbf{G}_t - \mathbf{G}_t^n\right)\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top.
\end{aligned} \tag{19}$$

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

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

We can see that Δ_t is symmetric matrix, and

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

B.1 Proof for Lemma 7

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

$$\begin{aligned}
\left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2 &\le (1-\eta\sigma_r)\left\|\mathbf{S}_t\mathbf{S}_t^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2+3\eta\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2^2,\\
\left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top\right\|_2 &\le (1-\eta\sigma_r)\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2,\\
\left\|\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top\right\|_2 &\le \left\|\mathbf{T}_t\mathbf{T}_t^\top\right\|_2\left(1-\eta\left\|\mathbf{T}_t\mathbf{T}_t^\top\right\|_2+2\eta\left\|\mathbf{D}_{\mathbf{T}}^*\right\|_2\right),
\end{aligned}\tag{20}$$

and the following non-expansion results:

$$\begin{aligned}
\left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\mathbf{S}_t^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2 &\le (1-\eta\sigma_r)\left\|\mathbf{S}_t\mathbf{S}_t^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2+\eta\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2^2,\\
\left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\mathbf{T}_t^\top\right\|_2 &\le \left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2,\\
\left\|\mathbf{S}_t\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top\right\|_2 &\le \left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2,\\
\left\|\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)\mathbf{T}_t^\top\right\|_2 &\le \left\|\mathbf{T}_t\mathbf{T}_t^\top\right\|_2+\eta\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2^2.
\end{aligned}\tag{21}$$

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

$$\begin{aligned}
\left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2 &\le \left(1-\frac{7}{10}\eta\sigma_r\right)D_t,\\
\left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\mathbf{S}_t^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2 &\le \left(1-\frac{9}{10}\eta\sigma_r\right)D_t,\\
\left\|\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)\mathbf{T}_t^\top\right\|_2 &\le \left(1+\frac{1}{10}\eta\sigma_r\right)D_t.
\end{aligned}\tag{22}$$

For the ease of the presentation we need to connect $\sqrt{\frac{kd\log d}{n}}$ with $\eta\sigma_r$ for the development of the proof. Since $\eta=\frac{1}{100\sigma_1}$ and $n>C_1k\kappa^2d\log^3 d\cdot\max(1,\sigma^2/\sigma_r^2)$, by choosing

$C_1 \geq 1000^2$, we have

$$\sqrt{\frac{kd\log d}{n}} \leq 0.1\eta\sigma_r. \tag{23}$$

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

$$\|\Delta_t\|_2 \leq \eta\sigma_r D_t + 4\sqrt{\frac{d\log d}{n}}\sigma. \tag{24}$$

Upper bound for $\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2$: According to equation (17), we have

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

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

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

and

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

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

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

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

$$\begin{aligned}\left\|\eta^2\mathbf{U}^\top\left(\mathbf{G}_t - \mathbf{G}_t^n\right)\left(\mathbf{G}_t - \mathbf{G}_t^n\right)^\top\mathbf{U}\right\|_2 \leq& \eta^2\left\|\Delta_t\mathbf{F}_t\mathbf{F}_t^\top\Delta_t\right\|_2 \\ \leq& \eta^2\left(\left\|\mathbf{S}_t\mathbf{S}_t^\top\right\|_2 + \left\|\mathbf{T}_t\mathbf{T}_t^\top\right\|_2 + 2\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2\right)\|\Delta_t\|_2^2 \\ \leq& \frac{1}{100}\eta\|\Delta_t\|_2^2,\end{aligned} \tag{25}$$

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

$$
\begin{aligned}
&\left\|\mathbf{S}_{t+1}\mathbf{S}_{t+1}^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 \\
&\quad\overset{(1)}{\leq} \underbrace{\left\|\mathcal{M}(\mathbf{S}_t)\mathcal{M}(\mathbf{S}_t)^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2}_{\text{I}} + \underbrace{2\eta \left\|\mathbf{D}_{\mathbf{S}}^*\right\|_2 \left\|\Delta_t\right\|_2}_{\text{V + VIII}} + \underbrace{\frac{1}{100}\eta \left\|\Delta_t\right\|_2^2}_{\text{II}} \\
&\qquad + \underbrace{2\eta \left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\mathbf{S}_t^T - \mathbf{D}_{\mathbf{S}}^*\right\|_2 \left\|\Delta_t\right\|_2}_{\text{III + VI}} + \underbrace{2\eta \left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\mathbf{T}_t^T\right\|_2 \left\|\Delta_t\right\|_2}_{\text{IV + VII}} \\
&\quad\overset{(2)}{\leq} \underbrace{\left(1 - \frac{7}{10}\eta\sigma_r\right)\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2}_{\text{I}} + \underbrace{\frac{1}{50}\left\|\Delta_t\right\|_2}_{\text{V + VIII}} + \underbrace{\frac{1}{100}\eta\left\|\Delta_t\right\|_2^2}_{\text{II}} + \underbrace{4\eta D_t \left\|\Delta_t\right\|_2}_{\text{III + VI + IV + VII}} \\
&\quad\overset{(3)}{\leq} \left(1 - \frac{7}{10}\eta\sigma_r\right)\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 + \frac{1}{10}\left\|\Delta_t\right\|_2 \\
&\quad\overset{(4)}{\leq} \left(1 - \frac{7}{10}\eta\sigma_r\right)\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 + \sqrt{\frac{kd\log d}{n}}D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma \\
&\quad\overset{(5)}{\leq} \left(1 - \frac{7}{10}\eta\sigma_r\right)\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 + 0.1\eta\sigma_r D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma
\end{aligned}
$$

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 $\frac{1}{100}\eta\left\|\Delta_t\right\|_2 < 0.0001$, and $4\eta D_t < 0.04$; inequality (4) is obtained by plugging in the relaxation of $\left\|\Delta_t\right\|_2$ (cf. equation (5)) and organizing according to D_t and σ; inequality (5) is obtained via the bound (24).

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

$$
\left\|\mathbf{S}_{t+1}\mathbf{S}_{t+1}^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 \leq \left(1 - \frac{7}{10}\eta\sigma_r\right)\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 + 0.1\eta\sigma_r D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma. \tag{26}
$$

This indicates a contraction with respect to D_t:

$$
\left\|\mathbf{S}_{t+1}\mathbf{S}_{t+1}^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 \leq \left(1 - \frac{6}{10}\eta\sigma_r\right)D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma. \tag{27}
$$

From the above result, we can verify that

$$
\left\|\mathbf{S}_{t+1}\mathbf{S}_{t+1}^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2 \leq \left(1 - \frac{6}{10}\eta\sigma_r\right)\left(D_t - 50\kappa\sqrt{\frac{d\log d}{n}}\sigma\right) + 50\kappa\sqrt{\frac{d\log d}{n}}\sigma. \tag{28}
$$

Upper bound for $\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2$: According to equation (18), we have

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

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

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

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

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

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

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

Collecting the above results, we find that

$$
\begin{aligned}
&\left\|\mathbf{S}_{t+1}\mathbf{T}_{t+1}^\top\right\|_2 \\
&\leq \underbrace{\left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)^\top\right\|_2}_{\mathrm{I}'} + \underbrace{\frac{1}{100}\eta\left\|\Delta_t\right\|_2^2}_{\mathrm{II}'} + \underbrace{\eta\left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2\left\|\Delta_t\right\|_2}_{\mathrm{III}'} + \underbrace{\eta\left\|\mathbf{D}_{\mathbf{S}}^*\right\|_2\left\|\Delta_t\right\|_2}_{\mathrm{V}'} \\
&\quad + \underbrace{\eta\left\|\mathcal{M}_{\mathbf{S}}(\mathbf{S}_t)\mathbf{T}_t^\top\right\|_2\left\|\Delta_t\right\|_2}_{\mathrm{IV}'} + \underbrace{\eta\left\|\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)\mathbf{S}_t^\top\right\|_2\left\|\Delta_t\right\|_2}_{\mathrm{VI}'} + \underbrace{\eta\left\|\mathcal{M}_{\mathbf{T}}(\mathbf{T}_t)\mathbf{T}_t^\top\right\|_2\left\|\Delta_t\right\|_2}_{\mathrm{VII}'} \\
&\overset{(1)}{\leq} \underbrace{(1-\eta\sigma_r)\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2}_{\mathrm{I}'} + \underbrace{\frac{1}{100}\eta\left\|\Delta_t\right\|_2^2}_{\mathrm{II}'} + \underbrace{\frac{1}{100}\left\|\Delta_t\right\|_2}_{\mathrm{V}} + \underbrace{5\eta D_t\left\|\Delta_t\right\|_2}_{\mathrm{III}'+\mathrm{IV}'+\mathrm{VI}'+\mathrm{VII}'} \\
&\overset{(2)}{\leq} (1-\eta\sigma_r)\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2 + \frac{1}{10}\left\|\Delta_t\right\|_2 \\
&\leq (1-\eta\sigma_r)\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2 + \sqrt{\frac{kd\log d}{n}}D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma \\
&\leq (1-\eta\sigma_r)\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2 + 0.1\eta\sigma_r D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma,
\end{aligned}
$$

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 $\frac{1}{100}\eta\left\|\Delta_t\right\|_2 < 0.001$ and $5\eta D_t < 0.05$. In summary, we have

$$
\left\|\mathbf{S}_{t+1}\mathbf{T}_{t+1}^\top\right\|_2 \leq (1-\eta\sigma_r)\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2 + 0.1\eta\sigma_r D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma. \tag{29}
$$

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

$$
\left\|\mathbf{S}_{t+1}\mathbf{T}_{t+1}^\top\right\|_2 \leq \left(1 - \frac{9}{10}\eta\sigma_r\right)D_t + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma. \tag{30}
$$

Given the above result, we can verify that

$$
\left\|\mathbf{S}_{t+1}\mathbf{T}_{t+1}^\top\right\|_2 \leq \left(1 - \frac{9}{10}\eta\sigma_r\right)\left(D_t - 50\kappa\sqrt{\frac{d\log d}{n}}\sigma\right) + 50\kappa\sqrt{\frac{d\log d}{n}}\sigma. \tag{31}
$$

Upper bound for $\left\|\mathbf{T}_t\mathbf{T}_t^\top\right\|_2$: According to equation (19) and similar deductions as in previous bounds for $\left\|\mathbf{S}_t\mathbf{S}_t^\top - \mathbf{D}_{\mathbf{S}}^*\right\|_2$ and $\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2$, we have

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

where the following expansions hold:

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

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

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

Assume that $\left\|\mathbf{T}_t\mathbf{T}_t^\top\right\|_2=zD_t$ for $0<z\le1$. Note that, z is not necessarily a constant. For notation simplicity we use the short hand that $\epsilon_{stat}=\sqrt{\frac{d\log d}{n}}\sigma$. With the choice of n and equation (5), we have $\left\|\Delta_t\right\|_2\le\eta\sigma_rD_t+4\epsilon_{stat}$. Therefore, we obtain that

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

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 $z=\frac{1+2\eta\left\|\mathbf{D}_{\mathbf{T}}^*\right\|_2}{2\eta D_t}>1$. Hence we can just safely plug-in $z=1$. Now, we arrange by organizing according to D_t and ϵ_{stat} and obtain that

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

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

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

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

Furthermore, from equations (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 reaches 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

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

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

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

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

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

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

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

where inequality (1) is deducted in the proof of Lemma 7; inequality (2) is by relaxing term I using Equation (20), relaxing $\left\|\Delta_t\right\|_2 \leq 51\epsilon_{stat}$, and grouping all other terms; inequality (3) is by the assumption that $\left\|\mathbf{D}_{\mathbf{T}}^*\right\|_2 \leq \epsilon_{stat}$; inequality (4) is by the choice of n such that $\epsilon_{stat} \leq 0.1$.

Putting all the results together, we obtain the conclusion of Lemma 12. ■

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

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

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

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

We 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 $\epsilon_0 = \sqrt{\frac{kd\log d}{n}}\sigma_r + \frac{4}{10}\sqrt{\frac{d\log d}{n}}\sigma$.

Thus, we have

$$\begin{aligned}\left\|\mathbf{S}_{t+1}\mathbf{S}_{t+1}^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2 &\le \left(1-\frac{7}{10}\eta\sigma_r\right)\left\|\mathbf{S}_t\mathbf{S}_t^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2+\epsilon_0\\ &\le \left(1-\frac{7}{10}\eta\sigma_r\right)^2\left\|\mathbf{S}_{t-1}\mathbf{S}_{t-1}^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2+\left(1+\left(1-\frac{7}{10}\eta\sigma_r\right)\right)\epsilon_0\\ &\le ...\\ &\le \left(1-\frac{7}{10}\eta\sigma_r\right)^{t+1}\left\|\mathbf{S}_0\mathbf{S}_0^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2+O(\eta^{-1}\sigma_r^{-1})\epsilon_0.\end{aligned}$$

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

$$\left\|\mathbf{S}_t\mathbf{S}_t^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2\le\left(1-\frac{7}{10}\eta\sigma_r\right)^t\left\|\mathbf{S}_0\mathbf{S}_0^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2+\epsilon_{comp}\le\left(1-\frac{7}{10}\eta\sigma_r\right)^t 0.1\sigma_r+\epsilon_{comp},$$

where the final inequality holds by simply plugging in the initialization condition.

After at most $t=\frac{1}{\log\frac{1}{1-0.005/\kappa}}\cdot\log\frac{1}{10000\sqrt{\frac{k\kappa^2 d\log d}{n}}}$ iterations, $\left\|\mathbf{S}_t\mathbf{S}_t^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2<\epsilon_{comp}$. Since $\frac{1}{\log\frac{1}{1-0.005/\kappa}}\le 1.1$, we further simplify this to $t>\log\frac{n}{k\kappa^2 d\log d}$. As a consequence, we claim that after $t=\left\lceil\log\frac{n}{k\kappa^2 d\log d}\right\rceil$ iterations, $\left\|\mathbf{S}_t\mathbf{S}_t^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2<C\epsilon_{comp}$ 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., $n^{z_1}>C_2 d\log^3 dk\kappa^2$ for some universal constant $z_1\in(0,1)$, then we should use $\Delta_t\le 0.5\eta\sigma_r D_t+\sqrt{\frac{d\log d}{n}}\sigma$ and the fact that $\sqrt{\frac{kd\log d}{n}}\ll\eta\sigma_r$ 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 $\left\|\mathbf{S}_t\mathbf{S}_t^\top-\mathbf{D}_{\mathbf{S}}^*\right\|_2$.

With the same argument, we also obtain $\left\|\mathbf{S}_t\mathbf{T}_t^\top\right\|_2<C\epsilon_{comp}$ after $t=\left\lceil\log\frac{n}{k\kappa^2 d\log d}\right\rceil$ iterations. Therefore, we obtain the conclusion of claim (a) in Theorem 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 $50\kappa\sqrt{\frac{d\log d}{n}}\sigma\le 0.05\sigma_r$ 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 α_i, ends at iteration $\alpha_{i+1} - 1$, and $D_{\alpha_{i+1}} \leq 0.5 D_{\alpha_i}$. We invoke Lemma 18 at α_i to establish equation (5) for all the iterations in i-th epoch. Finally, we take a union bound over all the epochs.

By definition, we have

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

From Lemma 15, we know that with probability at least $1 - \exp(-C)$,

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

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

At iteration α_i, note that $\left\|\mathbf{F}_{\alpha_i}\mathbf{F}_{\alpha_i}^T - \mathbf{X}^*\right\|_2 \leq \left\|\mathbf{S}_{\alpha_i}\mathbf{S}_{\alpha_i}^T - \mathbf{D}_{\mathbf{S}}^*\right\|_2 + \left\|\mathbf{T}_{\alpha_i}\mathbf{T}_{\alpha_i}^T - \mathbf{D}_{\mathbf{T}}^*\right\|_2 + 2\left\|\mathbf{S}_{\alpha_i}\mathbf{T}_{\alpha_i}^T\right\|_2 \leq 4D_{\alpha_i} + \|\mathbf{D}_{\mathbf{T}}^*\|_2 < 5D_{\alpha_i}$. By Lemma 18 we know that, with probability at least $1 - \exp(-C)$, we have

$$\sup_{\|\mathbf{X}\|_2 \leq 5D_{\alpha_i}} \frac{1}{n}\sum_i^n \langle \mathbf{A}_i, \mathbf{X}\rangle \mathbf{A}_i - \mathbf{X} \leq 5\sqrt{\frac{k^2 d\log d}{n}} D_{\alpha_i}.$$

Therefore, we find that

$$\begin{aligned}\|\Delta_{\alpha_i}\|_2 \leq& 5\sqrt{\frac{k^2 d\log d}{n}} D_{\alpha_i} + \sqrt{\frac{d\log d}{n}}\sigma \\ \leq& 0.5\eta\sigma_r D_{\alpha_i} + \sqrt{\frac{d\log d}{n}}\sigma\end{aligned}$$

where the second inequality is by the choice of n. Hence equation (5) is satisfied at iteration α_i. For notation simplicity, we define $A_t = D_t - 50\kappa\sqrt{\frac{d\log d}{n}}\sigma$. Invoking Lemma 7, we have

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

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

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

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

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

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

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 $A_t \leq \frac{4}{\eta t + \frac{4}{A_0}}$. To see this, we have

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

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

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

$$\max\left\{\left\|\mathbf{S}_t\mathbf{S}_t^T - \mathbf{D}_{\mathbf{S}}^*\right\|_2, \left\|\mathbf{T}_t\mathbf{T}_t^T\right\|_2, \left\|\mathbf{S}_t\mathbf{T}_t^T\right\|_2\right\} \leq \Theta\left(\epsilon_{stat}\right). \tag{33}$$

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 ϵ_{stat}, we need $T = \Theta\left(\log(1/\epsilon_{stat})\right) = \Theta\left(\log n\right)$. Equation (32) holds for all epochs with probability at least $1 - Td^{-c} > 1 - d^{-c+1}$ for some universal constant c, since $T = \Theta\left(\log n\right)$.

If n is large, i.e., $n^{z_1} > C_2 d\log^3 dk\kappa^2$ 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 $1-T/\exp(n^{z_1}) > 1-d^{-c}$ for some universal constant c, since $T < \log n$. If $t > \Theta\left(\frac{1}{\eta\epsilon_{stat}}\right)$, we can show using above argument that after $\Theta\left(\frac{1}{\eta\epsilon_{stat}}\right)$ number of iterations equation (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 $t \geq \Theta\left(\frac{1}{\eta\epsilon_{stat}}\right)$ number of iterations, by triangular inequality, and the assumption that $\|\mathbf{D}_{\mathbf{T}}^*\|_2 \leq \epsilon_{stat}$, we have $\left\|\mathbf{F}_t\mathbf{F}_t^\top - \mathbf{X}^*\right\|_2 \leq \Theta(\epsilon_{stat})$. 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Appendix D. Concentration bounds

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

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

for any matrix $F \in \mathbb{R}^{d*k}$ such that $\|\mathbf{F}\mathbf{F}^\top - \mathbf{X}^*\|_2 \le R$ 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 $\{\mathbf{X}_k\}$ of independent, random, self-adjoint matrices with dimension d. Assume that each random matrix satisfies*

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

Then, for all $t \ge 0$,

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k \mathbf{X}_k\right) \ge t\right) \le d\cdot\exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right) \quad \textit{where} \quad \sigma^2 := \left\|\sum_k \mathbb{E}(\mathbf{X}_k^2)\right\|_2. \tag{34}$$

Lemma 14. *Let* $\mathbf{A}$ *be a symmetric random matrix in* $\mathbb{R}^{d*d}$*, with the upper triangle entries* $(i \geq j)$ *being independently sampled from an identical sub-Gaussian distribution whose mean is* 0 *and variance proxy is* 1*. Let* ϵ *follows* $N(0, \sigma)$*. Then*

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

Proof

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

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

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

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

Applying the union bound to the above concentration results leads to

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

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

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

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

Lemma 15. ***(Lemma 8 re-stated)*** *Let* $\mathbf{A}_i$ *be symmetric random matrices in* $\mathbb{R}^{d*d}$*, with the upper triangle entries* $(i \geq j)$ *being independently sampled from an identical sub-Gaussian distribution whose mean is* 0 *and variance proxy is* 1*. Let* ϵ_i *follows* $N(0, \sigma)$*. Then*

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

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

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

Hence, we obtain

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

From the matrix Bernstein bound (Wainwright, 2019), we find that

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

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

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

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

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

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

Lemma 16. *Let* $\mathbf{A}$ *be a symmetric random matrix in* $\mathbb{R}^{d*d}$*, with the upper triangle entries* $(i \geq j)$ *being independently sampled from an identical sub-Gaussian distribution whose mean is* 0 *and variance proxy is* 1*. Let* $\mathbf{U}$ *be a deterministic symmetric matrix of the same dimension. Then, for some universal constant* C_1, C_2*, we have*

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

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

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

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

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

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

$$\mathbb{P}\left(\mathbf{x}^\top\left(\langle\mathbf{A},\mathbf{U}\rangle\mathbf{A}-\mathbf{U}\right)\mathbf{x} \geq t\,\|\mathbf{U}\|_F\right) \leq \exp\left(-\frac{t}{C_1}\right). \tag{35}$$

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

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

Now we fix ϵ to be $1/4$. Then, for equation (35) we take union bound over $\mathcal{V}$ and we have

$$\mathbb{P}\left(\max_{\mathbf{x}\in\mathcal{V}} \mathbf{x}^\top \left(\langle \mathbf{A}, \mathbf{U}\rangle \mathbf{A} - \mathbf{U}\right)\mathbf{x} \geq t \left\|\mathbf{U}\right\|_F\right) \leq |\mathcal{V}| \exp\left(-\frac{t}{C_1}\right), \quad \text{for} \quad \mathrm{t} > \mathrm{C}_2$$

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

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

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

■

Lemma 17. *Let $\mathbf{A}_i$ be a symmetric random matrix of dimension d by d, with the upper triangle entries ($i \geq j$) being independently sampled from an identical sub-Gaussian distribution whose mean is 0 and variance proxy is 1. Let $\mathbf{U}$ be a deterministic symmetric matrix of the same dimension. Then as long as $n > C_1 d\log^3 d$ for some universal $C_1, C_2 > 10$, we have*

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

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

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_i^n \left(\langle \mathbf{A}_i, \mathbf{U}\rangle \mathbf{A}_i - \mathbf{U}\right)\right\|_2 \leq \frac{1}{\kappa \log d\sqrt{kC_2}} \left\|\mathbf{U}\right\|_F\right) \geq 1 - \exp\left(-n^{z_2}\right). \tag{37}$$

Proof Following Lemma 13, we want to first bound the second order moment of the random matrices. Since $\mathbb{E}\langle \mathbf{A}_i, \mathbf{U}\rangle \mathbf{A}_i = \mathbf{U}$ and $\mathbf{U}$ has no randomness, we have

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

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

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

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

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

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

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

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

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

Then, the following inequality holds:

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

where C_1 is a universal constant inherited from Lemma 16 and

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

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

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

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

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

then plugging in $t = \sqrt{\frac{d\log d}{C_2 d\log^3 dk\kappa^2}}\|\mathbf{U}\|_F = \frac{1}{\kappa\log d\sqrt{kC_2}}\|\mathbf{U}\|_F$, we have

$$\begin{aligned}
&\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n}\sum_i^n\left(\langle\mathbf{A}_i,\mathbf{U}\rangle\mathbf{A}_i-\mathbf{U}\right)\right)\geq\frac{1}{\kappa\log d\sqrt{kC_2}}\|\mathbf{U}\|_F\right)\\
\leq& d\cdot\exp\left(\frac{-3\left(\frac{1}{\kappa\log d\sqrt{kC_2}}\|\mathbf{U}\|_F\right)^2 n}{18\|\mathbf{U}\|_F^2+2C_1 d\log d\|\mathbf{U}\|_F\left(\frac{1}{\kappa\log d\sqrt{kC_2}}\|\mathbf{U}\|_F\right)}\right)\\
=& d\cdot\exp\left(\frac{-3n}{18\left(\kappa\log d\sqrt{kC_2}\right)^2+2C_1 d\log d\left(\kappa\log d\sqrt{kC_2}\right)}\right)\\
\leq&\exp\left(\frac{-n}{C_3 n^{z_1}}\right)\\
\leq&\exp\left(-n^{z_2}\right)\qquad\text{(for some universal constant } z_2\in(0,1)).
\end{aligned}$$

In summary, we reach the conclusion of the lemma. ■

Lemma 18. ***(Lemma 10 re-stated)*** *Let $\mathbf{A}_i$ be a symmetric random matrix of dimension d by d. Its upper triangle entries ($i\geq j$) are independently sampled from an identical sub-Gaussian distribution whose mean is 0 and variance proxy is 1. If $\mathbf{U}$ symmetric is of rank k and is in a bounded spectral norm ball of radius R (i.e. $\|\mathbf{U}\|_2\leq R$), then we have*

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

and

$$\mathbb{P}\left(\sup_{\mathbf{U}:\|\mathbf{U}\|_2\leq R}\left\|\frac{1}{n}\sum_i^n\left(\langle\mathbf{A}_i,\mathbf{U}\rangle\mathbf{A}_i-\mathbf{U}\right)\right\|_2\leq\sqrt{\frac{d\log d}{n}}kR\right)\geq 1-\exp\left(-C_2 kd\right).\tag{39}$$

Proof

By the standard symmetrization argument,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{U}:\|\mathbf{U}\|_2\leq R}\left\|\frac{1}{n}\sum_i^n\left(\langle\mathbf{A}_i,\mathbf{U}\rangle\mathbf{A}_i-\mathbf{U}\right)\right\|_2\geq t\right)\\
\leq& 2\mathbb{P}\left(\sup_{\mathbf{U}:\|\mathbf{U}\|_2\leq R}\left\|\frac{1}{n}\sum_i^n\tau_i\left(\langle\mathbf{A}_i,\mathbf{U}\rangle\mathbf{A}_i\right)\right\|_2>t/2\right)\\
=& 2\mathbb{P}\left(\sup_{\mathbf{U}:\|\mathbf{U}\|_2\leq R}\sup_{\mathbf{x}:\|\mathbf{x}\|_2\leq 1}\frac{1}{n}\sum_i^n\tau_i\left\langle\mathbf{A}_i,\mathbf{U}\right\rangle\left\langle\mathbf{A}_i,\mathbf{x}\mathbf{x}^\top\right\rangle>t/2\right),
\end{aligned}$$

where τ_is are independent Rademacher random variable. Note that $\langle\mathbf{A}_i,\mathbf{U}\rangle$ is sub-Gaussian with Orcliz norm $\mathcal{O}\left(\|\mathbf{U}\|_F\right)$ and $\left\langle\mathbf{A}_i,\mathbf{x}\mathbf{x}^\top\right\rangle$ is sub-Gaussian with Orcliz norm $\mathcal{O}(1)$. Hence the product is hence sub-exponential, with Orcliz norm $\mathcal{O}\left(\|\mathbf{U}\|_F\right)$.

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* $\mathbf{X}$ *be a centered sub-exponential random variable. Then, for* t *such that* $t \leq c/\left\|\mathbf{X}\right\|_{\psi_1}$*, one has*

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

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

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\sup_{\mathbf{U}:\left\|\mathbf{U}\right\|_2\leq R}\sup_{\mathbf{x}:\left\|\mathbf{x}\right\|_2\leq 1}\frac{\lambda}{n}\sum_i^n \tau_i\left\langle\mathbf{A}_i,\mathbf{U}\right\rangle\left\langle\mathbf{A}_i,\mathbf{x}\mathbf{x}^\top\right\rangle\right)\right]\\
&\quad=\mathbb{E}\left[\exp\left(\sup_{\mathbf{U}:\left\|\mathbf{U}\right\|_2\leq 1}\sup_{\mathbf{x}:\left\|\mathbf{x}\right\|_2\leq 1}\frac{\lambda R}{n}\sum_i^n \tau_i\left\langle\mathbf{A}_i,\mathbf{U}\right\rangle\left\langle\mathbf{A}_i,\mathbf{x}\mathbf{x}^\top\right\rangle\right)\right]\\
&\quad\leq\exp\left(\frac{C_1\lambda^2 kR^2}{n}+C_2(k+1)d\right),
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{U}:\left\|\mathbf{U}\right\|_2\leq R}\sup_{\mathbf{x}:\left\|\mathbf{x}\right\|_2\leq 1}\frac{1}{n}\sum_i^n \tau_i\left\langle\mathbf{A}_i,\mathbf{U}\right\rangle\left\langle\mathbf{A}_i,\mathbf{x}\mathbf{x}^\top\right\rangle > t/2\right)\\
&\qquad\leq\exp\left(\frac{C_1\lambda^2 kR^2}{n}+C_2(k+1)d-\lambda t/2\right).
\end{aligned}$$

Let $\lambda=\frac{nt}{4C_1kR^2}$ and choose $t>C_3\sqrt{\frac{kd}{n}}\sqrt{k}R$. Then

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{U}:\left\|\mathbf{U}\right\|_2\leq R}\left\|\frac{1}{n}\sum_i^n\left(\langle\mathbf{A}_i,\mathbf{U}\rangle\mathbf{A}_i-\mathbf{U}\right)\right\|_2\geq t\right)\\
&\quad\leq 2\mathbb{P}\left(\sup_{\mathbf{U}:\left\|\mathbf{U}\right\|_2\leq R}\sup_{\mathbf{x}:\left\|\mathbf{x}\right\|_2\leq 1}\frac{1}{n}\sum_i^n \tau_i\left\langle\mathbf{A}_i,\mathbf{U}\right\rangle\left\langle\mathbf{A}_i,\mathbf{x}\mathbf{x}^\top\right\rangle > t/2\right)\\
&\quad\leq\exp\left(-\frac{C_4nt^2}{kR^2}+C_5kd\right)\\
&\quad\leq\exp\left(-C_6kd\right)
\end{aligned}$$

as long as we choose the constant C_3 large enough.

■

References

S. Balakrishnan, M. J. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45:77–120, 2017.

Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582, 2016a.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. *arXiv preprint arXiv:1605.07221*, 2016b.

Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Samuel Burer and Renato DC Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005.

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

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Yudong Chen, Ali Jalali, Sujay Sanghavi, and Constantine Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7): 4324–4337, 2013.

Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu. Singularity, misspecification, and the convergence rate of EM. *Annals of Statistics*, 48:3161–3182, 2020a.

Raaz Dwivedi, Nhat Ho, Koulik Khamaru, Martin J Wainwright, Michael I Jordan, and Bin Yu. Sharp analysis of Expectation-Maximization for weakly identifiable models. In *AISTATS*, 2020b.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

David Gross, Yi-Kai Liu, Steven T Flammia, Stephen Becker, and Jens Eisert. Quantum state tomography via compressed sensing. *Physical review letters*, 105(15):150401, 2010.

Moritz Hardt. Understanding alternating minimization for matrix completion. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 651–660. IEEE, 2014.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional M-estimation. In *Neural Information Processing Systems (NIPS)*, 2014a.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. *Advances in neural information processing systems*, 27, 2014b.

Amir Kalev, Robert L Kosut, and Ivan H Deutsch. Quantum tomography protocols with positivity are compressed sensing protocols. *npj Quantum Information*, 1(1):1–6, 2015.

Vladimir Koltchinskii, Karim Lounici, Alexandre B Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

Jeongyeol Kwon, Nhat Ho, and Constantine Caramanis. On the minimax optimality of the EM algorithm for learning two-component mixed linear regression. *AISTATS*, 2021.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

Jianhao Ma and Salar Fattahi. Global convergence of sub-gradient method for robust matrix recovery: Small initialization, noisy measurements, and over-parameterization. *Journal of Machine Learning Research*, 24(96):1–84, 2023.

Sahand Negahban and Martin J Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably. *SIAM Journal on Imaging Sciences*, 11(4):2165–2204, 2018.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Mahdi Soltanolkotabi, Dominik Stöger, and Changzhi Xie. Implicit balancing and regularization: Generalization and convergence guarantees for overparameterized asymmetric matrix sensing. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5140–5142. PMLR, 2023.

Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Advances in Neural Information Processing Systems*, 34:23831–23843, 2021.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag, New York, NY, 2000.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027v7*, 2011.

Roman Vershynin. *High Dimensional Probability. An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Andrew E Waters, Aswin C Sankaranarayanan, and Richard Baraniuk. Sparcs: Recovering low-rank and sparse matrices from compressive measurements. In *Advances in neural information processing systems*, pages 1089–1097, 2011.

Xingyu Xu, Yandi Shen, Yuejie Chi, and Cong Ma. The power of preconditioning in overparameterized low-rank matrix sensing. In *International Conference on Machine Learning*, pages 38611–38654. PMLR, 2023.

Gavin Zhang, Hong-Ming Chiu, and Richard Y Zhang. Fast and minimax optimal estimation of low-rank matrices via non-convex gradient descent. *arXiv preprint arXiv:2305.17224*, 2023.

Jialun Zhang and Richard Zhang. How many samples is a good initial point worth in low-rank matrix recovery? *Advances in Neural Information Processing Systems*, 33, 2020.

Jialun Zhang, Salar Fattahi, and Richard Y Zhang. Preconditioned gradient descent for over-parameterized nonconvex matrix factorization. *Advances in Neural Information Processing Systems*, 34:5985–5996, 2021.

Richard Y. Zhang, Somayeh Sojoudi, and Javad Lavaei. Sharp restricted isometry bounds for the inexistence of spurious local minima in nonconvex matrix recovery. *Journal of Machine Learning Research*, 20:1–34, 2019.

Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. *Advances in Neural Information Processing Systems*, 28:109–117, 2015.

Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.