RANK-1 MATRIX COMPLETION WITH GRADIENT DESCENT AND SMALL RANDOM INITIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The nonconvex formulation of matrix completion problem has received significant attention in recent years due to its affordable complexity compared to the convex formulation. Gradient descent (GD) is the simplest yet efficient baseline algorithm for solving nonconvex optimization problems. The success of GD has been witnessed in many different problems in both theory and practice when it is combined with random initialization. However, previous works on matrix completion require either careful initialization or regularizer to prove the convergence of GD. In this work, we study the rank-1 symmetric matrix completion and prove that GD converges to the ground truth when small random initialization is used. We show that in logarithmic amount of iterations, the trajectory enters the region where local convergence occurs. We provide an upper bound on the initialization size that is sufficient to guarantee the convergence and show that a larger initialization can be used as more samples are available. We observe that implicit regularization effect of GD plays a critical role in the analysis, and for the entire trajectory, it prevents each entry from becoming much larger than the others.

1 INTRODUCTION

Recovering a low-rank matrix from a number of linear measurements lies at the heart of many statistical learning problems. Depending on the structure of the matrix and linear measurements, it reduces to various problems such as phase retrieval (Chen et al., 2019), blind deconvolution (Ahmed et al., 2014), and matrix sensing (Tu et al., 2016). Matrix completion (Candès & Recht, 2009) is also one such type of problems where each measurement provides one entry of the matrix, and the goal is to recover the low-rank matrix from a partial, usually very sparse, observation of the entries. One of the most notable applications of matrix completion is collaborative filtering (F. Gleich & Lim, 2011), which aims to predict preferences of users to items based on a highly incomplete observation of user-item ratings. There are also a number of different applications such as principal component analysis (Candès et al., 2011), image reconstruction (Hu et al., 2019), to just name a few.

Extensive amount of work has been dedicated to provide an efficient recovery algorithm for matrix completion with theoretical guarantees. The convex relaxation based nuclear norm minimization (Candès & Recht, 2009; Candès & Tao, 2010) was the first algorithm proved to recover the matrix with near optimal sample complexity. Despite its theoretical success, the convex algorithm was found hard to be used in practical scenarios due to its unaffordable computational complexity and memory size. Hence, the nonconvex formulation of matrix completion with quadratic loss has received significant attention in recent years. Many different algorithms were proposed for the nonconvex problem, and their convergence toward the ground truth was analyzed. Examples include optimization on Grassmann manifolds (Keshavan et al., 2010), alternating minimization (Jain et al., 2013), projected gradient descent (Chen & Wainwright, 2015), gradient descent with regularizer (Sun & Luo, 2016), and (vanilla) gradient descent (Ma et al., 2020; Chen et al., 2020).

Gradient descent (GD) has served as a baseline algorithm for solving nonconvex optimization. However, the convergence of GD to global minimizers is not guaranteed, and it can take exponential time to escape saddle points (Du et al., 2017). Nevertheless, GD with random initialization was shown to recover the global minimum successfully in many different problems such as phase retrieval (Chen et al., 2019), matrix sensing (Li et al., 2018), matrix factorization (Ye & Du, 2021), and training of neural networks (Du et al., 2019). Previous works on matrix completion (Ma et al., 2020; Chen et al., 2020) proved the convergence of GD under the spectral initialization that locates the initial point in the local region of the minima. However, the role of random initialization when solving matrix completion with GD is not fully understood yet, although its success is observed in practice. Hence, we aim to answer the following question:

Can GD with random initialization solve the nonconvex matrix completion problem?

We answer this question with affirmative showing that GD with small random initialization converges to the ground truth successfully for rank-1 symmetric matrix completion. In the analysis, we use vanilla GD that does not modify GD in any way such as regularizer or truncation does. We also characterize the entire trajectory that GD follows by showing that the trajectory is well approximated by the fully observed case. The small initialization plays a critical role in analyzing the trajectory of early stages, where the randomly initialized vector is almost orthogonal to the first eigenvector of ground truth matrix. We provide a bound on the required initialization size for the algorithm to converge, and our bound suggests that one can use a larger initialization to improve the convergence speed as more samples are provided. However, in any case, GD with small random initialization takes only logarithmic amount of time to reach the point where local convergence can start. To the best of our knowledge, this is the first result on matrix completion that proves the convergence of vanilla GD without any carefully designed initialization. Although our result is restricted to the rank-1 case, we believe that this work provides an important evidence toward understanding more general rank-r case.

Related Works This work is motivated by the recent success of small initialization in matrix factorization and matrix sensing. It was first conjectured in Gunasekar et al. (2017) that small enough step sizes and initialization lead GD to converge to the minimum nuclear norm solution of a full-dimensional matrix sensing problem. The conjecture was proved in Li et al. (2018) for the fully overparameterized matrix sensing under the standard restricted isometry property (RIP). A recent study of Stöger & Soltanolkotabi (2021) provided more general results by showing that the early iterations of GD with small initialization has spectral bias. Many other works such as Arora et al. (2019); Razin & Cohen (2020); Li et al. (2021) also studied how GD or gradient flow with small initialization implicitly force the recovered matrix to become low-rank. However, the recovery guarantee for matrix completion was not provided by any work.

For the matrix sensing where RIP holds, the loss function has benign geometry globally that it does not contain any spurious local minima or non-strict saddle points (Bhojanapalli et al., 2016). In case of matrix completion, a similar result was obtained but with a regularizer that penalizes the matrices with large row (Ge et al., 2016). Controlling the norm of each row (absolute value of each entry in case of rank-1) is the biggest hurdle in the analysis of matrix completion. In the local convergence analysis of Ma et al. (2020), it was proved that gradient descent implicitly regularizes the largest ℓ_2-norm of the rows of error matrices, showing that explicit regularization is unnecessary. In this work, we also prove that such an implicit regularization is induced by GD if it starts from a point with small size. We show that the trajectory is close to the fully observed case in both ℓ_2 and ℓ_∞-norm. Hence, the trajectory is confined to the region where it has benign geometry, and GD can converge without any explicit regularizer.

Notations We denote vectors with a lowercase bold letter and matrices with an uppercase bold letter. The components or entries of them are written without bold. For any vector $\boldsymbol{x}$, its ℓ_2 and ℓ_∞-norm are denoted by $\|\boldsymbol{x}\|_2$ and $\|\boldsymbol{x}\|_\infty$, respectively. Also, for a vector $\boldsymbol{u}^\star$ that will be defined later, we denote the component of $\boldsymbol{x}$ that is orthogonal to $\boldsymbol{u}^\star$ as $\boldsymbol{x}_\perp$, i.e. $\boldsymbol{x}_\perp = \boldsymbol{x} - \boldsymbol{u}^\star \boldsymbol{u}^{\star\top}\boldsymbol{x}$. Asymptotic dependencies are denoted with the standard big O notations or with the symbols, $\lesssim$, $\asymp$ and $\gtrsim$.

2 Problem Formulation

The matrix completion problem aims to reconstruct a low-rank matrix from partially observed entries. In this work, we focus on the case where the ground truth matrix, denoted by $\boldsymbol{M}^\star \in \mathbb{R}^{n\times n}$, is a rank-1 positive semidefinite matrix. Hence, the ground truth matrix is decomposed as $\boldsymbol{M}^\star = \lambda^\star \boldsymbol{u}^\star \boldsymbol{u}^{\star\top}$ with $\lambda^\star > 0$ and $\|\boldsymbol{u}^\star\|_2 = 1$. We define $\boldsymbol{x}^\star = \sqrt{\lambda^\star}\boldsymbol{u}^\star$ so that $\boldsymbol{M}^\star = \boldsymbol{x}^\star \boldsymbol{x}^{\star\top}$. We further add the standard incoherence assumption, $\|\boldsymbol{u}^\star\|_\infty = \sqrt{\frac{\mu}{n}}$ for some constant $\mu > 0$. We consider a noiseless random sampling model that is also symmetric as $\boldsymbol{M}^\star$. Each entry in the diagonal and upper (or lower) triangular part of $\boldsymbol{M}^\star$ is revealed independently with probability $0 < p \leq 1$. Formally, we get as an observation of the matrix $\boldsymbol{M}^\circ$ whose (i, j)th entry is $\frac{1}{p}\delta_{ij}M^\star_{ij}$, where $[\delta_{ij}]_{1\leq i\leq j\leq n}$ are independent Bernoulli random variables with probability p and $\delta_{ji} = \delta_{ij}$. We denote the set of observed entries as $\Omega = \{(i, j) \mid \delta_{ij} = 1\}$, and we define an operator $\mathcal{P}_\Omega$ on matrices that makes the entries not contained in Ω zero. (e.g. $\boldsymbol{M}^\circ = \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}^\star)$)

To recover the matrix, we find $\boldsymbol{x} \in \mathbb{R}^n$ that minimizes the following nonconvex loss function that is the sum of squared differences on observed entries.

$$f(\boldsymbol{x}) := \frac{1}{4p}\sum_{(i,j)\in\Omega}(x_i x_j - x_i^\star x_j^\star)^2 \tag{1}$$

We apply vanilla GD to solve the optimization problem starting from a small randomly initialized vector $\boldsymbol{x}^{(0)}$. Each entry of $\boldsymbol{x}^{(0)}$ is sampled independently from the Gaussian distribution $\mathcal{N}\left(0, \frac{1}{n}\beta_0^2\right)$. The norm of $\boldsymbol{x}^{(0)}$ is expected to be β_0. The update rule of GD is written as

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \eta\nabla f\left(\boldsymbol{x}^{(t)}\right) = \boldsymbol{x}^{(t)} - \frac{\eta}{p}\mathcal{P}_\Omega\left(\boldsymbol{x}^{(t)}\boldsymbol{x}^{(t)\top}\right)\boldsymbol{x}^{(t)} + \eta\boldsymbol{M}^\circ\boldsymbol{x}^{(t)}, \tag{2}$$

where $\eta > 0$ is a constant step size.

We define F to be the loss function f when all entries of $\boldsymbol{M}^\star$ are observed, i.e.

$$F(\boldsymbol{x}) := \frac{1}{4}\left\|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{M}^\star\right\|_{\mathrm{F}}^2.$$

Also, we define $\widetilde{\boldsymbol{x}}^{(t)}$ as the trajectory of GD when it is applied to F with the same initial point $\boldsymbol{x}^{(0)}$. $\widetilde{\boldsymbol{x}}^{(t)}$ is a trajectory of the fully observed case. To be specific, it evolves with

$$\widetilde{\boldsymbol{x}}^{(t+1)} = \widetilde{\boldsymbol{x}}^{(t)} - \eta\nabla F(\widetilde{\boldsymbol{x}}^{(t)}) = \widetilde{\boldsymbol{x}}^{(t)} - \eta\left\|\widetilde{\boldsymbol{x}}^{(t)}\right\|_2^2\widetilde{\boldsymbol{x}}^{(t)} + \eta\boldsymbol{M}^\star\widetilde{\boldsymbol{x}}^{(t)}, \quad \widetilde{\boldsymbol{x}}^{(0)} = \boldsymbol{x}^{(0)}. \tag{3}$$

We lastly introduce the so-called *leave-one-out* sequences. They were major ingredient when controlling the ℓ_∞-norm in Ma et al. (2020). We also use them for similar purpose. For each $l \in [n]$, let us define an operator $\mathcal{P}_\Omega^{(l)}$ such that $\mathcal{P}_\Omega^{(l)}(\boldsymbol{X})$ is equal to $\boldsymbol{X}$ on the lth row/column and equal to $\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{X})$ otherwise. The lth leave-one-out sequence, $\boldsymbol{x}^{(t,l)}$, evolves with

$$\boldsymbol{x}^{(t+1,l)} = \boldsymbol{x}^{(t,l)} - \eta\mathcal{P}_\Omega^{(l)}\left(\boldsymbol{x}^{(t,l)}\boldsymbol{x}^{(t,l)\top}\right)\boldsymbol{x}^{(t,l)} + \eta\boldsymbol{M}^{(l)}\boldsymbol{x}^{(t,l)}, \quad \boldsymbol{x}^{(0,l)} = \boldsymbol{x}^{(0)}, \tag{4}$$

where $\boldsymbol{M}^{(l)} = \mathcal{P}_\Omega^{(l)}(\boldsymbol{M}^\star)$.

3 Main Results

In this section, we present our main results. The first main result is about the global convergence of GD with small random initialization.

Theorem 1. *Let the initial point $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ be sampled from the Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \frac{1}{n}\beta_0^2\boldsymbol{I})$ and $\boldsymbol{x}^{(t)}$ be updated with* (2)*. Suppose that a sufficiently small step size with $\eta\lambda^\star < 0.1$ is used and the sample complexity satisfies $n^2 p \gtrsim \mu^2 n\log^{22} n$. Then, there exists $T^\star = (1 + o(1))\frac{1}{\eta\lambda^\star}\log\frac{\sqrt{n}}{\beta_0}$ such that*

$$\min\left\{\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^\star\right\|_2, \left\|\boldsymbol{x}^{(t)} + \boldsymbol{x}^\star\right\|_2\right\} \lesssim \mu\frac{1}{\sqrt{\log n}}\|\boldsymbol{x}^\star\|_2, \tag{5}$$

$$\min\left\{\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^\star\right\|_\infty, \left\|\boldsymbol{x}^{(t)} + \boldsymbol{x}^\star\right\|_\infty\right\} \lesssim \mu\frac{1}{\sqrt{\log n}}\|\boldsymbol{x}^\star\|_\infty, \tag{6}$$

$$\max_{1\leq l\leq n}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t,l)}\right\|_2 \lesssim \mu\frac{1}{\sqrt{\log n}}\|\boldsymbol{x}^\star\|_\infty, \tag{7}$$

$$\max_{1\leq l\leq n}\left|(\boldsymbol{x}^{(t,l)} - \boldsymbol{x}^\star)_l\right| \lesssim \mu\frac{1}{\sqrt{\log n}}\|\boldsymbol{x}^\star\|_\infty \tag{8}$$

hold at $t = T^\star$ *with probability at least* $1 - o(1)$ *if a sufficiently small initialization with*

$$n^{-C} \lesssim \beta_0 \lesssim \sqrt{\lambda^\star} \sqrt[4]{\frac{np}{\log^{24} n}} \frac{1}{\sqrt[4]{n}} \tag{9}$$

is used, where $C > 0$ *is a large constant.*

We did not do our best to optimize the $\log$ factors, and most of them can be reduced with more delicate analysis. Several remarks on Theorem 1 are in order.

Global Convergence Theorem 1 proves that the trajectory of gradient descent eventually enters the local region of global minimizers $\pm\boldsymbol{x}^\star$ in the sense of both ℓ_2 and ℓ_∞-norm starting from a small random initialization. Combined with the result of Ma et al. (2020), GD starts to converge linearly to either $\boldsymbol{x}^\star$ or $-\boldsymbol{x}^\star$ after $t = T^\star$, completing the proof for global convergence.

Leave-one-out Sequence In order to apply the local convergence result of Ma et al. (2020), in addition to (5) and (6), the existence of leave-one-out sequences $\{\boldsymbol{x}^{(t,l)}\}_{l\in[n]}$ that satisfy (7) and (8) is necessary. Leave-one-out sequences also play critical role and appear naturally in the proof of Theorem 1.

Step Size We can assume that $\lambda^\star$ does not scale with n because a proper scaling on $\boldsymbol{M}^\star$ would achieve it. Then, according to the condition $\eta\lambda^\star < 0.1$, a vanishingly small step size is unnecessary for the convergence of GD, but a small, constant-sized step size is sufficient.

Sample Complexity The required sample complexity for Theorem 1 to hold is optimal up to logarithmic factor compared to the statistical lower bound of $\Omega(n \log n)$.

Convergence Time Considering that β_0^{-1} is at most polynomial in n, only $O(\log n)$ iterations are required for GD to enter the local region. It requires $O(\log(\frac{1}{\epsilon}))$ more iterations to achieve ϵ-accuracy in the local region, and thus, the overall iteration complexity is given by $O(\log n) + O(\log(\frac{1}{\epsilon}))$.

Initialization Size Although small initialization allows us to prove the global convergence of GD, a larger initialization is preferred because the convergence time, $T^\star$, is inversely proportional to β_0. When the sample complexity is optimal, i.e. $n^2 p \asymp n\,\mathrm{poly}(\log n)$, a bound on the initialization size provided by Theorem 1 reads $n^{-\frac{1}{4}}$ ignoring the log factors. However, as more samples are provided, we are allowed to use larger initialization to reduce the convergence time. When the sample complexity satisfies $n^2 p \asymp n^{1+a}$, the bound reads $n^{-\frac{1}{4}(1-a)}$ ignoring the log factors. The bound becomes near constant as a approaches 1, namely the fully observed case, and this agrees with the previous result that small initialization is unnecessary for the fully observed case (Ye & Du, 2021).

The next main result is about the trajectory of GD before it enters the local region. The theorem states that for all $t \leq T^\star$, $\boldsymbol{x}^{(t)}$ stays close to the fully observed case $\widetilde{\boldsymbol{x}}^{(t)}$ in both ℓ_2 and ℓ_∞-norm.

Theorem 2. *Suppose that the conditions of Theorem 1 hold and* $T^\star$ *be defined as in Theorem 1. Then, for all* $t \leq T^\star$*, we have*

$$\left\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \lesssim \frac{1}{\sqrt{\log n}} \left\|\widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \tag{10}$$

$$\left\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_\infty \lesssim \frac{1}{\sqrt{\log n}} \left\|\widetilde{\boldsymbol{x}}^{(t)}\right\|_\infty \tag{11}$$

with probability at least $1 - o(1)$.

Trajectory of GD The sequence $\widetilde{\boldsymbol{x}}^{(t)}$ is a linear combination of $\boldsymbol{x}^{(0)}$ and $\boldsymbol{u}^\star$, and it is easy to analyze how $\widetilde{\boldsymbol{x}}^{(t)}$ evolves. By showing that $\boldsymbol{x}^{(t)}$ stays close to $\widetilde{\boldsymbol{x}}^{(t)}$ for all iterations, we not only show the convergence of GD with small initialization as in Theorem 1 but also characterize the exact trajectory that GD follows by Theorem 2.

Implicit Regularization One can prove that $\widetilde{\boldsymbol{x}}^{(t)}$ is incoherent up to logarithmic factor throughout the whole iterations, and from (10) and (11), incoherence of $\boldsymbol{x}^{(t)}$ is bounded by that of $\widetilde{\boldsymbol{x}}^{(t)}$. Hence, Theorem 2 shows that the incoherence of $\boldsymbol{x}^{(t)}$ is *implicitly* controlled by GD without any regularizer. This is an improvement over the previous result on global convergence of GD for matrix completion (Ge et al., 2016), where explicit regularizer was employed to control ℓ_∞-norm of $\boldsymbol{x}^{(t)}$, although small initialization was not used in that work.

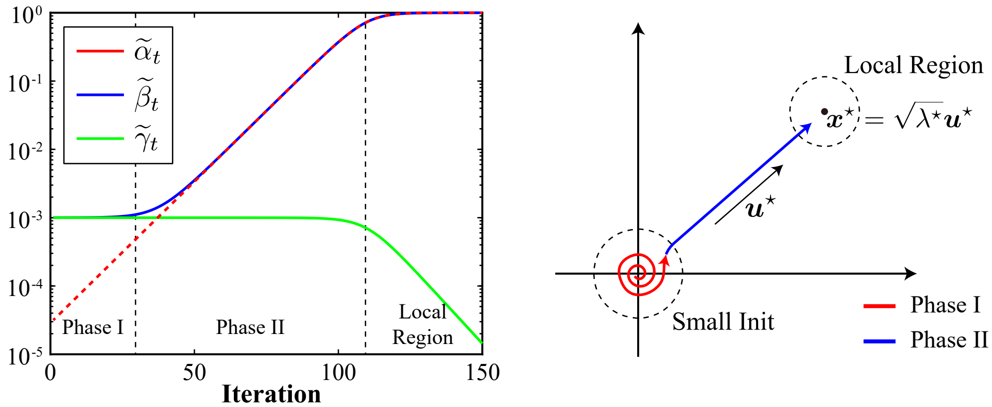

Figure 1: (left) Evolution of the quantities $\widetilde{\alpha}_t$, $\widetilde{\beta}_t$, $\widetilde{\gamma}_t$ simulated with $\widetilde{\alpha}_0 = \frac{1}{\sqrt{n}}\beta_0$, $\beta_0 = \frac{1}{n}$, $\lambda^\star = 1$, and $n = 1000$. (right) Trajectory of $\widetilde{\boldsymbol{x}}^{(t)}$. It starts from a point that is close to the origin due to small initialization. It finds the direction $\boldsymbol{u}^\star$ while rotating around the origin, and once it finds the direction, it keeps the direction and gets expanded.

4 Fully Observed Case and Proof Sketch

Before we explain about the proof of Theorems 1 and 2, we describe the trajectory of fully observed case. We characterize $\widetilde{\boldsymbol{x}}^{(t)}$ with three variables: $\widetilde{\alpha}_t = \left|\boldsymbol{u}^{\star\top}\widetilde{\boldsymbol{x}}^{(t)}\right|$, $\widetilde{\beta}_t = \|\widetilde{\boldsymbol{x}}^{(t)}\|_2$, $\widetilde{\gamma}_t = \|\widetilde{\boldsymbol{x}}^{(t)}_\perp\|_2$. According to (3), the three variables evolve with

$$\widetilde{\alpha}_{t+1} = (1 - \eta\widetilde{\beta}_t^2 + \eta\lambda^\star)\widetilde{\alpha}_t, \quad \widetilde{\gamma}_{t+1} = (1 - \eta\widetilde{\beta}_t^2)\widetilde{\gamma}_t, \quad \widetilde{\beta}_t^2 = \widetilde{\alpha}_t^2 + \widetilde{\gamma}_t^2.$$

At $t = 0$, due to random initialization, the initial vector is almost orthogonal to $\boldsymbol{u}^\star$, and we have $\widetilde{\alpha}_0 \approx \frac{1}{\sqrt{n}}\beta_0$ and $\widetilde{\gamma}_0 \approx \widetilde{\beta}_0 = \beta_0$. Also, due to small initialization, the term $\eta\widetilde{\beta}_t^2$ is ignorable until $\widetilde{\beta}_t$ becomes sufficiently large, so $\widetilde{\alpha}_t$ increases exponentially with the rate $1 + \eta\lambda^\star$ while $\widetilde{\gamma}_t$ stays still. Hence, at the early iterations where $(1 + \eta\lambda^\star)^t$ is still much less than $\sqrt{n}$, $\widetilde{\beta}_t$ is kept to its initial value β_0, while the trajectory becomes more parallel to $\boldsymbol{u}^\star$ because $\widetilde{\alpha}_t$ increases. When $(1 + \eta\lambda^\star)^t$ gets much larger than $\sqrt{n}$, the trajectory becomes almost parallel to $\boldsymbol{u}^\star$ in that $\widetilde{\beta}_t \approx \widetilde{\alpha}_t \gg \widetilde{\gamma}_t$. Until $\widetilde{\beta}_t$ reaches $\frac{1}{\log n}$, we can consider $\widetilde{\alpha}_t$ as increasing in a rate $(1 + \eta\lambda^\star)$, and it takes about $\frac{1}{\log(1+\eta\lambda^\star)} \log \frac{\sqrt{n}}{\beta_0}$ steps to reach this point. After this, we can no longer ignore $\widetilde{\beta}_t^2$, and $\widetilde{\alpha}_t$ increases with slower rate as $\widetilde{\beta}_t$ increases, but we can show that $\widetilde{\beta}_t^2$ becomes sufficiently close to $\lambda^\star$ within $\log\log n$ iterations. Finally, local convergence to $\boldsymbol{u}^\star$ occurs in that $\widetilde{\alpha}_t$ approaches $\lambda^\star$ and $\widetilde{\gamma}_t$ decreases exponentially with the rate $(1 - \eta\lambda^\star)$. We plotted the actual behavior of quantities $\widetilde{\alpha}_t$, $\widetilde{\beta}_t$, $\widetilde{\gamma}_t$ on the left side of Fig. 1.

We define the iterates before $\widetilde{\alpha}_t$ reaches $\frac{1}{\sqrt{np}}\beta_0$ as Phase I, and the next iterates before $\widetilde{\beta}_t^2$ reaches $\lambda^\star\left(1 - \frac{1}{\log n}\right)$ as Phase II. We analyze the trajectory of $\boldsymbol{x}^{(t)}$ in each phase separately. In Phase I, we take advantage of the small random initialization to show that the deviation of $\boldsymbol{x}^{(t)}$ from $\widetilde{\boldsymbol{x}}^{(t)}$ does not increase much and is kept to $\sqrt{\frac{1}{np}}$ times the norms of $\boldsymbol{x}^{(t)}$. In Phase II, we show that $\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}$ expands at a constant rate $(1 + \eta\lambda^\star)$. The next two sections give main lemmas that will be used to prove Theorems 1 and 2.

5 Phase I: Finding Direction

We provide detailed results and proof ideas for Phase I in this section. Our main goal is to analyze the deviation of $\boldsymbol{x}^{(t)}$ from $\widetilde{\boldsymbol{x}}^{(t)}$. First, if we look at the update equations (2) and (3), the second term is proportional to the third power of $\left\|\boldsymbol{x}^{(t)}\right\|_2$, while the other terms linearly depend on $\left\|\boldsymbol{x}^{(t)}\right\|_2$.

Hence, the second term becomes almost ignorable due to small initialization. Without the second terms, the difference between $\boldsymbol{x}^{(t)}$ and $\widetilde{\boldsymbol{x}}^{(t)}$ at $t = 1$ is $\eta(\boldsymbol{M}^\circ - \boldsymbol{M}^\star)\boldsymbol{x}^{(0)}$. From concentration inequalities, one can show that the ℓ_2 and ℓ_∞-norms of $\boldsymbol{x}^{(1)} - \widetilde{\boldsymbol{x}}^{(1)}$ are bounded by

$$\left\|\eta(\boldsymbol{M}^\circ - \boldsymbol{M}^\star)\boldsymbol{x}^{(0)}\right\|_2 \lesssim \mu\sqrt{\frac{\log n}{np}}\beta_0 \quad \text{and} \quad \left\|\eta(\boldsymbol{M}^\circ - \boldsymbol{M}^\star)\boldsymbol{x}^{(0)}\right\|_\infty \lesssim \sqrt{\frac{\mu \log^2 n}{np}}\frac{\beta_0}{\sqrt{n}},$$

respectively. Thus, the norms of $\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}$ are about $\frac{1}{\sqrt{np}}$ times smaller than those of $\widetilde{\boldsymbol{x}}^{(t)}$ at $t = 1$.

Due to the third terms of (2) and (3), the norms of $\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}$ can increase exponentially at a rate $(1+\eta\lambda^\star)$ in the worst case where $\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}$ is parallel to $\boldsymbol{u}^\star$. In such case, the norms of $\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}$ would be larger than those of $\widetilde{\boldsymbol{x}}^{(t)}$ at the end of Phase I because those of $\widetilde{\boldsymbol{x}}^{(t)}$ remain still in Phase I. However, we overcome this issue by proving that the bounds increase at most *polynomially* with respect to t, and because t is at most $O(\log n)$, the bounds remain $\frac{1}{\sqrt{np}}$ times smaller than the norms of $\widetilde{\boldsymbol{x}}^{(t)}$ throughout Phase I.

Lemma 3. *Let T_1 be the largest t such that $(1+\eta\lambda^\star)^t \le \sqrt{\frac{\log^{21} n}{np}}\sqrt{n}$. Suppose that the conditions of Theorem 1 hold. Then, for all $t \le T_1$, we have*

$$\left\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \lesssim \mu\sqrt{\frac{\log n}{np}}\beta_0 t, \tag{12}$$

$$\left\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_\infty \lesssim \sqrt{\frac{\mu^3 \log^2 n}{np}}\frac{\beta_0}{\sqrt{n}}t^2 \tag{13}$$

with probability at least $1 - o(1)$.

T_1 is defined to be the end of Phase I. Lemma 3 proves Theorem 2 for Phase I.

5.1 Proof of (12)

We first demonstrate how to obtain the ℓ_2-norm bound of Lemma 3. Let us define a sequence $\widehat{\boldsymbol{x}}^{(t)}$ that is updated with

$$\widehat{\boldsymbol{x}}^{(t+1)} = \widehat{\boldsymbol{x}}^{(t)} - \eta\left\|\widetilde{\boldsymbol{x}}^{(t)}\right\|_2^2\widehat{\boldsymbol{x}}^{(t)} + \eta\boldsymbol{M}^\circ\widehat{\boldsymbol{x}}^{(t)}, \quad \widehat{\boldsymbol{x}}^{(0)} = \boldsymbol{x}^{(0)}. \tag{14}$$

Note that the norm of $\widetilde{\boldsymbol{x}}^{(t)}$ is used in the second term of (14). The update equation of $\widehat{\boldsymbol{x}}^{(t)}$ differs on the third term compared to $\widetilde{\boldsymbol{x}}^{(t)}$ and on the second term compared to $\boldsymbol{x}^{(t)}$. We use $\widehat{\boldsymbol{x}}^{(t)}$ as a proxy for bounding $\left\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_2$. We first show that $\left\|\widehat{\boldsymbol{x}}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_2$ increases at most linearly with respect to t.

Lemma 4. *For all $t \le T_1$, we have*

$$\left\|\widehat{\boldsymbol{x}}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \lesssim \mu\sqrt{\frac{\log n}{np}}\beta_0 t \tag{15}$$

with probability at least $1 - o(1)$ if $n^2 p \gtrsim \mu^2 n \log^8 n$ and $\eta\lambda^\star < 0.1$.

The proof of this lemma is based on the fact that $\widehat{\boldsymbol{x}}^{(t)}$ is a product between $\boldsymbol{x}^{(0)}$ and a matrix polynomial of $\boldsymbol{I}$, $\boldsymbol{M}^\circ$, while $\widetilde{\boldsymbol{x}}^{(t)}$ is a product between $\boldsymbol{x}^{(0)}$ and a matrix polynomial of $\boldsymbol{I}$, $\boldsymbol{M}^\star$. We prove the lemma by comparing the two matrix polynomials. We remark that Lemma 4 holds regardless of the small initialization, but it relies on the randomness of $\boldsymbol{x}^{(0)}$.

Because $\boldsymbol{x}^{(t)}$ and $\widehat{\boldsymbol{x}}^{(t)}$ differ only on the second term, their initial difference is proportional to β_0^3. More precisely, it is $\frac{1}{\sqrt{np}}\beta_0^3$. We show that the difference increases exponentially at a rate $(1+\eta\lambda^\star)$.

Lemma 5. *If (13) holds for all $t \le T_1$, we have*

$$\left\|\boldsymbol{x}^{(t)} - \widehat{\boldsymbol{x}}^{(t)}\right\|_2 \lesssim \sqrt{\frac{\log^6 n}{np}}(1+\eta\lambda^\star)^t\beta_0^3 \tag{16}$$

for all $t \le T_1$ with probability at least $1 - o(1)$.

Due to the small initialization, one can prove that $\left\|\boldsymbol{x}^{(t)}-\widehat{\boldsymbol{x}}^{(t)}\right\|_2$ is ignorable compared to $\left\|\widehat{\boldsymbol{x}}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}\right\|_2$ in Phase I, and (12) is proved by (15). Note that the bound on β_0, (9), in Theorem 1, is obtained from here.

5.2 PROOF OF (13)

We control the lth component of $\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}$ with the help of lth leave-one-out sequence. The leave-one-out sequences have two important properties. First, because they are defined without only one row/column, they are extremely close to $\boldsymbol{x}^{(t)}$, and at $t=1$, $\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}\right\|_2$ is about $\frac{1}{\sqrt{np}}\frac{1}{\sqrt{n}}$. Second, the lth component of the lth leave-one-out sequence evolves similar to that of $\widetilde{\boldsymbol{x}}^{(t)}$, and it is easy to analyze. With these two properties, we bound the lth component of $\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}$ as

$$\left|\left(\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}\right)_l\right| \le \left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}\right\|_2 + \left|\left(\boldsymbol{x}^{(t,l)}-\widetilde{\boldsymbol{x}}^{(t)}\right)_l\right|. \tag{17}$$

We claim that both $\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}\right\|_2$ and $\left|\left(\boldsymbol{x}^{(t,l)}-\widetilde{\boldsymbol{x}}^{(t)}\right)_l\right|$ increase at most polynomially with respect to t from the initial scale $\frac{1}{\sqrt{np}}\frac{\beta_0}{\sqrt{n}}$.

Lemma 6. *For all $t \le T_1$, we have*

$$\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}\right\|_2 \lesssim \mu\sqrt{\frac{\log n}{np}}\frac{\beta_0}{\sqrt{n}}t \tag{18}$$

$$\left|\left(\boldsymbol{x}^{(t,l)}-\widetilde{\boldsymbol{x}}^{(t)}\right)_l\right| \lesssim \sqrt{\frac{\mu^3\log^2 n}{np}}\frac{\beta_0}{\sqrt{n}}t^2, \tag{19}$$

with probability at least $1-o(1)$.

As explained for $\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}$, due to the third terms of (2) and (14), $\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}$ can also increase exponentially with the rate $(1+\eta\lambda^\star)$ in the worst case where $\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}$ is parallel to $\boldsymbol{u}^\star$. This is contrary to our result (18) that $\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}\right\|_2$ increases only linearly. We show that $\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}$ is maintained almost orthogonal to $\boldsymbol{u}^\star$ in Phase I, and thus the worst case does not happen.

Lemma 7. *For all $l \in [n]$ and $t \le T_1$, we have*

$$\left|\boldsymbol{u}^{(l)\top}(\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)})\right| \lesssim \mu\sqrt{\frac{\log^2 n}{np}}\frac{\beta_0}{n}(1+\eta\lambda^\star)^t$$

with probability at least $1-o(1)$, where $\mathbf{u}^{(l)}$ is the first eigenvector of $\boldsymbol{M}^{(l)}$.

Note that $\boldsymbol{u}^{(l)}$ is almost parallel to $\boldsymbol{u}^\star$. The $\boldsymbol{u}^{(l)}$ component of $\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}$ is initialized at the order $\frac{1}{\sqrt{np}}\frac{\beta_0}{n}$, which is $\frac{1}{\sqrt{n}}$ times smaller than $\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}\right\|_2$. Although it is increased exponentially, from the definition of T_1, the $\boldsymbol{u}^{(l)}$ component is maintained much smaller than $\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}\right\|_2$ in Phase I.

One can show that $\left|\left(\boldsymbol{x}^{(t,l)}-\widetilde{\boldsymbol{x}}^{(t)}\right)_l\right|$ increases by $\left\|\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}\right\|_2\|\boldsymbol{u}^\star\|_\infty$ in each step, and summation of the bound (12) gives (19). Finally, (13) is obtained by putting (18) and (19) into (17).

6 PHASE II: EXPANSION

In the next phase, we show that the bounds obtained in Phase I are increased at a rate $(1+\eta\lambda^\star)$.

Lemma 8. *Let T_2 be the largest t such that $\widetilde{\beta}_t^2 \le \lambda^\star\left(1-\frac{1}{\log n}\right)$. Then, for all $T_1 < t \le T_2$, we have*

$$\left\|\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \lesssim \mu\sqrt{\frac{\log^3 n}{np}}\beta_0(1+\eta\lambda^\star)^{t-T_1}, \tag{20}$$

$$\left\|\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}\right\|_{\infty} \lesssim \sqrt{\frac{\mu^3 \log^6 n}{np}} \frac{\beta_0}{\sqrt{n}}(1+\eta\lambda^\star)^{t-T_1}, \tag{21}$$

$$\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}\right\|_{2} \lesssim \mu\sqrt{\frac{\log^4 n}{np}} \frac{\beta_0}{\sqrt{n}}(1+\eta\lambda^\star)^{t-T_1}, \tag{22}$$

$$\left|\left(\boldsymbol{x}^{(t,l)}-\widetilde{\boldsymbol{x}}^{(t)}\right)_l\right| \lesssim \sqrt{\frac{\mu^3 \log^6 n}{np}} \frac{\beta_0}{\sqrt{n}}(1+\eta\lambda^\star)^{t-T_1}, \tag{23}$$

with probability at least $1-o(1)$.

T_2 is defined to be the end of Phase II. We explain how Lemma 8 leads to Theorem 2 in Phase II. We first focus on (20) and (10). We can divide Phase II into three parts according to the behavior of $\|\widetilde{\boldsymbol{x}}^{(t)}\|_2$. First, $\|\widetilde{\boldsymbol{x}}^{(t)}\|_2$ is maintained to β_0 until $(1+\eta\lambda^\star)^t$ becomes $\sqrt{n}$, or $(1+\eta\lambda^\star)^{t-T_1}$ becomes $\sqrt{np}$. In this part, although the bounds increase exponentially with the rate $(1+\eta\lambda^\star)$, $\frac{1}{\sqrt{np}}$ factor that was already present in (12) of Phase I compensates this increment. At the end of the first part, $\|\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}\|_2$ is smaller than $\|\widetilde{\boldsymbol{x}}^{(t)}\|_2$ by some log factors. Next, $\|\widetilde{\boldsymbol{x}}^{(t)}\|_2$ increases with the rate $(1+\eta\lambda^\star)$ until it reaches $\frac{1}{\log n}$. Because both $\|\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}\|_2$ and $\|\widetilde{\boldsymbol{x}}^{(t)}\|_2$ increase with $(1+\eta\lambda^\star)$, the ratio between them is maintained in the second part. Lastly, in the remaining iterations, $\|\widetilde{\boldsymbol{x}}^{(t)}\|_2$ increases with $(1-\eta\widetilde{\beta}_t^2+\eta\lambda^\star)$ at each step, and the increment becomes smaller as it converges to $\sqrt{\lambda^\star}$. Hence, as in the first part, $\|\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}\|_2$ increases faster than $\|\widetilde{\boldsymbol{x}}^{(t)}\|_2$. However, as explained in Section 4, the length of this part is at most $\log\log n$, so the ratio between $\|\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}\|_2$ and $\|\widetilde{\boldsymbol{x}}^{(t)}\|_2$ increases only by some log factors. We prove that the log factors already present at the end of the second part compensates this, and finally (10) holds for all t in Phase II. A similar argument can be used to prove that the bounds for $\|\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}\|_\infty$, $\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}\|_2$, and $|(\boldsymbol{x}^{(t,l)}-\widetilde{\boldsymbol{x}}^{(t)})_l|$ are less than $\|\widetilde{\boldsymbol{x}}^{(t)}\|_\infty$ by some log factors throughout Phase II.

At the end of Phase II, $\widetilde{\boldsymbol{x}}^{(t)}$ is very close to $\pm\boldsymbol{x}^\star$ in both ℓ_2 and ℓ_∞-norms, so one can substitute $\widetilde{\boldsymbol{x}}^{(t)}$ of Lemma 8 with $\pm\boldsymbol{x}^\star$ to prove Theorem 1.

7 SIMULATION

In this section, we provide some simulation results that support our theoretical findings.

Trajectory of GD With the dimension $n=5000$, we constructed the ground truth vector $\boldsymbol{u}^\star$ by sampling it from the Gaussian distribution $\mathcal{N}(\mathbf{0},\frac{1}{n}\boldsymbol{I})$ and normalized it to have unit norm. We let $\lambda^\star=1$ so that the matrix $\boldsymbol{M}^\star$ is given by $\boldsymbol{u}^\star\boldsymbol{u}^{\star\top}$, and we randomly sampled the matrix symmetrically with sampling rate $p=0.1$. The initialization size was set to $\beta_0=\frac{1}{n}$ and step size of 0.1 was used for GD. Fig. 2 represents one trial of the experiment, but similar graphs were obtained in every repetition of the experiment. The evolution of some important quantities such as $\|\boldsymbol{x}^{(t)}\|_2$ and $|\boldsymbol{u}^{\star\top}\boldsymbol{x}^{(t)}|$ are depicted on the left side of Fig. 2. As observed in the fully observed case, the signal component $|\boldsymbol{u}^{\star\top}\boldsymbol{x}^{(0)}|$ increases with the rate $(1+\eta\lambda^\star)$ until it gets close to $\sqrt{\lambda^\star}$, and local convergence to $\boldsymbol{x}^\star$ occurs, where $\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^\star\|_2$ decreases exponentially. On the right side of Fig. 2, we describe the deviation of $\boldsymbol{x}^{(t)}$ from $\widetilde{\boldsymbol{x}}^{(t)}$ in both ℓ_2 and ℓ_∞-norms. The solid lines represent norms of $\boldsymbol{x}^{(t)}$ and the dotted lines represent those of $\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}$. We could observe that there is a gap between solid and dotted lines throughout whole iterations. Hence, $\boldsymbol{x}^{(t)}$ stays close to the trajectory of fully observed case as we proved in Theorem 2.

Small Initialization In the next experiment, we assessed the importance of small initialization for convergence of GD. We constructed the matrix $\boldsymbol{M}^\star$ as in the first experiment, but with $n=500$ this time. We measured the distance between $\boldsymbol{x}^{(t)}$ and $\boldsymbol{x}^\star$ at $t=\frac{1}{\log(1+\eta\lambda^\star)}\log\frac{\sqrt{n}}{\beta_0}+100$ and averaged it over 1000 trials. We repeated the experiment while changing the initialization size from 10^0 to 10^{-9} and the sampling probability from 0.01 to 0.04 with step size 0.1. The result is summarized to Fig. 3. In all sampling probabilities, the small initialization improves the convergence of GD.

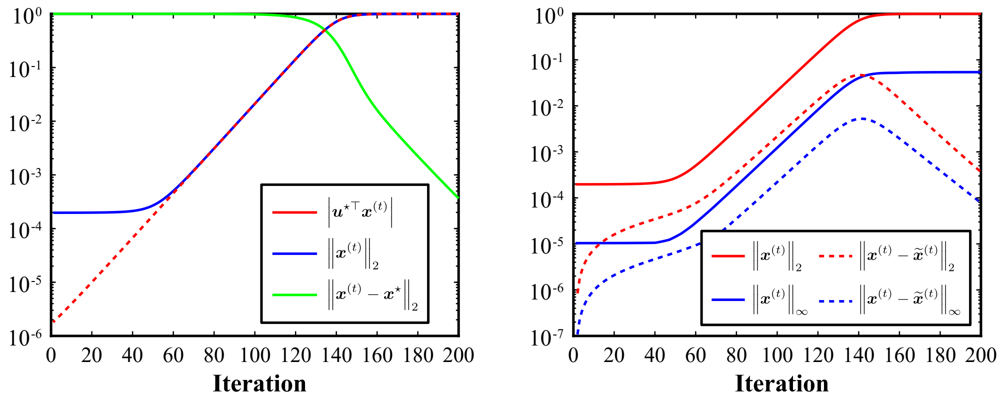

Figure 2: (left) Evolution of the quantities $\|\boldsymbol{x}^{(t)}\|$, $|\boldsymbol{u}^{\star\top}\boldsymbol{x}^{(t)}|$ that behaves similar to $\widetilde{\alpha}_t$, $\widetilde{\beta}_t$, respectively, and $\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^{\star}\|_2$ that shows local convergence. (right) Comparison between the norms of $\boldsymbol{x}^{(t)}$ and $\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}$.

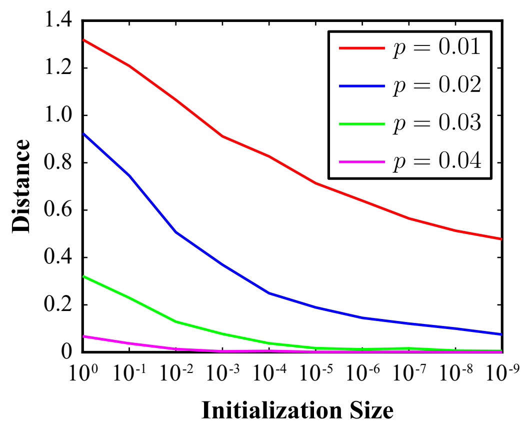

Figure 3: Convergence of GD with respect to initialization size and sampling probability.

Also, the performance starts to saturate at much larger initialization size as sampling probability increases, and this agrees with our finding (9) that a larger initialization is possible as more samples are available.

8 Conclusion

In this paper, we showed that for rank-1 symmetric matrix completion, GD can converge to the ground truth starting from a small random initialization. The result is interesting because the loss function of matrix completion does not have benign geometry globally if no regularizer is applied. Our result does not use any explicit regularizer and only rely on the implicit regularizing effect of GD. We expect that the results provided in this paper shed light on understanding the more general rank-r case, and we keep this as our future work. It would also be interesting to investigate the tightness of our bound on initialization size, which reads $n^{-\frac{1}{4}}$ when optimal number of samples are provided.

References

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics*, 48(3):1452–1474, 2020.

Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2014.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 7413–7424, 2019.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, 2016.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, 2011.

Ji Chen, Dekai Liu, and Xiaodong Li. Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization. *IEEE Transactions on Information Theory*, 66(9):5806–5841, 2020.

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.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1-2): 5–37, 2019.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685, 2019.

Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, 2017.

Justin Eldridge, Mikhail Belkin, and Yusu Wang. Unperturbed: spectral analysis beyond davis-kahan. In *Proceedings of Algorithmic Learning Theory*, pp. 321–358, 2018.

David F. Gleich and Lek-heng Lim. Rank aggregation via nuclear norm minimization. In *In Proceedings of the 17th ACM international conference on Knowledge discovery and data mining*, pp. 60–68. ACM, 2011.

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

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, 2017.

Yue Hu, Xiaohan Liu, and Mathews Jacob. A generalized structured low-rank matrix completion algorithm for mr image recovery. *IEEE Transactions on Medical Imaging*, 38(8):1841–1851, 2019.

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

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

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*, pp. 2–47, 2018.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *In International Conference on Learning Representations*, 2021.

Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20(3):451–632, 2020.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. In *Advances in Neural Information Processing Systems*, pp. 21174–21187, 2020.

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

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

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*, pp. 964–973, 2016.

Tian Ye and Simon S Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 1429–1439, 2021.

Detailed proofs for the results explained in the main text are provided in this supplementary. We say that an event happens with *very high probability* if it happens with probability at least $1 - \frac{1}{n^C}$ for some large constant $C > 0$. We say that an event happens with *high probability* if it happens with probability at least $1 - o(1)$. A union of $\mathrm{poly}(n)$ number of events that hold with very high probability still happens with high probability. For a matrix $\boldsymbol{A}$, we denote the spectral norm with $\|\boldsymbol{A}\|$ and maximum absolute value of the entries as $\|\boldsymbol{A}\|_\infty$. Also, the largest ℓ_2-norm of rows of $\boldsymbol{A}$ is denoted as $\|\boldsymbol{A}\|_{2,\infty}$.

A Spectral Analysis

We introduce some spectral bounds related to the matrices $\boldsymbol{M}^\circ$ and $\boldsymbol{M}^{(l)}$.

Lemma 9. *If $n^2 p \gtrsim n \log n$, we have*

$$\|\boldsymbol{M}^\circ - \boldsymbol{M}^\star\| \lesssim \lambda^\star \mu \sqrt{\frac{\log n}{np}}$$

with high probability.

Lemma 10. *If $n^2 p \gtrsim \mu n \log n$, for all $l \in [n]$, we have*

$$\left\|\boldsymbol{M}^\circ - \boldsymbol{M}^{(l)}\right\| \lesssim \lambda_1^\star \sqrt{\frac{r\mu}{np}}$$

with high probability.

The following bounds on eigenvectors are the results of Abbe et al. (2020).

Lemma 11. *If $n^2 p \gtrsim \mu^2 n \log n$, we have*

$$\|\boldsymbol{u}^\circ - \boldsymbol{u}^\star\|_2 \lesssim \mu \sqrt{\frac{\log n}{np}} \tag{24}$$

$$\|\boldsymbol{u}^\circ - \boldsymbol{u}^\star\|_\infty \lesssim \sqrt{\frac{\mu^3 \log n}{np}} \frac{1}{\sqrt{n}} \tag{25}$$

$$\left\|\boldsymbol{u}^\circ - \boldsymbol{u}^{(l)}\right\|_2 \lesssim \mu \sqrt{\frac{\log n}{np}} \frac{1}{\sqrt{n}} \tag{26}$$

with high probability.

Next, we state bounds on the eigenvalues of $\boldsymbol{M}^\circ$ and $\boldsymbol{M}^{(l)}$. The first eigenvalues of $\boldsymbol{M}^\circ$ and $\boldsymbol{M}^{(l)}$ are denoted as λ° and $\lambda^{(l)}$, respectively. The following lemma is derived from Lemmas 9 and 10 and Weyl's Theorem.

Lemma 12. *If $n^2 p \gtrsim \mu^2 n \log n$, we have*

$$|\lambda^\circ - \lambda^\star| \lesssim \lambda^\star \mu \sqrt{\frac{\log n}{np}} \tag{27}$$

$$\left|\lambda^{(l)} - \lambda^\star\right| \lesssim \lambda^\star \mu \sqrt{\frac{\log n}{np}} \tag{28}$$

for all $l \in [n]$ with high probability.

B Initialization

We introduce some results related to the initialization vector $\boldsymbol{x}^{(0)}$ in this section. Because $\boldsymbol{x}^{(0)}$ is sampled from a gaussian distribution, the results in this section are derived from basic properties of the gaussian distribution.

Lemma 13. *The size of $\boldsymbol{x}^{(0)}$ is in the range*

$$\beta_0\left(1-c\sqrt{\frac{\log n}{n}}\right) \le \left\|\boldsymbol{x}^{(0)}\right\|_2 \le \beta_0\left(1+c\sqrt{\frac{\log n}{n}}\right)$$

with high probability for some constant $c > 0$. The ℓ_∞-norm of $\boldsymbol{x}^{(0)}$ satisfies

$$\left\|\boldsymbol{x}^{(0)}\right\|_\infty \lesssim \sqrt{\log n}\frac{\beta_0}{\sqrt{n}}.$$

Lemma 14. *For a fixed vector $\boldsymbol{u}$,*

$$\frac{1}{\sqrt{\log n}}\frac{\beta_0}{\sqrt{n}}\|\boldsymbol{u}\|_2 \lesssim \left|\boldsymbol{u}^\top \boldsymbol{x}^{(0)}\right|$$

holds with high probability and

$$\left|\boldsymbol{u}^\top \boldsymbol{x}^{(0)}\right| \lesssim \sqrt{\log n}\frac{\beta_0}{\sqrt{n}}\|\boldsymbol{u}\|_2$$

holds with very high probability.

Lemma 14 implies that the $\boldsymbol{u}^\star$ component of $\boldsymbol{x}^{(0)}$ is in the range

$$\frac{1}{\sqrt{\log n}}\frac{\beta_0}{\sqrt{n}} \lesssim \left|\boldsymbol{u}^{\star\top}\boldsymbol{x}^{(0)}\right| \lesssim \sqrt{\log n}\frac{\beta_0}{\sqrt{n}}.$$

C Fully Observed Case

We provide some lemmas related to $\widetilde{\boldsymbol{x}}^{(t)}$ in this section. We first note that $\widetilde{\boldsymbol{x}}^{(t)}$ is explicitly written as

$$\widetilde{\boldsymbol{x}}^{(t)} = \prod_{s=1}^{t}(1-\eta\widetilde{\beta}_s^2)\boldsymbol{x}^{(0)} + \prod_{s=1}^{t}(1-\eta\widetilde{\beta}_s^2+\eta\lambda^\star)(\boldsymbol{u}^{\star\top}\boldsymbol{x}^{(0)})\boldsymbol{u}^\star.$$

Let us define T_2' as the last t such that $\widetilde{\beta}_t^2 \le \frac{\lambda^\star}{\log n \log\log n}$. Then, for all $t \le T_2'$, we have

$$\prod_{s=1}^{t}(1-\eta\widetilde{\beta}_s^2+\eta\lambda^\star) = (1+o(1))(1+\eta\lambda^\star)^t, \quad \prod_{s=1}^{t}(1-\eta\widetilde{\beta}_s^2) = 1+o(1)$$

because $\left(1+\frac{1}{\log n \log\log n}\right)^{\log n} = 1+o(1)$. The following lemma provides both upper and lower bounds for norms of $\widetilde{\boldsymbol{x}}^{(t)}$.

Lemma 15. *For all t, we have*

$$\frac{1}{\sqrt{\log n}}\left(1+\frac{(1+\eta\lambda^\star)^{\min\{t,T_2'\}}}{\sqrt{n}}\right)\beta_0 \lesssim \left\|\widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \lesssim \sqrt{\log n}\left(1+\frac{(1+\eta\lambda^\star)^t}{\sqrt{n}}\right)\beta_0$$

$$\frac{1}{\sqrt{\log n}}\left(1+\frac{(1+\eta\lambda^\star)^{\min\{t,T_2'\}}}{\sqrt{n}}\right)\frac{\beta_0}{\sqrt{n}} \lesssim \left\|\widetilde{\boldsymbol{x}}^{(t)}\right\|_\infty \lesssim \sqrt{\mu\log n}\left(1+\frac{(1+\eta\lambda^\star)^t}{\sqrt{n}}\right)\frac{\beta_0}{\sqrt{n}}$$

The lower bounds do not increase after $t = T_2'$. We have the tighter bounds

$$\left\|\widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \asymp \beta_0, \quad \frac{1}{\sqrt{n}}\beta_0 \le \left\|\widetilde{\boldsymbol{x}}^{(t)}\right\|_\infty \lesssim \sqrt{\log n}\frac{\beta_0}{\sqrt{n}} \tag{29}$$

in Phase I.

Until T_2', $\widetilde{\beta}_t$ increases with the rate $(1+\eta\lambda^\star)$, and hence $T_2' \asymp \frac{1}{\log(1+\eta\lambda^\star)}\log\frac{\sqrt{n}}{\beta_0}$. How much iterations will be required for it to reach $\sqrt{\lambda^\star}\sqrt{1-\frac{1}{\log n}}$ after T_2'? The following lemma proves that $O(\log\log n)$ iterations are required to reach the end of Phase II after T_2'. Note that $\alpha_t \approx \beta_t$ when $t \ge T_2'$, so both of them evolve with the recursive equation $x_{t+1} = (1-\eta x_t^2+\eta\lambda^\star)x_t$.

Lemma 16. *Let x_t evolves with the recursive equation*

$$x_{t+1} = (1 - \eta x_t^2 + \eta\lambda^\star)x_t, \quad x_0^2 = \frac{\lambda^\star}{\log n \log\log n}.$$

Then, at $t = \frac{6\log\log n}{\log(1+\eta\lambda^\star)}$, we have

$$x_t \geq \sqrt{\lambda^\star}\sqrt{1 - \frac{1}{\log n}},$$

and for all $t \leq \frac{6\log\log n}{\log(1+\eta\lambda^\star)}$, we have

$$\frac{(1+\eta\lambda^\star)^t x_0}{x_t} \lesssim \sqrt{\log^{11} n}.$$

Proof. For all $i \geq 1$, let N_i be the last t such that $\lambda^\star - x_t^2 \geq \frac{\lambda^\star}{e^i}$. Then, we have

$$\lambda^\star - x_{N_i}^2 \geq \frac{\lambda^\star}{e^i} > \lambda^\star - x_{N_i+1}^2. \tag{30}$$

Let $i \geq 2$. For all $N_{i-1} < t \leq N_i$,

$$1 + \frac{\eta\lambda^\star}{e^i} \leq \frac{x_{t+1}}{x_t},$$

and it implies

$$\left(1 + \frac{\eta\lambda^\star}{e^i}\right)^{N_i - N_{i-1} - 1} x_{N_{i-1}+1} \leq x_{N_i}.$$

From the lower and upper bounds provided by (30), we have

$$\sqrt{\lambda^\star}\sqrt{1 - \frac{1}{e^{i-1}}}\left(1 + \frac{\eta\lambda^\star}{e^i}\right)^{N_i - N_{i-1} - 1} \leq \sqrt{\lambda^\star}\sqrt{1 - \frac{1}{e^i}},$$

$$\left(1 + \frac{\eta\lambda^\star}{e^i}\right)^{2(N_i - N_{i-1} - 1)} \leq \frac{e^i - 1}{e^i}\frac{e^{i-1}}{e^{i-1} - 1} = \frac{e^{i-1} - \frac{1}{e}}{e^{i-1} - 1} \leq 1 + \frac{1}{e^{i-1}}.$$

Taking $\log$ on both sides and using the inequality $\frac{1}{2}x < \log(1+x) < x$ that holds for $0 < x < 1$, we get

$$N_i - N_{i-1} \leq 1 + \frac{1}{2}\frac{\log\left(1 + \frac{1}{e^{i-1}}\right)}{\log\left(1 + \frac{\eta\lambda^\star}{e^i}\right)} \leq 1 + \frac{e}{\eta\lambda^\star}.$$

For $t \leq N_1$, we have

$$1 + \frac{\eta\lambda^\star}{e} \leq \frac{x_{t+1}}{x_t},$$

and thus

$$\sqrt{\lambda^\star}\sqrt{1 - \frac{1}{e}} \geq x_{N_1} \geq \left(1 + \frac{\eta\lambda^\star}{e}\right)^{N_1} x_0 = \left(1 + \frac{\eta\lambda^\star}{e}\right)^{N_1}\sqrt{\frac{\lambda^\star}{\log n \log\log n}}.$$

Taking $\log$ on both sides we get

$$N_1 \leq 3\frac{\log\log n}{\eta\lambda^\star}$$

Hence, we have

$$\begin{aligned} N_{\log\log n} + 1 &\leq \left(1 + \frac{e}{\eta\lambda^\star}\right)(\log\log n - 1) + N_1 + 1 \\ &\leq \left(1 + \frac{e}{\eta\lambda^\star}\right)(\log\log n - 1) + 3\frac{\log\log n}{\eta\lambda^\star} + 1 \\ &\leq \frac{6\log\log n}{\log(1+\eta\lambda^\star)} \end{aligned}$$

but at $t = N_{\log\log n} + 1$, it holds that

$$x_t^2 > \lambda^\star\left(1 - \frac{1}{\log n}\right).$$

□

Hence, we have $T_2 \lesssim \frac{1}{\log(1+\eta\lambda^\star)} \log \frac{\sqrt{n}}{\beta_0} + \log\log n$. At the end of Phase II, $\widetilde{\boldsymbol{x}}^{(t)}$ is sufficiently close to $\boldsymbol{x}^\star$ in the sense that

$$\min\left\{\left\|\widetilde{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^\star\right\|_2, \left\|\widetilde{\boldsymbol{x}}^{(t)} + \boldsymbol{x}^\star\right\|_2\right\} \lesssim \frac{1}{\sqrt{\log n}} \|\boldsymbol{x}^\star\|_2, \tag{31}$$

$$\min\left\{\left\|\widetilde{\boldsymbol{x}}^{(t)} - \boldsymbol{x}^\star\right\|_\infty, \left\|\widetilde{\boldsymbol{x}}^{(t)} + \boldsymbol{x}^\star\right\|_\infty\right\} \lesssim \frac{1}{\sqrt{\log n}} \|\boldsymbol{x}^\star\|_\infty. \tag{32}$$

D Phase I

D.1 Proof of Lemma 4

Let us rewrite the update equations (3) and (14) as

$$\widetilde{\boldsymbol{x}}^{(t+1)} = \left(\boldsymbol{I} - \eta\widetilde{\beta}_t^2 + \eta \boldsymbol{M}^\star\right)\widetilde{\boldsymbol{x}}^{(t)},$$
$$\widehat{\boldsymbol{x}}^{(t+1)} = \left(\boldsymbol{I} - \eta\widetilde{\beta}_t^2 + \eta \boldsymbol{M}^\circ\right)\widehat{\boldsymbol{x}}^{(t)},$$

where $\widetilde{\beta}_t = \left\|\widetilde{\boldsymbol{x}}^{(t)}\right\|_2^2$. Then, $\widehat{\boldsymbol{x}}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}$ is a product between $\boldsymbol{x}^{(0)}$ and $P^{(t)}(\boldsymbol{I}, \boldsymbol{M}^\star, \boldsymbol{H})$, which is a matrix polynomial of $\boldsymbol{I}, \boldsymbol{M}^\star, \boldsymbol{H}$, where $\boldsymbol{H} = \boldsymbol{M}^\circ - \boldsymbol{M}^\star$.

$$P^{(t)}(\boldsymbol{I}, \boldsymbol{M}^\star, \boldsymbol{H}) := \left(\prod_{s=1}^{t}\left((1-\eta\widetilde{\beta}_s^2)\boldsymbol{I} + \eta\boldsymbol{M}^\star + \eta\boldsymbol{H}\right) - \prod_{s=1}^{t}\left((1-\eta\widetilde{\beta}_s^2)\boldsymbol{I} + \eta\boldsymbol{M}^\star\right)\right) \tag{33}$$

$$\widehat{\boldsymbol{x}}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)} = P^{(t)}(\boldsymbol{I}, \boldsymbol{M}^\star, \boldsymbol{H})\boldsymbol{x}^{(0)} \tag{34}$$

We classify the terms that appear after expanding the matrix polynomial $P^{(t)}(\boldsymbol{I}, \boldsymbol{M}^\star, \boldsymbol{H})$ into two types; 1) the terms that contain $\boldsymbol{H}$ but not $\boldsymbol{M}^\star$, 2) the terms that contain both $\boldsymbol{H}$ and $\boldsymbol{M}^\star$. We define $P_1^{(t)}(\boldsymbol{I}, \boldsymbol{H})$ to be a matrix polynomial of $\boldsymbol{I}$ and $\boldsymbol{H}$, which is equal to summation of the first type, and it is explicitly written as

$$P_1^{(t)}(\boldsymbol{I}, \boldsymbol{H}) = \prod_{s=1}^{t}\left((1-\eta\widetilde{\beta}_s^2)\boldsymbol{I} + \eta\boldsymbol{H}\right) - \prod_{s=1}^{t}(1-\eta\widetilde{\beta}_s^2)\boldsymbol{I}.$$

We correspondingly define $P_2^{(t)}(\boldsymbol{I}, \boldsymbol{M}^\star, \boldsymbol{H})$ to be summation of the second type, and it is equal to

$$P_2^{(t)}(\boldsymbol{I}, \boldsymbol{M}^\star, \boldsymbol{H}) = P^{(t)}(\boldsymbol{I}, \boldsymbol{M}^\star, \boldsymbol{H}) - P_1^{(t)}(\boldsymbol{I}, \boldsymbol{H}).$$

For $x, y \in \mathbb{R}$, we define $P_1^{(t)}(x, y)$ as the value that is obtained by substituting x, y instead of $\boldsymbol{I}, \boldsymbol{H}$, respectively. For example, $P_1^{(t)}(1, 2) = \prod_{s=1}^t (1 - \eta\widetilde{\beta}_s^2 + 2\eta) - \prod_{s=1}^t (1-\eta\widetilde{\beta}_s^2)$. For $x, y, z \in \mathbb{R}$, $P_2^{(t)}(x, y, z)$ is defined in a similar manner.

We bound the contribution of each type separately because the triangle inequality gives

$$\left\|\widehat{\boldsymbol{x}}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \le \left\|P_1^{(t)}(\boldsymbol{I}, \boldsymbol{H})\boldsymbol{x}^{(0)}\right\|_2 + \left\|P_2^{(t)}(\boldsymbol{I}, \boldsymbol{M}^\star, \boldsymbol{H})\boldsymbol{x}^{(0)}\right\|_2.$$

Every term in $P_1^{(t)}(\boldsymbol{I}, \boldsymbol{H})$ is $\boldsymbol{H}^s$ times a constant. We have $\left\|\boldsymbol{H}^s\boldsymbol{x}^{(0)}\right\|_2 \le \|\boldsymbol{H}\|^s \left\|\boldsymbol{x}^{(0)}\right\|_2$, and hence with triangle inequality

$$\left\|P_1^{(t)}(\boldsymbol{I}, \boldsymbol{H})\boldsymbol{x}^{(0)}\right\|_2 \le P_1^{(t)}(1, \eta\|\boldsymbol{H}\|)\beta_0.$$

If $n^2 p \gtrsim \mu^2 n \log^3 n$, we can further bound $P_1^{(t)}(1, \|\boldsymbol{H}\|)$ as

$$\begin{aligned} P_1^{(t)}(1, \|\boldsymbol{H}\|) &= \prod_{s=1}^{t}(1-\eta\widetilde{\beta}_s^2 + \eta\|\boldsymbol{H}\|) - \prod_{s=1}^{t}(1-\eta\widetilde{\beta}_s^2) \\ &= \prod_{s=1}^{t}(1-\eta\widetilde{\beta}_s^2)\left(\prod_{s=1}^{t}\left(1 + \frac{\eta\|\boldsymbol{H}\|}{1-\eta\widetilde{\beta}_s^2}\right) - 1\right) \end{aligned}$$

$$\leq \prod_{s=1}^{t}(1-\eta\widetilde{\beta}_s^2)\left(\left(1+\frac{\eta}{1-\eta\lambda^\star}\|\boldsymbol{H}\|\right)^t-1\right)$$
$$\leq \prod_{s=1}^{t}(1-\eta\widetilde{\beta}_s^2)\left(\exp\left(\frac{\eta}{1-\eta\lambda^\star}\|\boldsymbol{H}\|t\right)-1\right)$$
$$\lesssim \eta\|\boldsymbol{H}\|t\prod_{s=1}^{t}(1-\eta\widetilde{\beta}_s^2)$$

The fourth and fifth lines are derived from an elementary inequality $1+x \leq e^x \leq 1+2x$, which holds for small $x>0$.

We can decompose every term of second type as a product of η, $\lambda^\star$, $\boldsymbol{u}^\star$, $\boldsymbol{H}^s\boldsymbol{u}^\star$, $\boldsymbol{u}^{\star\top}\boldsymbol{H}^s\boldsymbol{x}^{(0)}$, $\boldsymbol{u}^{\star\top}\boldsymbol{H}^s\boldsymbol{u}^\star$, and $\boldsymbol{u}^{\star\top}\boldsymbol{x}^{(0)}$. We describe this with some examples.

$$(\eta\boldsymbol{H})^{s_1}(\eta\boldsymbol{M}^\star)(\eta\boldsymbol{H})^{s_2}\boldsymbol{x}^{(0)} = \eta^{s_1+s_2+1}\lambda^\star(\boldsymbol{H}^{s_1}\boldsymbol{u}^\star)(\boldsymbol{u}^{\star\top}\boldsymbol{H}^{s_2}\boldsymbol{x}^{(0)})$$
$$(\eta\boldsymbol{H})^{s}(\eta\boldsymbol{M}^\star)\boldsymbol{x}^{(0)} = \eta^{s+1}(\boldsymbol{H}^{s}\boldsymbol{u}^\star)(\boldsymbol{u}^{\star\top}\boldsymbol{x}^{(0)})$$
$$(\eta\boldsymbol{M}^\star)(\eta\boldsymbol{H})^{s}(\eta\boldsymbol{M}^\star)\boldsymbol{x}^{(0)} = \eta^{s+2}\lambda^{\star 2}\boldsymbol{u}^\star(\boldsymbol{u}^{\star\top}\boldsymbol{H}\boldsymbol{u}^\star)(\boldsymbol{u}^{\star\top}\boldsymbol{x}^{(0)})$$
$$(\eta\boldsymbol{M}^\star)(\eta\boldsymbol{H})^{s}\boldsymbol{x}^{(0)} = \eta^{s+1}\lambda^\star\boldsymbol{u}^\star(\boldsymbol{u}^{\star\top}\boldsymbol{H}^{s}\boldsymbol{x}^{(0)})$$

The terms $\boldsymbol{H}^x\boldsymbol{u}^\star$ and $\boldsymbol{u}^{\star\top}\boldsymbol{H}^s\boldsymbol{u}^\star$ are bounded with

$$\|\boldsymbol{H}^s\boldsymbol{u}^\star\|_2 \leq \|\boldsymbol{H}\|^s, \quad \left|\boldsymbol{u}^{\star\top}\boldsymbol{H}^s\boldsymbol{u}^\star\right| \leq \|\boldsymbol{H}\|^s. \tag{35}$$

For the terms that contain $\boldsymbol{x}^{(0)}$, we can apply Lemma 14 because the sampling is independent from the initialization. We have

$$\left|\boldsymbol{u}^{\star\top}\boldsymbol{x}^{(0)}\right| \leq \sqrt{\frac{\log n}{n}}\beta_0, \quad \left|\boldsymbol{u}^{\star\top}\boldsymbol{H}^s\boldsymbol{x}^{(0)}\right| \leq \sqrt{\frac{\log n}{n}}\|\boldsymbol{H}^s\boldsymbol{u}^\star\|\beta_0 \leq \sqrt{\frac{\log n}{n}}\|\boldsymbol{H}\|^s\beta_0 \tag{36}$$

for all $s \leq T_0 \lesssim \log n$ with high probability.

For every term of second type that includes s_1 times of $\eta\boldsymbol{M}^\star$ and s_2 times of $\eta\boldsymbol{H}$, the bounds (35) and (36) imply that ℓ_2-norm of the term multiplied by $\boldsymbol{x}^{(0)}$ is at most

$$(\eta\lambda^\star)^{s_1}(\eta\|\boldsymbol{H}\|)^{s_2}\sqrt{\frac{\log n}{n}}\beta_0.$$

Hence, similar to the first type, we have

$$\left\|P_2^{(t)}(\boldsymbol{I},\boldsymbol{M}^\star,\boldsymbol{H})\boldsymbol{x}^{(0)}\right\|_2 \leq P_2^{(t)}(1,\eta\lambda^\star,\eta\|\boldsymbol{H}\|)\sqrt{\frac{\log n}{n}}\beta_0.$$

If $n^2p \gtrsim \mu^2 n\log^3 n$, we can further bound $P_2^{(t)}(1,\eta\lambda^\star,\eta\|\boldsymbol{H}\|)$ as

$$P_2^{(t)}(1,\eta\lambda^\star,\eta\|\boldsymbol{H}\|) = \prod_{s=1}^{t}(1-\eta\beta_s^2+\eta\lambda^\star+\eta\|\boldsymbol{H}\|) - \prod_{s=1}^{t}\left(1-\eta\beta_s^2+\eta\lambda^\star\right) - P_1^{(t)}(1,\eta\|\boldsymbol{H}\|)$$
$$\leq \prod_{s=1}^{t}(1-\eta\beta_s^2+\eta\lambda^\star+\eta\|\boldsymbol{H}\|) - \prod_{s=1}^{t}\left(1-\eta\beta_s^2+\eta\lambda^\star\right)$$
$$\leq \left(\prod_{s=1}^{t}(1-\eta\beta_s^2+\eta\lambda^\star)\right)\left(\prod_{s=1}^{t}\left(1+\frac{\eta}{1-\eta\beta_s^2+\eta\lambda^\star}\|\boldsymbol{H}\|\right)-1\right)$$
$$\leq \left(\prod_{s=1}^{t}(1-\eta\beta_s^2+\eta\lambda^\star)\right)\left((1+\eta\|\boldsymbol{H}\|)^t-1\right)$$
$$\lesssim \eta\|\boldsymbol{H}\|t\prod_{s=1}^{t}(1-\eta\beta_s^2+\eta\lambda^\star).$$

Combining all, we have

$$\begin{aligned}\left\|\widehat{\boldsymbol{x}}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}\right\|_2 &\lesssim \eta\|\boldsymbol{H}\|t\left(\prod_{s=1}^{t}(1-\eta\beta_s^2)+\sqrt{\frac{\log n}{n}}\prod_{s=1}^{t}(1-\eta\beta_s^2+\eta\lambda^\star)\right)\beta_0\\ &\leq \eta\|\boldsymbol{H}\|t\left(1+\sqrt{\frac{\log n}{n}}(1+\eta\lambda^\star)^{T_0}\right)\beta_0\\ &\lesssim \eta\|\boldsymbol{H}\|t\left(1+\sqrt{\frac{\log^8 n}{np}}\right)\beta_0\\ &\lesssim \eta\lambda^\star\mu\sqrt{\frac{\log n}{np}}\beta_0 t\end{aligned}$$

if $n^2p \gtrsim \mu^2 n\log^8 n$.

D.2 Proof of Lemmas 5 to 7

We prove Lemmas 5 to 7 all together in an inductive manner.

Lemma 17. *For all $t \leq T_1$, we have*

$$\left\|\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \lesssim \mu\sqrt{\frac{\log n}{np}}\beta_0 t, \tag{37}$$

$$\left\|\boldsymbol{x}^{(t)}-\widetilde{\boldsymbol{x}}^{(t)}\right\|_\infty \lesssim \sqrt{\frac{\mu^3\log^2 n}{np}}\frac{\beta_0}{\sqrt{n}}t^2, \tag{38}$$

$$\left\|\boldsymbol{x}^{(t)}-\widehat{\boldsymbol{x}}^{(t)}\right\|_2 \lesssim \sqrt{\frac{\log^3 n}{np}}(1+\eta\lambda^\star)^t\beta_0^3, \tag{39}$$

$$\left\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)}\right\|_2 \lesssim \mu\sqrt{\frac{\log n}{np}}\frac{\beta_0}{\sqrt{n}}t, \tag{40}$$

$$\left|\boldsymbol{u}^{(l)\top}(\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)})\right| \lesssim \mu\sqrt{\frac{\log^2 n}{np}}(1+\eta\lambda^\star)^t\frac{\beta_0}{n}, \tag{41}$$

$$\left|\left(\boldsymbol{x}^{(t,l)}-\widetilde{\boldsymbol{x}}^{(t)}\right)_l\right| \lesssim \sqrt{\frac{\mu^3\log^2 n}{np}}\frac{\beta_0}{\sqrt{n}}t^2, \tag{42}$$

with high probability.

Before we start the proof, we introduce some notations. For $\boldsymbol{x}\in\mathbb{R}^n$, let us define

$$\|\boldsymbol{x}\|_{2,i}=\sqrt{\frac{1}{p}\sum_{j=1}^{n}\delta_{ij}x_j^2},\quad \boldsymbol{I}_{\boldsymbol{x}}=\frac{1}{\|\boldsymbol{x}\|_2^2}\operatorname{diag}(\|\boldsymbol{x}\|_{2,1}^2,\cdots,\|\boldsymbol{x}\|_{2,n}^2).$$

$\|\boldsymbol{x}\|_{2,i}$ is the ℓ_2-norm of $\boldsymbol{x}$ estimated with sampling of the ith row. With this notation, we can write the gradient of f as

$$\nabla f(\boldsymbol{x})=\|\boldsymbol{x}\|_2^2\boldsymbol{I}_{\boldsymbol{x}}\boldsymbol{x}-\boldsymbol{M}^\circ\boldsymbol{x}.$$

The function g is defined as

$$g(\boldsymbol{x})=\frac{1}{4p}\left\|\mathcal{P}_\Omega(\boldsymbol{x}\boldsymbol{x}^\top)\right\|_{\mathrm{F}}^2,$$

and its gradient satisfies

$$\nabla f(\boldsymbol{x})=\nabla g(\boldsymbol{x})-\boldsymbol{M}^\circ\boldsymbol{x}.$$

The hessian of $g(\boldsymbol{x})$ is equal to

$$\nabla^2 g(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 \boldsymbol{I}_{\boldsymbol{x}} + \frac{2}{p}\mathcal{P}_{\Omega}\big(\boldsymbol{x}\boldsymbol{x}^\top\big).$$

The base case $t = 0$ for induction hypotheses (37) to (42) trivially hold because all three sequences $\boldsymbol{x}^{(t)}$, $\widehat{\boldsymbol{x}}^{(t)}$, $\widetilde{\boldsymbol{x}}^{(t)}$ start from the same point. Now, we assume that the hypotheses hold up to the tth iteration and show that they hold at the $(t+1)$st iteration. For brevity, we drop the superscript (t) from $\boldsymbol{x}^{(t)}$, $\boldsymbol{x}^{(t,l)}$, $\widehat{\boldsymbol{x}}^{(t)}$, $\widetilde{\boldsymbol{x}}^{(t)}$ and denote them as $\boldsymbol{x}$, $\boldsymbol{x}^{(l)}$, $\widehat{\boldsymbol{x}}$, $\widetilde{\boldsymbol{x}}$, respectively. Also, recall that T_1 is defined to be the last t such that $(1+\eta\lambda^\star)^t \le \sqrt{\frac{\log^{21} n}{np}}\sqrt{n}$, and the magnitude of initialization satisfies $\beta_0^2 \lesssim \lambda^\star\sqrt{\frac{np}{\log^{24} n}}\frac{1}{\sqrt{n}}$ so that $(1+\eta\lambda^\star)^t\beta_0^2 = \frac{\lambda^\star}{\sqrt{\log^3 n}}$.

(39) at $(t+1)$ We first decompose $\boldsymbol{x}^{(t+1)} - \widehat{\boldsymbol{x}}^{(t+1)}$ as

$$\begin{aligned}
\boldsymbol{x}^{(t+1)} - \widehat{\boldsymbol{x}}^{(t+1)} &= (\boldsymbol{I} + \eta\boldsymbol{M}^\circ)(\boldsymbol{x} - \widehat{\boldsymbol{x}}) - \eta\|\boldsymbol{x}\|_2^2\boldsymbol{I}_{\boldsymbol{x}}\boldsymbol{x} + \eta\|\widetilde{\boldsymbol{x}}\|_2^2\widehat{\boldsymbol{x}} \\
&= \Big(\boldsymbol{I} - \eta\|\widetilde{\boldsymbol{x}}\|_2^2\boldsymbol{I} + \eta\boldsymbol{M}^\circ\Big)(\boldsymbol{x} - \widehat{\boldsymbol{x}}) - \eta\Big(\|\boldsymbol{x}\|_2^2\boldsymbol{I}_{\boldsymbol{x}} - \|\widetilde{\boldsymbol{x}}\|_2^2\boldsymbol{I}\Big)\boldsymbol{x}
\end{aligned}$$

With the help of Lemma 25, we bound the maximum entry of a diagonal matrix $\|\boldsymbol{x}\|_2^2\boldsymbol{I}_{\boldsymbol{x}} - \|\widetilde{\boldsymbol{x}}\|_2^2\boldsymbol{I}$.

$$\begin{aligned}
\max_{i\in[n]}\Big|\|\boldsymbol{x}\|_{2,i}^2 - \|\widetilde{\boldsymbol{x}}\|_2^2\Big| &\lesssim n\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_\infty\|\boldsymbol{x} + \widetilde{\boldsymbol{x}}\|_\infty + \sqrt{\frac{\log n}{p}}\|\widetilde{\boldsymbol{x}}\|_2\|\widetilde{\boldsymbol{x}}\|_\infty + \frac{\log n}{p}\|\widetilde{\boldsymbol{x}}\|_\infty^2 \\
&\lesssim \sqrt{\frac{\log^2 n}{np}}\beta_0^2 t^2 + \sqrt{\frac{\log^2 n}{np}}\beta_0^2 + \frac{\log^2 n}{np}\beta_0^2 \lesssim \sqrt{\frac{\log^2 n}{np}}\beta_0^2(t^2+1)
\end{aligned}$$

Hence, we have

$$\begin{aligned}
\Big\|\boldsymbol{x}^{(t+1)} - \widehat{\boldsymbol{x}}^{(t+1)}\Big\|_2 &\le \Big(1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2 + \eta\lambda^\circ\Big)\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2 + c_1\sqrt{\frac{\log^3 n}{np}}\beta_0^3(t^2+1) \\
&\le \left(1 + \eta\lambda^\star + c\eta\lambda^\star\sqrt{\frac{\log n}{np}}\right)\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2 + c\sqrt{\frac{\log^3 n}{np}}\beta_0^3(t^2+1)
\end{aligned}$$

for some universal constant $c > 0$, where (27) was used in the second line. From an analysis on the recursive equation

$$x_{t+1} = \left(1 + \eta\lambda^\star + c_1\eta\lambda^\star\sqrt{\frac{\log n}{np}}\right)x_t + c_1\sqrt{\frac{\log^3 n}{np}}\beta_0^3(t^2+1), \quad x_0 = 0$$

we can prove that

$$\Big\|\boldsymbol{x}^{(t+1)} - \widehat{\boldsymbol{x}}^{(t+1)}\Big\|_2 \le \frac{8c_1}{(\eta\lambda^\star)^3}\sqrt{\frac{\log^3 n}{np}}\beta_0^3(1+\eta\lambda^\star)^{t+1}.$$

(40) at $(t+1)$ We decompose $\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)}$ as

$$\begin{aligned}
\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)} &= (1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2)(\boldsymbol{x} - \boldsymbol{x}^{(l)}) - 2\eta\underbrace{\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^\top(\boldsymbol{x} - \boldsymbol{x}^{(l)})}_{①} \\
&\quad - \eta\underbrace{\int_0^1\Big(\nabla^2 g(\boldsymbol{x}(\tau)) - \Big(\|\widetilde{\boldsymbol{x}}\|_2^2\boldsymbol{I} + 2\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^\top\Big)\Big)(\boldsymbol{x} - \boldsymbol{x}^{(l)})\mathrm{d}\tau}_{②} \\
&\quad - \eta\underbrace{\left(\frac{1}{p}\mathcal{P}_{\Omega_l}\Big(\boldsymbol{x}^{(t,l)}\boldsymbol{x}^{(t,l)\top}\Big) - \mathcal{P}_l\Big(\boldsymbol{x}^{(t,l)}\boldsymbol{x}^{(t,l)\top}\Big)\right)\boldsymbol{x}^{(t,l)}}_{③}
\end{aligned}$$

$$+\eta\underbrace{\lambda^{(l)}\boldsymbol{u}^{(l)}\boldsymbol{u}^{(l)\top}(\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)})}_{④}+\eta\underbrace{\left(\boldsymbol{M}^{\circ}-\lambda^{(l)}\boldsymbol{u}^{(l)}\boldsymbol{u}^{(l)\top}\right)(\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)})}_{⑤}+\eta\underbrace{(\boldsymbol{M}^{\circ}-\boldsymbol{M}^{(l)})\boldsymbol{x}^{(t,l)}}_{⑥},$$

where $\boldsymbol{x}^{(l)}(\tau)=\boldsymbol{x}^{(l)}+\tau(\boldsymbol{x}-\boldsymbol{x}^{(l)})$. ① is easily bounded by

$$\left\|①\right\|_2\le\|\widetilde{\boldsymbol{x}}\|_2^2\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_2\lesssim\beta_0^2\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_2\lesssim\lambda^\star\sqrt{\frac{1}{\log^{24}n}}\sqrt{\frac{np}{n}}\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_2.$$

From Lemma 27 and (29), for all $0\le\tau\le1$, we have

$$\left\|\nabla^2 g(\boldsymbol{x}^{(l)}(\tau))-\left(\|\widetilde{\boldsymbol{x}}\|_2^2\boldsymbol{I}+2\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^\top\right)\right\|\lesssim n\|\boldsymbol{x}(\tau)-\widetilde{\boldsymbol{x}}\|_\infty(\|\boldsymbol{x}(\tau)\|_\infty+\|\widetilde{\boldsymbol{x}}\|_\infty)+\sqrt{\frac{\log^3 n}{np}}\beta_0^2 \quad (43)$$

From the definition of $\boldsymbol{x}(\tau)$, we have

$$\left\|\boldsymbol{x}^{(l)}(\tau)-\widetilde{\boldsymbol{x}}\right\|_\infty\le(1-\tau)\left\|\boldsymbol{x}^{(l)}-\boldsymbol{x}\right\|_2+\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_\infty\lesssim\sqrt{\frac{\mu^3\log^5 n}{np}}\frac{\beta_0}{\sqrt{n}},$$

where the last inequality is from the induction hypotheses (38), (40), and the fact that $t\le T_0\lesssim\log n$. Inserting this bound back to (43), we get

$$\left\|\nabla^2 g(\boldsymbol{x}^{(l)}(\tau))-\left(\|\widetilde{\boldsymbol{x}}\|_2^2\boldsymbol{I}+2\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^\top\right)\right\|\lesssim\sqrt{\frac{\mu^3\log^6 n}{np}}\beta_0^2, \quad (44)$$

which also implies

$$\left\|②\right\|_2\lesssim\sqrt{\frac{\mu^3\log^6 n}{np}}\beta_0^2\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_2\lesssim\lambda^\star\sqrt{\frac{\mu^3}{n\log^{18}n}}\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_2.$$

It is implied from Lemma 28 that

$$\left\|③\right\|_2\le\left\|\boldsymbol{x}^{(l)}\right\|_\infty^2\sqrt{\frac{\log n}{p}}\left\|\boldsymbol{x}^{(l)}\right\|_2\lesssim\sqrt{\frac{\log^3 n}{np}}\frac{\beta_0^3}{\sqrt{n}}\lesssim\lambda^\star\sqrt{\frac{1}{n\log^{21}n}}\frac{\beta_0}{\sqrt{n}}$$

A bound on ④ follows from the induction hypothesis (41).

$$\left\|④\right\|_2\le\lambda^{(l)}\left|\boldsymbol{u}^{(l)\top}(\boldsymbol{x}^{(t)}-\boldsymbol{x}^{(t,l)})\right|\lesssim\lambda^\star\mu\sqrt{\frac{\log^2 n}{np}}\frac{\beta_0}{n}(1+\eta\lambda^\star)^t\lesssim\lambda^\star\mu\frac{\sqrt{\log^{23}n}}{np}\frac{\beta_0}{\sqrt{n}}$$

The second largest eigenvalue of $\boldsymbol{M}^{(l)}$ is at most $\left\|\boldsymbol{M}^{(l)}-\boldsymbol{M}^\star\right\|$ by Weyl's Theorem, and from Lemma 10 and Lemma 9, we have

$$\left\|\boldsymbol{M}^{(l)}-\boldsymbol{M}^\star\right\|\lesssim\left\|\boldsymbol{M}^{(l)}-\boldsymbol{M}^{\circ}\right\|+\|\boldsymbol{M}^{\circ}-\boldsymbol{M}^\star\|\lesssim\lambda^\star\mu\sqrt{\frac{\log n}{np}}.$$

Hence, we get

$$\left\|⑤\right\|_2\le\left(\left\|\boldsymbol{M}^{\circ}-\boldsymbol{M}^{(l)}\right\|+\left\|\boldsymbol{M}^{(l)}-\lambda^{(l)}\boldsymbol{u}^{(l)}\boldsymbol{u}^{(l)\top}\right\|\right)\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_2\lesssim\lambda^\star\mu\sqrt{\frac{\log n}{np}}\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_2.$$

Lastly, we apply Lemma 28 again to get

$$\left\|⑥\right\|_2\le\lambda^\star\mu\sqrt{\frac{\log n}{np}}\frac{\beta_0}{\sqrt{n}}.$$

There exists a universal constant $c>0$ such that

$$2\left\|①\right\|_2+\left\|②\right\|_2+\left\|⑤\right\|_2\le\frac{c\lambda^\star\mu}{\log^2 n}\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_2$$

and

$$\left\|③\right\|_2 + \left\|④\right\|_2 + \left\|⑥\right\|_2 \le c\lambda^\star \mu \sqrt{\frac{\log n}{np}} \frac{\beta_0}{\sqrt{n}}.$$

if $n^2 p \gtrsim \mu^2 n \log^{22} n$. Combining all, we have

$$\begin{aligned}\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)}\right\|_2 &\le \left(1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2 + \frac{c\eta\lambda^\star}{\log^2 n}\right)\left\|\boldsymbol{x} - \boldsymbol{x}^{(l)}\right\|_2 + c\eta\lambda^\star\mu\sqrt{\frac{\log n}{np}}\frac{\beta_0}{\sqrt{n}} \\ &\le \left(1 + \frac{c\eta\lambda^\star}{\log^2 n}\right)\left\|\boldsymbol{x} - \boldsymbol{x}^{(l)}\right\|_2 + c\eta\lambda^\star\mu\sqrt{\frac{\log n}{np}}\frac{\beta_0}{\sqrt{n}}.\end{aligned}$$

An analysis on the recursive equation

$$x_{t+1} = \left(1 + \frac{c\eta\lambda^\star}{\log^2 n}\right)x_t + c\eta\lambda^\star\mu\sqrt{\frac{\log n}{np}}\frac{\beta_0}{\sqrt{n}}, \quad x_0 = 0$$

gives the desired bound

$$\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)}\right\|_2 \le 2c\eta\lambda^\star\mu\sqrt{\frac{\log n}{np}}\frac{\beta_0}{\sqrt{n}}.$$

(41) at $(t+1)$ We decompose $\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)}$ as

$$\begin{aligned}&\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)} \\ &= (\boldsymbol{x} - \eta\nabla f(\boldsymbol{x})) - \left(\boldsymbol{x}^{(l)} - \eta\nabla f^{(l)}(\boldsymbol{x}^{(l)})\right) \\ &= (\boldsymbol{x} - \eta\nabla g(\boldsymbol{x})) - \left(\boldsymbol{x}^{(l)} - \eta\nabla g^{(l)}(\boldsymbol{x}^{(l)})\right) + \eta\left(\boldsymbol{M}^\circ\boldsymbol{x} - \boldsymbol{M}^{(l)}\boldsymbol{x}^{(l)}\right) \\ &= (\boldsymbol{x} - \eta\nabla g(\boldsymbol{x})) - \left(\boldsymbol{x}^{(l)} - \eta\nabla g(\boldsymbol{x}^{(l)})\right) - \eta\left(\nabla g(\boldsymbol{x}^{(l)}) - \nabla g^{(l)}(\boldsymbol{x}^{(l)})\right) \\ &\quad + \eta\left(\boldsymbol{M}^\circ - \boldsymbol{M}^{(l)}\right)\boldsymbol{x} + \eta\boldsymbol{M}^{(l)}(\boldsymbol{x} - \boldsymbol{x}^{(l)}) \\ &= \int_0^1 (\boldsymbol{I} - \eta\nabla^2 g(\boldsymbol{x}^{(l)}(\tau))(\boldsymbol{x} - \boldsymbol{x}^{(l)})\,\mathrm{d}\tau - \eta\left(\frac{1}{p}\mathcal{P}_{\Omega_l}\left(\boldsymbol{x}^{(l)}\boldsymbol{x}^{(l)\top}\right) - \mathcal{P}_l\left(\boldsymbol{x}^{(l)}\boldsymbol{x}^{(l)\top}\right)\right)\boldsymbol{x}^{(l)} \\ &\quad + \eta\boldsymbol{M}^{(l)}(\boldsymbol{x} - \boldsymbol{x}^{(l)}) + \eta\left(\boldsymbol{M}^\circ - \boldsymbol{M}^{(l)}\right)(\boldsymbol{x} - \boldsymbol{x}^{(l)}) + \eta\left(\boldsymbol{M}^\circ - \boldsymbol{M}^{(l)}\right)\boldsymbol{x}^{(l)} \\ &= (1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2)(\boldsymbol{x} - \boldsymbol{x}^{(l)}) - 2\eta\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^\top(\boldsymbol{x} - \boldsymbol{x}^{(l)}) \\ &\quad - \eta\int_0^1 \left(\nabla^2 g(\boldsymbol{x}^{(l)}(\tau)) - \left(\|\widetilde{\boldsymbol{x}}\|_2^2\boldsymbol{I} + 2\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^\top\right)\right)(\boldsymbol{x} - \boldsymbol{x}^{(l)})\mathrm{d}\tau \\ &\quad - \eta\left(\frac{1}{p}\mathcal{P}_{\Omega_l}\left(\boldsymbol{x}^{(l)}\boldsymbol{x}^{(l)\top}\right) - \mathcal{P}_l\left(\boldsymbol{x}^{(l)}\boldsymbol{x}^{(l)\top}\right)\right)\boldsymbol{x}^{(l)} \\ &\quad + \eta\boldsymbol{M}^{(l)}(\boldsymbol{x} - \boldsymbol{x}^{(l)}) + \eta\left(\boldsymbol{M}^\circ - \boldsymbol{M}^{(l)}\right)(\boldsymbol{x} - \boldsymbol{x}^{(l)}) + \eta\left(\boldsymbol{M}^\circ - \boldsymbol{M}^{(l)}\right)\boldsymbol{x}^{(l)},\end{aligned}$$

where $\boldsymbol{x}^{(l)}(\tau) = \boldsymbol{x}^{(l)} + \tau(\boldsymbol{x} - \boldsymbol{x}^{(l)})$. Then, we take inner product with $\boldsymbol{u}^{(l)}$ on both sides.

$$\begin{aligned}\boldsymbol{u}^{(l)\top}\left(\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)}\right) &= \left(1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2\right)\boldsymbol{u}^{(l)\top}(\boldsymbol{x} - \boldsymbol{x}^{(l)}) - 2\eta\underbrace{\boldsymbol{u}^{(l)\top}\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^\top\left(\boldsymbol{x} - \boldsymbol{x}^{(l)}\right)}_{①} \\ &\quad - \eta\underbrace{\int_0^1 \boldsymbol{u}^{(l)\top}\left(\nabla^2 g(\boldsymbol{x}(\tau)) - \left(\|\widetilde{\boldsymbol{x}}\|_2^2\boldsymbol{I} + 2\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^\top\right)\right)(\boldsymbol{x} - \boldsymbol{x}^{(l)})\mathrm{d}\tau}_{②} \\ &\quad - \eta\underbrace{\boldsymbol{u}^{(l)\top}\left(\frac{1}{p}\mathcal{P}_{\Omega_l}\left(\boldsymbol{x}^{(l)}\boldsymbol{x}^{(l)\top}\right) - \mathcal{P}_l\left(\boldsymbol{x}^{(l)}\boldsymbol{x}^{(l)\top}\right)\right)\boldsymbol{x}^{(l)}}_{③}\end{aligned}$$

$$+\eta\lambda^{(l)}\boldsymbol{u}^{(l)\top}(\boldsymbol{x}-\boldsymbol{x}^{(l)})+\eta\underbrace{\boldsymbol{u}^{(l)\top}\left(\boldsymbol{M}^{\circ}-\boldsymbol{M}^{(l)}\right)(\boldsymbol{x}-\boldsymbol{x}^{(l)})}_{④}+\eta\underbrace{\boldsymbol{u}^{(l)\top}\left(\boldsymbol{M}^{\circ}-\boldsymbol{M}^{(l)}\right)\boldsymbol{x}^{(l)}}_{⑤}$$

For the term ①, we have

$$\left|\boldsymbol{u}^{(l)\top}\widetilde{\boldsymbol{x}}\right|\leq\left|\boldsymbol{u}^{(l)\top}\boldsymbol{x}^{(0)}\right|+(1+\eta\lambda^{\star})^{t}\left|\boldsymbol{u}^{\star\top}\boldsymbol{x}^{(0)}\right|\left|\boldsymbol{u}^{(l)\top}\boldsymbol{u}^{\star}\right|\lesssim\sqrt{\frac{\log n}{n}}(1+\eta\lambda^{\star})^{t}\beta_{0},$$

and thus,

$$\begin{aligned}|①|\leq\left|\boldsymbol{u}^{(l)\top}\widetilde{\boldsymbol{x}}\right|\|\widetilde{\boldsymbol{x}}\|_{2}\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_{2}&\lesssim\sqrt{\frac{\log n}{n}}(1+\eta\lambda^{\star})^{t}\beta_{0}^{2}\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_{2}\\&\lesssim\lambda^{\star}\mu\sqrt{\frac{\log^{2}n}{np}}\frac{\beta_{0}^{3}}{n}(1+\eta\lambda^{\star})^{t}t\\&\lesssim\lambda^{\star}\mu\sqrt{\frac{\log n}{np}}\frac{\beta_{0}}{n}.\end{aligned}$$

The definition of Phase I was used to bound $(1+\eta\lambda^{\star})^{t}t$ in deriving the last line. We use (44) to get

$$\begin{aligned}|②|&\lesssim\int_{0}^{1}\left\|\nabla^{2}g(\boldsymbol{x}(\tau))-\left(\|\widetilde{\boldsymbol{x}}\|_{2}^{2}\boldsymbol{I}+2\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^{\top}\right)\right\|\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_{2}\left\|\boldsymbol{u}^{(l)}\right\|_{2}\mathrm{d}\tau\\&\lesssim\sqrt{\frac{\mu^{3}\log^{6}n}{np}}\beta_{0}^{2}\cdot\mu\frac{\sqrt{\log n}}{np}\frac{\beta_{0}}{\sqrt{n}}t\lesssim\lambda^{\star}\sqrt{\frac{\mu^{5}}{np\log^{15}n}}\frac{\beta_{0}}{n}.\end{aligned}$$

We apply Lemma 29 to ③ to yield

$$|③|\lesssim\left\|\boldsymbol{x}^{(l)}\right\|_{\infty}^{2}\sqrt{\frac{\log n}{p}}\left(\left\|\boldsymbol{u}^{(l)}\right\|_{2}\left\|\boldsymbol{x}^{(l)}\right\|_{\infty}+\left\|\boldsymbol{x}^{(l)}\right\|_{2}\left\|\boldsymbol{u}^{(l)}\right\|_{\infty}\right)\lesssim\sqrt{\frac{\mu\log^{4}n}{np}}\frac{\beta_{0}^{3}}{n}\lesssim\lambda^{\star}\mu\sqrt{\frac{\mu}{np\log^{20}n}}\frac{\beta_{0}}{n}$$

Cauchy-Schwartz inequality is first applied to ④ to get

$$|④|\leq\left\|(\boldsymbol{M}^{\circ}-\boldsymbol{M}^{(l)})\boldsymbol{u}^{(l)}\right\|_{2}\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_{2}.$$

Lemma 28 gives a bound on the magnitude of $(\boldsymbol{M}^{\circ}-\boldsymbol{M}^{(l)})\boldsymbol{u}^{(l)}$, and thus, we have

$$|④|\leq\left\|(\boldsymbol{M}^{\circ}-\boldsymbol{M}^{(l)})\boldsymbol{u}^{(l)}\right\|_{2}\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_{2}\lesssim\lambda^{\star}\mu\sqrt{\frac{\log n}{np}}\frac{1}{\sqrt{n}}\left\|\boldsymbol{x}-\boldsymbol{x}^{(l)}\right\|_{2}\lesssim\lambda^{\star}\mu\frac{\log^{2}n}{np}\frac{\beta_{0}}{n}.$$

For the term ⑤, we again use Lemma 29 to get

$$|⑤|\leq\lambda^{\star}\mu\sqrt{\frac{\log^{2}n}{np}}\frac{\beta_{0}}{n}.$$

Combining all, there exists a universal constant $c>0$ such that

$$2|①|+|②|+|③|+|④|+|⑤|\leq c\lambda^{\star}\mu\sqrt{\frac{\log^{2}n}{np}}\frac{\beta_{0}}{n}.$$

Finally, we have

$$\begin{aligned}&\left|\boldsymbol{u}^{(l)\top}\left(\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t+1,l)}\right)\right|\\&\leq\left(1-\eta\|\widetilde{\boldsymbol{x}}\|_{2}^{2}+\eta\lambda^{(l)}\right)\left|\boldsymbol{u}^{(l)\top}(\boldsymbol{x}-\boldsymbol{x}^{(l)})\right|+c_{3}\eta\lambda^{\star}\mu\sqrt{\frac{\log^{2}n}{np}}\frac{\beta_{0}}{n}\\&\leq\left(1+\eta\lambda^{\star}+c\eta\lambda^{\star}\mu\sqrt{\frac{\log n}{np}}\right)\left|\boldsymbol{u}^{(l)\top}(\boldsymbol{x}-\boldsymbol{x}^{(l)})\right|+c_{3}\eta\lambda^{\star}\mu\sqrt{\frac{\log^{2}n}{np}}\frac{\beta_{0}}{n},\end{aligned}$$

where the last line used (28). An analysis on the recursive equation

$$x_{t+1} = \left(1 + \eta\lambda^\star + \eta\lambda^\star\mu\sqrt{\frac{\log n}{np}}\right)x_t + c_3\eta\lambda^\star\mu\sqrt{\frac{\log^2 n}{np}}\frac{\beta_0}{n}, \quad x_0 = 0$$

gives the bound

$$\left|\boldsymbol{u}^{(l)\top}\left(\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)}\right)\right| \le 2c_3\mu\sqrt{\frac{\log^2 n}{np}}\frac{\beta_0}{n}(1+\eta\lambda^\star)^{t+1}.$$

(42) at $(t+1)$ We decompose $(\boldsymbol{x}^{(t+1,l)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l$ as

$$(\boldsymbol{x}^{(t+1,l)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l = \left(1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2\right)(\boldsymbol{x}^{(l)} - \widetilde{\boldsymbol{x}})_l + \eta\lambda^\star\boldsymbol{u}^{\star\top}(\boldsymbol{x}^{(l)} - \widetilde{\boldsymbol{x}})u_l^\star + \eta\left(\|\widetilde{\boldsymbol{x}}\|_2^2 - \left\|\boldsymbol{x}^{(l)}\right\|_2^2\right)x_l^{(l)},$$

and this implies

$$\left|(\boldsymbol{x}^{(t+1,l)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l\right| \le \left(1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2\right)\left|(\boldsymbol{x}^{(l)} - \widetilde{\boldsymbol{x}})_l\right| + \eta\lambda^\star\left\|\boldsymbol{x}^{(l)} - \widetilde{\boldsymbol{x}}\right\|_2\|\boldsymbol{u}^\star\|_\infty + \eta\|\widetilde{\boldsymbol{x}}\|_2\left\|\boldsymbol{x}^{(l)} - \widetilde{\boldsymbol{x}}\right\|_2\|\widetilde{\boldsymbol{x}}\|_\infty.$$

From (37) and (40), we have

$$\left\|\boldsymbol{x}^{(l)} - \widetilde{\boldsymbol{x}}\right\|_2 \le \left\|\boldsymbol{x}^{(l)} - \boldsymbol{x}\right\|_2 + \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_2 \lesssim \mu\sqrt{\frac{\log n}{np}}\beta_0 t,$$

and hence,

$$\begin{aligned}
\left|(\boldsymbol{x}^{(t+1,l)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l\right| &\le \left|(\boldsymbol{x}^{(l)} - \widetilde{\boldsymbol{x}})_l\right| + c\eta\left(\lambda^\star\sqrt{\frac{\mu}{n}} + \sqrt{\frac{\log n}{n}}\beta_0^2\right)\mu\sqrt{\frac{\log^2 n}{np}}\beta_0 t \\
&\le \left|(\boldsymbol{x}^{(l)} - \widetilde{\boldsymbol{x}})_l\right| + c\eta\lambda^\star\sqrt{\frac{\mu^3\log^2 n}{np}}\frac{\beta_0}{\sqrt{n}}t
\end{aligned}$$

for some universal constant $c > 0$. By induction, similar result holds for all smaller t, and we have

$$\left|(\boldsymbol{x}^{(t+1,l)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l\right| \le c_2\eta\lambda^\star\sqrt{\frac{\mu^3\log^2 n}{np}}\frac{\beta_0}{\sqrt{n}}\sum_{s=1}^{t}s \le c_2\eta\lambda^\star\sqrt{\frac{\mu^3\log^2 n}{np}}\frac{\beta_0}{\sqrt{n}}(t+1)^2$$

(37) at $(t+1)$ We can obtain this through the combination of (39) and Lemma 4.

$$\begin{aligned}
\left\|\boldsymbol{x}^{(t+1)} - \widetilde{\boldsymbol{x}}^{(t+1)}\right\|_2 &\le \left\|\boldsymbol{x}^{(t+1)} - \widehat{\boldsymbol{x}}^{(t+1)}\right\|_2 + \left\|\widehat{\boldsymbol{x}}^{(t+1)} - \widetilde{\boldsymbol{x}}^{(t+1)}\right\|_2 \\
&\lesssim \sqrt{\frac{\log^3 n}{np}}(1+\eta\lambda^\star)^{t+1}\beta_0^3 + \mu\sqrt{\frac{\log n}{np}}t\beta_0 \\
&\lesssim \sqrt{\frac{\log^3 n}{np}}(1+\eta\lambda^\star)^{T_0}\beta_0^3 + \mu\sqrt{\frac{\log n}{np}}T_0\beta_0 \\
&\lesssim \sqrt{\frac{1}{np}}\beta_0 + \mu\sqrt{\frac{\log^3 n}{np}}\beta_0 \lesssim \mu\sqrt{\frac{\log^3 n}{np}}\beta_0
\end{aligned}$$

(38) at $(t+1)$ The lth component of $\boldsymbol{x}^{(t+1)} - \widetilde{\boldsymbol{x}}^{(t+1)}$ is bounded by

$$\begin{aligned}
\left|(\boldsymbol{x}^{(t+1)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l\right| &\le \left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)}\right\|_\infty + \left|(\boldsymbol{x}^{(t+1,l)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l\right| \\
&\le \left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)}\right\|_2 + \left|(\boldsymbol{x}^{(t+1,l)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l\right| \\
&\lesssim \sqrt{\frac{\mu^3\log^2 n}{np}}\frac{\beta_0}{\sqrt{n}}t^2
\end{aligned}$$

This completes the proof of Lemma 17.

At $t = T_1$ It is implied from Lemma 17 and the definition of T_1 that at $t = T_1$, there exists a constant $C_0 > 0$ such that

$$\begin{aligned}
\left\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_2 &\le C_0\mu\sqrt{\frac{\log^3 n}{np}}\beta_0, \\
\left\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_\infty &\le C_0\sqrt{\frac{\mu^3\log^6 n}{np}}\frac{\beta_0}{\sqrt{n}}, \\
\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t,l)}\right\|_2 &\le C_0\mu\sqrt{\frac{\log^3 n}{np}}\frac{\beta_0}{\sqrt{n}}, \\
\left|\left(\boldsymbol{x}^{(t,l)} - \widetilde{\boldsymbol{x}}^{(t)}\right)_l\right| &\le C_0\sqrt{\frac{\mu^3\log^6 n}{np}}\frac{\beta_0}{\sqrt{n}}.
\end{aligned} \tag{45}$$

These bounds serve as a base case for the induction of the next part.

E PHASE II

This section is mostly devoted to the proof of Lemma 8.

Proof of Theorems 1 and 2 We first explain how Theorems 1 and 2 are derived from Lemma 8. We first focus on $\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\|_2$. From the definition of T_1, we have

$$\left\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \lesssim \mu\sqrt{\frac{1}{\log^{15} n}}\frac{\beta_0}{\sqrt{n}}(1+\eta\lambda^\star)^t.$$

From the lower bounds of Lemma 15, for all $t \le T_2'$, we have

$$\left\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \lesssim \mu\sqrt{\frac{1}{\log^{15} n}}\left(1 + \frac{(1+\eta\lambda^\star)^t}{\sqrt{n}}\right)\beta_0 \lesssim \mu\sqrt{\frac{1}{\log^{14} n}}\left\|\widetilde{\boldsymbol{x}}^{(t)}\right\|_2,$$

and this proves (10) of Theorem 2 for $t \le T_2'$. Now, for $T_2' < t \le T_2$, the bound increases with the rate $(1+\eta\lambda^\star)$ from the bound of $t = T_2'$ as

$$\left\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \lesssim \mu\sqrt{\frac{1}{\log^{14} n}}\left\|\widetilde{\boldsymbol{x}}^{(T_2')}\right\|_2(1+\eta\lambda^\star)^{t-T_2'}.$$

We apply the result of Lemma 16 to obtain

$$\left\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \lesssim \mu\sqrt{\frac{1}{\log^{14} n}}\sqrt{\log^{11} n}\left\|\widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \lesssim \mu\sqrt{\frac{1}{\log n}}\left\|\widetilde{\boldsymbol{x}}^{(t)}\right\|_2, \tag{46}$$

which proves (10) of Theorem 2 for $T_2' < t \le T_2$. If we combine this with (31), we are able to prove (5) of Theorem 1. Going through a similar way with (21), (22), and (23), we can complete the proof of Theorems 1 and 2.

Proof of Lemma 8 Before we start the proof, we define a function G as

$$G(\boldsymbol{x}) = \frac{1}{4}\left\|\boldsymbol{x}\boldsymbol{x}^\top\right\|_{\mathrm{F}}^2.$$

The gradient of G satisfies

$$\nabla F(\boldsymbol{x}) = \nabla G(\boldsymbol{x}) - \boldsymbol{M}^\star\boldsymbol{x}.$$

Now, we assume that the hypotheses hold up to the tth iteration and show that they hold at the $(t+1)$st iteration. For brevity, we drop the superscript (t) from $\boldsymbol{x}^{(t)}$, $\boldsymbol{x}^{(t,l)}$, $\widetilde{\boldsymbol{x}}^{(t)}$ and denote them as $\boldsymbol{x}$, $\boldsymbol{x}^{(l)}$, $\widetilde{\boldsymbol{x}}$, respectively.

(20) at $(t+1)$ We decompose $\boldsymbol{x}^{(t+1)} - \widetilde{\boldsymbol{x}}^{(t+1)}$ as

$$
\begin{aligned}
&\boldsymbol{x}^{(t+1)} - \widetilde{\boldsymbol{x}}^{(t+1)} \\
&= (\boldsymbol{x} - \eta \nabla f(\boldsymbol{x})) - (\widetilde{\boldsymbol{x}} - \eta \nabla F(\widetilde{\boldsymbol{x}})) \\
&= (\boldsymbol{x} - \eta \nabla g(\boldsymbol{x})) - (\widetilde{\boldsymbol{x}} - \eta \nabla G(\widetilde{\boldsymbol{x}})) + \eta \left(\boldsymbol{M}^{\circ}\boldsymbol{x} - \boldsymbol{M}^{\star}\widetilde{\boldsymbol{x}}\right) \\
&= (\boldsymbol{x} - \eta \nabla g(\boldsymbol{x})) - (\widetilde{\boldsymbol{x}} - \eta \nabla g(\widetilde{\boldsymbol{x}})) - \eta \left(\nabla g(\widetilde{\boldsymbol{x}}) - \nabla G(\widetilde{\boldsymbol{x}})\right) + \eta \boldsymbol{M}^{\star}(\boldsymbol{x} - \widetilde{\boldsymbol{x}}) + \eta(\boldsymbol{M}^{\circ} - \boldsymbol{M}^{\star})\boldsymbol{x} \\
&= \int_0^1 (\boldsymbol{I} - \eta \nabla^2 g(\boldsymbol{x}(\tau)))(\boldsymbol{x} - \widetilde{\boldsymbol{x}})\, \mathrm{d}\tau - \eta \|\widetilde{\boldsymbol{x}}\|_2^2 \left(\boldsymbol{I}_{\widetilde{\boldsymbol{x}}} - \boldsymbol{I}\right)\widetilde{\boldsymbol{x}} + \eta \boldsymbol{M}^{\star}(\boldsymbol{x} - \widetilde{\boldsymbol{x}}) + \eta(\boldsymbol{M}^{\circ} - \boldsymbol{M}^{\star})\boldsymbol{x} \\
&= \underbrace{\left((1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2)\boldsymbol{I} - 2\eta \widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^{\top} + \eta \boldsymbol{M}^{\star}\right)(\boldsymbol{x} - \widetilde{\boldsymbol{x}})}_{:=①} - \eta \underbrace{\int_0^1 \left(\nabla^2 g(\boldsymbol{x}(\tau)) - \left(\|\widetilde{\boldsymbol{x}}\|_2^2 \boldsymbol{I} + 2\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^{\top}\right)\right)(\boldsymbol{x} - \widetilde{\boldsymbol{x}})\mathrm{d}\tau}_{:=②} \\
&\quad - \eta \underbrace{\|\widetilde{\boldsymbol{x}}\|_2^2 \left(\boldsymbol{I}_{\widetilde{\boldsymbol{x}}} - \boldsymbol{I}\right)\widetilde{\boldsymbol{x}}}_{:=③} + \eta \underbrace{(\boldsymbol{M}^{\circ} - \boldsymbol{M}^{\star})\boldsymbol{x}}_{:=④},
\end{aligned}
$$

where $\boldsymbol{x}(\tau) = \widetilde{\boldsymbol{x}} + \tau(\boldsymbol{x} - \widetilde{\boldsymbol{x}})$. For the term ①, we require a bound on

$$\left\|(1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2)\boldsymbol{I} - 2\eta \widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^{\top} + \eta \boldsymbol{M}^{\star}\right\|.$$

If we write $\widetilde{\boldsymbol{x}}$ as $\widetilde{\alpha}_t \boldsymbol{u}^{\star} + \widetilde{\boldsymbol{x}}_{\perp}$, we have $\widetilde{\alpha}_t^2 \le \lambda^{\star}$ and $\|\widetilde{\boldsymbol{x}}_{\perp}\|_2 \lesssim \beta_0$. Then, we have

$$
\begin{aligned}
&\left\|(1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2)\boldsymbol{I} - 2\eta \widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^{\top} + \eta \boldsymbol{M}^{\star}\right\| \\
&= \left\|(1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2)\boldsymbol{I} + \eta(\lambda^{\star} - 2\widetilde{\alpha}_t^2)\boldsymbol{u}^{\star}\boldsymbol{u}^{\star\top} - 2\eta(\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^{\top} - \widetilde{\alpha}_t^2 \boldsymbol{u}^{\star}\boldsymbol{u}^{\star\top})\right\| \\
&\le \left\|(1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2)\boldsymbol{I} + \eta(\lambda^{\star} - 2\widetilde{\alpha}_t^2)\boldsymbol{u}^{\star}\boldsymbol{u}^{\star\top}\right\| + 2\eta\left\|\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^{\top} - \widetilde{\alpha}_t^2 \boldsymbol{u}^{\star}\boldsymbol{u}^{\star\top}\right\| \\
&\le (1 + \eta\lambda^{\star}) + 2\eta(2\alpha_t\|\widetilde{\boldsymbol{x}}_{\perp}\|_2 + \|\widetilde{\boldsymbol{x}}_{\perp}\|_2^2) \\
&\le 1 + \eta\lambda^{\star} + \frac{c}{\log^2 n}
\end{aligned}
$$

for some constant $c > 0$. This implies the desired bound

$$\left\|①\right\|_2 \le \left(1 + \eta\lambda^{\star} + \frac{c}{\log^2 n}\right)\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_2.$$

For all $0 \le \tau \le 1$, we have $\|\boldsymbol{x}(\tau) - \widetilde{\boldsymbol{x}}\|_\infty \le \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_\infty$, and going through the same process we derived (46), we have

$$\left\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_\infty \lesssim \lambda^{\star}\mu\sqrt{\frac{1}{\log^3 n}}\sqrt{\frac{\mu}{n}}.$$

for all t in Phase II. Hence, by Lemma 27, we have

$$\left\|\nabla^2 g(\boldsymbol{x}(\tau)) - \left(\|\widetilde{\boldsymbol{x}}\|_2^2 \boldsymbol{I} + 2\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^{\top}\right)\right\| \lesssim n\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_\infty(\|\boldsymbol{x}\|_\infty + \|\widetilde{\boldsymbol{x}}\|_\infty) + \lambda^{\star}\mu\sqrt{\frac{\log n}{np}} \lesssim \lambda^{\star}\mu^2\sqrt{\frac{1}{\log^3 n}}.$$

This gives

$$\left\|②\right\|_2 \lesssim \lambda^{\star}\mu^2\sqrt{\frac{1}{\log^3 n}}\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_2. \tag{47}$$

For the term ③, we use Lemma 25 to obtain

$$\left\|③\right\|_2 \lesssim \lambda^{\star}\sqrt{\frac{\mu\log n}{np}}\|\widetilde{\boldsymbol{x}}\|_2.$$

Lastly, the term ④ is bounded with

$$\left\|④\right\|_2 \lesssim \|\boldsymbol{M}^{\circ} - \boldsymbol{M}^{\star}\|_2\|\boldsymbol{x}\|_2 \lesssim \lambda^{\star}\mu\sqrt{\frac{\log n}{np}}\|\widetilde{\boldsymbol{x}}\|_2.$$

Combining all, there exists a universal constant $c > 0$ such that

$$\|①\|_2 + \|②\|_2 \le \left(1 + \eta\lambda^\star + \frac{c}{\log^{3/2} n}\right) \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_2,$$
$$\|③\|_2 + \|④\|_2 \le c\lambda^\star \mu \sqrt{\frac{\log n}{np}} \|\widetilde{\boldsymbol{x}}\|_2,$$

and we have

$$\left\|\boldsymbol{x}^{(t+1)} - \widetilde{\boldsymbol{x}}^{(t+1)}\right\|_2 \le \left(1 + \eta\lambda^\star + \frac{c}{\log^{3/2} n}\right) \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_2 + c\eta\lambda^\star \mu \sqrt{\frac{\log n}{np}} \|\widetilde{\boldsymbol{x}}\|_2.$$

Because $\|\widetilde{\boldsymbol{x}}\|_2$ is about β_0 and it can grow in a rate at most $(1 + \eta\lambda^\star)$, it holds that

$$\|\widetilde{\boldsymbol{x}}\|_2 \lesssim (1 + \eta\lambda^\star)^{t-T_1} \beta_0. \tag{48}$$

If we write

$$\left\|\boldsymbol{x}^{(t)} - \widetilde{\boldsymbol{x}}^{(t)}\right\|_2 \le C_t \mu \sqrt{\frac{\log^4 n}{np}} \beta_0 (1 + \eta\lambda^\star)^{t-T_1},$$

we finally have

$$\left\|\boldsymbol{x}^{(t+1)} - \widetilde{\boldsymbol{x}}^{(t+1)}\right\|_2 \le \left(\left(1 + \eta\lambda^\star + \frac{c}{\log^{3/2} n}\right) C_t + \frac{c\eta\lambda^\star}{\log^{3/2} n}\right) \mu \sqrt{\frac{\log^4 n}{np}} \beta_0 (1 + \eta\lambda^\star)^{t-T_1}.$$

Hence, if we let

$$C_{t+1} = \left(1 + \eta\lambda^\star + \frac{c}{\log^{3/2} n}\right) C_t + \frac{c\eta\lambda^\star}{\log^{3/2} n},$$

we have $C_t \le 2C_{T_1}$ for all t in Phase II. This proves (20) at $(t+1)$.

(22) at $(t+1)$ Similar to the proof of (40), we have the decomposition

$$\begin{aligned}
\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)} = & \underbrace{(1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2 - 2\eta\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^\top)(\boldsymbol{x} - \boldsymbol{x}^{(l)})}_{:=①} \\
& - \eta \underbrace{\int_0^1 \left(\nabla^2 g(\boldsymbol{x}^{(l)}(\tau)) - \left(\|\widetilde{\boldsymbol{x}}\|_2^2 \boldsymbol{I} + 2\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^\top\right)\right)(\boldsymbol{x} - \boldsymbol{x}^{(l)}) \mathrm{d}\tau}_{:=②} \\
& - \eta \underbrace{\left(\frac{1}{p}\mathcal{P}_{\Omega_l}\left(\boldsymbol{x}^{(l)}\boldsymbol{x}^{(l)\top}\right) - \mathcal{P}_l\left(\boldsymbol{x}^{(l)}\boldsymbol{x}^{(l)\top}\right)\right)\boldsymbol{x}^{(l)}}_{:=③} \\
& + \eta\boldsymbol{M}^\star(\boldsymbol{x} - \boldsymbol{x}^{(l)}) + \eta \underbrace{(\boldsymbol{M}^\circ - \boldsymbol{M}^\star)(\boldsymbol{x} - \boldsymbol{x}^{(l)})}_{:=④} + \eta \underbrace{\left(\boldsymbol{M}^\circ - \boldsymbol{M}^{(l)}\right)\boldsymbol{x}^{(l)}}_{⑤},
\end{aligned}$$

where $\boldsymbol{x}^{(l)}(\tau) = \boldsymbol{x}^{(l)} + \tau(\boldsymbol{x} - \boldsymbol{x}^{(l)})$. Both of the terms ① and ② can be bounded similar to ① and ② of $\boldsymbol{x}^{(t+1)} - \widetilde{\boldsymbol{x}}^{(t+1)}$ as

$$\|①\|_2 \le \left(1 + \eta\lambda^\star + \frac{c}{\log^2 n}\right) \left\|\boldsymbol{x} - \boldsymbol{x}^{(l)}\right\|_2,$$
$$\|②\|_2 \lesssim \lambda^\star \mu^2 \sqrt{\frac{1}{\log^3 n}} \left\|\boldsymbol{x} - \boldsymbol{x}^{(l)}\right\|_2.$$

for some constant $c > 0$. For the terms ③ and ⑤, we use Lemma 28 to obtain

$$\|③\|_2 \lesssim \sqrt{\frac{\log n}{p}} \left\|\boldsymbol{x}^{(l)}\right\|_2 \left\|\boldsymbol{x}^{(l)}\right\|_\infty^2 \lesssim \lambda^\star \mu \sqrt{\frac{\log n}{np}} \frac{1}{\sqrt{n}} \|\widetilde{\boldsymbol{x}}\|_2,$$

$$\left\|⑤\right\|_2 \lesssim \lambda^\star \mu \sqrt{\frac{\log n}{np}} \frac{1}{\sqrt{n}} \|\widetilde{\boldsymbol{x}}\|_2.$$

From Lemma 9, the term ④ is bounded as

$$\left\|④\right\|_2 \le \|\boldsymbol{M}^\circ - \boldsymbol{M}^\star\| \left\|\boldsymbol{x} - \boldsymbol{x}^{(l)}\right\|_2 \lesssim \lambda^\star \mu \sqrt{\frac{\log n}{np}} \left\|\boldsymbol{x} - \boldsymbol{x}^{(l)}\right\|_2.$$

Combining all with (48), there exists a universal constant $c > 0$ such that

$$\left\|①\right\|_2 + \left\|②\right\|_2 + \left\|④\right\|_2 \le \left(1 + \eta\lambda^\star + \frac{c}{\log^{3/2} n}\right) \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_2,$$

$$\left\|③\right\|_2 + \left\|⑤\right\|_2 \le c\lambda^\star \mu \sqrt{\frac{\log n}{np}} \frac{\beta_0}{\sqrt{n}} (1 + \eta\lambda^\star)^{t - T_1}.$$

Hence, we have

$$\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)}\right\|_2 \le \left(1 + \eta\lambda^\star + \frac{c}{\log^{3/2} n}\right) \left\|\boldsymbol{x} - \boldsymbol{x}^{(l)}\right\|_2 + c\lambda^\star \mu \sqrt{\frac{\log n}{np}} \frac{\beta_0}{\sqrt{n}} (1 + \eta\lambda^\star)^{t - T_1},$$

and if we write

$$\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t,l)}\right\|_2 \le C_t \mu \sqrt{\frac{\log^4 n}{np}} \frac{\beta_0}{\sqrt{n}} (1 + \eta\lambda^\star)^{t - T_1},$$

we finally have

$$\left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)}\right\|_2 \le \left(\left(1 + \eta\lambda^\star + \frac{c}{\log^{3/2} n}\right) C_t + \frac{c\eta\lambda^\star}{\log^{3/2} n}\right) \mu \sqrt{\frac{\log^4 n}{np}} \frac{\beta_0}{\sqrt{n}} (1 + \eta\lambda^\star)^{t - T_1}.$$

Hence, if we let

$$C_{t+1} = \left(1 + \eta\lambda^\star + \frac{c}{\log^{3/2} n}\right) C_t + \frac{c\eta\lambda^\star}{\log^{3/2} n},$$

we have $C_t \le 2C_{T_1}$ for all t in Phase II. This proves (22) at $(t+1)$.

(23) at $(t+1)$ We use the same bound

$$\left|(\boldsymbol{x}^{(t+1,l)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l\right| \le \left(1 - \eta\|\widetilde{\boldsymbol{x}}\|_2^2\right) \left|(\boldsymbol{x}^{(l)} - \widetilde{\boldsymbol{x}})_l\right| + \eta(\lambda^\star \|\boldsymbol{u}^\star\|_\infty + \|\widetilde{\boldsymbol{x}}\|_2 \|\widetilde{\boldsymbol{x}}\|_\infty) \left\|\boldsymbol{x}^{(l)} - \widetilde{\boldsymbol{x}}\right\|_2.$$

that was used in the proof of (42). From (20) and (22), we have

$$\left\|\boldsymbol{x}^{(l)} - \widetilde{\boldsymbol{x}}\right\|_2 \le \left\|\boldsymbol{x}^{(l)} - \boldsymbol{x}\right\|_2 + \|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_2 \lesssim \mu \sqrt{\frac{\log^4 n}{np}} \beta_0 (1 + \eta\lambda^\star)^{t - T_1},$$

and also it holds that $\lambda^\star \|\boldsymbol{u}^\star\|_\infty + \|\widetilde{\boldsymbol{x}}\|_2 \|\widetilde{\boldsymbol{x}}\|_\infty \lesssim \lambda^\star \sqrt{\frac{\mu}{n}}$. Hence, there exists a universal constant $c > 0$ such that

$$(\lambda^\star \|\boldsymbol{u}^\star\|_\infty + \|\widetilde{\boldsymbol{x}}\|_2 \|\widetilde{\boldsymbol{x}}\|_\infty) \left\|\boldsymbol{x}^{(l)} - \widetilde{\boldsymbol{x}}\right\|_2 \le c\lambda^\star \sqrt{\frac{\mu^3 \log^5 n}{np}} \frac{\beta_0}{\sqrt{n}} (1 + \eta\lambda^\star)^{t - T_1}.$$

Finally, we have

$$\left|(\boldsymbol{x}^{(t+1,l)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l\right| \le \left|(\boldsymbol{x}^{(l)} - \widetilde{\boldsymbol{x}})_l\right| + c\eta\lambda^\star \sqrt{\frac{\mu^3 \log^5 n}{np}} \frac{\beta_0}{\sqrt{n}} (1 + \eta\lambda^\star)^{t - T_1},$$

and an analysis on the recursive equation

$$x_{t+1} = x_t + c\eta\lambda^\star \sqrt{\frac{\mu^3 \log^5 n}{np}} \frac{\beta_0}{\sqrt{n}} (1 + \eta\lambda^\star)^{t - T_1}, \quad x_{T_0} = c\sqrt{\frac{\mu^3 \log^6 n}{np}} \frac{\beta_0}{\sqrt{n}}$$

gives the desired bound

$$\left|(\boldsymbol{x}^{(t+1,l)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l\right| \le 2c\eta\lambda^\star \sqrt{\frac{\mu^3 \log^6 n}{np}} \frac{\beta_0}{\sqrt{n}} (1 + \eta\lambda^\star)^{t+1 - T_1}.$$

(21) at $(t+1)$ The lth component of $\boldsymbol{x}^{(t+1)} - \widetilde{\boldsymbol{x}}^{(t+1)}$ is bounded by

$$\begin{aligned}\left|(\boldsymbol{x}^{(t+1)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l\right| &\le \left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)}\right\|_\infty + \left|(\boldsymbol{x}^{(t+1,l)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l\right| \\ &\le \left\|\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t+1,l)}\right\|_2 + \left|(\boldsymbol{x}^{(t+1,l)} - \widetilde{\boldsymbol{x}}^{(t+1)})_l\right| \\ &\lesssim \sqrt{\frac{\mu^3 \log^6 n}{np}} \frac{\beta_0}{\sqrt{n}} (1+\eta\lambda^\star)^{t+1-T_0}\end{aligned}$$

F Technical Lemmas

We introduce some technical lemmas in this section. Most of them are the results of classical concentration inequalities.

Theorem 18 (Matrix Bernstein Inequality)**.** *Let $\{\boldsymbol{X}_i\}$ be $n \times n$ independent symmetric random matrices. Assume that each random matrix satisfies $\mathbb{E}\,\boldsymbol{X}_i = \boldsymbol{0}$ and $\|\boldsymbol{X}_i\| \le L$ almost surely. Then, for all $\tau \ge 0$, we have*

$$\mathbb{P}\left[\left\|\sum_i \boldsymbol{X}_i\right\| \ge \tau\right] \le n \exp\left(\frac{-\tau^2/2}{V + L\tau/3}\right),$$

where $V = \left\|\sum_i \mathbb{E}(\boldsymbol{X}_i^2)\right\|$.

Corollary 19 (Matrix Bernstein Inequality)**.** *Let $\{\boldsymbol{X}_i\}$ be $n \times n$ independent symmetric random matrices. Assume that each random matrix satisfies $\mathbb{E}\,\boldsymbol{X}_i = \boldsymbol{0}$ and $\|\boldsymbol{X}_i\| \le L$ almost surely. Then, with high probability, we have*

$$\left\|\sum_i \boldsymbol{X}_i\right\| \lesssim \sqrt{V \log n} + L \log n,$$

where $V = \left\|\sum_i \mathbb{E}(\boldsymbol{X}_i^2)\right\|$.

Lemma 20. *For any fixed matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$, we have*

$$\left\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}) - \boldsymbol{M}\right\| \lesssim \sqrt{\frac{n\log n}{p}} \|\boldsymbol{M}\|_\infty + \frac{\log n}{p}\|\boldsymbol{M}\|_\infty$$

with high probability.

Proof. We decompose the matrix into the sum of independent symmetric matrices.

$$\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}) - \boldsymbol{M} = \sum_{i<j}\left(\frac{\delta_{ij}}{p} - 1\right) M_{ij}(\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top) + \sum_i \left(\frac{\delta_{ii}}{p} - 1\right) M_{ii}\boldsymbol{e}_i\boldsymbol{e}_i^\top$$

We calculate L and V of Corollary 19. We have $L \le \frac{1}{p}\|\boldsymbol{M}\|_\infty$ because

$$\begin{aligned}\left\|\left(\frac{\delta_{ij}}{p} - 1\right) M_{ij}(\boldsymbol{e}_i\boldsymbol{e}_j^\top + \boldsymbol{e}_j\boldsymbol{e}_i^\top)\right\| &\le \frac{1}{p}\|\boldsymbol{M}\|_\infty, \\ \left\|\left(\frac{\delta_{ii}}{p} - 1\right) M_{ii}\boldsymbol{e}_i\boldsymbol{e}_i^\top\right\| &\le \frac{1}{p}\|\boldsymbol{M}\|_\infty.\end{aligned}$$

We also have the following bound on V.

$$V = \frac{1-p}{p}\left\|\sum_{i,j} M_{ij}^{\star 2}\boldsymbol{e}_i\boldsymbol{e}_i^\top\right\| \le \frac{n}{p}\|\boldsymbol{M}\|_\infty^2$$

Hence, Corollary 19 implies the desired result. □

We can prove Lemma 9 by applying Lemma 20 to $\boldsymbol{M}^\star$ and using $\|\boldsymbol{M}^\star\|_\infty = \lambda^\star \frac{\mu}{n}$.

We introduce classical Bernstein inequality and the results obtained from it.

Theorem 21 (Bernstein Inequality)**.** *Let $\{X_i\}$ be independent random variables. Assume that each random variable satisfies $\mathbb{E}\, X_i = 0$ and $|X_i| \le L$ almost surely. Then, for all $\tau \ge 0$, we have*

$$\mathbb{P}\left[\left|\sum_i X_i\right| \ge \tau\right] \le 2\exp\left(\frac{-\tau^2/2}{V + L\tau/3}\right),$$

where $V = \sum_i \mathbb{E}\big[X_i^2\big]$.

Corollary 22 (Bernstein Inequality)**.** *Let $\{X_i\}$ be independent random variables. Assume that each random variable satisfies $\mathbb{E}\, X_i = 0$ and $|X_i| \le L$ almost surely. Then, with high probability, we have*

$$\left|\sum_i X_i\right| \lesssim \sqrt{V \log n} + L \log n,$$

where $V = \sum_i \mathbb{E}\big[X_i^2\big]$.

Lemma 23. *Let and $\{X_i\}$ be independent Bernoulli random variables with expectation p. Then, for any fixed vector $\boldsymbol{a}$, we have*

$$\left|\sum_i \left(\frac{X_i}{p} - 1\right) a_i\right| \lesssim \sqrt{\frac{\log n}{p}}\|\boldsymbol{a}\|_2 + \frac{\log n}{p}\|\boldsymbol{a}\|_\infty$$

with high probability.

Proof. We can apply Corollary 22 with $L = \frac{1}{p}\|\boldsymbol{a}\|_\infty$ and $V = \frac{1-p}{p}\|\boldsymbol{a}\|_2^2$. □

Lemma 24. *If $n^2 p \gtrsim \mu n \log n$, we have*

$$\|\boldsymbol{M}^\circ\|_{2,\infty} \lesssim \lambda^\star \sqrt{\frac{\mu}{np}}$$

with high probability.

Proof. Let us consider ℓ_2-norm of the ith row of $\boldsymbol{M}^\circ$.

$$\begin{aligned}
\|\boldsymbol{M}^\circ_{i*}\|_2^2 &= \lambda^{\star 2} u_i^{\star 2} \sum_j \frac{1}{p^2}\delta_{ij} u_j^{\star 2} \\
&\le \frac{1}{p}\lambda^{\star 2}\|\boldsymbol{u}^\star\|_\infty^2 \left(\|\boldsymbol{u}^\star\|_2^2 + \left(\sum_j \frac{1}{p}\delta_{ij} u_j^{\star 2} - \|\boldsymbol{u}^\star\|_2^2\right)\right) \\
&\lesssim \frac{\lambda^{\star 2}\mu}{np}\left(1 + \sqrt{\frac{\log n}{np}}\right) \lesssim \frac{\lambda^{\star 2}\mu}{np}
\end{aligned}$$

The third line follows from Lemma 23. □

Proof of Lemma 10. The spectral norm of a symmetric matrix that has nonzero entries only on the lth row/column is bounded by twice of the norm of its lth row. Hence,

$$\begin{aligned}
\left\|\boldsymbol{M}^\circ - \boldsymbol{M}^{(l)}\right\| &\le 2\left\|(\boldsymbol{M}^\circ - \boldsymbol{M}^{(l)})_{l*}\right\|_2 = 2\|(\boldsymbol{M}^\circ - \boldsymbol{M}^\star)_{l*}\|_2 \\
&\lesssim \|\boldsymbol{M}^\circ\|_{2,\infty} + \|\boldsymbol{M}^\star\|_{2,\infty} \lesssim \lambda^\star\sqrt{\frac{\mu}{np}},
\end{aligned}$$

where the last inequality follows from Lemma 24. □

Lemma 25. *Let $\boldsymbol{y}$ be a vector that is independent from the sampling. Then, if $n^2 p \gtrsim n \log n$, we have*

$$\max_{i\in[n]}\left|\|\boldsymbol{x}\|_{2,i}^2 - \|\boldsymbol{y}\|_2^2\right| \lesssim n\|\boldsymbol{x} - \boldsymbol{y}\|_\infty(\|\boldsymbol{x}\|_\infty + \|\boldsymbol{y}\|_\infty) + \sqrt{\frac{\log n}{p}}\|\boldsymbol{y}\|_2\|\boldsymbol{y}\|_\infty + \frac{\log n}{p}\|\boldsymbol{y}\|_\infty^2$$

with very high probability.

Proof. Let us fix i and decompose the difference as

$$\|\boldsymbol{x}\|_{2,i}^2 - \|\boldsymbol{y}\|_2^2 = \frac{1}{p}\sum_{j=1}^n \delta_{ij}(x_j^2 - y_j^2) + \sum_{j=1}^n \left(\frac{\delta_{ij}}{p} - 1\right) y_j^2.$$

The first term is bounded as

$$\left|\frac{1}{p}\sum_{j=1}^n \delta_{ij}(x_j^2 - y_j^2)\right| \le \|\boldsymbol{x}-\boldsymbol{y}\|_\infty \|\boldsymbol{x}+\boldsymbol{y}\|_\infty \frac{1}{p}\sum_{j=1}^n \delta_{ij} \lesssim n\|\boldsymbol{x}-\boldsymbol{y}\|_\infty \|\boldsymbol{x}+\boldsymbol{y}\|_\infty,$$

and the second term is bounded as

$$\left|\sum_{j=1}^n \left(\frac{\delta_{ij}}{p} - 1\right) y_j^2\right| \lesssim \sqrt{\frac{\log n}{p}}\|\boldsymbol{y}\|_2\|\boldsymbol{y}\|_\infty + \frac{\log n}{p}\|\boldsymbol{y}\|_\infty^2$$

by Lemma 23. □

Lemma 26. *Let $\boldsymbol{y}$ be a vector that is independent from the sampling. Then, if $n^2 p \gtrsim n\log n$, we have*

$$\left\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{x}\boldsymbol{x}^\top) - \boldsymbol{y}\boldsymbol{y}^\top\right\| \lesssim n\|\boldsymbol{x}-\boldsymbol{y}\|_\infty(\|\boldsymbol{x}\|_\infty + \|\boldsymbol{y}\|_\infty) + \sqrt{\frac{n\log n}{p}}\|\boldsymbol{y}\|_\infty^2$$

with very high probability.

Proof. We have the following sequence of inequalities

$$\begin{aligned}
\left\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{x}\boldsymbol{x}^\top) - \boldsymbol{y}\boldsymbol{y}^\top\right\| &\le \left\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{x}\boldsymbol{x}^\top) - \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{y}\boldsymbol{y}^\top)\right\| + \left\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{y}\boldsymbol{y}^\top) - \boldsymbol{y}\boldsymbol{y}^\top\right\| \\
&\le \|\boldsymbol{x}\boldsymbol{x}^\top - \boldsymbol{y}\boldsymbol{y}^\top\|_\infty \left\|\frac{1}{p}\mathcal{P}_\Omega(\mathbf{1}\mathbf{1}^\top)\right\| + \left\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{y}\boldsymbol{y}^\top) - \boldsymbol{y}\boldsymbol{y}^\top\right\| \\
&\lesssim \|\boldsymbol{x}-\boldsymbol{y}\|_\infty(\|\boldsymbol{x}\|_\infty + \|\boldsymbol{y}\|_\infty)\left\|\frac{1}{p}\mathcal{P}_\Omega(\mathbf{1}\mathbf{1}^\top)\right\| + \left\|\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{y}\boldsymbol{y}^\top) - \boldsymbol{y}\boldsymbol{y}^\top\right\| \\
&\lesssim n\|\boldsymbol{x}-\boldsymbol{y}\|_\infty(\|\boldsymbol{x}\|_\infty + \|\boldsymbol{y}\|_\infty) + \sqrt{\frac{n\log n}{p}}\|\boldsymbol{y}\|_\infty^2,
\end{aligned}$$

where the second line is derived from a basic inequality $\|\boldsymbol{A}\| \le |||\boldsymbol{A}|||$ that holds for any matrix $\boldsymbol{A}$, and the last line follows by applying Lemma 20 to $\mathbf{1}\mathbf{1}^\top$ and $\boldsymbol{y}\boldsymbol{y}^\top$. □

Lemma 27. *Let $\boldsymbol{y}$ be a vector that is independent from the sampling. Then, if $n^2 p \gtrsim n\log n$, we have*

$$\begin{aligned}
\left\|\nabla^2 g(\boldsymbol{x}) - \left(\|\boldsymbol{y}\|_2^2 \boldsymbol{I} + 2\boldsymbol{y}\boldsymbol{y}^\top\right)\right\| &\lesssim n\|\boldsymbol{x}-\boldsymbol{y}\|_\infty(\|\boldsymbol{x}\|_\infty + \|\boldsymbol{y}\|_\infty) \\
&\quad + \sqrt{\frac{\log n}{p}}\|\boldsymbol{y}\|_2\|\boldsymbol{y}\|_\infty + \frac{\log n}{p}\|\boldsymbol{y}\|_\infty^2 + \sqrt{\frac{n\log n}{p}}\|\boldsymbol{y}\|_\infty^2
\end{aligned}$$

Proof. This follows directly from Lemmas 25 and 26. □

Let us define an operator $\mathcal{P}_{\Omega_l}$ such that an entry of $\mathcal{P}_{\Omega_l}(\boldsymbol{X})$ is equal to that of $\boldsymbol{X}$ if it is contained both in the lth row/column and Ω, and otherwise 0. We also define an operator $\mathcal{P}_l$ that makes the entries outside the lth row/column zero. Then, we have

$$\mathcal{P}_\Omega^{(l)}(\boldsymbol{X}) - \mathcal{P}(\boldsymbol{X}) = \frac{1}{p}\mathcal{P}_\Omega^{(l)}(\boldsymbol{X}) - \mathcal{P}_l(\boldsymbol{X}).$$

The following lemma was also introduced in Ma et al. (2020), but we include the proof for completeness.

Lemma 28. *Suppose that a matrix $\boldsymbol{M}$ and a vector $\boldsymbol{v}$ are independent from sampling of the lth row/column. If $n^2p \gtrsim n\log n$, we have*

$$\left\|\left(\frac{1}{p}\mathcal{P}_{\Omega_l}(\boldsymbol{M}) - \mathcal{P}_l(\boldsymbol{M})\right)\boldsymbol{v}\right\|_2 \lesssim \|\boldsymbol{M}\|_\infty\left(\sqrt{\frac{\log n}{p}}\|\boldsymbol{v}\|_2 + \frac{\log n}{p}\|\boldsymbol{v}\|_\infty + \sqrt{\frac{n}{p}}\|\boldsymbol{v}\|_\infty\right)$$

with high probability.

Proof. If consider the contribution of lth term and the other terms separately, we have

$$\begin{aligned}\left\|\left(\frac{1}{p}\mathcal{P}_{\Omega_l}(\boldsymbol{M}) - \mathcal{P}_l(\boldsymbol{M})\right)\boldsymbol{v}\right\|_2 &\leq \left|\sum_{j=1}^n\left(\frac{\delta_{lj}}{p}-1\right)M_{lj}v_j\right| + |v_l|\sqrt{\sum_{i=1}^n\left(\frac{\delta_{il}}{p}-1\right)^2 M_{il}^2} \\ &\leq \|\boldsymbol{M}\|_\infty\left(\left|\sum_{j=1}^n\left(\frac{\delta_{lj}}{p}-1\right)v_j\right| + \|\boldsymbol{v}\|_\infty\sqrt{\sum_{i=1}^n\left(\frac{\delta_{il}}{p}-1\right)^2}\right)\end{aligned}$$

From Lemma 23, we have

$$\left|\sum_{j=1}^n\left(\frac{\delta_{lj}}{p}-1\right)v_j\right| \lesssim \sqrt{\frac{\log n}{p}}\|\boldsymbol{v}\|_2 + \frac{\log n}{p}\|\boldsymbol{v}\|_\infty$$

with high probability. Regarding the second term, notice that

$$\sum_{i=1}^n\left(\frac{\delta_{il}}{p}-1\right)^2 = n + \left(\frac{1}{p}-2\right)\sum_{i=1}^n\frac{\delta_{il}}{p}.$$

Lemma 23 implies that $\sum_{i=1}^n \frac{\delta_{il}}{p} \asymp n$ with high probability if $n^2p \gtrsim n\log n$. Hence, we have

$$\sum_{i=1}^n\left(\frac{\delta_{il}}{p}-1\right)^2 \lesssim \frac{n}{p},$$

and this finishes the proof. □

Lemma 29. *Let $\boldsymbol{M}$ be a matrix and $\boldsymbol{v}$, $\boldsymbol{w}$ be vectors that are independent from sampling of the lth row/column. Then, if $n^2p \gtrsim n\log n$, we have*

$$\begin{aligned}&\left|\boldsymbol{w}^\top\left(\frac{1}{p}\mathcal{P}_{\Omega_l}(\boldsymbol{M}) - \mathcal{P}_l(\boldsymbol{M})\right)\boldsymbol{v}\right| \\ &\lesssim \|\boldsymbol{M}\|_\infty\left(\sqrt{\frac{\log n}{p}}(\|\boldsymbol{v}\|_2\|\boldsymbol{w}\|_\infty + \|\boldsymbol{w}\|_2\|\boldsymbol{v}\|_\infty) + \frac{\log n}{p}\|\boldsymbol{v}\|_\infty\|\boldsymbol{w}\|_\infty\right)\end{aligned}$$

Proof. We can consider the lth row and lth column separately by

$$\begin{aligned}\left|\boldsymbol{w}^\top\left(\frac{1}{p}\mathcal{P}_{\Omega_l}(\boldsymbol{M}) - \mathcal{P}_l(\boldsymbol{M})\right)\boldsymbol{v}\right| &\leq \left|v_l\sum_i\left(\frac{\delta_{lj}}{p}-1\right)M_{il}w_i\right| + \left|w_l\sum_j\left(\frac{\delta_{lj}}{p}-1\right)M_{lj}v_j\right| \\ &\leq \|\boldsymbol{v}\|_\infty\left|\sum_i\left(\frac{\delta_{lj}}{p}-1\right)M_{il}w_i\right| + \|\boldsymbol{w}\|_\infty\left|\sum_j\left(\frac{\delta_{lj}}{p}-1\right)M_{lj}v_j\right|\end{aligned}$$

If we apply Lemma 23 to each summation, we get the desired result. □