TOWARDS RESOLVING THE IMPLICIT BIAS OF GRADIENT DESCENT FOR MATRIX FACTORIZATION: GREEDY LOW-RANK LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Matrix factorization is a simple and natural test-bed to investigate the implicit regularization of gradient descent. Gunasekar et al. (2018) conjectured that Gradient Flow with infinitesimal initialization converges to the solution that minimizes the nuclear norm, but a series of recent papers argued that the language of norm minimization is not sufficient to give a full characterization for the implicit regularization. In this work, we provide theoretical and empirical evidence that for depth-2 matrix factorization, gradient flow with infinitesimal initialization is mathematically equivalent to a simple heuristic rank minimization algorithm, Greedy Low-Rank Learning, under some reasonable assumptions. This generalizes the rank minimization view from previous works to a much broader setting and enables us to construct counter-examples to refute the conjecture from Gunasekar et al. (2018). We also extend the results to the case where depth ≥ 3, and we show that the benefit of being deeper is that the above convergence has a much weaker dependence over initialization magnitude so that this rank minimization is more likely to take effect for initialization with practical scale.

1 INTRODUCTION

There are usually far more learnable parameters in deep neural nets than the number of training data, but still deep learning works well on real-world tasks. Even with explicit regularization, the model complexity of state-of-the-art neural nets is so large that they can fit randomly labeled data easily Zhang et al. (2017). Towards explaining the mystery of generalization, we must understand what kind of implicit regularization does Gradient Descent (GD) impose during training. Ideally, we are hoping for a nice mathematical characterization on how GD constrains the set of functions that can be expressed by a trained neural net.

As a direct analysis for deep neural nets could be quite hard, a line of works turned to study the implicit regularization on simpler problems to get inspirations, for example, low-rank matrix factorization, a fundamental problem in machine learning and information process. Given a set of observations about an unknown matrix $\boldsymbol{W}^* \in \mathbb{R}^{d\times d}$ of rank $r^* \ll d$, one needs to find a low-rank solution $\boldsymbol{W}$ that is compatible with the given observations. Examples include matrix sensing, matrix completion, phase retrieval, robust principal component analysis, just to name a few (see Chi et al. 2019 for a survey). When $\boldsymbol{W}^*$ is symmetric and positive semidefinite, one way to solve all these problems is to parameterize $\boldsymbol{W}$ as $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{U}^\top$ for $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ and optimize $\mathcal{L}(\boldsymbol{U}) := \frac{1}{2}f(\boldsymbol{U}\boldsymbol{U}^\top)$, where $f(\cdot)$ is some empirical risk function depending on the observations, and r is the rank constraint. In theory, if the rank constraint is too loose, the solutions do not have to be low-rank and we may fail to recover $\boldsymbol{W}^*$. However, even in the case where the rank is unconstrained (i.e., $r = d$), GD with small initialization can still get good performance in practice. This empirical observation reveals that the implicit regularization of GD exists even in this simple matrix factorization problem, but its mechanism is still on debate. Gunasekar et al. (2017) proved that Gradient Flow (GD with infinitesimal step size, a.k.a., GF) with infinitesimal initialization finds the minimum nuclear norm solution in a special case of matrix sensing, and further conjectured this holds in general.

Conjecture 1.1 (Gunasekar et al. 2017, informal)**.** With sufficiently small initialization, GF converges to the minimum nuclear norm solution of matrix sensing.

Subsequently, Arora et al. (2019a) challenged this view by arguing that a simple mathematical norm may not be a sufficient language for characterizing implicit regularization. One example illustrated in Arora et al. (2019a) is regarding matrix sensing with a single observation. They showed that GD with small initialization enhances the growth of large singular values of the solution and attenuates that of smaller ones. This enhancement/attenuation effect encourages low-rank, and it is further intensified with depth in deep matrix factorization (i.e., GD optimizes $f(\boldsymbol{U}_1 \cdots \boldsymbol{U}_L)$ for $L \geq 2$). However, these are not captured by the nuclear norm alone. Gidel et al. (2019); Gissin et al. (2020) further exploited this idea and showed in the special case of full-observation matrix sensing that GF learns solutions with gradually increasing rank. Razin and Cohen (2020) showed in a simple class of matrix completion problems that GF decreases the rank along the trajectory while any norm grows towards infinity. More aggressively, they conjectured that the implicit regularization can be explained by rank minimization rather than norm minimization.

Our Contributions. In this paper, we move one further step towards resolving the implicit regularization in the matrix factorization problem. Our theoretical results show that GD performs rank minimization via a greedy process in a broader setting. Specifically, we provide theoretical evidence that GF with infinitesimal initialization is in general mathematically equivalent to another algorithm called *Greedy Low-Rank Learning* (GLRL). At a high level, GLRL is a greedy algorithm that performs optimization under a rank constraint and relaxes the constraint by 1 whenever the optimization objective stops decreasing. As a by-product, we refute Conjecture 1.1 by demonstrating that counterexamples exist (Example 5.9).

We also extend our results to deep matrix factorization Section 6, where we show that GF with infinitesimal identity initialization converges to GLRL before GLRL reaches its first stationary point. This result confirms the intuition achieved on toy models (Gissin et al., 2020), that benefits of depth in matrix factorization is to encourage rank minimization even for initialization with a relatively larger scale, and thus it is more likely to happen in practice. This shows that describing the implicit regularization using GLRL is more expressive than using the language of norm minimization. We validate all our results with experiments in Appendix B.

2 Related Works

Norm Minimization. The view of norm minimization, or the closely related view of margin maximization, has been explored in different settings. Besides the nuclear norm minimization for matrix factorization (Gunasekar et al., 2017) discussed in the introduction, previous works have also studied the norm minimization/margin maximization for linear regression (Wilson et al., 2017; Soudry et al., 2018a;b; Nacson et al., 2019b;c; Ji and Telgarsky, 2019b), deep linear neural nets (Ji and Telgarsky, 2019a; Gunasekar et al., 2018), homogeneous neural nets (Nacson et al., 2019a; Lyu and Li, 2020), ultra-wide neural nets (Jacot et al., 2018; Arora et al., 2019b; Chizat and Bach, 2020).

Small Initialization and Rank Minimization. The initialization scale can greatly influence the implicit regularization. A sufficiently large initialization can make the training dynamics fall into the lazy training regime defined in Chizat et al. (2019) and diminish test accuracy. Using small initialization is particularly important to bias gradient descent to low-rank solutions for matrix factorization, as empirically observed in Gunasekar et al. (2017). Arora et al. (2019a); Gidel et al. (2019); Gissin et al. (2020); Razin and Cohen (2020) studied how gradient flow with small initialization encourages low-rank in simple settings, as discussed in the introduction. Li et al. (2018) proved recovery guarantees for gradient flow solving matrix sensing under Restricted Isometry Property (RIP), but the proof cannot be generalized easily to the case without RIP. Belabbas (2020) made attempts to prove that gradient flow is approximately rank-1 in the very early phase of training, but it does not exclude the possibility that the approximation error explodes later and gradient flow is not converging to low-rank solutions. Compared to these works, the current paper studies how GF encourages low-rank in a much broader setting.

3 Background

Notations. For two matrices $\boldsymbol{A}, \boldsymbol{B}$, $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \mathrm{Tr}(\boldsymbol{A}\boldsymbol{B}^\top)$ denotes the inner product of $\boldsymbol{A}$ and $\boldsymbol{B}$. We use $\|\boldsymbol{A}\|_{\mathrm{F}}$ to denote the Frobenius norm of $\boldsymbol{A}$ and use $\|\boldsymbol{A}\|_*$ to denote the nuclear norm. For a matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$, we use $\lambda_1(\boldsymbol{A}), \cdots, \lambda_d(\boldsymbol{A})$ to denote the eigenvalues of $\boldsymbol{A}$ in decreasing order. We use $\boldsymbol{A} \succeq \boldsymbol{B}$ or $\boldsymbol{B} \preceq \boldsymbol{A}$ to stand for $\boldsymbol{A} - \boldsymbol{B}$ is a positive semidefinite (PSD) matrix.

Matrix Factorization. Matrix factorization problem asks one to optimize $\mathcal{L}(\boldsymbol{U}, \boldsymbol{V}) := \frac{1}{2} f(\boldsymbol{U}\boldsymbol{V}^\top)$ among $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{d\times r}$, where $f : \mathbb{R}^{d\times d} \to \mathbb{R}$ is a convex function. A notable example is matrix sensing. There is an unknown rank-r^* matrix $\boldsymbol{W}^* \in \mathbb{R}^{d\times d}$ with $r^* \ll d$. You have m measurements, $\boldsymbol{X}_1, \dots, \boldsymbol{X}_m \in \mathbb{R}^{d\times d}$, and you can observe $y_i := \langle \boldsymbol{X}_i, \boldsymbol{W}^* \rangle$. The goal is to reconstruct $\boldsymbol{W}^*$. The function $f(\boldsymbol{W})$ here is defined by $f(\boldsymbol{W}) := \frac{1}{2}\sum_{i=1}^m \left(\langle \boldsymbol{W}, \boldsymbol{X}_i\rangle - y_i\right)^2$. Matrix completion is another notable example. It can be seen as a special case of matrix sensing in which every measurement has the form $\boldsymbol{X}_i = \boldsymbol{e}_{p_i}\boldsymbol{e}_{q_i}^\top$, where $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_d\}$ stands for the standard basis (i.e., exactly one entry is observed through each measurement).

For technical simplicity, in this paper we focus on the symmetric case as in previous works (Gunasekar et al., 2017). Given a convex $\mathcal{C}^3$-smooth function $f(\cdot)$, we parameterize $\boldsymbol{W}$ as $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{U}^\top$ for $\boldsymbol{U} \in \mathbb{R}^{d\times r}$ and optimize $\mathcal{L}(\boldsymbol{U}) = \frac{1}{2} f(\boldsymbol{U}\boldsymbol{U}^\top)$ to search for a low-rank solution $\boldsymbol{W}$ among PSD matrices. Note that this adaptation is not restrictive: let $\boldsymbol{U}' = \left[\begin{smallmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{smallmatrix}\right] \in \mathbb{R}^{2d\times r}$, $f'\left(\left[\begin{smallmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{smallmatrix}\right]\right) = f(\boldsymbol{B})$, then $f(\boldsymbol{U}\boldsymbol{V}^\top) = f'(\boldsymbol{U}'\boldsymbol{U}'^\top)$, and thus solving the symmetric case is sufficient to solve all other cases. We assume WLOG throughout this paper that $f(\boldsymbol{W}) = f(\boldsymbol{W}^\top)$; otherwise, we can set $f'(\boldsymbol{W}) = \frac{1}{2}\left(f(\boldsymbol{W}) + f(\boldsymbol{W}^\top)\right)$ so that $f'(\boldsymbol{W}) = f'(\boldsymbol{W}^\top)$ and $\mathcal{L}(\boldsymbol{U}) = \frac{1}{2} f'(\boldsymbol{U}\boldsymbol{U}^\top)$ are unaffected. This assumption also makes $\nabla f(\boldsymbol{W})$ symmetric for every symmetric $\boldsymbol{W}$.

Overparameterized Matrix Factorization. We particularly focus on the overparameterized case, where $r = d$, namely no rank constraint for the matrix $\boldsymbol{W}$. Our focus in this paper is to analyze the implicit regularization of GF:

$$\frac{d\boldsymbol{U}}{dt} = -\nabla\mathcal{L}(\boldsymbol{U}) = -\nabla f(\boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U}. \tag{1}$$

Let $\boldsymbol{W}(t) = \boldsymbol{U}(t)\boldsymbol{U}(t)^\top \in \mathbb{R}^{d\times d}$. Then the following end-to-end dynamics holds for $\boldsymbol{W}(t)$:

$$\frac{d\boldsymbol{W}}{dt} = -\boldsymbol{W}\nabla f(\boldsymbol{W}) - \nabla f(\boldsymbol{W})\boldsymbol{W}. \tag{2}$$

We use $\phi(\boldsymbol{W}_0, t)$ to denote the matrix $\boldsymbol{W}(t)$ in (2) when $\boldsymbol{W}(0) = \boldsymbol{W}_0$, where $\boldsymbol{W}_0$ is a PSD matrix.

4 Warmup Examples

Before introducing our main results, we illustrate how GD performs greedy learning using two warmup examples.

Linearization Around the Origin. In general, for a loss function $\mathcal{L}(\boldsymbol{U}) = \frac{1}{2} f(\boldsymbol{U}\boldsymbol{U}^\top)$, we can always apply Taylor expansion $f(\boldsymbol{W}) \approx f(\boldsymbol{0}) + \langle \boldsymbol{W}, \nabla f(\boldsymbol{0})\rangle$ around the origin to approximate it with a linear function. This motivates us to study the linear case: $f(\boldsymbol{W}) := f_0 - \langle \boldsymbol{W}, \boldsymbol{Q}\rangle$ for some symmetric matrix $\boldsymbol{Q}$. In this case, the matrix $\boldsymbol{U}$ follows the ODE $\frac{d\boldsymbol{U}}{dt} = \boldsymbol{Q}\boldsymbol{U}$, which can be understood as a continuous version of the classical power iteration method for solving the top eigenvector. Let $\boldsymbol{Q} := \sum_{i=1}^d \mu_i \boldsymbol{v}_i\boldsymbol{v}_i^\top$ be the eigendecomposition of $\boldsymbol{Q}$, where $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_d$ and $\boldsymbol{v}_1, \dots, \boldsymbol{v}_d$ are orthogonal to each other. Then we can write the solution as:

$$\boldsymbol{U}(t) = e^{t\boldsymbol{Q}}\boldsymbol{U}(0) = \left(\sum\nolimits_{i=1}^d e^{\mu_i t}\boldsymbol{v}_i\boldsymbol{v}_i^\top\right)\boldsymbol{U}(0). \tag{3}$$

When $\mu_1 > \mu_2$, the ratio between $e^{\mu_1 t}$ and $e^{\mu_i t}$ for $i \neq 1$ increases exponentially fast. As $t \to +\infty$, $\boldsymbol{U}(t)$ and $\boldsymbol{W}(t)$ become approximately rank-1 as long as $\boldsymbol{v}_i^\top\boldsymbol{U}(0) \neq \boldsymbol{0}$, i.e.,

$$\lim_{t\to\infty} e^{-\mu_1 t}\boldsymbol{U}(t) = \boldsymbol{v}_1\boldsymbol{v}_1^\top\boldsymbol{U}(0), \qquad \lim_{t\to\infty} e^{-2\mu_1 t}\boldsymbol{W}(t) = (\boldsymbol{v}_1^\top\boldsymbol{W}(0)\boldsymbol{v}_1)\boldsymbol{v}_1\boldsymbol{v}_1^\top. \tag{4}$$

The analysis for the simple linear case reveals that GD encourages low-rank through a process similar to power iteration. However, $f(\boldsymbol{W})$ is non-linear in general, and the linear approximation is close to $f(\boldsymbol{W})$ only if $\boldsymbol{W}$ is very small. With sufficiently small initialization, we can imagine that GD still resembles the above power iteration in the early phase of the optimization. But what if $\boldsymbol{W}(t)$ grows to be so large that the linear approximation is far from the actual $f(\boldsymbol{W})$?

Full-observation Matrix Sensing. To understand the dynamics of GD when the linearization fails, we now consider a well-studied special case (Gissin et al., 2020): $\mathcal{L}(\boldsymbol{U}) = \frac{1}{2} f(\boldsymbol{U}\boldsymbol{U}^\top), f(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{W} - \boldsymbol{W}^*\|_{\mathrm{F}}^2$ for some unknown PSD matrix $\boldsymbol{W}^*$. GF in this case can be written as:

$$\frac{d\boldsymbol{U}}{dt} = (\boldsymbol{W}^* - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{U}, \qquad \frac{d\boldsymbol{W}}{dt} = (\boldsymbol{W}^* - \boldsymbol{W})\boldsymbol{W} + \boldsymbol{W}(\boldsymbol{W}^* - \boldsymbol{W}). \tag{5}$$

Let $\boldsymbol{W}^* := \sum_{i=1}^d \mu_i \boldsymbol{v}_i\boldsymbol{v}_i^\top$ be the eigendecomposition of $\boldsymbol{W}^*$. Our previous analysis shows that the dynamics is approximately $\frac{d\boldsymbol{U}}{dt} = \boldsymbol{W}^*\boldsymbol{U}$ in the early phase and thus encourages low-rank.

To get a sense for the later phases, we simplify the setting by specifying $\boldsymbol{U}(0) = \sqrt{\alpha}\boldsymbol{I}$ for a small number α. We can write $\boldsymbol{W}(0)$ and $\boldsymbol{W}^*$ as diagonal matrices $\boldsymbol{W}(0) = \mathrm{diag}(\alpha, \alpha, \cdots, \alpha), \boldsymbol{W}^* = \mathrm{diag}(\mu_1, \mu_2, \cdots, \mu_d)$ with respect to the basis $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_d$. It is easy to see that $\boldsymbol{W}(t)$ is always a diagonal matrix, since the time derivatives of non-diagonal coordinates stay 0 during training. Let $\boldsymbol{W}(t) = \mathrm{diag}(\sigma_1(t), \sigma_2(t), \cdots, \sigma_d(t))$, then $\sigma_i(t)$ satisfies the dynamical equation $\frac{d}{dt}\sigma_i(t) = 2\sigma_i(t)(\mu_i - \sigma_i(t))$, and thus $\sigma_i(t) = \frac{\alpha\mu_i}{\alpha + (\mu_i - \alpha)e^{-2\mu_i t}}$. This shows that every $\sigma_i(t)$ increases from α to μ_i over time. As $\alpha \to 0$, every $\sigma_i(t)$ has a sharp transition from near 0 to near μ_i at time roughly $(\frac{1}{2\mu_i} + o(1)) \log \frac{1}{\alpha}$, which can be seen from the following limit:

$$\lim_{\alpha \to 0} \sigma_i\left((\tfrac{1}{2\mu_i} + c)\log(1/\alpha)\right) = \lim_{\alpha \to 0} \frac{\alpha\mu_i}{\alpha + (\mu_i - \alpha)\alpha^{1+2c\mu_i}} = \begin{cases} 0 & c \in (-\frac{1}{2\mu_i}, 0), \\ \mu_i & c \in (0, +\infty). \end{cases}$$

This means for every $q \in (\frac{1}{2\mu_i}, \frac{1}{2\mu_{i+1}})$ for $i = 1, \ldots, d-1$ (or $q \in (\frac{1}{2\mu_i}, +\infty)$ for $i = d$), $\lim_{\alpha \to 0} \boldsymbol{W}(q \log(1/\alpha)) = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_i, 0, 0, \cdots, 0)$. Therefore, when the initialization is sufficiently small, GF learns each component of $\boldsymbol{W}^*$ one by one, according to the relative order of eigenvalues. At a high level, this shows a greedy nature of GD: GD starts learning with simple models; whenever it underfits, it increases the model complexity (which is rank in our case). This is also called *sequential learning* or *incremental learning* in Gidel et al. (2019); Gissin et al. (2020).

However, it is unclear how and why this sequential learning/incremental learning can occur in general. Through the first warmup example, we may understand why GD learns a rank-1 matrix in the early phase, but does GD always learn solutions with rank $2, 3, 4, \ldots$ sequentially? If true, what is the mechanism behind this? The current paper answers the questions by providing both theoretical and empirical evidence that the greedy learning behavior does occur in general with a similar reason as for the first warmup example.

5 Greedy Low-Rank Learning (GLRL)

In this section, we present a trajectory-based analysis for the implicit bias of GF on matrix factorization. Our main result is that GF with infinitesimal initialization is generically the same as that of a simple greedy algorithm, *Greedy Low-Rank Learning* (GLRL, Algorithm 1).

The GLRL algorithm consists of several phases, numbered from 1. In phase r, GLRL increases the rank constraint to r and optimizes $\mathcal{L}(\boldsymbol{U}_r) := \frac{1}{2}f(\boldsymbol{U}_r\boldsymbol{U}_r^\top)$ among $\boldsymbol{U}_r \in \mathbb{R}^{d\times r}$ via GD until it reaches a stationary point $\boldsymbol{U}_r(\infty)$. A natural way to initialize GD is to choose the convergent solution $\boldsymbol{U}_{r-1}(\infty) \in \mathbb{R}^{d\times(r-1)}$ from the last phase and append a zero column to make it have r columns. However, it is still a stationary point of $\mathcal{L}(\cdot)$ and GD cannot escape from it without perturbation. GLRL resolves this issue by adding a column vector so that the loss decreases the fastest: GLRL first chooses a (normalized) top eigenvector $\boldsymbol{u}_r$ of $-\nabla f(\boldsymbol{W}_r)$ where $\boldsymbol{W}_r := \boldsymbol{U}_{r-1}(\infty)\boldsymbol{U}_{r-1}^\top(\infty)$, and then constructs $\boldsymbol{U}_r(0) \leftarrow [\boldsymbol{U}_{r-1}(\infty)\ \ \sqrt{\epsilon}\boldsymbol{u}_r] \in \mathbb{R}^{d\times r}$ as the initialization of phase r. Here $\epsilon > 0$ is a parameter that controls the magnitude of the new column. When $-\nabla f(\boldsymbol{W}_r) \preceq \boldsymbol{0}$, such descent direction would not exist and the algorithm returns $\boldsymbol{W}_r$.

Algorithm 1: Greedy Low-Rank Learning

parameter : step size $\eta > 0$; small $\epsilon > 0$
$\boldsymbol{W}_0 \leftarrow \boldsymbol{0} \in \mathbb{R}^{d\times d}$, and $\boldsymbol{U}_0(\infty) \in \mathbb{R}^{d\times 0}$
for $r = 1, 2, \ldots$ **do**
  $\boldsymbol{u}_r \leftarrow$ top unit eigenvector of $-\nabla f(\boldsymbol{W}_{r-1})$
  $\boldsymbol{U}_r(0) \leftarrow [\boldsymbol{U}_{r-1}(\infty)\ \ \sqrt{\epsilon}\boldsymbol{u}_r] \in \mathbb{R}^{d\times r}$
  for $t = 0, 1, \ldots$ **do**
    $\boldsymbol{U}_r(t+1) \leftarrow \boldsymbol{U}_r(t) - \eta\nabla\mathcal{L}(\boldsymbol{U}_r(t))$
  $\boldsymbol{W}_r \leftarrow \boldsymbol{U}_r(\infty)\boldsymbol{U}_r^\top(\infty)$ [a]
  if $\lambda_1(-\nabla f(\boldsymbol{W}_r)) \le 0$ **then**
    return $\boldsymbol{W}_r$

[a] In practice, we approximate the infinite time limit by running sufficiently many steps until convergence.

Taking the learning rate $\eta \to 0$, we can define the (limiting) trajectory of GLRL. The goal is to show that the trajectory of GLRL is very close to that of GF with infinitesimal initialization. Recall that $\phi(\boldsymbol{W}_0, t)$ stands for the solution $\boldsymbol{W}(t)$ in (2) when $\boldsymbol{W}(0) = \boldsymbol{W}_0$.

Definition 5.1 (Trajectory of GLRL)**.** Let $\overline{\boldsymbol{W}}_{0,\epsilon} := \boldsymbol{0}$. For every $r \ge 1$, if $\overline{\boldsymbol{W}}_{r-1,\epsilon}$ exists and is not a local minimizer of f, we define $\boldsymbol{W}^{\mathrm{G}}_{r,\epsilon}(t) := \phi(\overline{\boldsymbol{W}}_{r-1,\epsilon} + \epsilon\boldsymbol{u}_{r,\epsilon}\boldsymbol{u}_{r,\epsilon}^\top, t)$, where $\boldsymbol{u}_{r,\epsilon}$ is a (normalized) top eigenvector of $\nabla f(\overline{\boldsymbol{W}}_{r-1,\epsilon})$. We define $\overline{\boldsymbol{W}}_{r,\epsilon} := \lim_{t\to+\infty} \boldsymbol{W}^{\mathrm{G}}_{r,\epsilon}(t)$ if the limit exists.

Throughout this paper, we always focus on the case where the top eigenvalue of every $\nabla f(\overline{\boldsymbol{W}}_{r-1,\epsilon})$ is unique. In this case, the trajectory of GLRL is unique, since the normalized top eigenvectors can only be $\pm\boldsymbol{u}_{r,\epsilon}$, and both of them lead to the same $\boldsymbol{W}^{\mathrm{G}}_{r,\epsilon}(t)$.

Comparison to existing greedy algorithms for rank-constrained optimization: The most related one to GLRL (Algorithm 1) is probably *Rank-1 Matrix Pursuit* (R1MP) proposed by Wang et al. (2014) for matrix completion, which was later generalized to general convex loss in Yao and Kwok (2016). R1MP maintains a set of rank-1 matrices as the basis, and in phase r, R1MP adds the same $\boldsymbol{u}_r\boldsymbol{u}_r^\top$ as defined in Algorithm 1 into its basis and solve $\min_{\boldsymbol{\alpha}} f(\sum_{i=1}^r \alpha_i \boldsymbol{u}_i\boldsymbol{u}_i^\top)$ for rank-r estimation. The main difference between R1MP and GLRL is that the optimization in each phase of R1MP is performed on the coefficients $\boldsymbol{\alpha}$, while the entire $\boldsymbol{U}_r$ evolves with GD in each phase of GLRL. In Figure 3, we provide empirical evidence that GLRL generalizes better than R1MP when ground truth is low-rank, although GLRL may have a higher computational cost depending on η, ϵ.

Similar to R1MP, Greedy Efficient Component Optimization (GECO, Shalev-Shwartz and Singer 2010) also chooses the r-th component of its basis as the top eigenvector of $-\nabla f(\boldsymbol{W}_r)$, while it solves $\min_{\boldsymbol{\beta}} f(\sum_{1\le i,j\le r} \beta_{ij}\boldsymbol{u}_i\boldsymbol{u}_j^\top)$ for the rank-r estimation. Khanna et al. (2017) provided convergence guarantee for GECO assuming strong convexity. Haeffele and Vidal (2019) proposed a local-descent meta algorithm, of which GLRL can be viewed as a specific realization.

5.1 The Limiting Trajectory: A General Theorem for Dynamical System

Before showing the equivalence between GF and GLRL, we first introduce our idea from a higher level of perspective by analyzing a general dynamical system.

$$\frac{d\boldsymbol{\theta}}{dt} = \boldsymbol{g}(\boldsymbol{\theta}).$$

We will use $\boldsymbol{\theta}$ to denote the vectorization of $\boldsymbol{W}$ in our analysis later, whose ODE is given in (2), but for now we just consider a general dynamical system. We use $\phi(\boldsymbol{\theta}_0, t)$ to denote the value of $\boldsymbol{\theta}(t)$ in the case of $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$. Let $\boldsymbol{J}(\boldsymbol{\theta})$ be the Jacobian matrix of $\boldsymbol{g}(\boldsymbol{\theta})$. We assume that $\boldsymbol{g}(\boldsymbol{\theta})$ is $\mathcal{C}^2$-smooth and $\phi(\boldsymbol{\theta}_0, t)$ exists for all $\boldsymbol{\theta}_0$ and t. For technical simplicity, we assume that $\boldsymbol{J}(\mathbf{0})$ is diagonalizable over $\mathbb{R}$. Let $\boldsymbol{J}(\mathbf{0}) := \tilde{\boldsymbol{V}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{V}}^{-1}$ be the eigendecomposition, where $\tilde{\boldsymbol{V}}$ is an invertible matrix and $\tilde{\boldsymbol{D}} = \mathrm{diag}(\tilde{\mu}_1, \dots, \tilde{\mu}_d)$ is the diagonal matrix consisting of the eigenvalues $\tilde{\mu}_1 \ge \tilde{\mu}_2 \ge \cdots \ge \tilde{\mu}_d$. Let $\tilde{\boldsymbol{V}} = (\tilde{\boldsymbol{v}}_1, \dots, \tilde{\boldsymbol{v}}_d)$ and $\tilde{\boldsymbol{V}}^{-1} = (\tilde{\boldsymbol{u}}_1, \dots, \tilde{\boldsymbol{u}}_d)^\top$, then $\tilde{\boldsymbol{u}}_i, \tilde{\boldsymbol{v}}_i$ are left and right eigenvectors associated with $\tilde{\mu}_i$ and $\tilde{\boldsymbol{u}}_i^\top\tilde{\boldsymbol{v}}_j = \delta_{ij}$. We can rewrite the eigendecomposition as $\boldsymbol{J}(\mathbf{0}) = \sum_{i=1}^d \tilde{\mu}_i\tilde{\boldsymbol{v}}_i\tilde{\boldsymbol{u}}_i^\top$.

In this formulation, GF with infinitesimal initialization can be denoted as $\phi(\boldsymbol{\theta}_0, t)$ for initialization $\boldsymbol{\theta}_0$ near $\mathbf{0}$. Note that $\boldsymbol{W} = \mathbf{0}$ is a critical point of (2). Motivated by this, we assume that $\mathbf{0}$ is a critical point of the dynamical system, i.e., $\boldsymbol{g}(\mathbf{0}) = \mathbf{0}$. We consider the case where the top eigenvalue $\tilde{\mu}_1$ is positive. This means the critical point $\boldsymbol{\theta} = \mathbf{0}$ is unstable, and in matrix factorization it corresponds to the case where $\mathbf{0}$ is a strict saddle point.

The key observation is that if the initialization is infinitesimal, the trajectory is almost uniquely determined. To be more precise, we need the following definition:

Definition 5.2. Let $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ and $\boldsymbol{u} \in \mathbb{R}^d$. We say that $\{\boldsymbol{\theta}_\alpha\}_{\alpha\in(0,1)}$ converges to $\boldsymbol{\theta}_0$ with positive alignment with $\boldsymbol{u}$ if $\lim\limits_{\alpha\to 0} \boldsymbol{\theta}_\alpha = \boldsymbol{\theta}_0$ and $\liminf\limits_{\alpha\to 0} \left\langle \frac{\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}_0}{\|\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}_0\|_2}, \boldsymbol{u} \right\rangle > 0$.

A special case is that the direction of $\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}_0$ converges, i.e., $\bar{\boldsymbol{\theta}} := \lim_{\alpha\to 0} \frac{\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}_0}{\|\boldsymbol{\theta}_\alpha - \boldsymbol{\theta}_0\|_2}$ exists. In this case, except a zero-measure set of $\bar{\boldsymbol{\theta}}$ on the unit sphere, any $\bar{\boldsymbol{\theta}}$ leads to positive alignment with either $\boldsymbol{u}$ or $-\boldsymbol{u}$, and thus any convergent sequence generically falls into either of these two categories.

The following theorem shows that if the initial point $\boldsymbol{\theta}_\alpha$ converges to $\mathbf{0}$ with positive alignment with $\tilde{\boldsymbol{u}}_1$ as $\alpha \to 0$, then the trajectory starting with $\boldsymbol{\theta}_\alpha$ converges to a unique trajectory $\boldsymbol{z}(t)$. One can easily show that for the other case that the initialization has positive alignment with $-\tilde{\boldsymbol{u}}_1$ the trajectory converges to $\boldsymbol{z}'(t) := \lim_{\alpha\to 0} \boldsymbol{z}'_\alpha(t)$, where $\boldsymbol{z}'_\alpha(t) := \phi(-\alpha\tilde{\boldsymbol{v}}_1, t + \frac{1}{\tilde{\mu}_1}\log\frac{1}{\alpha})$. This is somewhat surprising: different initial points should lead to very different trajectories, but our analysis shows that generically there are only two types of trajectories for infinitesimal initialization. We will soon see how this theorem helps in our analysis for matrix factorization.

Theorem 5.3. *Let $\boldsymbol{z}_\alpha(t) := \phi(\alpha\tilde{\boldsymbol{v}}_1, t + \frac{1}{\tilde{\mu}_1}\log\frac{1}{\alpha})$ for every $\alpha > 0$, then $\forall t \in \mathbb{R}$, $\boldsymbol{z}(t) := \lim_{\alpha\to 0}\boldsymbol{z}_\alpha(t)$ exists and is also a solution of (2), i.e., $\boldsymbol{z}(t) = \phi(\boldsymbol{z}(0), t)$. If $\boldsymbol{\delta}_\alpha$ converges to $\mathbf{0}$ with positive alignment with $\tilde{\boldsymbol{u}}_1$ as $\alpha \to 0$, then $\forall t \in \mathbb{R}$, there is a constant $C > 0$ such that for every sufficiently small α,*

$$\left\|\phi\left(\boldsymbol{\delta}_\alpha, t + \frac{1}{\tilde{\mu}_1}\log\frac{1}{\langle\boldsymbol{\delta}_\alpha, \tilde{\boldsymbol{u}}_1\rangle}\right) - \boldsymbol{z}(t)\right\|_2 \le C\cdot\|\boldsymbol{\delta}_\alpha\|_2^{\frac{\tilde{\gamma}}{\tilde{\mu}_1+\tilde{\gamma}}}, \tag{6}$$

where $\tilde{\gamma} := \tilde{\mu}_1 - \tilde{\mu}_2 > 0$ is the eigenvalue gap.

Proof sketch. The main idea is to linearize the dynamics near origin as we have done for the first warmup example. For sufficiently small $\boldsymbol{\theta}$, by Taylor expansion of $\boldsymbol{g}(\boldsymbol{\theta})$, the dynamics is approximately $\frac{d\boldsymbol{\theta}}{dt} \approx \boldsymbol{J}(\mathbf{0})\boldsymbol{\theta}$, which can be understood as a continuous version of power iteration. If the linear approximation is exact, then $\boldsymbol{\theta}(t) = e^{t\boldsymbol{J}(0)}\boldsymbol{\theta}(0)$. For large enough t, $e^{t\boldsymbol{J}(0)} = \sum_{i=1}^d e^{\tilde{\mu}_i t}\tilde{\boldsymbol{v}}_i\tilde{\boldsymbol{u}}_i^\top = e^{\tilde{\mu}_1 t}\tilde{\boldsymbol{v}}_1\tilde{\boldsymbol{u}}_1^\top + O(e^{\tilde{\mu}_2 t})$. Therefore, as long as the initial point $\boldsymbol{\theta}(0)$ has a positive inner product with $\tilde{\boldsymbol{u}}_1$, $\boldsymbol{\theta}(t)$ should be very close to $\epsilon\tilde{\boldsymbol{v}}_1$ for some $\epsilon > 0$, and the rest of the trajectory after t should be close to the trajectory starting from $\epsilon\tilde{\boldsymbol{v}}_1$. However, here is a tradeoff: we should choose time t to be long enough so that the power iteration takes effect; but if the time t is so long that the norm of $\boldsymbol{\theta}(t)$ reaches a constant scale, then the linearization fails unavoidably. Nevertheless, if the initialization scale is sufficiently small, we show via a careful error analysis that there is always a suitable choice of t_0 such that $\boldsymbol{\theta}(t_0)$ is well approximated by $\epsilon\tilde{\boldsymbol{v}}_1$ and the difference between $\boldsymbol{\theta}(t_0 + t)$ and $\phi(\epsilon\tilde{\boldsymbol{v}}_1, t)$ is bounded as well. We defer the details to Appendix C. □

Note that the assumption that $\boldsymbol{J}(\mathbf{0})$ is diagonalizable could be restrictive in some cases. But this is only a technical assumption for simplifying the analysis. One can easily generalize the theorem to the case where $\boldsymbol{J}(\mathbf{0})$ is non-diagonalizable by replacing eigendecomposition with Jordan normal form.

5.2 Equivalence Between GD and GLRL: Rank-One Case

Now we establish the equivalence between GF and GLRL in the first phase. The main idea is to apply Theorem 5.3 on (2). For this, we need the following lemma on the eigenvalues and eigenvectors.

Lemma 5.4. *Let $\boldsymbol{g}(\boldsymbol{W}) := -\boldsymbol{W}\nabla f(\boldsymbol{W}) - \nabla f(\boldsymbol{W})\boldsymbol{W}$ and $\boldsymbol{J}(\boldsymbol{W})$ be its Jacobian. Then $\boldsymbol{J}(\mathbf{0})$ is symmetric and thus diagonalizable. Let $-\nabla f(\mathbf{0}) = \sum_{i=1}^d \mu_i \boldsymbol{u}_{1[i]}\boldsymbol{u}_{1[i]}^\top$ be the eigendecomposition of the symmetric matrix $-\nabla f(\mathbf{0})$, where $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_d$. Then $\boldsymbol{J}(\mathbf{0})$ has the form:*

$$\boldsymbol{J}(\mathbf{0})[\boldsymbol{\Delta}] = \sum_{i=1}^d\sum_{j=1}^d(\mu_i + \mu_j)\left\langle\boldsymbol{\Delta}, \boldsymbol{u}_{1[i]}\boldsymbol{u}_{1[j]}^\top\right\rangle\boldsymbol{u}_{1[i]}\boldsymbol{u}_{1[j]}^\top,$$

where $\boldsymbol{J}(\mathbf{0})[\boldsymbol{\Delta}]$ stands for the resulting matrix produced by left-multiplying $\boldsymbol{J}(\mathbf{0})$ to the vectorization of $\boldsymbol{\Delta}$. For every pair of $1 \le i \le j \le d$, $\mu_i + \mu_j$ is an eigenvalue of $\boldsymbol{J}(\mathbf{0})$ and $\boldsymbol{u}_{1[i]}\boldsymbol{u}_{1[j]}^\top + \boldsymbol{u}_{1[j]}\boldsymbol{u}_{1[i]}^\top$ is a corresponding eigenvector. All the other eigenvalues are 0.

We simplify the notation by letting $\boldsymbol{u}_1 := \boldsymbol{u}_{1[1]}$. A direct corollary of Lemma 5.4 is that $\boldsymbol{u}_1\boldsymbol{u}_1^\top$ is the top eigenvector of $\boldsymbol{J}(\mathbf{0})$. According to Theorem 5.3, now there are only two types of trajectories, which correspond to infinitesimal initialization $\boldsymbol{W}_\alpha \to \mathbf{0}$ with positive alignment with $\boldsymbol{u}_1\boldsymbol{u}_1^\top$ or $-\boldsymbol{u}_1\boldsymbol{u}_1^\top$. As the initialization must be PSD, $\boldsymbol{W}_\alpha \to \mathbf{0}$ cannot have positive alignment with $-\boldsymbol{u}_1\boldsymbol{u}_1^\top$. For the former case, we have the following theorem which states that $\boldsymbol{W}_{r,\epsilon}^{\mathrm{G}}(t)$ and $\phi(\boldsymbol{W}_\alpha, t)$ are pointwise close to each other after shifting by an time offset. For this, we only assume that $\mathbf{0}$ is a saddle point and has an eigenvalue gap. In the full observation case, this assumption is satisfied easily if the ground-truth matrix has distinct eigenvalues.

Assumption 5.5. Let $\mu_1 := \lambda_1(-\nabla f(\mathbf{0})), \mu_2 := \lambda_2(-\nabla f(\mathbf{0}))$. Then $\mu_1 > \max\{\mu_2, 0\}$.

Theorem 5.6. *Under Assumption 5.5, the following limit exists for every $t \in (-\infty, +\infty)$.*

$$\boldsymbol{W}_1^{\mathrm{G}}(t) := \lim_{\epsilon\to 0}\boldsymbol{W}_{1,\epsilon}^{\mathrm{G}}\left(\frac{1}{2\mu_1}\log\frac{1}{\epsilon} + t\right) = \lim_{\epsilon\to 0}\phi\left(\epsilon\boldsymbol{u}_1\boldsymbol{u}_1^\top, \frac{1}{2\mu_1}\log\frac{1}{\epsilon} + t\right) \tag{7}$$

Let $\boldsymbol{W}_\alpha \in \boldsymbol{R}^{d\times d}$ be PSD matrices converging to $\mathbf{0}$ with positive alignment with $\boldsymbol{u}_1\boldsymbol{u}_1^\top$ as $\alpha \to 0$, that is, $\lim_{\alpha\to 0}\boldsymbol{W}_\alpha = \mathbf{0}$ and there exists $\alpha_0, q > 0$ such that $\left\langle\boldsymbol{W}_\alpha, \boldsymbol{u}_1\boldsymbol{u}_1^\top\right\rangle \ge q\|\boldsymbol{W}_\alpha\|_{\mathrm{F}}$ for all $\alpha < \alpha_0$. Then for every $t \in (-\infty, +\infty)$, there is a constant $C > 0$ (independent of α) such that

$$\left\|\phi\left(\boldsymbol{W}_\alpha, \frac{1}{2\mu_1}\log\frac{1}{\left\langle\boldsymbol{W}_\alpha, \boldsymbol{u}_1\boldsymbol{u}_1^\top\right\rangle} + t\right) - \boldsymbol{W}_1^{\mathrm{G}}(t)\right\|_{\mathrm{F}} \le C\|\boldsymbol{W}_\alpha\|_{\mathrm{F}}^{\frac{\tilde{\gamma}}{2\mu_1+\tilde{\gamma}}}. \tag{8}$$

for every sufficiently small α, where $\tilde{\gamma} := 2\mu_1 - \max\{\mu_1 + \mu_2, 0\}$.

It is worth to note that $\boldsymbol{W}_1^{\mathrm{G}}(t)$ has rank ≤ 1 for any $t \in (-\infty, +\infty)$, since every $\boldsymbol{W}_{1,\epsilon}^{\mathrm{G}}(t)$ has rank ≤ 1 and the set of matrices with rank ≤ 1 is closed. This matches with the first warmup example: GD does start learning with rank-1 solutions. Interestingly, when $\boldsymbol{W}_1^{\mathrm{G}}(t)$ converges to a local minimizer, which is definitely rank-1, the following theorem shows that GF also converge to the rank-1 solution.

Assumption 5.7. $f(\boldsymbol{W})$ is locally analytic at each point.

Theorem 5.8. *Under Assumptions 5.5 and 5.7, if* $\left\|\boldsymbol{W}_1^{\mathrm{G}}(t)\right\|_{\mathrm{F}}$ *is bounded for* $t \geq 0$*, then* $\overline{\boldsymbol{W}}_1 := \lim_{t\to+\infty} \boldsymbol{W}_1^{\mathrm{G}}(t)$ *exists. Further, if* $\overline{\boldsymbol{W}}_1$ *is a local minimizer of* $f(\,\cdot\,)$ *among PSD matrices, then for PSD matrices* $\boldsymbol{W}_\alpha \in \boldsymbol{R}^{d\times d}$ *converging to* $\mathbf{0}$ *with positive alignment with* $\boldsymbol{u}_1\boldsymbol{u}_1^\top$ *as* $\alpha \to 0$*, it holds that* $\lim_{\alpha\to 0}\lim_{t\to+\infty}\phi(\boldsymbol{W}_\alpha, t) = \overline{\boldsymbol{W}}_1$.

Assumption 5.7 is a natural assumption, since $f(\,\cdot\,)$ in most notable problems in matrix factorization is quadratic or polynomial (e.g., matrix sensing, matrix completion). In general, it is unlikely for a gradient-based optimization process to get stuck into saddle points (Lee et al., 2017; Panageas et al., 2019). Thus, we should expect to see in general that GLRL finds the rank-1 solution if the problem is feasible with rank-1 solution. This means at least for this subclass of problems, the implicit regularization of GD is rather unrelated with norm minimization. Below is a concrete example:

Example 5.9 (Counter example of Conjecture 1.1, Gunasekar et al. 2017)**.** Theorem 5.8 enables us to construct counterexamples of the *implicit nuclear norm regularization* conjecture in (Gunasekar et al., 2017). The idea is to construct a problem where every rank-1 stationary point of $\mathcal{L}(\boldsymbol{U})$ (i.e., $\nabla\mathcal{L}(\boldsymbol{U}) = \mathbf{0}$ and $\boldsymbol{U}$ is rank-1) is a global minimizer but none of them is minimizing the nuclear norm. Below we give a concrete matrix completion problem which meets the above requirement. Let the observations be $\Omega = \{(1,3),(1,4),(2,3),(3,1),(3,2),(4,1)\}$ and the ground-truth matrix be $\boldsymbol{M}$, where "?" and the numbers stand for the unobserved and observed respectively. Formally the optimization problem is $\mathcal{L}(\boldsymbol{U}) = \frac{1}{2}f(\boldsymbol{U}\boldsymbol{U}^\top), f(\boldsymbol{W}) = \frac{1}{2}\sum_{(i,j)\in\Omega}(W_{ij} - M_{ij})^2$.

$$\boldsymbol{M} = \begin{bmatrix} ? & ? & 1 & R \\ ? & ? & R & ? \\ 1 & R & ? & ? \\ R & ? & ? & ? \end{bmatrix}, \boldsymbol{M}_{\mathrm{norm}} = \begin{bmatrix} R & 1 & 1 & R \\ 1 & R & R & 1 \\ 1 & R & R & 1 \\ R & 1 & 1 & R \end{bmatrix}, \boldsymbol{M}_{\mathrm{rank}} = \begin{bmatrix} 1 & R & 1 & R \\ R & R^2 & R & R^2 \\ 1 & R & 1 & R \\ R & R^2 & R & R^2 \end{bmatrix}.$$

Here $R > 1$ is a large constant, e.g., $R = 100$. The minimum nuclear norm solution is the rank-2 matrix $\boldsymbol{M}_{\mathrm{norm}}$, which has $\|\boldsymbol{M}_{\mathrm{norm}}\|_* = 4R$ (which is 400 when $R = 100$). $\boldsymbol{M}_{\mathrm{rank}}$ is a rank-1 solution with much larger nuclear norm, $\|\boldsymbol{M}_{\mathrm{norm}}\|_* = 2R^2 + 2$ (which is 20002 when $R = 100$). We can verify that $f(\,\cdot\,)$ satisfies Assumptions 5.5 and 5.7 and $\boldsymbol{W}_1^{\mathrm{G}}(t)$ converges to the rank-1 solution $\boldsymbol{M}_{\mathrm{rank}}$. Therefore, GF with infinitesimal initialization converges to $\boldsymbol{M}_{\mathrm{rank}}$ rather than $\boldsymbol{M}_{\mathrm{norm}}$, which refutes the conjecture in (Gunasekar et al., 2017). See Appendix I for a formal statement.

5.3 Equivalence between GD and GLRL: General Case

Theorem 5.6 showed that for any *fixed* time t, the rank-1 GLRL trajectory approximates GF with small initialization, that is, $\boldsymbol{W}_1^{\mathrm{G}}(t) = \lim_{\alpha\to 0}\widehat{\boldsymbol{W}}_\alpha(t)$, where $\widehat{\boldsymbol{W}}_\alpha(t) := \phi(\boldsymbol{W}_\alpha, \frac{1}{2\mu_1}\log\frac{1}{\langle \boldsymbol{W}_\alpha, \boldsymbol{u}_1\boldsymbol{u}_1^\top\rangle} + t)$.

However, in general it does not hold for infinite time, i.e., $\boldsymbol{W}_1^{\mathrm{G}}(\infty) \neq \lim_{\alpha\to 0}\widehat{\boldsymbol{W}}_\alpha(\infty)$, unless the prerequisite in Theorem 5.8 is satisfied, i.e., $\overline{\boldsymbol{W}}_1 = \boldsymbol{W}_1^{\mathrm{G}}(\infty)$ is a local minimizer of f. This is because of the well-known result that GD only converges to local minimizers (Lee et al., 2016; 2017). We adapt Theorem 2 of Lee et al. (2017) (Theorem E.6) to the setting of GF (Theorem E.7), and it yields the following result (Theorem 5.10), whose proof is deferred to Appendix E.4.

Theorem 5.10. *Let* $f : \mathbb{R}^{d\times d} \to \mathbb{R}$ *be a convex* $\mathcal{C}^2$*-smooth function. (1). All stationary points of* $\mathcal{L} : \mathbb{R}^{d\times d} \to \mathbb{R}, \mathcal{L}(\boldsymbol{U}) = \frac{1}{2}f(\boldsymbol{U}\boldsymbol{U}^\top)$ *are either strict saddles or global minimizers; (2). For any random initialization, GF* (1) *converges to strict saddles of* $\mathcal{L}(\boldsymbol{U})$ *with probability* 0.

Therefore, for convex f such as matrix sensing and completion, suppose f has no rank-1 minimizer, then no matter how small α is, $\widehat{\boldsymbol{W}}_\alpha(\infty)$ (if exists) is a global minimizer and thus away from $\overline{\boldsymbol{W}}_1$, which is rank-1. In other words, $\boldsymbol{W}_1^{\mathrm{G}}(t)$ only describes the limiting trajectory of GF in the first phase, i.e., when GF goes from near $\mathbf{0}$ to near $\overline{\boldsymbol{W}}_1$. However, GF eventually escapes from $\overline{\boldsymbol{W}}_1$ after sufficiently long time (which depends on α) and this part is not described by $\boldsymbol{W}_1^{\mathrm{G}}(t)$.

To understand the trajectory of GF escaping from $\overline{\boldsymbol{W}}_1$, a priori, we need to know how GF approaches $\overline{\boldsymbol{W}}_1$. Interestingly, using a similar argument for Theorem 5.3, Theorem 5.11 shows that generically

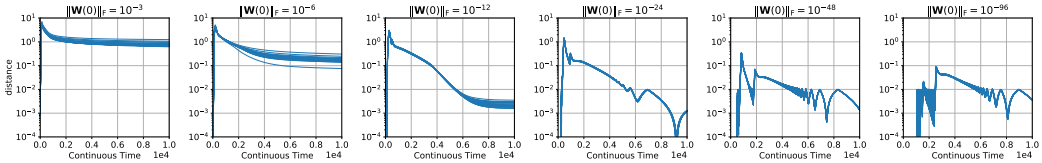

Figure 1: The trajectory of depth-2 GD, $\boldsymbol{W}_{\mathrm{GD}}(t)$, converges to GLRL's trajectory, $\boldsymbol{W}_{\mathrm{GLRL}}(t)$ as initialization scale goes to 0. We plotted the distance $\mathrm{dist}(t) = \min_{t' \in \mathcal{T}} \left\|\boldsymbol{W}_{\mathrm{GD}}(t) - \boldsymbol{W}_{\mathrm{GLRL}}(t')\right\|_{\mathrm{F}}$, where $\mathcal{T}$ is a carefully chosen small subset of $\mathbb{R}^+$ (for efficient computation), while 0.02-covers the entire trajectory of GLRL, that is, $\max_t \min_{t' \in \mathcal{T}} \left\|\boldsymbol{W}_{\mathrm{GLRL}}(t) - \boldsymbol{W}_{\mathrm{GLRL}}(t')\right\|_{\mathrm{F}} \leq 0.02$ ($\|\boldsymbol{W}^*\|_{\mathrm{F}} = 20$). The oscillations when $\mathrm{dist} \leq 0.02$ is probably caused by discretization. For each $\|\boldsymbol{W}(0)\|_{\mathrm{F}}$, we run 20 random seeds and plot them separately. The ground truth $\boldsymbol{W}^* \in \mathbb{R}^{20\times 20}$ is randomly generated and of rank-3. 30% entries are observed. See more inAppendix B.1.

GF only escapes from the direction of $\boldsymbol{v}_1\boldsymbol{v}_1^\top$, where $\boldsymbol{v}_1$ is the (unique) top eigenvector of $-\nabla f(\overline{\boldsymbol{W}}_1)$, and thus the limiting trajectory is exactly that of GLRL in phase 2 until GF getting close to $\overline{\boldsymbol{W}}_2$. If $\overline{\boldsymbol{W}}_2$ is still not a local minimizer of $f(\,\cdot\,)$ among all PSD matrices (but it is a local minimizer among rank-2 PSD matrices), then the above process continues and repeats, until for some k, $\overline{\boldsymbol{W}}_k$ is a local minimizer. Here by "generically" we hide some technical assumptions and they can be found in Appendix F. See Figure 1 and Figure 2 for experimental verification of the equivalence between GD and GLRL. We end this section with the following characterization of GF:

Theorem 5.11 (Theorem E.4, informal)**.** *Let $\overline{\boldsymbol{W}}$ be a critical point of (2) satisfying that for some $k \geq 0$, $\overline{\boldsymbol{W}}$ is a local minimizer among PSD matrices with rank $\leq k$ but $\overline{\boldsymbol{W}}$ is not a local minimizer among all PSD matrices. Let $-\nabla f(\overline{\boldsymbol{W}}) = \sum_{i=1}^d \mu_i \boldsymbol{v}_i\boldsymbol{v}_i^\top$ be the eigendecomposition of $-\nabla f(\overline{\boldsymbol{W}})$. Under some mild conditions on the eigenvalues of $-\nabla f$ and $\nabla^2\mathcal{L}$, if there is a time $T_\alpha \in (-\infty, +\infty)$ for every α so that $\phi(\boldsymbol{W}_\alpha, T_\alpha)$ converges to $\overline{\boldsymbol{W}}$ with positive alignment with $\boldsymbol{v}_1\boldsymbol{v}_1^\top$ as $\alpha \to 0$, then for every fixed t, $\lim_{\alpha\to 0} \phi(\boldsymbol{W}_\alpha, T_\alpha + \frac{1}{2\mu_1}\log\frac{1}{\langle \phi(\boldsymbol{W}_\alpha, T_\alpha), \boldsymbol{v}_1\boldsymbol{v}_1^\top\rangle} + t)$ exists and is equal to $\boldsymbol{W}^{\mathrm{G}}(t) := \lim_{\epsilon\to 0} \phi(\overline{\boldsymbol{W}} + \epsilon\boldsymbol{v}_1\boldsymbol{v}_1^\top, \frac{1}{2\mu_1}\log\frac{1}{\epsilon} + t)$.*

Characterization of the trajectory of GF. Generically, the trajectory of GF with small initialization can be split into K phases by $K+1$ stationary points $\{\overline{\boldsymbol{W}}_i\}_{i=0}^K$ ($\overline{\boldsymbol{W}}_0 = \boldsymbol{0}$), where in phase i GF escapes from $\overline{\boldsymbol{W}}_{i-1}$ in the direction of the top principal component of $-\nabla f(\overline{\boldsymbol{W}}_{i-1})$ and gets close to $\overline{\boldsymbol{W}}_i$. Here, $\{\overline{\boldsymbol{W}}_i\}_{i=0}^{K-1}$ corresponds to K strict saddles of $\mathcal{L}(\,\cdot\,)$ and $\overline{\boldsymbol{W}}_K$ is the global minimizer. The smaller the initialization is, the longer GF stays around each saddle. Moreover, $\{\overline{\boldsymbol{W}}_i\}_{i=0}^K$ could be computed via evaluating $\{\overline{\boldsymbol{W}}_{i,\epsilon}\}_{i=0}^K$ in GLRL (Definition 5.1) with sufficiently small ϵ.

6 Benefits of Depth: A View from GLRL

In this section, we consider matrix factorization problems with depth $L \geq 3$. Our goal is to understand the effect of the depth-L parametrization $\boldsymbol{W} = \boldsymbol{U}_1\boldsymbol{U}_2\cdots\boldsymbol{U}_L$ on the implicit bias — how does depth encourage GF to find low rank solutions? We take the standard assumption in existing analysis for the end-to-end dynamics that the weight matrices have a balanced initialization, i.e. $\boldsymbol{U}_i^\top(0)\boldsymbol{U}_i(0) = \boldsymbol{U}_{i+1}(0)\boldsymbol{U}_{i+1}^\top(0)$, $\forall 1 \leq i \leq L-1$. Arora et al. (2018) showed that if $\{\boldsymbol{U}_i\}_{i=1}^L$ is balanced at initialization, then we have the following end-to-end dynamics. Similar to the depth-2 case, we use $\phi(\boldsymbol{W}(0), t)$ to denote $\boldsymbol{W}(t)$, where

$$\frac{d\boldsymbol{W}}{dt} = -\sum_{i=0}^{L-1} (\boldsymbol{W}\boldsymbol{W}^\top)^{\frac{i}{L}} \nabla f(\boldsymbol{W}) (\boldsymbol{W}^\top\boldsymbol{W})^{1-\frac{i+1}{L}}. \tag{9}$$

The lemma below is the foundation of our analysis for the deep case, which greatly simplifies Equation (9). Due to the space limit, we defer its derivations and applications into appendix.

Lemma 6.1. *If $\boldsymbol{W}(t)$ is a symmetric solution of (9), then for $\boldsymbol{M}(t) := \boldsymbol{W}(t)^{2/L}$, we have*

$$\frac{\mathrm{d}\boldsymbol{M}}{\mathrm{d}t} = -\nabla f(\boldsymbol{M}^{L/2})\boldsymbol{M}^{L/2} - \boldsymbol{M}^{L/2}\nabla f(\boldsymbol{M}^{L/2}). \tag{10}$$

Our main result, Theorem 6.2, gives a characterization of the limiting trajectory for deep matrix factorization with infinitesimal identity initialization. Here $\overline{\boldsymbol{W}}(t) := \lim_{\alpha\to 0} \boldsymbol{W}_\alpha^{\mathrm{G}}(t)$ is the trajectory of deep GLRL , where $\boldsymbol{W}_\alpha^{\mathrm{G}}(t) := \phi(\alpha\boldsymbol{e}_1\boldsymbol{e}_1^\top, \frac{\alpha^{-(1-\frac{1}{P})}}{2\mu_1(P-1)} + t)$ (See Algorithm 2). The dynamics for general initialization is more complicated. Please see discussions in Appendices B.7 and B.8.

Theorem 6.2. *Let $P = \frac{L}{2}$, $L \geq 3$. Suppose $\lambda_1(\nabla f(\boldsymbol{0})) > \lambda_2(\nabla f(\boldsymbol{0}))$, then $\forall t \in \mathbb{R}$, we have:*

$$\left\|\phi\left(\alpha\boldsymbol{I}, \frac{\alpha^{-(1-1/P)}}{2\mu_1(P-1)} + t\right) - \overline{\boldsymbol{W}}(t)\right\|_2 = O(\alpha^{\frac{1}{P(P+1)}}), \tag{11}$$

and for any $2 \leq k \leq d$,

$$\lambda_k\left(\phi\left(\alpha\boldsymbol{I}, \frac{\alpha^{-(1-1/P)}}{2\mu_1(P-1)} + t\right)\right) = O(\alpha). \tag{12}$$

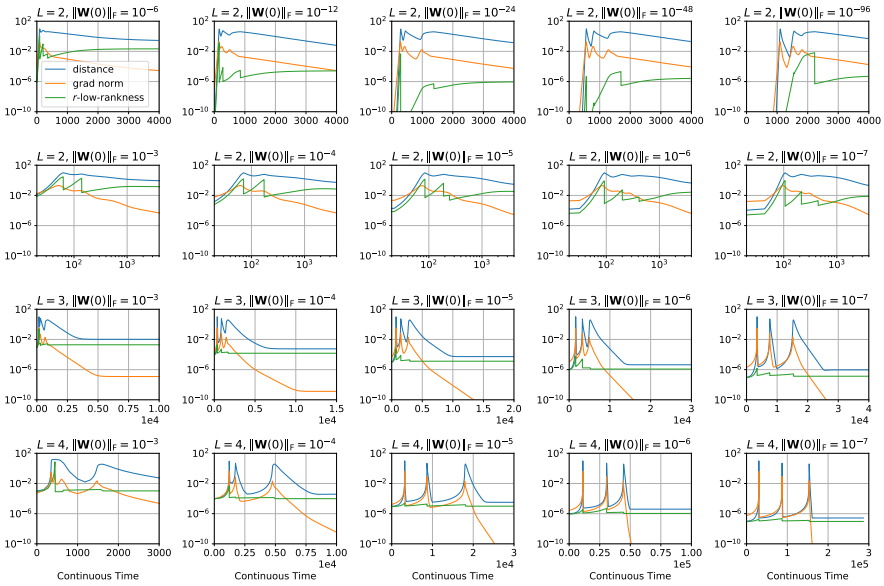

Figure 2: GF passes through the same set of saddles as GLRL when initialization is small, and gets much closer to the saddles when $L \geq 3$. Depth-2 GF requires a much smaller initialization scale to maintain small low-rankness. Here the ground truth matrix $\boldsymbol{W}_*$ has size of 20×20 and rank of 3 as stated in Appendix B.1. Thus GLRL has 3 phases with saddles $\{\overline{\boldsymbol{W}}_i\}_{i=0}^3$, where $\overline{\boldsymbol{W}}_0 = \mathbf{0}$ and $\overline{\boldsymbol{W}}_3 = \boldsymbol{W}^*$. For each setting, we plot the distance between the current step of GD and the closest saddle of GLRL, $\|\boldsymbol{W}_{GD}(t) - \boldsymbol{W}_r^*\|_F$, the F-norm of full gradient $\left\|\nabla_{\boldsymbol{U}_{1:L}} \mathcal{L}(\boldsymbol{U}_{1:L})\right\|_F$ and r-low-rankness of $\boldsymbol{W}_{GD}(t)$, where $r := \arg\min_{0 \leq i \leq 3} \|\boldsymbol{W}_{GD}(t) - \boldsymbol{W}_i^*\|_F$.

Definition 6.3 (r-low-rankness)**.** For matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$, we define the *$r$-low-rankness* of $\boldsymbol{M}$ as $\sqrt{\sum_{i=r+1}^{d} \sigma_i^2(\boldsymbol{M})}$, where $\sigma_i(\boldsymbol{M})$ is the i-th largest singular value of $\boldsymbol{M}$.

So how does depth encourage GF to find low-rank solution? When the ground truth is low rank, say rank-k, our experiments (Figure 2) suggest that deep matrix sensing finds solutions with smaller k-low-rankness, thus achieving better generalization. Intuitively, this is contradictory to what Theorem 6.2 suggests, i.e. , the convergence rate of deep GLRL at constant time gets slower as the depth increases. However, it turns out the uniform upper bound for distance between GF and GLRL is not the ideal metric for the eventual k-low-rankness of learnt solution. Below we will illustrate why *the r-low-rankness of GF within each phase r* is a better metric and how they could be different.

Suppose f admits a unique local minimizer $\boldsymbol{W}_0$ among rank-1 PSD matrices, and we run GF from $\alpha \boldsymbol{I}$ for both depth-2 and depth-L cases. Intuitively, the 1-low-rankness of the depth-2 solution is $O(\alpha^{1-\frac{\mu_2}{\mu_1}})$, which can be seen from the second warmup example in Section 4. For the depth-L solution, though it may diverge from the trajectory of deep GLRL more than the depth-2 solution does, its 1-low-rankness is only $O(\alpha)$. In Theorem 6.4, we show that around $\boldsymbol{W}_0$ there is a basin of rank-1 matrices where GF converges to $\boldsymbol{W}_0$ geometrically. Moreover, starting from a full rank matrix $O(\alpha)$ close to the basin, GF still can converge to a solution $O(\alpha)$ close to $\boldsymbol{W}_0$.

Theorem 6.4. *Under the same settings as Theorem 6.2, suppose $\overline{\boldsymbol{W}}(\infty)$ exists, we have*

$$\inf_{t \in \mathbb{R}} \left\|\phi(\alpha \boldsymbol{I}, t) - \overline{\boldsymbol{W}}(\infty)\right\|_F = O(\alpha). \tag{13}$$

Interpretation for the advantage of depth with multiple phases. For depth-2 GLRL, the low-rankness is raised to some power smaller than 1 per phase, which depends on the eigenvalue. For deep GLRL, we show the low-rankness is only multiplied by some constant for the first phase (note $\mathbf{0}$ is also a stationary point) and speculate it to be true for later phases. This conjecture is supported by our experiments, see Figure 2. Interestingly, our theory and experiments (Figure 6) suggest that while **being deep is good for generalization, being deeper may not be better**, because the growth of low rankness is constantly multiplicative for the first phase for $L \geq 3$. While the above theoretic result is only for identity initialization, Theorem B.1 shows the dynamics of GF (9) converges point-wise when $L \to \infty$ with a suitable time rescaling. (See Figure 4 for an experimental verification)

7 Conclusion

In this work, we connect gradient descent to GLRL to explain the success of using gradient descent to find low-rank solutions in the matrix factorization problem. This enables us to construct counterexamples to the implicit nuclear norm conjecture in Gunasekar et al. (2017). Taking the view of GLRL can also help us understand the benefits of depth.

REFERENCES

Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

Sanjeev Arora, N Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *35th International Conference on Machine Learning*, 2018.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7411–7422. Curran Associates, Inc., 2019a.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8139–8148. Curran Associates, Inc., 2019b.

Mohamed Ali Belabbas. On implicit regularization: Morse functions and applications to matrix factorization. *arXiv preprint arXiv:2001.04264*, 2020.

Jeff Bezanson, Stefan Karpinski, Viral B Shah, and Alan Edelman. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*, 2012.

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

Lénaïc Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. volume 125 of *Proceedings of Machine Learning Research*, pages 1305–1338. PMLR, 09–12 Jul 2020.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 2937–2947. Curran Associates, Inc., 2019.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1329–1338, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3196–3206. Curran Associates, Inc., 2019.

Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. In *International Conference on Learning Representations*, 2020.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6151–6159. Curran Associates, Inc., 2017.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9482–9491. Curran Associates, Inc., 2018.

Benjamin D. Haeffele and René Vidal. Structured Low-Rank Matrix Factorization: Global Optimality, Algorithms, and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 42(6):1468–1482, 2019.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Overview of mini-batch gradient descent. URL https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.

Jean-Baptiste Hiriart-Urruty and A. S. Lewis. The clarke and michel-penot subdifferentials of the eigenvalues of a symmetric matrix. *Comput. Optim. Appl.*, 13(1-3):13–23, 1999. doi: 10.1023/A:1008644520093.

Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8571–8580. Curran Associates, Inc., 2018.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019a.

Ziwei Ji and Matus Telgarsky. A refined primal-dual analysis of the implicit bias. *arXiv preprint arXiv:1906.04540*, 2019b.

Rajiv Khanna, Ethan Elenberg, Alexandros G Dimakis, and Sahand Negahban. On approximation guarantees for greedy low rank optimization. *arXiv preprint arXiv:1703.02721*, 2017.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 2–47. PMLR, 06–09 Jul 2018.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Mor Shpigel Nacson, Suriya Gunasekar, Jason Lee, Nathan Srebro, and Daniel Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4683–4692, Long Beach, California, USA, 09–15 Jun 2019a. PMLR.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 3420–3428. PMLR, 16–18 Apr 2019b.

Mor Shpigel Nacson, Nathan Srebro, and Daniel Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 3051–3059. PMLR, 16–18 Apr 2019c.

Ioannis Panageas, Georgios Piliouras, and Xiao Wang. First-order methods almost always avoid saddle points: The case of vanishing step-sizes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6474–6483. Curran Associates, Inc., 2019.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Lawrence Perko. *Differential equations and dynamical systems*, volume 7. Springer Science & Business Media, 2013.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *arXiv preprint arXiv:2005.06398*, 2020.

Shai Shalev-Shwartz and Yoram Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. *Machine learning*, 80(2-3):141–163, 2010.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018a.

Daniel Soudry, Elad Hoffer, and Nathan Srebro. The implicit bias of gradient descent on separable data. In *International Conference on Learning Representations*, 2018b.

Zheng Wang, Ming-Jun Lai, Zhaosong Lu, Wei Fan, Hasan Davulcu, and Jieping Ye. Rank-one matrix pursuit for matrix completion. In *International Conference on Machine Learning*, pages 91–99, 2014.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4148–4158. Curran Associates, Inc., 2017.

Quanming Yao and James Tin Yau Kwok. Greedy learning of generalized low-rank models. In *IJCAI International Joint Conference on Artificial Intelligence*, 2016.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

Stanisław Łojasiewicz. Ensembles semi-analytiques. *IHES notes*, 1965.

A DEEP GREEDY LOW-RANK LEARNING (DEEP GLRL) ALGORITHM

Algorithm 2: Deep Greedy Low-Rank Learning (Deep GLRL)

parameter : step size $\eta > 0$; small $\epsilon > 0$
$\epsilon' \leftarrow \epsilon^{1/L}$, $\mathcal{L}(\boldsymbol{U}_1, \cdots, \boldsymbol{U}_L) := f(\boldsymbol{W}_1 \cdots \boldsymbol{W}_L)$.
$\boldsymbol{W}_0 \leftarrow \boldsymbol{0} \in \mathbb{R}^{d\times d}$, and $\boldsymbol{U}_{0,1}(\infty), \ldots, \boldsymbol{U}_{0,L}(\infty) \in \mathbb{R}^{d\times 0}$ are empty matrices
for $r = 1, 2, 3, \ldots$ **do**
  let $\boldsymbol{u}_r$ be a top (unit) eigenvector of $-\nabla f(\boldsymbol{W}_{r-1})$
  $\boldsymbol{U}_{r,1}(0) \leftarrow [\boldsymbol{U}_{r-1,1}(\infty) \quad \epsilon' \boldsymbol{u}_r] \in \mathbb{R}^{d\times r}$
  $\boldsymbol{U}_{r,k}(0) \leftarrow \begin{bmatrix} \boldsymbol{U}_{r-1,k}(\infty) & \boldsymbol{0} \\ \boldsymbol{0} & \epsilon' \end{bmatrix} \in \mathbb{R}^{r\times r}$ for all $2 \le k \le L-1$
  $\boldsymbol{U}_{r,L}(0) \leftarrow \begin{bmatrix} \boldsymbol{U}_{r-1,L}(\infty) \\ \epsilon' \boldsymbol{u}_r^\top \end{bmatrix} \in \mathbb{R}^{r\times d}$
  for $t = 0, 1, \ldots$ **do**
    $\boldsymbol{U}_{r,i}(t+1) \leftarrow \boldsymbol{U}_{r,i}(t) - \eta \nabla_{\boldsymbol{U}_i} \mathcal{L}(\boldsymbol{U}_{r,1}(t), \cdots, \boldsymbol{U}_{r,L}(t)), \forall 1 \le i \le L.$
  $\boldsymbol{W}_r \leftarrow \boldsymbol{U}_{r,1}(\infty) \cdots \boldsymbol{U}_{r,L}(\infty)$
  if $\lambda_1(-\nabla f(\boldsymbol{W}_r)) \le 0$ **then**
    return $\boldsymbol{W}_r$

B EXPERIMENTS

B.1 GENERAL SETUP

The ground-truth matrix $\boldsymbol{W}_*$ is low-rank by construction: We sample a random orthogonal matrix $\boldsymbol{U}$, a diagonal matrix $\boldsymbol{S}$ with Frobenius norm $\|\boldsymbol{S}\|_{\mathrm{F}} = 1$ and set $\boldsymbol{W}_* = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top$. Each measurement $\boldsymbol{X}$ in $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ is generated by sampling two one-hot vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ uniformly and $\boldsymbol{X} = \frac{1}{2}\boldsymbol{u}\boldsymbol{v}^\top + \frac{1}{2}\boldsymbol{v}\boldsymbol{u}^\top$.

In Figures 1, 2, 3 and 5 to 7, the ground truth matrix $\boldsymbol{W}_*$ has shape of 20×20 and rank of 3, where $\lambda_1(\boldsymbol{W}^*) = 17.41$, $\lambda_2(\boldsymbol{W}^*) = 8.85$, $\lambda_4(\boldsymbol{W}^*) = 4.31$ $p = 0.3$ and $\lambda_1(-\nabla f(\boldsymbol{0})) = 6.23, \lambda_2(-\nabla f(\boldsymbol{0})) = 5.41$. $p = 0.3$ is used for generating measurements, except $p = 0.25$ in Figure 3.

Initialization To enforce balanceness of gradient descent's trajectory, we initialize the matrices as follows: Given expected Frobenius norm $\tilde{\epsilon} = \|\boldsymbol{W}\|_{\mathrm{F}}$, we sample L orthogonal matrices $\boldsymbol{U}_{1,\ldots,L}$, and sample a diagonal matrix $\boldsymbol{D}$ with expected Frobenius norm $\tilde{\epsilon}$, then gradient descent's initial parameters are $\boldsymbol{W}_i = \boldsymbol{U}_i \boldsymbol{D}^{1/L} \boldsymbol{U}_{i+1}^\top$ where $\boldsymbol{U}_{L+1} = \boldsymbol{U}_1$. In this way we can guarantee that the end-to-end matrix $\boldsymbol{W} = \boldsymbol{W}_1 \cdots \boldsymbol{W}_L$ is symmetric and the initialization is balanced. To make a fair comparison with gradient descent, GLRL's initialization scale ϵ is given by $\epsilon = \tilde{\epsilon}^{\frac{1}{L}} d^{-\frac{L+1}{2L}}$, so that the initialized parameters have the same scale when depth is the same. Note that GLRL is deterministic and GD depends on only the initialization.

Gradient Descent We discretize the time to simulate gradient flow. As gradient flow gets stuck at a saddle point for most of the time, we use RMSprop (Hinton et al.) with small learning rate for faster convergence. We remark that unlike many implementations (e.g., in PyTorch), our learning rate (discretized time interval) is the same for all parameters. We're doing full gradient descent instead of stochastic gradient descent.

GLRL In Figures 1, 2, 3 and 6, the GLRL's trajectory is obtained by running Algorithm 1 with $\epsilon = 10^{-4}$ and $\eta = 10^{-3}$. The stopping criterion is that if the loop has been iterated for 10^7 times.

The code is written in Julia (Bezanson et al., 2012), PyTorch (Paszke et al., 2019).

B.2 EXPERIMENTAL EQUIVALENCE BETWEEN GLRL AND GRADIENT DESCENT

Here we provide experimental evidence supporting our theoretical claims about the equivalence between GLRL and GF for both cases, $L = 2$ and $L \geq 3$.

In Figure 1, we show the distance from every point on GF (simulated by GD) from random initialization is close to the trajectory of GLRL. The distance for deep case are much smaller than that of depth-2 case.

In Figure 2, we first run GLRL and obtain its saddle points. We also define the distance of a matrix $\boldsymbol{W}$ to saddles passed by GLRL to be $\min_{0 \leq i \leq 3} \left\| \boldsymbol{W} - \overline{\boldsymbol{W}}_i \right\|_{\mathrm{F}}$, where $\{\overline{\boldsymbol{W}}_i\}_{i=0}^{3}$ are the saddles.

B.3 HOW WELL DOES GLRL WORK?

We compare GLRL with gradient descent (with not-so-small initialization), nuclear norm minimization and R1MP (Wang et al., 2014). We use CVXPY (Diamond and Boyd, 2016; Agrawal et al., 2018) for finding the nuclear norm solution. The results are shown in Figure 3. GLRL can fully recover the ground truth, while others have difficulty doing so.

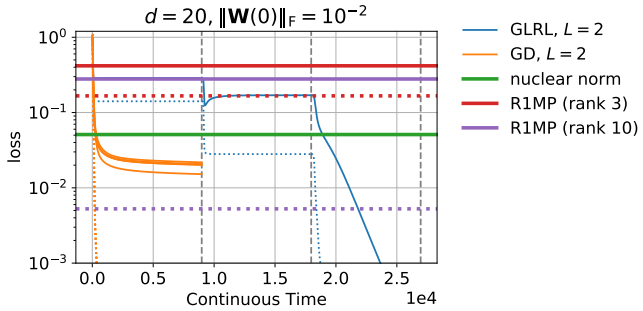

Figure 3: GD with small initialization outperforms R1MP and minimal nuclear norm solution on synthetic data with low-rank ground truth. Solid (dotted) curves correspond to test (training) loss. Here the loss $\ell(\boldsymbol{W}) := \frac{1}{d^2} \left\| \boldsymbol{W} - \boldsymbol{W}^* \right\|_{\mathrm{F}}^2$ and $\ell(\mathbf{0}) = 1$. We run 10 random seeds for GD and plot them separately (most of them overlap). More details in Appendix B.1.

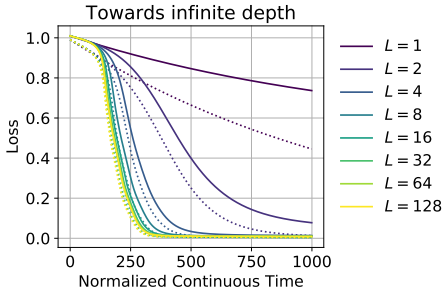

Figure 4: The marginal value of being deeper. The trajectory of GD converges when depth goes to infinity. Solid (dotted) curves correspond to test (train) loss. The x-axis stands for the normalized continuous time t (multiplied by L). More details in Appendix B.6.

B.4 HOW DOES INITIALIZATION AFFECT THE CONVERGENCE RATE TO THE RANK-1 GLRL TRAJECTORY?

We use the general setting in Appendix B.1. In these experiments, we use constant learning rate of 10^{-5}. The reference matrix $\boldsymbol{W}_{\text{ref}}$ is obtained by running the first stage of GLRL with $\|\boldsymbol{W}(0)\|_{\mathrm{F}} = 10^{-48}$ and we pick one matrix in the trajectory with $\|\boldsymbol{W}_{\text{ref}}\|_{\mathrm{F}}$ about 0.6.

For every $\epsilon = 10^i, i \in \{-1, -2, -3, -4, -5\}$, we run both gradient descent and the first stage of GLRL with $\|\boldsymbol{W}(0)\|_{\mathrm{F}} = \epsilon$. For gradient descent, we use random initialization so $\|\boldsymbol{W}(0)\|_{\mathrm{F}}$ is full rank w.p. 1. The distance of a trajectory to $\boldsymbol{W}_{\text{ref}}$ is defined as $\min_{t \geq 0} \|\boldsymbol{W}(t) - \boldsymbol{W}_{\text{ref}}\|_{\mathrm{F}}$. In practice, as we discretized time to simulate gradient flow, we check every t during simulation to compute the distance. As a result, the estimation might be inaccurate when a trajectory is really close to $\boldsymbol{W}_{\text{ref}}$.

The result is shown at Figure 5. We observe that GLRL trajectories are closer to the reference matrix $\boldsymbol{W}_{\text{ref}}$ by magnitudes. Thus the take home message here is that GLRL is in general a more computational efficient method to simulate the trajectory of GF (GD) with infinitesimal initialization, as one can start GLRL with a much larger initialization, while still maintaining high precision.

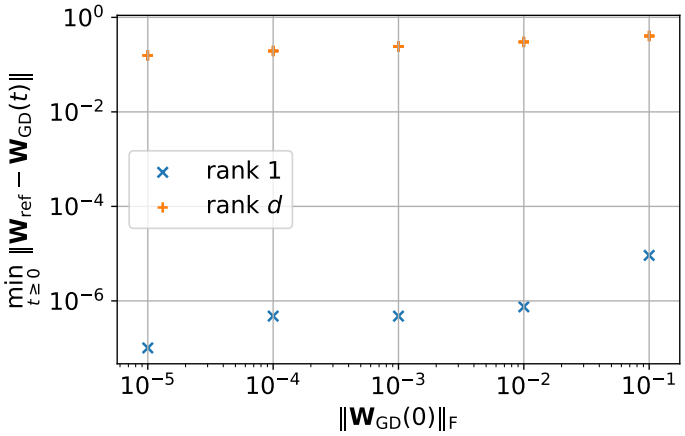

Figure 5: Using $\epsilon \boldsymbol{v}_1 \boldsymbol{v}_1^\top$ (denoted by 'rank 1') as initialization makes GD much closer to GLRL compared to using random initialization (denoted by 'rank d'), where $\boldsymbol{v}_1$ is the top eigenvector of $-\nabla f(\mathbf{0})$. We take a fixed reference matrix on the trajectory of GLRL with constant norm and plot the distance of GD with each initialization to it respectively. See more details at Appendix B.4.

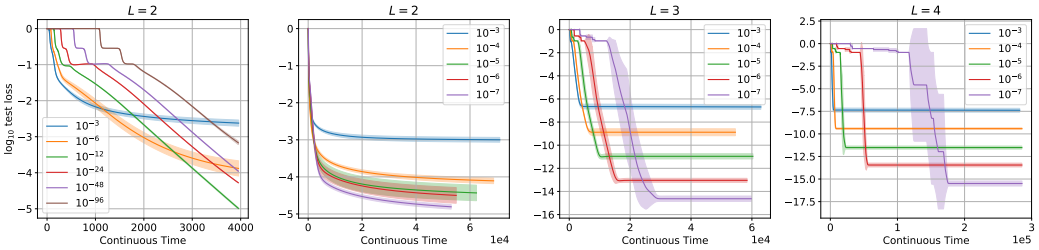

Figure 6: Deep matrix factorization encourages GF to find low rank solutions at a much practical initialization scale, e.g. 10^{-3}. Here the ground truth is rank-3. For each setting, we run 5 different random seeds. The shaded area indicates one standard deviation. We can see that performance of GD is quite robust to its initialization.

B.5 Benefit of Depth: polynomial vs exponential dependence on initialization

To verify the our theory in Section 6 , we run gradient descent with different depth and initialization. The results are shown in Figure 6. We can see that as the initialization becomes smaller and smaller, the trajectories of all depth converge to GLRL. However, a depth-2 model requires exponentially small initialization, while deeper models require polynomial small initialization, though it takes much longer to converge.

B.6 The marginal value of being deeper.

Theorem B.1 shows that the end-to-end dynamics (Equation (14)) converges point-wise while $L \to \infty$ if the product of learning rate and depth, ηL, is fixed as constant. Interestingly, Equation (14) also allows us to simulate the dynamics of $\boldsymbol{W}(t)$ for all depths L while the computation time is independent of L. In Figure 4, we compare the effect of depth while fixing the initialization and ηL. We can see that deeper models converge faster. The difference between $L = 1, 2$, and 4 is large, while difference among $L \geq 16$ is marginal.

Theorem B.1. *Suppose $\boldsymbol{W} = \tilde{\boldsymbol{U}}\tilde{\Sigma}\tilde{\boldsymbol{V}}^\top$ is the SVD decomposition of $\boldsymbol{W}$ where $\tilde{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_d)$. The dynamics of L-layer linear net is the following, $\circ$ denotes the entry-wise multiplication:*

$$\frac{\mathrm{d}\boldsymbol{W}}{\mathrm{d}t} = -L\tilde{\boldsymbol{U}}\left(\left(\tilde{\boldsymbol{U}}^\top \nabla f(\boldsymbol{W})\tilde{\boldsymbol{V}}\right) \circ \boldsymbol{K}^{(L)}\right)\tilde{\boldsymbol{V}}^\top, \tag{14}$$

where $K_{i,i}^{(L)} = \sigma_i^{2-2/L}$, $K_{i,j}^{(L)} = \frac{\sigma_i^2 - \sigma_j^2}{L\sigma_i^{2/L} - L\sigma_j^{2/L}}$ *for* $i \neq j$.

Proof of Theorem B.1. We start from Equation (9):

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{W}}{\mathrm{d}t} &= -\sum_{l=0}^{L-1} (\boldsymbol{W}\boldsymbol{W}^\top)^{\frac{l}{L}} \nabla f(\boldsymbol{W}) (\boldsymbol{W}^\top \boldsymbol{W})^{\frac{L-1-l}{L}} \\ &= -\sum_{l=0}^{L-1} \tilde{\boldsymbol{U}} \tilde{\Sigma}^{\frac{2l}{L}} \tilde{\boldsymbol{U}}^\top \nabla f(\boldsymbol{W}) \tilde{\boldsymbol{V}} \tilde{\Sigma}^{\frac{2(L-1-l)}{L}} \tilde{\boldsymbol{V}} \\ &= -L\tilde{\boldsymbol{U}} \left[L^{-1} \sum_{l=0}^{L-1} \tilde{\Sigma}^{\frac{2l}{L}} (\tilde{\boldsymbol{U}}^\top \nabla f(\boldsymbol{W}) \tilde{\boldsymbol{V}}) \tilde{\Sigma}^{\frac{2(L-1-l)}{L}} \right] \tilde{\boldsymbol{V}}.\end{aligned}$$

Note that $\tilde{\Sigma}$ is diagonal, so

$$\tilde{\Sigma}^{\frac{2l}{L}} (\tilde{\boldsymbol{U}}^\top \nabla f(\boldsymbol{W}) \tilde{\boldsymbol{V}}) \tilde{\Sigma}^{\frac{2(L-1-l)}{L}} = (\tilde{\boldsymbol{U}}^\top \nabla f(\boldsymbol{W}) \tilde{\boldsymbol{V}}) \circ \boldsymbol{H}^{(l)},$$

where $\boldsymbol{H}_{i,j}^{(l)} = \sigma_i^{\frac{2l}{L}} \sigma_j^{\frac{2(L-1-l)}{L}}$. Therefore

$$\begin{aligned}L^{-1} \sum_{l=0}^{L-1} \tilde{\Sigma}^{\frac{2l}{L}} (\tilde{\boldsymbol{U}}^\top \nabla f(\boldsymbol{W}) \tilde{\boldsymbol{V}}) \tilde{\Sigma}^{\frac{2(L-1-l)}{L}} &= L^{-1} \sum_{l=0}^{L-1} (\tilde{\boldsymbol{U}}^\top \nabla f(\boldsymbol{W}) \tilde{\boldsymbol{V}}) \circ \boldsymbol{H}^{(l)} \\ &= (\tilde{\boldsymbol{U}}^\top \nabla f(\boldsymbol{W}) \tilde{\boldsymbol{V}}) \circ \boldsymbol{K}^{(L)},\end{aligned}$$

where $\boldsymbol{K}^{(L)} = L^{-1} \sum_{l=0}^{L-1} \boldsymbol{H}^{(l)}$. Hence

$$\frac{\mathrm{d}\boldsymbol{W}}{\mathrm{d}t} = -L\tilde{\boldsymbol{U}} \left[(\tilde{\boldsymbol{U}}^\top \nabla f(\boldsymbol{W}) \tilde{\boldsymbol{V}}) \circ \boldsymbol{K}^{(L)} \right] \tilde{\boldsymbol{V}}.$$

The entries of $\boldsymbol{K}^{(L)}$ can be directly calculated by

$$K_{i,j}^{(L)} = L^{-1} \sum_{l=0}^{L-1} \sigma_i^{\frac{2l}{L}} \sigma_j^{\frac{2(L-1-l)}{L}} = \begin{cases} \sigma_i^{2-2/L}, & i = j, \\ \frac{\sigma_i^2 - \sigma_j^2}{L\sigma_i^{2/L} - L\sigma_j^{2/L}}, & i \neq j. \end{cases}$$

□

Corollary B.2. *As* $L \to \infty$, $\boldsymbol{K}^{(L)}$ *converges to* $\boldsymbol{K}^*$, *where* $K_{i,i}^* = \sigma_i^2$, $K_{i,j}^* = \frac{\sigma_i^2 - \sigma_j^2}{\ln \sigma_i^2 - \ln \sigma_j^2}$ *for* $i \neq j$.

Experiment details We follow the general setting in Appendix B.1. The ground truth $\boldsymbol{W}^*$ is different but is generated in the same manner and has the same shape of 20×20 and $p = 0.3$ is used for observation generation. We directly apply Equation (14), in which we compute $\tilde{\boldsymbol{V}}$ and $\tilde{\boldsymbol{U}}$ through SVD, to simulate the trajectory together with a constant learning rate of $\frac{10^{-3}}{L}$ for depth L. $\boldsymbol{W}(0)$ is sampled from $10^{-3} \times \mathcal{N}(0, \boldsymbol{I}^{d \times d})$.

B.7 Escaping direction for deep matrix factorization

Recall that we only prove that the identity matrix will escape in the direction of the top eigenvector with infinitesimal initialization. The main burden for us to generalize this proof to general initialization is that we don't know how to analyze the early phase dynamics of Equation (10), i.e., the analytical solution of Equation (15) is difficult to compute, when $L \geq 3$. Intuitionally, the direction that the infinitesimal initialization escapes $\mathbf{0}$ is exactly $\widehat{\boldsymbol{M}}(\infty) := \lim_{t \to \infty} \frac{\boldsymbol{M}(t)}{\|\boldsymbol{M}(t)\|_{\mathrm{F}}}$, where $\boldsymbol{M}(t)$ is the solution of Equation (15). Showing $\widehat{\boldsymbol{M}}(\infty) = \boldsymbol{v}_1 \boldsymbol{v}_1^\top$ is a critical step in our analysis towards convergence to GLRL.

$$\frac{\mathrm{d}\boldsymbol{M}}{\mathrm{d}t} = -\nabla f(\mathbf{0}) \boldsymbol{M}^{L/2} - \boldsymbol{M}^{L/2} \nabla f(\mathbf{0}). \tag{15}$$

However, unlike the depth-2 case, $\widehat{\boldsymbol{M}}(\infty)$ can be different from $\boldsymbol{v}_1\boldsymbol{v}_1^\top$ even if $\boldsymbol{v}_1^\top\boldsymbol{M}(0)\boldsymbol{v}_1 > 0$. We here give an example for diagonal $\boldsymbol{M}(0)$ and $\nabla f(\mathbf{0})$ at Appendix B.8. Nevertheless, we still conjecture that except for a zero measure set of $\boldsymbol{M}(0)$, $\widehat{\boldsymbol{M}}(\infty) = \boldsymbol{v}_1\boldsymbol{v}_1^\top$, based on the following theoretical and experimental evidences:

- If $\boldsymbol{v}_1^\top\boldsymbol{M}(0)\boldsymbol{v}_1 > 0$ and $\mathrm{rank}(\boldsymbol{M}(0)) = 1$, we prove that $\widehat{\boldsymbol{M}}(\infty) = \boldsymbol{v}_1\boldsymbol{v}_1^\top$. (See Theorem B.3)
- For the counter-example, we show experimentally, even with perturbation of only magnitude 10^{-5}, $\widehat{\boldsymbol{M}}(\infty) = \boldsymbol{v}_1\boldsymbol{v}_1^\top$. The results are shown at Figure 7. The y-axis indicates $\langle \boldsymbol{v}_1, \boldsymbol{u}_1(t)\rangle$ where $\boldsymbol{u}_1(t)$ is the top eigenvector of $\boldsymbol{M}(t)$. As $\|\boldsymbol{W}(t)\|_{\mathrm{F}}$ becomes larger, $\boldsymbol{u}_1(t)$ aligns better with $\boldsymbol{v}_1$, which means the noise helps $\boldsymbol{M}$ escaping from $\boldsymbol{v}_1$. The larger the noise is, the faster $\boldsymbol{u}_1(t)$ converges to $\boldsymbol{v}_1$.

Theorem B.3 (rank-1 initialization escapes along the top eigenvector)**.** *When* $\mathrm{rank}(\boldsymbol{M}(0)) = 1$, $\lim_{t\to\infty} \frac{\boldsymbol{M}(t)}{\|\boldsymbol{M}(t)\|_F} = \boldsymbol{v}_1\boldsymbol{v}_1^\top$, *if* $\boldsymbol{v}_1^\top\boldsymbol{M}(0)\boldsymbol{v}_1 > 0$.

Proof of Theorem B.3. Let $\boldsymbol{M}(0) = \boldsymbol{v}_1\boldsymbol{v}_1^\top$, and $\boldsymbol{v}(t) \in \mathbb{R}^d$ be the solution of

$$\frac{\mathrm{d}\boldsymbol{v}(t)}{\mathrm{d}t} = \|\boldsymbol{v}(t)\|_2^{L-2}\,\nabla f(\mathbf{0})\boldsymbol{v}(t), \qquad \boldsymbol{v}(0) = \boldsymbol{v}_1. \tag{16}$$

It's easy to check that $\boldsymbol{M}(t) = \boldsymbol{v}(t)\boldsymbol{v}^\top(t)$ is the solution of Equation (15), because

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{M}}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}\boldsymbol{v}^\top + \boldsymbol{v}\frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t}^\top &= -\nabla f(\mathbf{0})\boldsymbol{M}(t)\,\|\boldsymbol{v}(t)\|_2^{L-2} - \boldsymbol{M}(t)\nabla f(\mathbf{0})\,\|\boldsymbol{v}(t)\|_2^{L-2}\\ &= -\nabla f(\mathbf{0})\boldsymbol{M}^{L/2} - \boldsymbol{M}^{L/2}\nabla f(\mathbf{0}).\end{aligned}$$

Let $\tau : \mathbb{R} \to \mathbb{R}$, such that $\tau(0) = 0$, and $\frac{\mathrm{d}\tau(t)}{\mathrm{d}t} = \|\boldsymbol{v}(t)\|_2^{L-2}$, we have

$$\mathrm{d}\boldsymbol{v} = -\nabla f(\mathbf{0})\boldsymbol{v}\,\|\boldsymbol{v}\|^{L-2}\,\mathrm{d}t = -\nabla f(\mathbf{0})\boldsymbol{v}\mathrm{d}\tau.$$

That is, under time rescaling $t \to \tau(t)$, the trajectory of $\boldsymbol{v}(t)$ is exactly the power iteration, regardless of L. □

B.8 The counter example for Escaping direction for deep matrix factorization

The counter example The $\nabla f(\mathbf{0}) = \mathrm{diag}(2, 0.9, 0.8, \ldots, 0.1) \in \mathbb{R}^{10\times 10}$ is diagonal. The $\boldsymbol{W}(0)$ is also diagonal and $\boldsymbol{W}(0)_{i,i} \sim \mathrm{Unif}[0.9, 1.1]\alpha$ for $i \in [10]\backslash 2$, $\boldsymbol{W}(0)_{2,2} = 16\alpha$, where $\alpha = 10^{-16}$ is a small constant. The depth is 4.

Lemma B.4. *With* $\nabla f(\mathbf{0})$ *and* $\boldsymbol{W}(0)$ *constructed above,* $\boldsymbol{v}_1\boldsymbol{M}(0)\boldsymbol{v}_1^\top > 0$ *and* $\widehat{\boldsymbol{M}}(\infty) \neq \boldsymbol{v}_1\boldsymbol{v}_1^\top$.

Proof of Lemma B.4. It's easy to check that $\boldsymbol{v}_1 = \boldsymbol{e}_1$, so $\boldsymbol{v}_1\boldsymbol{M}(0)\boldsymbol{v}_1^\top > 0$. Next we'll prove that $\widehat{\boldsymbol{M}}(\infty) \neq \boldsymbol{v}_1\boldsymbol{v}_1^\top$.

As both $\boldsymbol{W}(0)$ and $\nabla f(\mathbf{0})$ are diagonal, $\boldsymbol{W}(t)$ is always diagonal and has dynamics

$$\frac{\mathrm{d}M(t)_{i,i}}{\mathrm{d}t} = -2\nabla f(\mathbf{0})_{i,i}M(t)_{i,i}^2, \quad \forall i \in [10],$$

therefore we have closed form of $\boldsymbol{M}(t)$:

$$M(t)_{i,i}^{-1} = M(0)_{i,i}^{-1} - 2\nabla f(\mathbf{0})_{i,i}t, \quad \forall i \in [10].$$

For $i \in [10]$, the time for $M(t)_{i,i}$ going to infinity is $(2M(0)_{i,i}\nabla f(\mathbf{0})_{i,i})^{-1}$. By simple calculation, $M(t)_{2,2}$ goes to infinity fastest, thus $\widehat{\boldsymbol{M}}(\infty) = \boldsymbol{e}_2\boldsymbol{e}_2^\top \neq \boldsymbol{v}_1\boldsymbol{v}_1^\top$. □

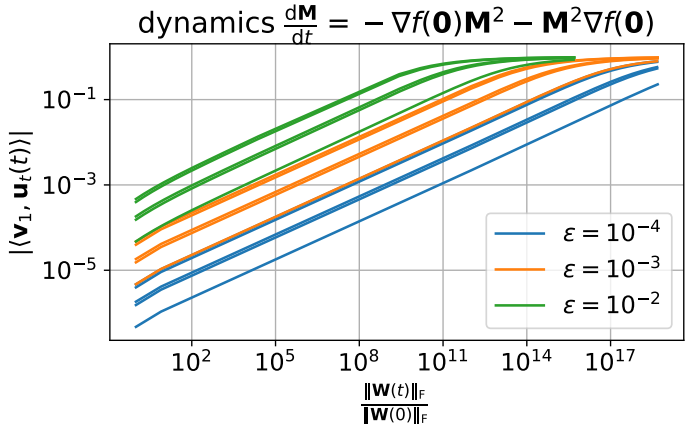

Figure 7: Dynamics of $\frac{\mathrm{d}\boldsymbol{M}}{\mathrm{d}t} = -\nabla f(\mathbf{0})\boldsymbol{M}^{L/2} - \boldsymbol{M}^{L/2}\nabla f(\mathbf{0})$ plotted, where $L = 4$, $\boldsymbol{u}_1(t)$ is the top eigenvector of $\boldsymbol{W}(t)$ and ϵ is the relative magnitude of noise. The initialization we use in this experiment is $\boldsymbol{W}_{\text{noise}}(0) = \boldsymbol{W}(0) + \frac{\alpha\epsilon}{2}(\boldsymbol{Z} + \boldsymbol{Z}^\top)$, where $\boldsymbol{W}(0)$ is what we construct at Appendix B.8, and $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{I}_{d^2\times d^2})$ is a matrix with entries sampled from i.i.d. Gaussians. We run 5 fixed random seeds (the noise matrix) for each ϵ. The trajectory of $\boldsymbol{W}$ is calculated by simulating gradient flow on $\boldsymbol{M}$ with small timestep and RMSprop (Hinton et al.) for faster convergence.

We remark that the scales of $\boldsymbol{W}(0)$ and $\nabla f(\mathbf{0})$ don't matter as in gradient flow, as scaling $\dot{\boldsymbol{W}}(0)$ is equivalent to scaling time (by Lemma B.5 below). And for this reason, the x-axis is the chosen as $\frac{\|\boldsymbol{W}(t)\|_F}{\|\boldsymbol{W}(0)\|_F}$, the relative growth rate.

Lemma B.5. *Suppose $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d$ is a P-homogeneous function, that is, $\boldsymbol{g}(\alpha\boldsymbol{\theta}) = \lambda^P \boldsymbol{g}(\alpha)$ for any $\alpha > 0$, and $\frac{\mathrm{d}\boldsymbol{\theta}'(t)}{\mathrm{d}t} = \boldsymbol{g}(\boldsymbol{\theta}'(t))$. Then $\alpha\boldsymbol{\theta}'(\alpha^{P-1}t)$ is the solution of*

$$\frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t} = \boldsymbol{g}(\boldsymbol{\theta}(t)), \qquad \boldsymbol{\theta}(0) = \alpha\boldsymbol{\theta}'(0). \tag{17}$$

Proof of Lemma B.5. Simply plug in $\boldsymbol{\theta}(t) = \alpha\boldsymbol{\theta}'(\alpha^{P-1}t)$, we have

$$\frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t} = \frac{\mathrm{d}\alpha\boldsymbol{\theta}'(\alpha^{P-1}t)}{\mathrm{d}t} = \alpha^P \frac{\mathrm{d}\boldsymbol{\theta}'(\alpha^{P-1}t)}{\mathrm{d}(\alpha^{P-1}t)} = \alpha^P \boldsymbol{g}(\boldsymbol{\theta}'(\alpha^{P-1}t)) = \boldsymbol{g}(\alpha\boldsymbol{\theta}'(\alpha^{P-1}t)) = \boldsymbol{g}(\boldsymbol{\theta}(t)). \tag{18}$$

□

C Proofs for Dynamical System

In this section, we prove Theorem 5.3 in Section 5.1. Note that it suffices to prove for the case that $\boldsymbol{J}(\mathbf{0})$ is diagonal.

Theorem C.1. *If $\boldsymbol{J}(\mathbf{0}) = \mathrm{diag}(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ is diagonal, then the statement in Theorem 5.3 holds.*

Proof for Theorem 5.3. We show how to prove Theorem 5.3 based on Theorem C.1. Let $\frac{d\boldsymbol{\theta}}{dt} = \boldsymbol{g}(\boldsymbol{\theta})$ be the dynamical system in Theorem 5.3. Let $\tilde{\boldsymbol{U}} = (\tilde{\boldsymbol{u}}_1, \ldots, \tilde{\boldsymbol{u}}_d)$, $\tilde{\boldsymbol{V}} = (\tilde{\boldsymbol{v}}_1, \ldots, \tilde{\boldsymbol{v}}_d)$ be the matrices formed by left and right eigenvectors. Now we define the following new dynamics by changing the basis:

$$\hat{\boldsymbol{\theta}}(t) = \tilde{\boldsymbol{U}}^\top \boldsymbol{\theta}(t)\tilde{\boldsymbol{V}}.$$

Then $\frac{d\hat{\boldsymbol{\theta}}(t)}{dt} = \hat{\boldsymbol{g}}(\hat{\boldsymbol{\theta}})$ for $\hat{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}) := \tilde{\boldsymbol{U}}^\top \boldsymbol{g}(\hat{\boldsymbol{\theta}})\tilde{\boldsymbol{V}}$, and the associated Jacobian matrix is $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}) := \tilde{\boldsymbol{U}}^\top \boldsymbol{J}(\hat{\boldsymbol{\theta}})\tilde{\boldsymbol{V}}$, and thus $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}}) = \mathrm{diag}(\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$.

Now we apply Theorem C.1 to $\hat{\boldsymbol{\theta}}(t)$. Then $\hat{\boldsymbol{z}}_\alpha(t) := \tilde{\boldsymbol{U}}^\top \boldsymbol{z}_\alpha(t)\tilde{\boldsymbol{V}}$ converges to the limit $\hat{\boldsymbol{z}}(t) := \lim_{\alpha\to 0} \hat{\boldsymbol{z}}_\alpha(t)$. This shows that the limit $\boldsymbol{z}(t) = \tilde{\boldsymbol{V}}\hat{\boldsymbol{z}}(t)\tilde{\boldsymbol{U}}^\top$ exists in Theorem 5.3. Now given $\boldsymbol{\delta}_\alpha$ converging to $\mathbf{0}$ with positive alignment with $\tilde{\boldsymbol{u}}_1$ as $\alpha \to 0$, we can define $\hat{\boldsymbol{\delta}}_\alpha := \tilde{\boldsymbol{U}}^\top \boldsymbol{\delta}(t)\tilde{\boldsymbol{V}}$, then $\hat{\boldsymbol{\delta}}_\alpha$ converges to $\mathbf{0}$ with positive alignment with $\boldsymbol{e}_1$, where $\boldsymbol{e}_1$ is the first vector in the standard basis and is also the top eigenvector of $\hat{\boldsymbol{J}}(\mathbf{0})$. Therefore, for every $t \in (-\infty, +\infty)$, there is a constant $C > 0$ such that

$$\left\| \tilde{\boldsymbol{U}}^\top \phi\left(\boldsymbol{\delta}_\alpha, t + \frac{1}{\tilde{\mu}_1}\log\frac{1}{\langle \boldsymbol{\delta}_\alpha, \tilde{\boldsymbol{u}}_1\rangle}\right)\tilde{\boldsymbol{V}} - \hat{\boldsymbol{z}}(t)\right\|_2 \le C \cdot \|\boldsymbol{\delta}_\alpha\|_2^{\frac{\tilde{\gamma}}{\tilde{\mu}_1+\tilde{\gamma}}} \tag{19}$$

for every sufficiently small α. As $\tilde{\boldsymbol{U}}^\top, \tilde{\boldsymbol{V}}$ are invertible, this directly implies Equation (6). □

Now we only need to prove Theorem C.1. Let $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d$ be the standard basis. Then $\tilde{\boldsymbol{u}}_1 = \tilde{\boldsymbol{v}}_1 = \boldsymbol{e}_1$ in this diagonal case. We only use $\boldsymbol{e}_1$ to stand for $\tilde{\boldsymbol{u}}_1$ and $\tilde{\boldsymbol{v}}_1$ in the rest of our analysis.

Let $R > 0$. Since $\boldsymbol{g}(\boldsymbol{\theta})$ is $\mathcal{C}^2$-smooth, there exists $\beta > 0$ such that $\|\boldsymbol{J}(\boldsymbol{\theta}) - \boldsymbol{J}(\boldsymbol{\theta}+\boldsymbol{h})\|_2 \le \beta\|\boldsymbol{h}\|_2$ for all $\|\boldsymbol{\theta}\|_2, \|\boldsymbol{\theta}+\boldsymbol{h}\|_2 \le R$. Then it can be proved by integration that $\|\boldsymbol{g}(\boldsymbol{\theta}+\boldsymbol{h}) - \boldsymbol{g}(\boldsymbol{\theta}) - \boldsymbol{J}(\boldsymbol{\theta})\boldsymbol{h}\|_2 \le \beta\|\boldsymbol{h}\|_2^2$ and $\|\boldsymbol{g}(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}+\boldsymbol{h})\|_2 \le \sup_{\xi\in(0,1)} \|\boldsymbol{J}(\boldsymbol{\theta}+\xi\boldsymbol{h})\|_2 \cdot \|\boldsymbol{h}\|_2$.

Let $\kappa := \beta/\tilde{\mu}_1$. We assume WLOG that $R \le 1/\kappa$.

Let $F(x) = \log x - \log(1+\kappa x)$. It is easy to seen that $F'(x) = \frac{1}{x+\kappa x^2}$ and $F(x)$ is an increasing function with range $(-\infty, \log(1/\kappa))$. We use $F^{-1}(y)$ to denote the inverse function of $F(x)$.

Define $T_\alpha(r) := \frac{1}{\tilde{\mu}_1}(F(r) - F(\alpha)) = \frac{1}{\tilde{\mu}_1}\left(\log\frac{r}{\alpha} - \log\frac{1+\kappa r}{1+\kappa\alpha}\right)$.

Lemma C.2. *For $\boldsymbol{\theta}(t) = \phi(\boldsymbol{\theta}_0, t)$ with $\|\boldsymbol{\theta}_0\|_2 \le \alpha$ and $t \le T_\alpha(r)$,*

$$\|\boldsymbol{\theta}(t)\|_2 \le \frac{1+\kappa r}{1+\kappa\alpha}\alpha \cdot e^{\tilde{\mu}_1 t} \le r.$$

Proof. $\|\boldsymbol{g}(\boldsymbol{\theta}) - \boldsymbol{J}(\mathbf{0})\boldsymbol{\theta}\|_2 = \|\boldsymbol{g}(\boldsymbol{\theta}) - \boldsymbol{g}(\mathbf{0}) - \boldsymbol{J}(\mathbf{0})\boldsymbol{\theta}\|_2 \le \beta\|\boldsymbol{\theta}\|_2^2$.

$$\frac{1}{2}\frac{d\|\boldsymbol{\theta}(t)\|_2^2}{dt} = \langle \boldsymbol{\theta}(t), \boldsymbol{g}(\boldsymbol{\theta}(t))\rangle \le \langle \boldsymbol{\theta}(t), \boldsymbol{J}(\mathbf{0})\boldsymbol{\theta}(t)\rangle + \beta\|\boldsymbol{\theta}(t)\|_2^3 \le \tilde{\mu}_1\|\boldsymbol{\theta}(t)\|_2^2 + \beta\|\boldsymbol{\theta}(t)\|_2^3.$$

So

$$\frac{d\|\boldsymbol{\theta}(t)\|_2}{dt} \le \tilde{\mu}_1\|\boldsymbol{\theta}(t)\|_2 + \beta\|\boldsymbol{\theta}(t)\|_2^2$$

Then

$$\frac{d}{dt}F(\|\boldsymbol{\theta}(t)\|_2) \le \tilde{\mu}_1.$$

So $F(\|\boldsymbol{\theta}(t)\|_2) \le F(\alpha) + \tilde{\mu}_1 t$. By definition of $T_\alpha(r)$, we then know that $\|\boldsymbol{\theta}(t)\|_2 \le r$ for all $t \le T_\alpha(r)$. So

$$\log\|\boldsymbol{\theta}(t)\|_2 \le F(\|\boldsymbol{\theta}(t)\|_2) + \log(1+\kappa r) \le F(\alpha) + \tilde{\mu}_1 t + \log(1+\kappa r).$$

Expending $F(\alpha)$ proves the lemma. □

Lemma C.3. *For $\boldsymbol{\theta}(t) = \phi(\boldsymbol{\theta}_0, t)$ with $\|\boldsymbol{\theta}_0\|_2 \le \alpha$ and $t \le T_\alpha(r)$, we have*

$$\boldsymbol{\theta}(t) = e^{t\boldsymbol{J}(\mathbf{0})}\boldsymbol{\theta}_0 + O(r^2).$$

Proof. Let $\hat{\boldsymbol{\theta}}(t) = e^{t\boldsymbol{J}(\mathbf{0})}\boldsymbol{\theta}_0$.

$$\begin{aligned}\frac{d}{dt}\|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t)\|_2 &\le \left\|\frac{d\boldsymbol{\theta}(t)}{dt} - \frac{d\hat{\boldsymbol{\theta}}(t)}{dt}\right\|_2 \\ &= \left\|\boldsymbol{g}(\boldsymbol{\theta}(t)) - \boldsymbol{J}(\mathbf{0})\hat{\boldsymbol{\theta}}(t)\right\|_2 \\ &\le \|\boldsymbol{g}(\boldsymbol{\theta}(t)) - \boldsymbol{J}(\mathbf{0})\boldsymbol{\theta}(t)\|_2 + \left\|\boldsymbol{J}(\mathbf{0})(\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t))\right\|_2 \\ &\le \beta\|\boldsymbol{\theta}(t)\|_2^2 + \tilde{\mu}_1\|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t)\|_2 \\ &\le \beta\left(\frac{1+\kappa r}{1+\kappa\alpha}\alpha\right)^2 \cdot e^{2\tilde{\mu}_1 t} + \tilde{\mu}_1\|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t)\|_2.\end{aligned}$$

By Grönwall's inequality, we have

$$\|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t)\|_2 \le \int_0^t \beta\left(\frac{1+\kappa r}{1+\kappa\alpha}\alpha\right)^2 \cdot e^{2\tilde{\mu}_1\tau}e^{\tilde{\mu}_1(t-\tau)}d\tau.$$

Evaluating the integral gives

$$\|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t)\|_2 \le \beta\left(\frac{1+\kappa r}{1+\kappa\alpha}\alpha\right)^2 e^{\tilde{\mu}_1 t} \cdot \frac{e^{\tilde{\mu}_1 t} - 1}{\tilde{\mu}_1} \le \kappa\left(\frac{1+\kappa r}{1+\kappa\alpha}\alpha \cdot e^{\tilde{\mu}_1 t}\right)^2 \le \kappa r^2,$$

which proves the lemma. □

Lemma C.4. *Let $\boldsymbol{\theta}(t) = \phi(\boldsymbol{\theta}_0, t), \hat{\boldsymbol{\theta}}(t) = \phi(\hat{\boldsymbol{\theta}}_0, t)$. If $\max\{\|\boldsymbol{\theta}_0\|_2, \|\hat{\boldsymbol{\theta}}_0\|_2\} \le \alpha$. For $t \le T_\alpha(r)$, we have*

$$\|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t)\|_2 \le e^{\tilde{\mu}_1 t + \kappa r}\|\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_0\|_2.$$

Proof. For $t \le T_\alpha(r)$,

$$\begin{aligned}\frac{d}{dt}\|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t)\|_2 &\le \left\|\frac{d\boldsymbol{\theta}(t)}{dt} - \frac{d\hat{\boldsymbol{\theta}}(t)}{dt}\right\|_2 \\ &= \left\|\boldsymbol{g}(\boldsymbol{\theta}(t)) - \boldsymbol{g}(\hat{\boldsymbol{\theta}}(t))\right\|_2 \\ &\le \sup_{\xi\in(0,1)}\left\|\boldsymbol{J}\left(\xi\boldsymbol{\theta}(t) + (1-\xi)\hat{\boldsymbol{\theta}}(t)\right)\right\|_2 \cdot \|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t)\|_2.\end{aligned}$$

By Lemma C.2, $\max\{\|\boldsymbol{\theta}(t)\|_2, \|\hat{\boldsymbol{\theta}}(t)\|_2\} \le \frac{1+\kappa r}{1+\kappa\alpha}\alpha \cdot e^{\tilde{\mu}_1 t}$ for all $t \le T_\alpha(r)$. By the smoothness assumption, the norm of Jacobian can be upper bounded by $\tilde{\mu}_1 + \beta \cdot \frac{1+\kappa r}{1+\kappa\alpha}\alpha \cdot e^{\tilde{\mu}_1 t}$. Thus,

$$\frac{d}{dt}\|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t)\|_2 \le \left(\tilde{\mu}_1 + \beta \cdot \frac{1+\kappa r}{1+\kappa\alpha}\alpha \cdot e^{\tilde{\mu}_1 t}\right)\|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t)\|_2.$$

This implies that

$$\log\frac{\|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t)\|_2}{\|\boldsymbol{\theta}(0) - \hat{\boldsymbol{\theta}}(0)\|_2} \le \int_0^t \left(\tilde{\mu}_1 + \beta \cdot \frac{1+\kappa r}{1+\kappa\alpha}\alpha \cdot e^{\tilde{\mu}_1\tau}\right)d\tau \le \tilde{\mu}_1 t + \kappa \cdot \frac{1+\kappa r}{1+\kappa\alpha}\alpha e^{\tilde{\mu}_1 t} \le \tilde{\mu}_1 t + \kappa r.$$

Therefore, $\|\boldsymbol{\theta}(t) - \hat{\boldsymbol{\theta}}(t)\|_2 \le e^{\tilde{\mu}_1 t + \kappa r}\|\boldsymbol{\theta}(0) - \hat{\boldsymbol{\theta}}(0)\|_2$. □

Lemma C.5. *For every $t \in (-\infty, +\infty)$, $\boldsymbol{z}(t)$ exists and $\boldsymbol{z}_\alpha(t)$ converges to $\boldsymbol{z}(t)$ in the following rate:*

$$|\boldsymbol{z}_\alpha(t) - \boldsymbol{z}(t)| = O(\alpha),$$

where O hides constants depending on $g(\boldsymbol{\theta})$ and t.

Proof. We prove the lemma in the cases of $t \in (-\infty, F(R)/\tilde{\mu}_1]$ and $t > F(R)/\tilde{\mu}_1$ respectively.

Case 1. Fix $t \in (-\infty, F(R)/\tilde{\mu}_1]$. Let $\tilde{\alpha}$ be the unique number such that $\frac{\tilde{\alpha}}{1+\kappa\tilde{\alpha}} = \alpha$ (i.e., $F(\tilde{\alpha}) = \log\alpha$). Let α' be an arbitrary number less than α. Let $t_0 := \frac{1}{\tilde{\mu}_1}\log\frac{\alpha}{\alpha'}$. Then

$t_0 = \frac{1}{\tilde{\mu}_1}\left(F(\tilde{\alpha}) - \log \alpha'\right) \le T_{\alpha'}(\tilde{\alpha})$. By Lemma C.3, we have

$$\|\phi\left(\alpha' \boldsymbol{e}_1, t_0\right) - \alpha \boldsymbol{e}_1\|_2 = \left\|\phi\left(\alpha' \boldsymbol{e}_1, t_0\right) - e^{t_0 \boldsymbol{J}(\boldsymbol{0})} \alpha' \boldsymbol{e}_1\right\|_2 = O(\tilde{\alpha}^2).$$

Let $r := F^{-1}(\tilde{\mu}_1 t) \le R$. Then $t + \frac{1}{\tilde{\mu}_1} \log \frac{1}{\alpha} = T_{\tilde{\alpha}}(r)$ if $\tilde{\alpha} < r$.

By Lemma C.2, $\|\phi\left(\alpha' \boldsymbol{e}_1, t_0\right)\|_2 \le \tilde{\alpha}$. Also, $\|\alpha \boldsymbol{e}_1\|_2 = \frac{\tilde{\alpha}}{1+\kappa\tilde{\alpha}} \le \tilde{\alpha}$. By Lemma C.4,

$$\begin{aligned}
\|\boldsymbol{z}_\alpha(t) - \boldsymbol{z}_{\alpha'}(t)\|_2 &= \left\|\phi\left(\alpha' \boldsymbol{e}_1, t + \frac{1}{\tilde{\mu}_1} \log \frac{1}{\alpha'}\right) - \phi\left(\alpha \boldsymbol{e}_1, t + \frac{1}{\tilde{\mu}_1} \log \frac{1}{\alpha}\right)\right\|_2 \\
&= \left\|\phi\left(\phi(\alpha' \boldsymbol{e}_1, t_0), t + \frac{1}{\tilde{\mu}_1} \log \frac{1}{\alpha}\right) - \phi\left(\alpha \boldsymbol{e}_1, t + \frac{1}{\tilde{\mu}_1} \log \frac{1}{\alpha}\right)\right\|_2 \\
&\le O(\tilde{\alpha}^2 \cdot e^{\tilde{\mu}_1 (t + \frac{1}{\tilde{\mu}_1} \log \frac{1}{\alpha}) + \kappa r}) \\
&\le O\left(\frac{\tilde{\alpha}^2}{\alpha}\right).
\end{aligned}$$

For α small enough, we have $\tilde{\alpha} = O(\alpha)$, so for any $\alpha' \in (0, \alpha)$,

$$\|\boldsymbol{z}_\alpha(t) - \boldsymbol{z}_{\alpha'}(t)\|_2 = O(\alpha).$$

This implies that $\{\boldsymbol{z}_\alpha(t)\}$ satisfies Cauchy's criterion for every t, and thus the limit $\boldsymbol{z}(t)$ exists for $t \le F(R)/\tilde{\mu}_1$. The convergence rate can be deduced by taking limits for $\alpha' \to 0$ on both sides.

Case 2. For $t = F(R)/\tilde{\mu}_1 + \tau$ with $\tau > 0$, $\phi(\boldsymbol{\theta}, \tau)$ is locally Lipschitz with respect to $\boldsymbol{\theta}$. So

$$\begin{aligned}
\|\boldsymbol{z}_\alpha(t) - \boldsymbol{z}_{\alpha'}(t)\|_2 &= \|\phi(\boldsymbol{z}_\alpha(F(R)/\tilde{\mu}_1), \tau) - \phi(\boldsymbol{z}_{\alpha'}(F(R)/\tilde{\mu}_1), \tau)\|_2 \\
&= O(\|\boldsymbol{z}_\alpha(F(R)/\tilde{\mu}_1) - \boldsymbol{z}_{\alpha'}(F(R)/\tilde{\mu}_1)\|_2) \\
&= O(\alpha),
\end{aligned}$$

which proves the lemma for $t > F(R)/\tilde{\mu}_1$. □

Proof for Theorem C.1. The existence of $\boldsymbol{z}(t)$ has already been proved in Lemma C.5. Due to continuity of $\phi(\cdot, t)$ for every $t \in \mathbb{R}$, we have

$$\boldsymbol{z}(t) = \lim_{\alpha \to 0} \boldsymbol{z}_\alpha(t) = \lim_{\alpha \to 0} \phi\left(\alpha \tilde{\boldsymbol{v}}_1, t + \frac{1}{\tilde{\mu}_1} \log \frac{1}{\alpha}\right) = \phi\left(\lim_{\alpha \to 0} \phi\left(\alpha \tilde{\boldsymbol{v}}_1 + \frac{1}{\tilde{\mu}_1} \log \frac{1}{\alpha}\right), t\right) = \phi\left(\boldsymbol{z}(0), t\right).$$

Now it's only left to prove (6). WLOG we can assume that $\|\boldsymbol{\delta}_\alpha\|_2$ is decreasing and $\frac{\alpha}{2} \le \|\boldsymbol{\delta}_\alpha\|_2 \le \alpha$ (otherwise we can do reparameterization). Then our goal becomes proving

$$\|\boldsymbol{\theta}_\alpha(t) - \boldsymbol{z}(t)\|_2 = O\left(\alpha^{\frac{\tilde{\gamma}}{\tilde{\mu}_1 + \tilde{\gamma}}}\right). \tag{20}$$

where $\boldsymbol{\theta}_\alpha(t) := \phi\left(\boldsymbol{\delta}_\alpha, t + \frac{1}{\tilde{\mu}_1} \log \frac{1}{\langle \boldsymbol{\delta}_\alpha, \boldsymbol{e}_1 \rangle}\right)$. We prove (20) in the cases of $t \in (-\infty, F(R)/\tilde{\mu}_1]$ and $t > F(R)/\tilde{\mu}_1$ respectively.

Case 1. Fix $t \in (-\infty, (F(R) + \log q)/\tilde{\mu}_1]$. Let $\tilde{\alpha}_1 = \alpha^{\frac{\tilde{\gamma}}{\tilde{\mu}_1 + \tilde{\gamma}}}$. Let $\alpha_1 := e^{F(\tilde{\alpha}_1)} = \frac{\tilde{\alpha}_1}{1+\kappa\tilde{\alpha}_1}$. Let $t_0 := \frac{1}{\tilde{\mu}_1}(F(\tilde{\alpha}_1) - \log \alpha) \le T_{\|\boldsymbol{\delta}(\alpha)\|_2}(\tilde{\alpha}_1)$. Let $q_\alpha := \left\langle \frac{\boldsymbol{\delta}(\alpha)}{\alpha}, \boldsymbol{e}_1 \right\rangle$. By Definition 5.2, there exists $q > 0$ such that $q_\alpha \ge q$ for all sufficiently small α.

$$\begin{aligned}
\|\phi\left(\boldsymbol{\delta}(\alpha), t_0\right) - \alpha_1 q_\alpha \boldsymbol{e}_1\|_2 &= \left\|\phi\left(\boldsymbol{\delta}(\alpha), t_0\right) - e^{t_0 \boldsymbol{J}(\boldsymbol{0})} \boldsymbol{\delta}(\alpha)\right\|_2 + \left\|\left(e^{t_0 \boldsymbol{J}(\boldsymbol{0})} - e^{\tilde{\mu}_1 t_0} \boldsymbol{e}_1 \boldsymbol{e}_1^\top\right) \boldsymbol{\delta}(\alpha)\right\|_2 \\
&= O(\tilde{\alpha}_1^2) + e^{\frac{\tilde{\mu}_2}{\tilde{\mu}_1}(F(\tilde{\alpha}_1) - \log \alpha)} \|\boldsymbol{\delta}(\alpha)\|_2 \\
&= O(\tilde{\alpha}_1^2) + \left(\frac{\alpha_1}{\alpha}\right)^{\frac{\tilde{\mu}_2}{\tilde{\mu}_1}} \|\boldsymbol{\delta}(\alpha)\|_2 \\
&= O(\tilde{\alpha}_1^2 + \alpha_1^{\tilde{\mu}_2/\tilde{\mu}_1} \alpha^{1 - \tilde{\mu}_2/\tilde{\mu}_1}) \\
&= O(\alpha_1^2).
\end{aligned}$$

Let $r := F^{-1}(\tilde{\mu}_1 t + \log \frac{1}{q_\alpha}) \le R$. Then $t + \frac{1}{\tilde{\mu}_1} \log \frac{1}{\alpha_1 q_\alpha} = T_{\tilde{\alpha}}(r)$ if $\tilde{\alpha} < r$. By Lemma C.2, $\|\phi(\boldsymbol{\delta}(\alpha), t_0)\|_2 \le \tilde{\alpha}_1$. Also, $\|\alpha_1 q_\alpha \boldsymbol{e}_1\|_2 \le \alpha_1 = \frac{\tilde{\alpha}_1}{1+\kappa\tilde{\alpha}_1} \le \tilde{\alpha}_1$. By Lemma C.4,

$$\begin{aligned}\|\boldsymbol{\theta}_\alpha(t) - \boldsymbol{z}_{\alpha_1}(t)\|_2 &\le \left\| \phi\left(\phi(\boldsymbol{\delta}(\alpha), t_0), t + \frac{1}{\tilde{\mu}_1}\log\frac{1}{\alpha_1 q_\alpha}\right) - \phi\left(\alpha_1 q_\alpha \boldsymbol{e}_1, t + \frac{1}{\tilde{\mu}_1}\log\frac{1}{\alpha_1 q_\alpha}\right)\right\|_2 \\ &= O\left(\alpha_1^2 \cdot e^{\tilde{\mu}_1\left(t + \frac{1}{\tilde{\mu}_1}\log\frac{1}{\alpha_1 q_\alpha}\right) + \kappa r}\right) \\ &= O(\alpha_1).\end{aligned}$$

Combining this with the convergence rate for $\boldsymbol{z}_{\alpha_1}(t)$, we have

$$\|\boldsymbol{\theta}_\alpha(t) - \boldsymbol{z}(t)\|_2 \le \|\boldsymbol{\theta}_\alpha(t) - \boldsymbol{z}_{\alpha_1}(t)\|_2 + \|\boldsymbol{z}_{\alpha_1}(t) - \boldsymbol{z}(t)\|_2 = O(\alpha_1).$$

Case 2. For $t = (F(R) + \log q)/\tilde{\mu}_1 + \tau$ with $\tau > 0$, $\phi(\boldsymbol{\theta}, \tau)$ is locally Lipschitz with respect to $\boldsymbol{\theta}$. So

$$\begin{aligned}\|\boldsymbol{\theta}_\alpha(t) - \boldsymbol{z}(t)\|_2 &= \|\phi(\boldsymbol{\theta}_\alpha((F(R) + \log q)/\tilde{\mu}_1), \tau) - \phi(\boldsymbol{z}((F(R) + \log q)/\tilde{\mu}_1), \tau)\|_2 \\ &= O(\|\boldsymbol{\theta}_\alpha((F(R) + \log q)/\tilde{\mu}_1) - \boldsymbol{z}((F(R) + \log q)/\tilde{\mu}_1)\|_2) \\ &= O(\alpha_1),\end{aligned}$$

which proves (20) for $t > (F(R) + \log q)/\tilde{\mu}_1$. □

D Eigenvalues of Jacobian and Hessians

In this section we analyze the eigenvalues of the Jacobian at critical points.

Let $\boldsymbol{U}_0 \in \mathbb{R}^{d\times k}$ be a rank-k critical point of $\mathcal{L}(\boldsymbol{U}) = \frac{1}{2} f(\boldsymbol{U}\boldsymbol{U}^\top)$, i.e., $\operatorname{rank}(\boldsymbol{U}_0) = k$ and $\nabla\mathcal{L}(\boldsymbol{U}_0) = \nabla f(\boldsymbol{U}_0\boldsymbol{U}_0^\top)\boldsymbol{U}_0 = \boldsymbol{0}$. This implies

$$\nabla f(\boldsymbol{W}_0)\boldsymbol{W}_0 = \left(\nabla f(\boldsymbol{U}_0\boldsymbol{U}_0^\top)\boldsymbol{U}_0\right)\boldsymbol{U}_0^\top = \boldsymbol{0} \tag{21}$$

for $\boldsymbol{W}_0 := \boldsymbol{U}_0\boldsymbol{U}_0^\top$, and thus $\boldsymbol{U}_0$ and $\boldsymbol{W}_0$ are fixed points of (1), (2), respectively.

For a real-valued or vector-valued function $F(\boldsymbol{W})$, we use $DF(\boldsymbol{W})[\boldsymbol{\Delta}]$, $D^2F(\boldsymbol{W})[\boldsymbol{\Delta}, \boldsymbol{\Delta}]$ to denote the first and second order directional derivatives of $F(\,\cdot\,)$ at $\boldsymbol{W}$. Then we have

$$D(-\nabla\mathcal{L})(\boldsymbol{U}_0)[\boldsymbol{\Delta}] = -\nabla f(\boldsymbol{U}_0\boldsymbol{U}_0^\top)\boldsymbol{\Delta} - D(\nabla f)(\boldsymbol{U}_0\boldsymbol{U}_0^\top)[\boldsymbol{U}_0\boldsymbol{\Delta}^\top + \boldsymbol{\Delta}\boldsymbol{U}_0^\top]\boldsymbol{U}_0, \tag{22}$$

and for $\boldsymbol{g}(\boldsymbol{W}) := -\nabla f(\boldsymbol{W})\boldsymbol{W} - \boldsymbol{W}\nabla f(\boldsymbol{W})$,

$$D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{\Delta}] = -\nabla f(\boldsymbol{W}_0)\boldsymbol{\Delta} - \boldsymbol{\Delta}\nabla f(\boldsymbol{W}_0) - \boldsymbol{W}_0 D(\nabla f)(\boldsymbol{W}_0)[\boldsymbol{\Delta}] - D(\nabla f)(\boldsymbol{W}_0)[\boldsymbol{\Delta}]\boldsymbol{W}_0,$$

$$\langle D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{\Delta}_1], \boldsymbol{\Delta}_2\rangle = -\langle \boldsymbol{\Delta}_1, \nabla f(\boldsymbol{W}_0)\boldsymbol{\Delta}_2 + \boldsymbol{\Delta}_2\nabla f(\boldsymbol{W}_0)\rangle - D^2 f(\boldsymbol{W}_0)[\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2\boldsymbol{W}_0 + \boldsymbol{W}_0\boldsymbol{\Delta}_2].$$

D.1 Eigenvalues at Origin

Proof for Lemma 5.4. For $\boldsymbol{W}_0 = \boldsymbol{0}$, we have

$$\begin{aligned}D\boldsymbol{g}(\boldsymbol{0})[\boldsymbol{\Delta}] &= -\nabla f(\boldsymbol{0})\boldsymbol{\Delta} - \boldsymbol{\Delta}\nabla f(\boldsymbol{0}) \\ \langle D\boldsymbol{g}(\boldsymbol{0})[\boldsymbol{\Delta}_1], \boldsymbol{\Delta}_2\rangle &= -\langle \boldsymbol{\Delta}_1, \nabla f(\boldsymbol{0})\boldsymbol{\Delta}_2 + \boldsymbol{\Delta}_2\nabla f(\boldsymbol{0})\rangle\end{aligned}$$

Then $D\boldsymbol{g}(\boldsymbol{W}_0)$ is symmetric since

$$\begin{aligned}\langle D\boldsymbol{g}(\boldsymbol{0})[\boldsymbol{\Delta}_2], \boldsymbol{\Delta}_1\rangle &= -\langle \boldsymbol{\Delta}_2, \nabla f(\boldsymbol{0})\boldsymbol{\Delta}_1\rangle - \langle \boldsymbol{\Delta}_2, \boldsymbol{\Delta}_1\nabla f(\boldsymbol{0})\rangle \\ &= -\langle \nabla f(\boldsymbol{0})\boldsymbol{\Delta}_2, \boldsymbol{\Delta}_1\rangle - \langle \boldsymbol{\Delta}_2\nabla f(\boldsymbol{0}), \boldsymbol{\Delta}_1\rangle \\ &= \langle D\boldsymbol{g}(\boldsymbol{0})[\boldsymbol{\Delta}_1], \boldsymbol{\Delta}_2\rangle.\end{aligned}$$

Let $-\nabla f(\boldsymbol{0}) = \sum_{i=1}^d \mu_i \boldsymbol{u}_{1[i]}\boldsymbol{u}_{1[i]}^\top$ be the eigendecomposition of the symmetric matrix $-\nabla f(\boldsymbol{0})$. For $\boldsymbol{\Delta} = \boldsymbol{u}_{1[i]}\boldsymbol{u}_{1[j]}^\top + \boldsymbol{u}_{1[j]}\boldsymbol{u}_{1[i]}^\top$, we have

$$D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{\Delta}] = \mu_i \boldsymbol{u}_{1[i]}\boldsymbol{u}_{1[j]}^\top + \mu_j \boldsymbol{u}_{1[j]}\boldsymbol{u}_{1[i]}^\top = (\mu_i + \mu_j)\boldsymbol{\Delta}.$$

So $\boldsymbol{u}_{1[i]}\boldsymbol{u}_{1[j]}^\top + \boldsymbol{u}_{1[j]}\boldsymbol{u}_{1[i]}^\top$ is an eigenvector of $D\boldsymbol{g}(\boldsymbol{0})$ associated with eigenvalue $\mu_i + \mu_j$. Note that $\{\boldsymbol{u}_{1[i]}\boldsymbol{u}_{1[j]}^\top + \boldsymbol{u}_{1[j]}\boldsymbol{u}_{1[i]}^\top : i, j \in [d]\}$ spans all the symmetric matrices, so these are all the eigenvectors in the space of symmetric matrices.

For every antisymmetric matrix $\boldsymbol{\Delta}$ (i.e., $\boldsymbol{\Delta} = -\boldsymbol{\Delta}^\top$), we have

$$D\boldsymbol{g}(\mathbf{0})[\boldsymbol{\Delta}] = D\boldsymbol{g}(\mathbf{0})[\boldsymbol{\Delta}^\top] = D\boldsymbol{g}(\mathbf{0})[-\boldsymbol{\Delta}].$$

So $D\boldsymbol{g}(\mathbf{0})[\boldsymbol{\Delta}] = \mathbf{0}$ and every antisymmetric matrix is an eigenvector associated with eigenvalue 0.

Since any matrix can be expressed as the sum of a symmetric matrix and an antisymmetric matrix, we have found all the eigenvalues. □

D.2 Eigenvalues at Local Minimizer

By (21), the symmetric matrices $-\nabla f(\boldsymbol{W}_0)$ and $\boldsymbol{W}_0$ commute, so they can be simultaneously diagonalizable, and as (21) implies that they have different column spans, we can have the following diagonalization:

$$-\nabla f(\boldsymbol{W}_0) = \sum_{i=1}^{d-k} \mu_i \boldsymbol{v}_i \boldsymbol{v}_i^\top, \qquad \boldsymbol{W}_0 = \sum_{i=d-k+1}^{d} \mu_i \boldsymbol{v}_i \boldsymbol{v}_i^\top. \tag{23}$$

First we prove the following lemma on the eigenvalues and eigenvectors of $-\nabla^2 \mathcal{L}$:

Lemma D.1. *For every $\boldsymbol{\Delta} \in \mathbb{R}^{d\times d}$, if*

$$\boldsymbol{U}_0 \boldsymbol{\Delta}^\top + \boldsymbol{\Delta} \boldsymbol{U}_0^\top = \mathbf{0} \tag{24}$$

then $\boldsymbol{\Delta}$ is an eigenvector of $-\nabla^2 \mathcal{L}(\boldsymbol{U}_0)$ associated with eigenvalue 0. Moreover, the solutions of (24) spans a linear space with dimension $\frac{k(k-1)}{2}$.

Proof. Suppose $\boldsymbol{U}_0 \boldsymbol{\Delta}^\top + \boldsymbol{\Delta} \boldsymbol{U}_0^\top = \mathbf{0}$, then $\boldsymbol{U}_0 \boldsymbol{\Delta}^\top = -\boldsymbol{\Delta} \boldsymbol{U}_0^\top$. This implies that the column space $\boldsymbol{\Delta}$ and $\boldsymbol{U}_0$ are the same, i.e., there is an invertible matrix $\boldsymbol{R} \in \mathbb{R}^{k\times k}$, such that $\boldsymbol{\Delta} = \boldsymbol{U}_0 \boldsymbol{R}$. Then

$$\begin{aligned} D(-\nabla \mathcal{L})(\boldsymbol{U}_0)[\boldsymbol{\Delta}] &= -\nabla f(\boldsymbol{U}_0 \boldsymbol{U}_0^\top) \boldsymbol{U}_0 \boldsymbol{R} - D(\nabla f)(\boldsymbol{U}_0 \boldsymbol{U}_0^\top)[\boldsymbol{U}_0 \boldsymbol{\Delta}^\top + \boldsymbol{\Delta} \boldsymbol{U}_0^\top] \boldsymbol{U}_0 \\ &= -\left(\nabla f(\boldsymbol{U}_0 \boldsymbol{U}_0^\top) \boldsymbol{U}_0\right) \boldsymbol{R} - D(\nabla f)(\boldsymbol{U}_0 \boldsymbol{U}_0^\top)[\mathbf{0}] \boldsymbol{U}_0 \\ &= \mathbf{0}. \end{aligned}$$

Replacing $\boldsymbol{\Delta}$ with $\boldsymbol{U}_0 \boldsymbol{R}$ in (24) gives $\boldsymbol{U}_0 (\boldsymbol{R} + \boldsymbol{R}^\top) \boldsymbol{U}_0^\top = \mathbf{0}$, which is equivalent to $\boldsymbol{R} = -\boldsymbol{R}^\top$ since $\boldsymbol{U}_0$ is full-rank. Since the dimension of $k \times k$ antisymmetric matrices is $\frac{k(k-1)}{2}$, the span spanned by the solutions of (24) also has dimension $\frac{k(k-1)}{2}$. □

Definition D.2 (Eigendecomposition of $-\nabla^2 \mathcal{L}(\boldsymbol{U}_0)$)**.** Let

$$-D^2 \mathcal{L}(\boldsymbol{U}_0)[\boldsymbol{\Delta}_1, \boldsymbol{\Delta}_2] = \sum_{p=1}^{kd} \xi_p \left\langle \boldsymbol{E}_p, \boldsymbol{\Delta}_1 \right\rangle \left\langle \boldsymbol{E}_p, \boldsymbol{\Delta}_2 \right\rangle$$

be the eigendecomposition of the Hessian matrix $-\nabla^2 \mathcal{L}(\boldsymbol{U}_0)$, where $\xi_1, \ldots, \xi_{kd} \in \mathbb{R}$ are eigenvalues, $\boldsymbol{E}_1, \ldots, \boldsymbol{E}_{kd} \in \mathbb{R}^{d\times k}$ are eigenvectors satisfying $\langle \boldsymbol{E}_p, \boldsymbol{E}_q \rangle = \mathbb{1}_{[p=q]}$. We enforce ξ_p to be 0 and $\boldsymbol{E}_p$ to be a solution of (24) for every $kd - \frac{k(k-1)}{2} < p \leq kd$.

Lemma D.3. *Let $\boldsymbol{A} \in \mathbb{R}^{D\times D}$ be a matrix. If $\{\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_K\}$ is a set of linearly independent left eigenvectors associated with eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_K$ and $\{\tilde{\boldsymbol{v}}_1, \ldots, \tilde{\boldsymbol{v}}_{D-K}\}$ is a set of linearly independent right eigenvectors associated with eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{D-K}$, and $\hat{\lambda}_i \neq \tilde{\lambda}_j$ for all $1 \leq i \leq k, 1 \leq j \leq D-K$, then $\boldsymbol{A}$ is diagonalizable.*

Proof. For every $\lambda \in \{\hat{\lambda}_1, \ldots, \hat{\lambda}_K\} =: \hat{\Lambda}$, the geometric multiplicity of λ is at least the number of $\hat{\lambda}_i \in \hat{\Lambda}$ that equals to λ, denoted as $\hat{n}_\lambda$. For every $\lambda \in \{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{D-K}\} =: \tilde{\Lambda}$, the geometric multiplicity of λ is at least the number of $\tilde{\lambda}_i \in \tilde{\Lambda}$ that equals to λ, denoted $\tilde{n}_\lambda$.

Summing up the geometric multiplicity for $\hat{\Lambda} \cup \tilde{\Lambda}$, we have

$$\sum_{\lambda \in \hat{\Lambda}} \hat{n}_\lambda + \sum_{\lambda \in \hat{\Lambda}} \hat{n}_\lambda = D.$$

Since $\hat{\Lambda}$ and $\tilde{\Lambda}$ are disjoint, these are all the eigenvalues, and the geometric multiplicity of each eigenvalues equals to its algebraic multiplicity. This proves that $\boldsymbol{A}$ is diagonalizable. □

Theorem D.4. *If $\mu_i + \mu_j \neq \xi_p$ for any $1 \leq i \leq j \leq d-k$ and $1 \leq p \leq kd - \frac{k(k-1)}{2}$, then $D\boldsymbol{g}(\boldsymbol{W}_0)$ is diagonalizable with the following 3 types of eigenvectors:*

1. *$\mu_i + \mu_j$ is an eigenvalue for every $1 \leq i \leq j \leq d-k$, and $\hat{\boldsymbol{U}}_{ij} := \boldsymbol{v}_i\boldsymbol{v}_j^\top + \boldsymbol{v}_j\boldsymbol{v}_i^\top$ is an associated left eigenvector.*
2. *ξ_p is an eigenvalue for every $1 \leq p \leq kd - \frac{k(k-1)}{2}$, and $\tilde{\boldsymbol{V}}_p := \boldsymbol{E}_p\boldsymbol{U}_0^\top + \boldsymbol{U}_0\boldsymbol{E}_p^\top$ is an associated right eigenvector.*
3. *0 is an eigenvalue, and any antisymmetric matrix is an associated right eigenvector.*

Proof of Theorem D.4. We first prove each item respectively, and then prove that $D\boldsymbol{g}(\boldsymbol{W}_0)$ is diagonalizable.

Proof for Item 1. For $\hat{\boldsymbol{U}}_{ij} = \boldsymbol{v}_i\boldsymbol{v}_j^\top + \boldsymbol{v}_j\boldsymbol{v}_i^\top$, we know that

$$\begin{aligned} -\nabla f(\boldsymbol{W}_0)\hat{\boldsymbol{U}}_{ij} + \hat{\boldsymbol{U}}_{ij}(-\nabla f(\boldsymbol{W}_0)) &= (\lambda_i + \lambda_j)\hat{\boldsymbol{U}}_{ij} \\ \hat{\boldsymbol{U}}_{ij}\boldsymbol{W}_0 &= \boldsymbol{0} \\ \boldsymbol{W}_0\hat{\boldsymbol{U}}_{ij} &= \boldsymbol{0} \end{aligned}$$

So we have

$$\left\langle D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{\Delta}_1], \hat{\boldsymbol{U}}_{ij} \right\rangle = \left\langle \boldsymbol{\Delta}_1, (\lambda_i + \lambda_j)\hat{\boldsymbol{U}}_{ij} \right\rangle - D^2 f(\boldsymbol{W}_0)[\boldsymbol{\Delta}_1, \boldsymbol{0}] = (\lambda_i + \lambda_j)\left\langle \boldsymbol{\Delta}_1, \hat{\boldsymbol{U}}_{ij} \right\rangle,$$

which shows that $\hat{\boldsymbol{U}}_{ij}$ is a left eigenvector associated with eigenvalue $\lambda_i + \lambda_j$.

Proof for Item 2. By definition of eigenvector, we have $D(-\nabla\mathcal{L})(\boldsymbol{U}_0)[\boldsymbol{E}_p] = \xi_p\boldsymbol{E}_p$, so

$$\xi_p\boldsymbol{E}_p = -\nabla f(\boldsymbol{U}_0\boldsymbol{U}_0^\top)\boldsymbol{E}_p - D(\nabla f)(\boldsymbol{U}_0\boldsymbol{U}_0^\top)[\boldsymbol{U}_0\boldsymbol{E}_p^\top + \boldsymbol{E}_p\boldsymbol{U}_0^\top]\boldsymbol{U}_0.$$

Right-multiplying both sides by $\boldsymbol{U}_0^\top$, we get

$$\begin{aligned} \xi_p\boldsymbol{E}_p\boldsymbol{U}_0^\top &= -\nabla f(\boldsymbol{W}_0)\boldsymbol{E}_p\boldsymbol{U}_0^\top - D(\nabla f)(\boldsymbol{W}_0)[\tilde{\boldsymbol{V}}_p]\boldsymbol{W}_0 \\ &= -\nabla f(\boldsymbol{W}_0)(\boldsymbol{E}_p\boldsymbol{U}_0^\top + \boldsymbol{U}_0\boldsymbol{E}_p^\top) - D(\nabla f)(\boldsymbol{W}_0)[\tilde{\boldsymbol{V}}_p]\boldsymbol{W}_0 \\ &= -\nabla f(\boldsymbol{W}_0)\tilde{\boldsymbol{V}}_p - D(\nabla f)(\boldsymbol{W}_0)[\tilde{\boldsymbol{V}}_p]\boldsymbol{W}_0, \end{aligned}$$

where the second equality uses the fact that $\nabla f(\boldsymbol{U}_0\boldsymbol{U}_0^\top)\boldsymbol{U}_0 = \boldsymbol{0}$ since $\boldsymbol{U}_0$ is a critical point. Summing $\xi_p\boldsymbol{E}_p\boldsymbol{U}_0^\top = -\nabla f(\boldsymbol{W}_0)\tilde{\boldsymbol{V}}_p - D(\nabla f)(\boldsymbol{W}_0)[\tilde{\boldsymbol{V}}_p]\boldsymbol{W}_0$ with its transpose gives

$$\begin{aligned} \xi_p\tilde{\boldsymbol{V}}_p &= -\nabla f(\boldsymbol{W}_0)\tilde{\boldsymbol{V}}_p - \tilde{\boldsymbol{V}}_p\nabla f(\boldsymbol{W}_0) - D(\nabla f)(\boldsymbol{W}_0)[\tilde{\boldsymbol{V}}_p]\boldsymbol{W}_0 - \boldsymbol{W}_0 D(\nabla f)(\boldsymbol{W}_0)[\tilde{\boldsymbol{V}}_p] \\ &= D\boldsymbol{g}(\boldsymbol{W}_0)[\tilde{\boldsymbol{V}}_p], \end{aligned}$$

which proves that $\tilde{\boldsymbol{V}}_p$ is a right eigenvector associated with eigenvalue ξ_p.

Proof for Item 3. Since $\nabla f(\boldsymbol{W})$ is symmetric, $\boldsymbol{g}(\boldsymbol{W})$ is also symmetric. For any $\boldsymbol{\Delta} = -\boldsymbol{\Delta}^\top$,

$$D\boldsymbol{g}(\boldsymbol{U})[\boldsymbol{\Delta}] = D\boldsymbol{g}(\boldsymbol{U})[\boldsymbol{\Delta}^\top] = D\boldsymbol{g}(\boldsymbol{U})[-\boldsymbol{\Delta}].$$

So $D\boldsymbol{g}(\boldsymbol{U})[\boldsymbol{\Delta}] = \boldsymbol{0}$ and $\boldsymbol{\Delta}$ is an eigenvector associated with eigenvalue 0.

Proof for Jacobian is diagonalizable. Let $\mathcal{S}$ be the space of symmetric matrices and $\mathcal{A}$ be the space of antisymmetric matrices. It is easy to see that $\mathcal{S}$ and $\mathcal{A}$ are orthogonal to each other, and $\mathcal{S}$ and $\mathcal{A}$ are invariant subspaces of $D\boldsymbol{g}(\boldsymbol{U})[\boldsymbol{\Delta}]$. Let $\boldsymbol{h} : \mathcal{S} \to \mathcal{S}, \boldsymbol{\Delta} \mapsto D\boldsymbol{g}(\boldsymbol{U})[\boldsymbol{\Delta}]$ be the linear map $D\boldsymbol{g}(\boldsymbol{U})[\boldsymbol{\Delta}]$ restricted on symmetric matrices. We only need to prove that $\boldsymbol{h}$ is diagonalizable.

It is easy to see that $\{\hat{\boldsymbol{U}}_{ij}\}$ are linearly independent to each other and thus spans a subspace of $\mathcal{S}$ with dimension $\frac{(d-k)(d-k+1)}{2}$. Suppose that there exists scalars α_p for $1 \leq p \leq kd - k(k-1)/2$, not all zero, such that $\sum_{p=1}^{kd-k(k-1)/2} \alpha_p\tilde{\boldsymbol{V}}_p = \boldsymbol{0}$. Then $\sum_{p=1}^{kd-k(k-1)/2} \boldsymbol{E}_p$ is should be a solution of (24), which contradicts to the linear independence of $\{\boldsymbol{E}_p\}_{1\leq p\leq kd}$. Therefore, $\{\tilde{\boldsymbol{V}}_p\}$ spans a subspace of

$\mathcal{S}$ with dimension $kd - \frac{k(k-1)}{2}$. Note that

$$\frac{(d-k)(d-k+1)}{2} + \left(kd - \frac{k(k-1)}{2}\right) = \frac{d(d+1)}{2} = \dim(\mathcal{S}).$$

By Lemma D.3, $\boldsymbol{h}$ and thus $D\boldsymbol{g}(\boldsymbol{U})$ is diagonalizable. □

E Proofs for the Depth-2 Case

E.1 Proof for Theorem 5.6

Proof for Theorem 5.6. We view $\boldsymbol{W}(t)$ as a dynamical system following Section 5.1. By Lemma 5.4, we have $\tilde{\mu}_1 = 2\mu_1$, $\tilde{\boldsymbol{u}}_1 = \tilde{\boldsymbol{v}}_1 = \boldsymbol{u}_1\boldsymbol{u}_1^\top$, $\tilde{\mu}_2 = \max\{\mu_1 + \mu_2, 0\}$, where μ_1, μ_2 is the largest and the second largest eigenvalues of $-\nabla f(\mathbf{0})$ and $\boldsymbol{u}_1$ is an eigenvector associated with μ_1. So $\tilde{\gamma} := \tilde{\mu}_1 - \tilde{\mu}_2 = \min\{2\mu_1, \mu_1 - \mu_2\}$. Then it is easy to translate Theorem 5.3 to Theorem 5.6. □

E.2 Proof for Theorem 5.8

The proof for Theorem 5.8 relies on the following Lemma on the gradient flow around a local minimizer:

Lemma E.1. *If $\bar{\boldsymbol{\theta}}$ is a local minimizer of $\mathcal{L}(\boldsymbol{\theta})$ and for all $\left\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\right\|_2 \le r$, $\boldsymbol{\theta}$ satisfies Łojasiewicz inequality:*

$$\|\nabla \mathcal{L}(\boldsymbol{\theta})\|_2 \ge c\left(\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\bar{\boldsymbol{\theta}})\right)^\mu$$

for some $\mu \in [1/2, 1)$, then the gradient flow $\boldsymbol{\theta}(t) = \phi(\boldsymbol{\theta}_0, t)$ converges to a point $\boldsymbol{\theta}_\infty$ near $\bar{\boldsymbol{\theta}}$ if $\boldsymbol{\theta}_0$ is close enough to $\bar{\boldsymbol{\theta}}$, and the distance can be bounded by $\|\boldsymbol{\theta}_\infty - \bar{\boldsymbol{\theta}}\|_2 = O\left(\|\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}\|_2^{2(1-\mu)}\right)$.

Proof. For every $t \ge 0$, if $\|\boldsymbol{\theta}(t) - \bar{\boldsymbol{\theta}}\|_2 \le r$,

$$\begin{aligned}
\frac{d}{dt}\left(\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}(\bar{\boldsymbol{\theta}})\right)^{1-\mu} &= (1-\mu)\left(\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}(\bar{\boldsymbol{\theta}})\right)^{-\mu} \cdot \left\langle \nabla\mathcal{L}, \frac{d\boldsymbol{\theta}}{dt}\right\rangle \\
&= -(1-\mu)\left(\mathcal{L}(\boldsymbol{\theta}(t)) - \mathcal{L}(\bar{\boldsymbol{\theta}})\right)^{-\mu} \cdot \|\nabla\mathcal{L}\|_2 \cdot \left\|\frac{d\boldsymbol{\theta}}{dt}\right\|_2 \\
&\le -(1-\mu)c\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|_2.
\end{aligned}$$

Therefore $\|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_0\|_2 \le \int_0^t \left\|\frac{d\boldsymbol{\theta}}{dt}\right\|_2 \le \frac{1}{(1-\mu)c}\mathcal{L}(\boldsymbol{\theta}_0)^{1-\mu} = O\left(\|\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}\|_2^{2(1-\mu)}\right)$. If we choose $\|\boldsymbol{\theta}(t) - \bar{\boldsymbol{\theta}}\|_2$ small enough, then $\|\boldsymbol{\theta}(t) - \bar{\boldsymbol{\theta}}\|_2 \le \|\boldsymbol{\theta}(t) - \boldsymbol{\theta}_0\|_2 + \|\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}\|_2 = O(\|\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}\|_2^{2(1-\mu)}) < r$, and thus $\int_0^{+\infty}\left\|\frac{d\boldsymbol{\theta}}{dt}\right\|_2$ is convergent and finite. This implies that $\boldsymbol{\theta}_\infty := \lim_{t\to+\infty}\boldsymbol{\theta}(t)$ exists and $\|\boldsymbol{\theta}_\infty - \bar{\boldsymbol{\theta}}\|_2 = O\left(\|\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}\|_2^{2(1-\mu)}\right)$. □

Proof for Theorem 5.8. As $\boldsymbol{W}_1^{\mathrm{G}}(t)$ satisfies (2), there exists $\boldsymbol{u}(t) \in \mathbb{R}^d$ such that $\boldsymbol{u}(t)\boldsymbol{u}(t)^\top = \boldsymbol{W}_1^{\mathrm{G}}(t)$ and $\boldsymbol{u}(t)$ satisfies (1). If $\boldsymbol{W}_1^{\mathrm{G}}(t)$ does not diverge to infinity, then so does $\boldsymbol{u}(t)$. This also implies that there is a limit point $\bar{\boldsymbol{u}}$.

Using a similar argument as in Lemma E.1, we can show that when $\boldsymbol{u}(t)$ is sufficiently close to $\bar{\boldsymbol{u}}$, the remaining length of the trajectory is finite. So $\lim_{t\to+\infty}\boldsymbol{u}(t) = \bar{\boldsymbol{u}}$. Therefore, $\overline{\boldsymbol{W}}_1 := \lim_{t\to+\infty}\boldsymbol{W}_1^{\mathrm{G}}(t) = \bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^\top$ exists.

If $\overline{\boldsymbol{W}}_1$ is a local minimizer of $f(\,\cdot\,)$, $\overline{\boldsymbol{U}} = (\bar{\boldsymbol{u}}, \mathbf{0}, \cdots, \mathbf{0}) \in \mathbb{R}^{d\times d}$ is also a local minimizer of $\mathcal{L}(\,\cdot\,)$. By analyticity of $f(\,\cdot\,)$, Łojasiewicz inequality holds for $\mathcal{L}(\,\cdot\,)$ around $\overline{\boldsymbol{U}}$ (Łojasiewicz, 1965), i.e., for all $\left\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\right\|_2 \le r$,

$$\|\nabla\mathcal{L}(\boldsymbol{U})\|_2 \ge c\left(\mathcal{L}(\boldsymbol{U}) - \mathcal{L}(\bar{\boldsymbol{u}}\bar{\boldsymbol{u}}^\top)\right)^\mu$$

for some $\mu \in [1/2, 1)$. For every $\epsilon > 0$, we can always find a time t_ϵ such that $\|\boldsymbol{u}(t_0) - \bar{\boldsymbol{u}}\|_2 \le \epsilon$. By Theorem 5.6, there also exists a number $\bar{\alpha}$ such that for all $\alpha < \bar{\alpha}$,

$$\left\|\boldsymbol{W}_\alpha(t_0) - \boldsymbol{W}_1^{\mathrm{G}}(t_0)\right\|_2 \le \epsilon,$$

where $\boldsymbol{W}_\alpha(t) := \phi\left(\boldsymbol{W}_\alpha, \frac{1}{2\mu_1}\log\frac{1}{\langle \boldsymbol{W}_\alpha, \boldsymbol{u}_1\boldsymbol{u}_1^\top\rangle} + t\right)$. For any $\boldsymbol{W}_\alpha(t_0)$, it is easy to construct a decomposition $\boldsymbol{W}_\alpha(t_0) := \boldsymbol{U}_\alpha(t_0)\boldsymbol{U}_\alpha(t_0)^\top$ such that $\left\|\boldsymbol{U}_\alpha(t_0) - \overline{\boldsymbol{U}}\right\|_2 = O(\epsilon)$, e.g., find an arbitrary decomposition and then right-multiply an orthogonal matrix so that the row vector with the largest norm aligns with the direction of $\bar{\boldsymbol{u}}$.

Taking $\epsilon \to 0$ and applying Lemma E.1 completes the proof. □

E.3 Proof for Theorem 5.11

Definition E.2. For a critical point $\overline{\boldsymbol{W}}$ of (2), we say that $\overline{\boldsymbol{W}}$ is a k-local-optimal saddle point if $\overline{\boldsymbol{W}}$ can be represented by $\overline{\boldsymbol{W}} = \overline{\boldsymbol{U}}\overline{\boldsymbol{U}}^\top$ for some $\overline{\boldsymbol{U}} \in \mathbb{R}^{d\times k}$ and $\overline{\boldsymbol{U}}$ is a local minimizer of $\mathcal{L}(\boldsymbol{U}) := \frac{1}{2}f(\boldsymbol{U}\boldsymbol{U}^\top)$, but $\overline{\boldsymbol{W}}$ itself is not a local minimizer of $f(\boldsymbol{W})$ among all PSD matrices.

For every k-local-optimal saddle point $\overline{\boldsymbol{W}}$, if $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_{d-k}$ are the eigenvalues of $-\nabla f(\overline{\boldsymbol{W}})$ (excluding k zeros as $-\nabla f(\overline{\boldsymbol{W}})$ must have rank $\le d-k$; see (23)), then $\mu_1 \ge 0$. Based on this, we define the following notion of non-degenerate k-local-optimal saddle point.

Definition E.3. We say that a k-local-optimal saddle point $\overline{\boldsymbol{W}}$ is non-degenerate if both of the following two conditions hold:

1. $\mu_1 > \max\{\mu_2, 0\}$;
2. $\mu_i + \mu_j \ne \xi_p$ for any $1 \le i \le j \le d-k$ and $1 \le p \le kd - \frac{k(k-1)}{2}$, where ξ_p is defined according to Definition D.2.

Theorem E.4. *Let $\overline{\boldsymbol{W}}$ be a non-degenerate k-local-optimal saddle point. If we can find time T_α for each α so that $\phi(\boldsymbol{W}_\alpha, T_\alpha)$ converges to $\overline{\boldsymbol{W}}$ with positive alignment with $\boldsymbol{v}_1\boldsymbol{v}_1^\top$ as $\alpha \to 0$, then for every fixed t, $\lim_{\alpha\to 0}\phi\left(\boldsymbol{W}_\alpha, T_\alpha + \frac{1}{2\mu_1}\log\frac{1}{\langle\phi(\boldsymbol{W}_\alpha, T_\alpha), \boldsymbol{v}_1\boldsymbol{v}_1^\top\rangle} + t\right)$ exists and is equal to $\boldsymbol{W}^{\mathrm{G}}(t) := \lim_{\epsilon\to 0}\phi(\overline{\boldsymbol{W}} + \epsilon\boldsymbol{v}_1\boldsymbol{v}_1^\top, \frac{1}{2\mu_1}\log\frac{1}{\epsilon} + t)$.*

Proof. By (21), the symmetric matrices $-\nabla f(\boldsymbol{W}_0)$ and $\boldsymbol{W}_0$ commute, so $-\nabla f(\boldsymbol{W}_0)$ has rank $\le d-k$. By Theorem D.4, $\boldsymbol{J}(\boldsymbol{W})$ is diagonalizable and has eigenvalues $\mu_i + \mu_j, \xi_p, 0$. Since $\overline{\boldsymbol{U}}$ is a local minimizer, $\xi_p \le 0$ for all p. Thus, if $\mu_1 > \max\{\mu_2, 0\}$, then $2\mu_1$ is the unique largest eigenvalue of $\boldsymbol{J}(\boldsymbol{W})$, and Theorem D.4 shows that $\boldsymbol{v}_1\boldsymbol{v}_1^\top$ is a left eigenvector associated with $2\mu_1$. Combining these with Theorem 5.3 completes the proof. □

E.4 Gradient Flow only finds minimizers for overparametrized matrix sensing (Proof for Theorem 5.10)

The proof for Theorem 5.10 is based on the following two theorems from the literature.

Theorem E.5 (Theorem 3.1 in Du and Lee 2018)**.** *Let $f : \mathbb{R}^{d\times d} \to \mathbb{R}$ be a $\mathcal{C}^2$ convex function. Then $\mathcal{L} : \mathbb{R}^{d\times k} \to \mathbb{R}$, $\mathcal{L}(\boldsymbol{U}) = f(\boldsymbol{U}\boldsymbol{U}^\top), k \ge d$ satisfies that (1). Every local minimizer of $\mathcal{L}$ is also a global minimizer; (2). All saddles are strict. Here saddles denote those stationary points whose hessian are not positive semi-definite (thus including local maximizers).* [1]

Theorem E.6 (Theorem 2 in Lee et al. 2017)**.** *Let $\boldsymbol{g}$ be a $\mathcal{C}^1$ mapping from $\mathcal{X} \to \mathcal{X}$ and $\det(D\boldsymbol{g}(x)) \ne 0$ for all $\boldsymbol{x} \in \mathcal{X}$. Then the set of initial points that converge to an unstable fixed point has measure zero, $\mu\left(\{\boldsymbol{x}_0 : \lim_{k\to\infty}\boldsymbol{g}^k(\boldsymbol{x}_0) \in \mathcal{A}_{\boldsymbol{g}}^*\}\right) = 0$, where $\mathcal{A}_{\boldsymbol{g}}^* = \{\boldsymbol{x} : \boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x}, \max_i|\lambda_i(D\boldsymbol{g}(\boldsymbol{x}))| > 1\}$.*

[1] Though the original theorem is proven for convex functions of form $\sum_{i=1}^n \ell(\boldsymbol{x}_i\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{x}_i^\top, y_i)$, where $\ell(\cdot,\cdot)$ is $\mathcal{C}^2$ convex for its first variable. By scrutinizing their proof, we can see the assumption can be relaxed to f is $\mathcal{C}^2$ convex.

Theorem E.7 (GF only finds minimizers, Continuous analog of Theorem E.6)**.** *Let $\boldsymbol{f}:\mathbb{R}^d\to\mathbb{R}^d$ be a $\mathcal{C}^1$-smooth function, and $\phi:\mathbb{R}^d\times\mathbb{R}\to\mathbb{R}^d$ be the solution of the following differential equation,*

$$\frac{\mathrm{d}\phi(\boldsymbol{x},t)}{\mathrm{d}t}=\boldsymbol{f}(\phi(\boldsymbol{x},t)),\quad \phi(\boldsymbol{x},0)=\boldsymbol{x},\quad \forall \boldsymbol{x}\in\mathbb{R}^d, t\in\mathbb{R}.$$

Then the set of initial points that converge to a unstable fixed point of $\boldsymbol{f}$ has measure zero, $\mu\left(\left\{\boldsymbol{x}_0:\lim_{t\to\infty}\phi(\boldsymbol{x}_0,t)\in\mathcal{U}_{\boldsymbol{f}}^*\right\}\right)=0$, where $\mathcal{U}_{\boldsymbol{f}}^*=\{\boldsymbol{x}:\boldsymbol{f}(\boldsymbol{x})=0,\lambda_1(D\boldsymbol{f}(\boldsymbol{x}))>0\}$.

Proof of Theorem E.7. By Theorem 1 in Section 2.3, Perko (2013), we know $\phi(\cdot,\cdot)$ is $\mathcal{C}^1$-smooth for both x,t. We let $\boldsymbol{g}(x)=\phi(x,1)$, then we know $\boldsymbol{g}^{-1}(x)=\phi(x,-1)$ and both $\boldsymbol{g},\boldsymbol{g}^{-1}$ are $\mathcal{C}^1$-smooth. Note that $D\boldsymbol{g}^{-1}(x)$ is the inverse matrix of $D\boldsymbol{g}(x)$. So both of the two matrices are invertible. Thus we can apply Theorem E.6 and we know $\mu\left(\{x_0:\lim_{k\to\infty}\boldsymbol{g}^k(x_0)\in\mathcal{A}_{\boldsymbol{g}}^*\}\right)=0$.

Note that if $\lim_{t\to\infty}\phi(x,t)$ exists, then $\lim_{k\to\infty}\boldsymbol{g}^k(\boldsymbol{x})=\lim_{t\to\infty}\phi(\boldsymbol{x},t)$. It remains to show that $\mathcal{U}_{\boldsymbol{f}}^*\subseteq\mathcal{A}_{\boldsymbol{g}}^*$. For $\boldsymbol{f}(\boldsymbol{x}_0)=\boldsymbol{0}$, we have $\phi(\boldsymbol{x}_0,t)=\boldsymbol{x}_0$ and thus $\boldsymbol{g}(\boldsymbol{x}_0)=\boldsymbol{x}_0$. Now it suffices to prove that $\lambda_1(D\boldsymbol{g}(\boldsymbol{x}_0))>1$. For every $t\in[0,1]$, by Corollary of Theorem 1 in Section 2.3, Perko (2013), we have $\frac{\partial}{\partial t}D\phi(\boldsymbol{x},t)=D\boldsymbol{f}(\phi(\boldsymbol{x},t))D\phi(\boldsymbol{x},t)$, $\forall\boldsymbol{x},t$. Thus,

$$\frac{\partial}{\partial t}D\phi(x_0,t)=D\boldsymbol{f}(\phi(x_0,t))D\phi(x_0,t)=D\boldsymbol{f}(x_0)D\phi(x_0,t).$$

Solving this ODE gives $D\boldsymbol{g}(\boldsymbol{x}_0)=D\phi(\boldsymbol{x},1)=e^{D\boldsymbol{f}(x_0)}D\phi(\boldsymbol{x},0)=e^{D\boldsymbol{f}(x_0)}$, where the last equality is because $D\phi(\boldsymbol{x},0)\equiv\boldsymbol{I}$, $\forall\boldsymbol{x}$. Combining this with $\lambda_1(D\boldsymbol{f}(x_0))>0$ by the definition of strict saddle point, we have $\lambda_1(D\boldsymbol{g}(x_0))>1$.

Thus we have $\mathcal{U}_{\boldsymbol{f}}^*:=\{\boldsymbol{x}_0:\boldsymbol{f}(\boldsymbol{x}_0)=\boldsymbol{0},\lambda_1(D\boldsymbol{f}(\boldsymbol{x}_0))>0\}\subseteq\mathcal{A}_{\boldsymbol{g}}^*$, which implies that $\{\boldsymbol{x}_0:\lim_{t\to\infty}\phi(\boldsymbol{x}_0,t)\in\mathcal{U}^*\}\subseteq\{\boldsymbol{x}_0:\lim_{k\to\infty}\boldsymbol{g}^k(\boldsymbol{x}_0)\in\mathcal{A}_{\boldsymbol{g}}^*\}$ □

Theorem 5.10. *Let $f:\mathbb{R}^{d\times d}\to\mathbb{R}$ be a convex $\mathcal{C}^2$-smooth function. (1). All stationary points of $\mathcal{L}:\mathbb{R}^{d\times d}\to\mathbb{R},\mathcal{L}(\boldsymbol{U})=\frac{1}{2}f(\boldsymbol{U}\boldsymbol{U}^\top)$ are either strict saddles or global minimizers; (2). For any random initialization, GF (1) converges to strict saddles of $\mathcal{L}(\boldsymbol{U})$ with probability* 0.

Proof of Theorem 5.10. For (1), by Theorem E.5, we immediately know all the stationary points of $\mathcal{L}(\,\cdot\,)$ are either global minimizers or strict saddles. (2) is just a direct consequence of Theorem E.7 by setting $\boldsymbol{f}$ in the above proof to $-\nabla\mathcal{L}$. □

F Equivalence Between GF and GLRL

Now we provide theoretical evidence that GF and GLRL are equivalent in the case where GLRL does not end in the first phase.

However, our argument is not completely rigorous. The difficulty here is that there could be many cases that can occur mathematically but rarely appear in practice. So we turn to cases that hold "generically". We use this word in the following situations:

1. We want to assume that GF converges to a local minimizer (i.e., GF does not get stuck on saddle points);
2. We want to assume that a k-local-optimal saddle point $\overline{\boldsymbol{W}}$ of $f(\boldsymbol{W})$ is non-degenerate (see Appendix E.3);
3. We want to assume that a sequence of PSD matrices converging to a PSD has positive alignment with $\boldsymbol{v}\boldsymbol{v}^\top$ for some fixed vector $\boldsymbol{v}$.

Theorem E.4 uncovers how GF with infinitesimal initialization generically behave. Let $\overline{\boldsymbol{W}}_0:=\boldsymbol{0}$. For every $r\geq 1$, if $\overline{\boldsymbol{W}}_{r-1}$ is a k-local-optimal saddle point, it is non-degenerate generically. Then we immediately know that the limiting trajectory $\boldsymbol{W}_r^{\mathrm{G}}(t)=\lim_{\epsilon\to 0}\phi(\overline{\boldsymbol{W}}_{r-1}+\epsilon\boldsymbol{v}_1\boldsymbol{v}_1^\top,\frac{1}{2\mu_1}\log\frac{1}{\epsilon}+t)$ exists, and it is generically the same as GF initialized sufficiently close to $\overline{\boldsymbol{W}}_{r-1}$. Thus, if there exists T_α ($\forall\alpha$) such that $\phi(\boldsymbol{W}_\alpha,T_\alpha)$ converges to $\overline{\boldsymbol{W}}_{r-1}$ with positive alignment with $\boldsymbol{v}_1\boldsymbol{v}_1^\top$ as $\alpha\to 0$, then the trajectory of GF after T_α is close to the trajectory of $\boldsymbol{W}_r^{\mathrm{G}}(t)$, and it can get arbitrarily close

to the convergent point $\overline{\boldsymbol{W}}_r := \lim_{t\to+\infty} \boldsymbol{W}_r^{\mathrm{G}}(t)$, which is generically a local minimizer among PSD matrices with rank $\leq r$.

If $\overline{\boldsymbol{W}}_r$ is indeed a local minimizer among all PSD matrices, then $\lambda_1(-\nabla f(\overline{\boldsymbol{W}}_r)) \leq 0$ and GLRL returns $\overline{\boldsymbol{W}}_r$. With a similar argument for Theorem 5.8, it can be shown that GF eventually converges to $\overline{\boldsymbol{W}}_r$. If not, $\overline{\boldsymbol{W}}_r$ is a k-local-optimal saddle point, and there is generically a suitable choice of T'_α so that $\phi(\boldsymbol{W}_\alpha, T'_\alpha)$ converges to $\overline{\boldsymbol{W}}_r$ with positive alignment with $\boldsymbol{v}'_1\boldsymbol{v}'^{\top}_1$, where $\boldsymbol{v}'_1$ is the top eigenvector of $-\nabla f(\overline{\boldsymbol{W}}_r)$. Then the process will continue until it finds a local minimizer among all PSD matrices.

The above statement establishes the equivalence between GLRL and GF with infinitesimal initialization generically. In our experiments, we also see that such equivalence holds for matrix completion with random data (Figure 1). Therefore, we conjecture that such equivalence is general, and towards this it requires a smoothed analysis on $f(\,\cdot\,)$. We leave it as a future work.

G Proofs for Deep Matrix Factorization

G.1 Preliminary Lemmas

Lemma G.1. *If $\boldsymbol{W}(0)$ is PSD, then $\boldsymbol{W}(t)$ is also PSD and $\mathrm{rank}(\boldsymbol{W}(t)) = \mathrm{rank}(\boldsymbol{W}(0))$ for any t.*

Proof of Lemma G.1. Note that we can always find a set of balanced $\boldsymbol{U}_i(t)$, such that $\boldsymbol{U}_1(t)\dots\boldsymbol{U}_L(t) = \boldsymbol{W}(t)$, $d_2 = d_3 = \cdots = d_L = \mathrm{rank}(\boldsymbol{W}(t))$ and write the dynamics of $\boldsymbol{W}(t)$ in the space of $\{\boldsymbol{U}_i\}_{i=1}^L$. Thus it's clear that for all t', $\mathrm{rank}(\boldsymbol{W}(t')) \leq \mathrm{rank}(\boldsymbol{W}(t))$. We can apply the same argument for t' and we know $\mathrm{rank}(\boldsymbol{W}(t)) \leq \mathrm{rank}(\boldsymbol{W}(t'))$. Thus $\mathrm{rank}(\boldsymbol{W}(t))$ is constant over time, and we denote it by k.

Since eigenvalues are continuous matrix functions, and $\forall t, \lambda_i(\boldsymbol{W}(t)),\ i \in [k] \neq 0$. Thus they cannot change their signs and $\boldsymbol{W}(t)$ must be PSD. □

Since $\boldsymbol{W}(t)$ is PSD, (9) can be rewritten as the following:

$$\frac{d\boldsymbol{W}}{dt} = -\sum_{i=0}^{L-1} \boldsymbol{W}^{\frac{2i}{L}} \nabla f(\boldsymbol{W}) \boldsymbol{W}^{2-\frac{2i+2}{L}}. \tag{25}$$

Proof for Lemma 6.1. Suppose $\boldsymbol{W}(t)$ is a symmetric solution of Equation (9). By Lemma G.1, we know $\boldsymbol{W}(t)$ also satisfies Equation (25). Now we let $\boldsymbol{R}(t)$ be the solution of the following ODE with $\boldsymbol{R}(0) := (\boldsymbol{W}(0))^{\frac{1}{L}}$. Note we don't define $\boldsymbol{R}(t)$ by $(\boldsymbol{W}(t))^{\frac{1}{L}}$.

$$\frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}t} = -\sum_{i=0}^{L-1} (-1)^i \boldsymbol{R}^i \nabla f(\boldsymbol{R}^L) \boldsymbol{R}^{L-1-i}. \tag{26}$$

The calculation below shows that $\boldsymbol{R}^L(t)$ also satisfies (25).

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{R}^L}{\mathrm{d}t} &= \sum_{j=0}^{L-1} \boldsymbol{R}^j \frac{\mathrm{d}\boldsymbol{R}}{\mathrm{d}t} \boldsymbol{R}^{L-1-j} = \sum_{j=0}^{L-1}\sum_{i=0}^{L-1} (-1)^i \boldsymbol{R}^{i+j} \nabla f(\boldsymbol{R}^L) \boldsymbol{R}^{2L-2-i-j} \\
&= \sum_{i=0}^{2L-2} \left(\sum_{j=0}^{i} (-1)^j \right) \boldsymbol{R}^i \nabla f(\boldsymbol{R}^L) \boldsymbol{R}^{2T-2-i} \\
&= \sum_{i=0}^{L-1} (\boldsymbol{R}^L)^{\frac{2i}{L}} \nabla f(\boldsymbol{R}^L) (\boldsymbol{R}^L)^{2-\frac{2+2i}{L}}.
\end{aligned}$$

Since $\boldsymbol{R}^L(0) = \boldsymbol{W}(0)$, by existence and uniqueness theorem, $\boldsymbol{R}^L(t) = \boldsymbol{W}(t), \forall t \in \mathbb{R}$. So

$$\frac{\mathrm{d}\boldsymbol{M}}{\mathrm{d}t} = \boldsymbol{R}\frac{\mathrm{d}\boldsymbol{R}}{dt} + \frac{\mathrm{d}\boldsymbol{R}}{dt}\boldsymbol{R} = -\nabla f(\boldsymbol{M}^{L/2})\boldsymbol{M}^{L/2} - \boldsymbol{M}^{L/2}\nabla f(\boldsymbol{M}^{L/2}),$$

which completes the proof. □

G.2 Proof for Theorem 6.2

Now we turn to prove Theorem 6.2.

Let Sym_d be the inner product space of $d \times d$ real symmetric matrix. Let $\lambda_m : \mathrm{Sym}_d \to \mathbb{R}, \boldsymbol{M} \mapsto \lambda_m(\boldsymbol{M})$ be the m-th largest eigenvalue of a symmetric matrix $\boldsymbol{M}$. The following theorem gives the Clarke's derivatives of the eigenvalue:

Theorem G.2 ((Hiriart-Urruty and Lewis, 1999))**.** *The subdifferential of the eigenvalue in the Clarke's sense is given below:*

$$\partial^\circ \lambda_m(\boldsymbol{M}) = \mathrm{conv}\{\boldsymbol{v}\boldsymbol{v}^\top : \boldsymbol{M}\boldsymbol{v} = \lambda_m(\boldsymbol{M})\boldsymbol{v}, \|\boldsymbol{v}\|_2 = 1\}.$$

Let $P = L/2$. Then (10) can be rewritten as

$$\frac{d\boldsymbol{M}}{dt} = -\left(\nabla f(\boldsymbol{M}^P)\boldsymbol{M}^P + \boldsymbol{M}^P \nabla f(\boldsymbol{M}^P)\right). \tag{27}$$

The following lemma about the growth rate of $\lambda_k(\boldsymbol{M})$ would be useful.

Lemma G.3. *Suppose $\boldsymbol{M}(t)$ satisfies Equation (27), we have for any $T' > T$, and $k \in [d]$,*

$$|\lambda_k(\boldsymbol{M}(T')) - \lambda_k(\boldsymbol{M}(T))| \le \int_{t=T}^{T'} 2\lambda_k(\boldsymbol{M}(t))^P \left\|\nabla f(\boldsymbol{M}^P(t))\right\|_2 \mathrm{d}t. \tag{28}$$

and

$$\frac{1}{P-1}\left|\lambda_k^{1-P}(\boldsymbol{M}(T')) - \lambda_k^{1-P}(\boldsymbol{M}(T))\right| \le \int_{t=T}^{T'} 2\left\|\nabla f(\boldsymbol{M}^P(t))\right\|_2 \mathrm{d}t. \tag{29}$$

Proof. Since $\lambda_k(\boldsymbol{M}(t))$ is locally Lipschitz in t, by Rademacher's theorem, we know $\lambda_k(\boldsymbol{M}(t))$ is differentiable almost everywhere, and the following holds

$$|\lambda_k(\boldsymbol{M}(T')) - \lambda_k(\boldsymbol{M}(T))| \le \int_{t=T}^{T'} \frac{\mathrm{d}\lambda_k(\boldsymbol{M}(t))}{\mathrm{d}t}\mathrm{d}t.$$

Thus,

$$\begin{aligned}\frac{d}{dt}\lambda_k(\boldsymbol{M}(t)) &\in \left\{\left\langle \boldsymbol{G}, \frac{d\boldsymbol{M}(t)}{dt}\right\rangle : \boldsymbol{G} \in \partial^\circ \lambda_k(\boldsymbol{M}(t))\right\} \\ &= \left\{-2\lambda_k(\boldsymbol{M}(t))^P \left\langle \boldsymbol{G}, \nabla f(\boldsymbol{M}(t)^P)\right\rangle : \boldsymbol{G} \in \partial^\circ \lambda_k(\boldsymbol{M}(t))\right\}\end{aligned}$$

Note that $\|\boldsymbol{G}\|_{\mathrm{F}} \le \|\boldsymbol{G}\|_* = 1$, we have $|\left\langle \boldsymbol{G}, \nabla f(\boldsymbol{M}(t)^P)\right\rangle| \le \left\|\nabla f(\boldsymbol{M}^P(t))\right\|_2$. More specifically, it's $\lambda_{\min}(\nabla f(\boldsymbol{M}^P(t))) \le \left\langle \boldsymbol{G}, \nabla f(\boldsymbol{M}(t)^P)\right\rangle \le \lambda_{\max}(\nabla f(\boldsymbol{M}^P(t)))$.

With a similar argument, we can prove Equation (29).

□

To prove Theorem 6.2, it suffices to consider the case that $\boldsymbol{M}(0) = \hat{\alpha}\boldsymbol{I}$ where $\hat{\alpha} := \alpha^{1/P}$. WLOG we can assume $-\nabla f(\boldsymbol{0}) = \mathrm{diag}(\mu_1, \ldots, \mu_d)$ by choosing a suitable standard basis. We use $\phi_m(\boldsymbol{M}_0, t)$ to denote the solution of $\boldsymbol{M}(t)$ when $\boldsymbol{M}(0) = \boldsymbol{M}_0$.

Let $R > 0$. Since $f(\cdot)$ is $\mathcal{C}^3$-smooth, there exists $\beta > 0$ such that $\|\nabla f(\boldsymbol{W}_1) - \nabla f(\boldsymbol{W}_2)\|_2 \le \beta \|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_2$ for all $\boldsymbol{W}_1, \boldsymbol{W}_2$ with $\|\boldsymbol{W}_1\|_2, \|\boldsymbol{W}_2\|_2 \le R$.

Let $\kappa = \beta/\mu_1$. We assume WLOG that $R \le \frac{1}{\kappa(P-1)}$.

Let $F_{\hat{\alpha}}(x) := \int_{x^{-(P-1)}}^{\hat{\alpha}^{-(P-1)}} \frac{dz}{1+\kappa z^{-P/(P-1)}}$. Then $F'_{\hat{\alpha}}(x) = \frac{(P-1)x^{-P}}{1+\kappa x^P} = \frac{P-1}{(1+\kappa x^P)x^P}$. We will use this function to bound norm growth. Let $g_{\hat{\alpha},c}(t) = \frac{1}{\hat{\alpha}^{-(P-1)} - \kappa(P-1)c - 2\mu_1(P-1)t}$. Define $T_{\hat{\alpha}}(r) =$

$\frac{\hat{\alpha}^{-(P-1)}-\kappa(P-1)r-r^{-(P-1)}}{2\mu_1(P-1)}$. It is easy to verify that $g_{\hat{\alpha},r}(T_{\hat{\alpha}}(r)) = r^{P-1}$. In section, we will use $\phi_{mm}(\boldsymbol{M}(0), t)$ to denote $\boldsymbol{M}(t)$, which is the solution of Equation (27).

Lemma G.4. *For any $x \in [\hat{\alpha}, R]$ we have*

$$\left(\hat{\alpha}^{-(P-1)} - x^{-(P-1)}\right) - F_{\hat{\alpha}}(x) \in [0, \kappa(P-1)x].$$

Proof. On the one hand, we have

$$\hat{\alpha}^{-(P-1)} - x^{-(P-1)} - F_{\hat{\alpha}}(x) = \int_{x^{-(P-1)}}^{\hat{\alpha}^{-(P-1)}} \left(1 - \frac{1}{1+\kappa z^{-P/(P-1)}}\right) dz \geq 0.$$

On the other hand,

$$\begin{aligned}
\hat{\alpha}^{-(P-1)} - x^{-(P-1)} - F_{\hat{\alpha}}(x) &= \int_{x^{-(P-1)}}^{\hat{\alpha}^{-(P-1)}} \frac{\kappa}{z^{P/(P-1)}+\kappa} dz \\
&\leq \kappa \int_{x^{-(P-1)}}^{\hat{\alpha}^{-(P-1)}} \frac{1}{z^{P/(P-1)}} dz \\
&= \kappa(P-1) \cdot \frac{-1}{z^{1/(P-1)}}\Bigg|_{x^{-(P-1)}}^{\hat{\alpha}^{-(P-1)}} \\
&\leq \kappa(P-1)x,
\end{aligned}$$

which completes the proof. □

Lemma G.5. *Let $\boldsymbol{M}_0$ be a PSD matrix with $\|\boldsymbol{M}_0\|_2 \leq 1$. For $\boldsymbol{M}(t) := \phi_m(\hat{\alpha}\boldsymbol{M}_0, t)$ and $t \leq T_{\hat{\alpha}}(c)$,*

$$\|\boldsymbol{M}(t)\|_2 \leq g_{\hat{\alpha},c}(t)^{\frac{1}{P-1}}.$$

Proof. Since $\left\|\nabla f(\boldsymbol{M}^P)\right\|_2 \leq \mu_1 + \beta\|\boldsymbol{M}\|_2^P$, by Lemma G.3, we have

$$\begin{aligned}
\|\boldsymbol{M}(t)\|_2 &\leq \|\boldsymbol{M}(0)\|_2 + \int_0^t 2(\mu_1 + \beta\|\boldsymbol{M}(\tau)\|_2^P)\|\boldsymbol{M}(\tau)\|_2^P d\tau \\
&= \hat{\alpha} + 2\mu_1(P-1)\int_0^t \frac{d\tau}{F'_{\hat{\alpha}}(\|\boldsymbol{M}(\tau)\|_2)}
\end{aligned}$$

So

$$F_{\hat{\alpha}}(\|\boldsymbol{M}(t)\|_2) \leq 2\mu_1(P-1)t.$$

If $\|\boldsymbol{M}(t)\|_2 < \hat{\alpha}$, then $\|\boldsymbol{M}(t)\|_2 \leq g_{\hat{\alpha},c}(t)^{\frac{1}{P-1}}$. If $\|\boldsymbol{M}(t)\|_2 \geq \hat{\alpha}$, then by Lemma G.4,

$$F_{\hat{\alpha}}(\|\boldsymbol{M}(t)\|_2) \leq 2\mu_1(P-1)T_{\hat{\alpha}}(c) = \hat{\alpha}^{-(P-1)} - \kappa(P-1)c - c^{-(P-1)} \leq F_{\hat{\alpha}}(c),$$

so $\|\boldsymbol{M}(t)\|_2 \leq c$ for all $t \leq T_{\hat{\alpha}}(c)$. Applying Lemma G.4 again, we have

$$\hat{\alpha}^{-(P-1)} - \|\boldsymbol{M}(t)\|_2^{-(P-1)} \leq F(\|\boldsymbol{M}(t)\|_2) + \kappa(P-1)c \leq 2\mu_1(P-1)t + \kappa(P-1)c,$$

which implies $\|\boldsymbol{M}(t)\|_2 \leq g_{\hat{\alpha},c}(t)^{\frac{1}{P-1}}$ by definition. □

Consider the following ODE:

$$\frac{d\widehat{\boldsymbol{M}}}{dt} = -\left(\nabla f(\boldsymbol{0})\widehat{\boldsymbol{M}}^P + \widehat{\boldsymbol{M}}^P \nabla f(\boldsymbol{0})\right).$$

We use $\hat{\phi}_m(\widehat{\boldsymbol{M}}_0, t)$ to denote the solution of $\widehat{\boldsymbol{M}}(t)$ when $\widehat{\boldsymbol{M}}(0) = \widehat{\boldsymbol{M}}_0$. For diagonal matrix $\widehat{\boldsymbol{M}}_0$, $\widehat{\boldsymbol{M}}(t)$ is also diagonal for any t, and it is easy to show that

$$\boldsymbol{e}_i^\top \widehat{\boldsymbol{M}}(t)\boldsymbol{e}_i = \begin{cases} \left(\frac{1}{\left(\hat{\alpha}\boldsymbol{e}_i^\top \widehat{\boldsymbol{M}}_0\boldsymbol{e}_i\right)^{-(P-1)} - 2\mu_i(P-1)t}\right)^{\frac{1}{P-1}} & \boldsymbol{e}_i^\top \widehat{\boldsymbol{M}}_0\boldsymbol{e}_i \neq 0, \\ 0 & \boldsymbol{e}_i^\top \widehat{\boldsymbol{M}}_0\boldsymbol{e}_i = 0. \end{cases} \tag{30}$$

The following lemma shows that the trajectory of $\boldsymbol{M}(t)$ is close to $\widehat{\boldsymbol{M}}(t)$.

Lemma G.6. *Let $\boldsymbol{M}_0$ be a diagonal PSD matrix with $\|\boldsymbol{M}_0\|_2 \leq 1$. For $\boldsymbol{M}(t) := \phi_m(\hat{\alpha}\boldsymbol{M}_0, t)$ and $\widehat{\boldsymbol{M}}(t) := \hat{\phi}_m(\hat{\alpha}\boldsymbol{M}_0, t)$, we have*

$$\|\boldsymbol{M}(T_{\hat{\alpha}}(r)) - \widehat{\boldsymbol{M}}(T_{\hat{\alpha}}(r))\|_2 = O(r^{P+1}).$$

Proof. We bound the difference $\boldsymbol{D} := \boldsymbol{M} - \widehat{\boldsymbol{M}}$ between $\boldsymbol{M}$ and $\widehat{\boldsymbol{M}}$. For any unit vector $\boldsymbol{v}$,

$$\begin{aligned}
\frac{d}{dt}(\boldsymbol{v}^\top \boldsymbol{D}\boldsymbol{v}) &= -2\left(\boldsymbol{v}^\top \nabla f(\boldsymbol{0})\left(\boldsymbol{M}^P - \widehat{\boldsymbol{M}}^P\right)\boldsymbol{v} + \boldsymbol{v}^\top\left(\nabla f(\boldsymbol{M}^P) - \nabla f(\boldsymbol{0})\right)\boldsymbol{M}^P\boldsymbol{v}\right) \\
&\leq 2\left(\|\nabla f(\boldsymbol{0})\|_2\|\boldsymbol{M}^P - \widehat{\boldsymbol{M}}^P\|_2 + \|\nabla f(\boldsymbol{M}^P) - \nabla f(\boldsymbol{0})\|_2\|\boldsymbol{M}^P\|_2\right) \\
&\leq 2\left(\lambda_1 P \max\{\|\boldsymbol{M}\|_2^{P-1}, \|\widehat{\boldsymbol{M}}\|_2^{P-1}\}\|\boldsymbol{D}\|_2 + \beta\|\boldsymbol{M}\|_2^{2P}\right),
\end{aligned}$$

This implies that

$$\begin{aligned}
\|\boldsymbol{D}(t)\|_2 &\leq \int_0^t 2\left(\lambda_1 P \max\{\|\boldsymbol{M}(\tau)\|_2^{P-1}, \|\widehat{\boldsymbol{M}}(\tau)\|_2^{P-1}\}\|\boldsymbol{D}(\tau)\|_2 + \beta\|\boldsymbol{M}(\tau)\|_2^{2P}\right)d\tau \\
&\leq \int_0^t 2\left(\lambda_1 P g_{\hat{\alpha},r}(\tau)\|\boldsymbol{D}(\tau)\|_2 + \beta g_{\hat{\alpha},r}(\tau)^{\frac{2P}{P-1}}\right)d\tau
\end{aligned}$$

So

$$\begin{aligned}
\|\boldsymbol{D}(T_{\hat{\alpha}}(r))\|_2 &\leq \int_0^{T_{\hat{\alpha}}(r)} 2\beta g_{\hat{\alpha},r}(t)^{\frac{2P}{P-1}} \exp\left(2\lambda_1 P\int_t^{T_{\hat{\alpha}}(r)} g_{\hat{\alpha},r}(\tau)d\tau\right)dt \\
&= \int_0^{T_{\hat{\alpha}}(r)} 2\beta g_{\hat{\alpha},r}(t)^{\frac{2P}{P-1}} \exp\left(\frac{P}{P-1}\ln\frac{g_{\hat{\alpha},r}(T_{\hat{\alpha}}(r))}{g_{\hat{\alpha},r}(t)}\right)dt \\
&= \int_0^{T_{\hat{\alpha}}(r)} 2\beta g_{\hat{\alpha},r}(t)^{\frac{P}{P-1}} g_{\hat{\alpha},r}(T_{\hat{\alpha}}(r))^{\frac{P}{P-1}}dt \\
&= 2\beta\cdot\frac{1}{2\mu_1}g_{\hat{\alpha},r}(T_{\hat{\alpha}}(r))^{\frac{1}{P-1}}\cdot g_{\hat{\alpha},r}(T_{\hat{\alpha}}(r))^{\frac{P}{P-1}} \\
&= \kappa g_{\hat{\alpha},r}(T_{\hat{\alpha}}(r))^{\frac{P+1}{P-1}} \\
&= \kappa r^{P+1}.
\end{aligned}$$

which proves the bound. □

Lemma G.7. *Let $\boldsymbol{M}(t) = \phi_m(\hat{\alpha}\boldsymbol{M}_0, t), \widetilde{\boldsymbol{M}}(t) = \phi_m(\hat{\alpha}\widetilde{\boldsymbol{M}}_0, t)$. If $\max\{\|\boldsymbol{M}_0\|_2, \|\widetilde{\boldsymbol{M}}_0\|_2\} \leq 1$. For $t \leq T_{\hat{\alpha}}(r)$, we have*

$$\|\boldsymbol{M}(t) - \widetilde{\boldsymbol{M}}(t)\|_2 \leq \left(\frac{r}{\hat{\alpha}}\right)^P e^{2\kappa r^P}\|\boldsymbol{M}(0) - \widetilde{\boldsymbol{M}}(0)\|_2.$$

Proof. Define $\boldsymbol{D}(t) = \boldsymbol{M}(t) - \widetilde{\boldsymbol{M}}(t)$. For any unit vector $\boldsymbol{v}$,

$$\begin{aligned}
\frac{d}{dt}(\boldsymbol{v}^\top \boldsymbol{D}\boldsymbol{v}) &= -2\left(\boldsymbol{v}^\top \nabla f(\boldsymbol{M}^P)\left(\boldsymbol{M}^P - \widetilde{\boldsymbol{M}}^P\right)\boldsymbol{v} + \boldsymbol{v}^\top\left(\nabla f(\boldsymbol{M}^P) - \nabla f(\widetilde{\boldsymbol{M}}^P)\right)\widetilde{\boldsymbol{M}}^P\boldsymbol{v}\right) \\
&\leq 2\left(\|\nabla f(\boldsymbol{M}^P)\|_2\|\boldsymbol{M}^P - \widetilde{\boldsymbol{M}}^P\|_2 + \beta\|\boldsymbol{M}^P - \widetilde{\boldsymbol{M}}^P\|_2\|\widehat{\boldsymbol{M}}_2^P\|_2\right) \\
&\leq 2\left(\mu_1 + 2\beta\|\widetilde{\boldsymbol{M}}\|_2^P\right)P\max\{\|\boldsymbol{M}\|_2^{P-1}, \|\widetilde{\boldsymbol{M}}\|_2^{P-1}\}\|\boldsymbol{D}\|_2
\end{aligned}$$

So

$$\begin{aligned}\|\boldsymbol{D}(T_{\hat{\alpha}}(r))\|_2 &\le \|\boldsymbol{D}(0)\|_2 \cdot \exp\left(\int_0^{T_{\hat{\alpha}}(r)} 2\left(\mu_1 + 2\beta\|\widetilde{\boldsymbol{M}}\|_2^P\right) P \max\{\|\boldsymbol{M}\|_2^{P-1}, \|\widetilde{\boldsymbol{M}}\|_2^{P-1}\} dt\right)\\ &\le \|\boldsymbol{D}(0)\|_2 \cdot \exp\left(2P\mu_1 \int_0^{T_{\hat{\alpha}}(r)} \left(1 + 2\kappa g_{\hat{\alpha},r}(t)^{\frac{P}{P-1}}\right) g_{\hat{\alpha},r}(t) dt\right)\\ &\le \|\boldsymbol{D}(0)\|_2 \cdot \exp\left(\frac{P}{P-1} \ln \frac{g_{\hat{\alpha},r}(T_{\hat{\alpha}}(r))}{g_{\hat{\alpha},r}(0)} + 2\kappa g_{\hat{\alpha},r}(T_{\hat{\alpha}}(r))^{\frac{P}{P-1}}\right)\\ &\le \|\boldsymbol{D}(0)\|_2 \left(\frac{r}{\hat{\alpha}}\right)^P e^{2\kappa r^P},\end{aligned}$$

which proves the bound. □

Let $\boldsymbol{M}_\alpha^{\mathrm{G}}(t) := \phi_m\left(\alpha \boldsymbol{e}_1\boldsymbol{e}_1^\top, \frac{\hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)} + t\right)$. Let $\overline{\boldsymbol{M}}(t) := \lim_{\alpha\to 0} \boldsymbol{M}_\alpha^{\mathrm{G}}(t)$.

Lemma G.8. *For every $t \in (-\infty, +\infty)$, $\overline{\boldsymbol{M}}(t)$ exists and $\boldsymbol{M}_{\hat{\alpha}}^{\mathrm{G}}(t)$ converges to $\overline{\boldsymbol{M}}(t)$ in the following rate:*

$$\left\|\boldsymbol{M}_{\hat{\alpha}}^{\mathrm{G}}(t) - \overline{\boldsymbol{M}}(t)\right\|_2 = O(\hat{\alpha}).$$

Proof. Let c be a sufficiently small constant. Let $\bar{T} := \frac{-\kappa(P-1)c - c^{-(P-1)}}{2\mu_1(P-1)}$. We prove this lemma in the cases of $t \in (-\infty, \bar{T}]$ and $t > \bar{T}$ respectively.

Case 1. Fix $t \in (-\infty, \bar{T}]$. Then $\frac{\hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)} + t \le T_{\hat{\alpha}}(c)$. Let $\tilde{\alpha}$ be the unique number such that $\kappa(P-1)\tilde{\alpha} + \tilde{\alpha}^{-(P-1)} = \hat{\alpha}^{-(P-1)}$. Let $\hat{\alpha}' < \hat{\alpha}$ be an arbitrarily small number. Let $t_0 := T_{\hat{\alpha}'}(\tilde{\alpha}) = \frac{(\hat{\alpha}')^{-(P-1)} - \hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)}$. By Lemma G.6 and (30), we have

$$\left\|\phi_m(\hat{\alpha}'\boldsymbol{e}_1\boldsymbol{e}_1^\top, t_0) - \hat{\alpha}\boldsymbol{e}_1\boldsymbol{e}_1^\top\right\|_2 \le \left\|\phi_m(\hat{\alpha}'\boldsymbol{e}_1\boldsymbol{e}_1^\top, t_0) - \hat{\phi}_m(\hat{\alpha}'\boldsymbol{e}_1\boldsymbol{e}_1^\top, t_0)\right\|_2 \le O(\tilde{\alpha}^{P+1}).$$

By Lemma G.5, $\|\phi_m(\hat{\alpha}'\boldsymbol{e}_1\boldsymbol{e}_1^\top, t_0)\|_2 \le \tilde{\alpha}$. Then by Lemma G.7, we have

$$\left\|\phi_m(\hat{\alpha}'\boldsymbol{e}_1\boldsymbol{e}_1^\top, t_0 + t) - \boldsymbol{\phi}(\hat{\alpha}\boldsymbol{e}_1\boldsymbol{e}_1^\top, t)\right\|_2 \le \left(\frac{c}{\tilde{\alpha}}\right)^P e^{2\kappa c^P} \cdot O(\tilde{\alpha}^{P+1}) = O(\tilde{\alpha}) = O(\hat{\alpha}).$$

This implies that $\{\boldsymbol{M}_{\hat{\alpha}}^{\mathrm{G}}(t)\}$ satisfies Cauchy's criterion for every t, and thus the limit $\overline{\boldsymbol{M}}(t)$ exists for $t \le \bar{T}$. The convergence rate can be deduced by taking limits for $\hat{\alpha}' \to 0$ on both sides.

Case 2. For $t = \bar{T} + \tau$ with $\tau > 0$, $\phi_m(\boldsymbol{M}, \tau)$ is locally Lipschitz with respect to $\boldsymbol{M}$. So

$$\begin{aligned}\left\|\boldsymbol{M}_{\hat{\alpha}}^{\mathrm{G}}(t) - \boldsymbol{M}_{\hat{\alpha}'}^{\mathrm{G}}(t)\right\|_2 &= \left\|\phi_m(\boldsymbol{M}_{\hat{\alpha}}^{\mathrm{G}}(\bar{T}), \tau) - \phi_m(\boldsymbol{M}_{\hat{\alpha}'}^{\mathrm{G}}(\bar{T}), \tau)\right\|_2\\ &= O(\left\|\boldsymbol{M}_{\hat{\alpha}}^{\mathrm{G}}(\bar{T}) - \boldsymbol{M}_{\hat{\alpha}'}^{\mathrm{G}}(\bar{T})\right\|_2)\\ &= O(\hat{\alpha}),\end{aligned}$$

which proves the lemma for $t > \bar{T}$. □

Theorem G.9. *For every $t \in (-\infty, +\infty)$, as $\alpha \to 0$, we have:*

$$\left\|\phi_m\left(\hat{\alpha}\boldsymbol{I}, \frac{\hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)} + t\right) - \overline{\boldsymbol{M}}(t)\right\|_2 = O(\hat{\alpha}^{\frac{1}{P+1}}), \tag{31}$$

and for any $2 \le k \le d$,

$$\lambda_k\left(\phi_m\left(\hat{\alpha}\boldsymbol{I}, \frac{\hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)} + t\right)\right) = O(\hat{\alpha}). \tag{32}$$

Proof. Let $\boldsymbol{M}_{\hat{\alpha}}(t) := \phi_m\left(\hat{\alpha}\boldsymbol{I}, \frac{\hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)} + t\right)$. Again we let c be a sufficiently small constant and $\bar{T} := \frac{-\kappa(P-1)c - c^{-(P-1)}}{2\mu_1(P-1)}$. We prove in the cases of $t \in (-\infty, \bar{T}]$ and $t > \bar{T}$ respectively.

Case 1. Fix $t \in (-\infty, \bar{T}]$. Let $\hat{\alpha}_1 := \hat{\alpha}^{\frac{1}{P+1}}$. Let $\tilde{\alpha}_1$ be the unique number such that $\kappa(P-1)\tilde{\alpha}_1 + \tilde{\alpha}_1^{-(P-1)} = \hat{\alpha}_1^{-(P-1)}$. Let $t_0 := T_{\hat{\alpha}}(\tilde{\alpha}_1) = \frac{\hat{\alpha}^{-(P-1)} - \hat{\alpha}_1^{-(P-1)}}{2\mu_1(P-1)}$. Then

$$\begin{aligned}\left\|\phi_m(\hat{\alpha}\boldsymbol{I}, t_0) - \hat{\alpha}_1 \boldsymbol{e}_1 \boldsymbol{e}_1^\top\right\|_2 &\le \left\|\phi_m(\hat{\alpha}\boldsymbol{I}, t_0) - \hat{\phi}_m(\hat{\alpha}\boldsymbol{I}, t_0)\right\|_2 + \left\|\hat{\phi}_m(\hat{\alpha}\boldsymbol{I}, t_0) - \hat{\alpha}_1 \boldsymbol{e}_1 \boldsymbol{e}_1^\top\right\|_2 \\ &= O(\tilde{\alpha}_1^{P+1} + \hat{\alpha}) \\ &= O(\hat{\alpha}).\end{aligned}$$

By Lemma G.5, $\|\phi_m(\hat{\alpha}'\boldsymbol{I}, t_0)\|_2 \le \tilde{\alpha}_1$. Then by Lemma G.7, we have

$$\begin{aligned}\left\|\boldsymbol{M}_{\hat{\alpha}}(t) - \boldsymbol{M}^{\mathrm{G}}_{\hat{\alpha}_1}(t)\right\|_2 &= \left\|\phi_m(\hat{\alpha}\boldsymbol{I}, t_0 + t) - \phi_m(\hat{\alpha}_1 \boldsymbol{e}_1 \boldsymbol{e}_1^\top, t)\right\|_2 \\ &\le \left(\frac{c}{\tilde{\alpha}_1}\right)^P e^{2\kappa c^P} \cdot O(\hat{\alpha}) = O(\hat{\alpha}^{\frac{1}{P+1}}).\end{aligned}$$

Combining this with the convergence rate for $\boldsymbol{M}^{\mathrm{G}}_{\hat{\alpha}_1}(t)$ proves the bound Equation (11).

For Equation (12), by Lemma G.3, we have

$$\begin{aligned}\lambda_k^{1-P}(\boldsymbol{M}_{\hat{\alpha}}(\bar{T})) - \lambda_k^{1-P}(\boldsymbol{M}_{\hat{\alpha}}(t_0)) &\le \int_{t=t_0}^{\bar{T}} 2(P-1)\left\|\nabla f(\boldsymbol{M}^P_{\hat{\alpha}}(t))\right\|_2 \mathrm{d}t \\ &\le \int_{t=t_0}^{\bar{T}} 2(P-1)(\mu + \beta\left\|\boldsymbol{M}_{\hat{\alpha}}(t)\right\|^P))\mathrm{d}t \\ &\le -2(P-1)\left(\mu_1(t - T_1) + \frac{\kappa}{2} \cdot g_{\hat{\alpha},c}(t)^{\frac{1}{P-1}}\right).\end{aligned} \tag{33}$$

By Lemma G.6, $\lambda_1(\boldsymbol{M}_{\hat{\alpha}}(\bar{T})) = \left\|\boldsymbol{M}_{\hat{\alpha}}(\bar{T})\right\|_2 = c + O(c^{P+1})$. For $k \ge 2$,

$$\begin{aligned}\lambda_k(\boldsymbol{M}_{\hat{\alpha}}(\bar{T}))^{-(P-1)} &\ge \Omega(\hat{\alpha}^{-(P-1)}) - 2(P-1)\left(\mu_1(\bar{T} - T_1) + \frac{\kappa}{2} \cdot c\right) \\ &\ge \Omega(\hat{\alpha}^{-(P-1)}) - \frac{\hat{\alpha}^{-\frac{P-1}{P+1}} - c^{-(P-1)}}{2\mu_1(P-1)} - O(c) \\ &\ge \Omega(\hat{\alpha}^{-(P-1)}).\end{aligned}$$

Thus $\lambda_k(\boldsymbol{M}_{\hat{\alpha}}(\bar{T})) \le O(\hat{\alpha})$.

Case 2. For $t = \bar{T} + \tau$ with $\tau > 0$, $\phi_m(\boldsymbol{M}, \tau)$ is locally Lipschitz with respect to $\boldsymbol{M}$. So

$$\begin{aligned}\left\|\boldsymbol{M}_{\hat{\alpha}}(t) - \boldsymbol{M}^{\mathrm{G}}_{\hat{\alpha}_1}(t)\right\|_2 &= \left\|\phi_m(\boldsymbol{M}_{\hat{\alpha}}(\bar{T}), \tau) - \phi_m(\boldsymbol{M}^{\mathrm{G}}_{\hat{\alpha}_1}(\bar{T}), \tau)\right\|_2 \\ &= O(\left\|\boldsymbol{M}_{\hat{\alpha}}(\bar{T}) - \boldsymbol{M}^{\mathrm{G}}_{\hat{\alpha}_1}(\bar{T})\right\|_2) = O(\hat{\alpha}^{\frac{1}{P+1}}),\end{aligned}$$

which proves the bound Equation (11).

For Equation (12), again by Lemma G.3, we have

$$\begin{aligned}&\lambda_k^{1-P}(\boldsymbol{M}_{\hat{\alpha}}(\bar{T})) - \lambda_k^{1-P}(\boldsymbol{M}_{\hat{\alpha}}(\bar{T} + \tau)) \\ \le &\int_{t=\bar{T}}^{\bar{T}+\tau} 2(P-1)\left\|\nabla f(\boldsymbol{M}^P_{\hat{\alpha}}(t))\right\|_2 \mathrm{d}t \\ \le &\int_{t=\bar{T}}^{\bar{T}+\tau} 2(P-1)(\beta\left\|\boldsymbol{M}^P_{\hat{\alpha}}(t) - (\boldsymbol{M}^{\mathrm{G}})^P(t)\right\|_2 + \left\|\nabla f\left((\boldsymbol{M}^{\mathrm{G}})^P(t)\right)\right\|_2)\mathrm{d}t \\ \le &\int_{t=\bar{T}}^{\bar{T}+\tau} 2(P-1)(O(\hat{\alpha}^{\frac{1}{1+P}}) + \left\|\nabla f\left((\boldsymbol{M}^{\mathrm{G}})^P(t)\right)\right\|_2)\mathrm{d}t \\ \le &O(1).\end{aligned} \tag{34}$$

Thus $\lambda_k^{1-P}(\boldsymbol{M}_{\hat{\alpha}}(\bar{T} + \tau)) = \Omega(\hat{\alpha}^{-(P-1)})$, that is, $\lambda_k(\boldsymbol{M}_{\hat{\alpha}}(\bar{T} + \tau)) = O(\hat{\alpha})$. □

Proof of Theorem 6.2. Note that

$$\left(\phi_m\left(\hat{\alpha}\boldsymbol{I}, \frac{\hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)}+t\right)\right)^P = \phi\left(\alpha\boldsymbol{I}, \frac{\hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)}+t\right),$$

and

$$\left(\overline{\boldsymbol{M}}(t)\right)^P = \overline{\boldsymbol{W}}(t).$$

By Theorem G.9, We have

$$\begin{aligned}
&\left\|\phi\left(\alpha\boldsymbol{I}, \frac{\alpha^{-(1-1/P)}}{2\mu_1(P-1)}+t\right) - \overline{\boldsymbol{W}}(t)\right\|_2 \\
\leq &\left\|\phi\left(\alpha\boldsymbol{I}, \frac{\hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)}+t\right) - \overline{\boldsymbol{W}}(t)\right\|_2 = \left\|\left(\phi_m\left(\hat{\alpha}\boldsymbol{I}, \frac{\hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)}+t\right)\right)^P - \left(\overline{\boldsymbol{M}}(t)\right)^P\right\|_2 \\
\leq &P\left\|\phi_m\left(\hat{\alpha}\boldsymbol{I}, \frac{\hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)}+t\right) - \overline{\boldsymbol{M}}(t)\right\|_2 \max\left(\left\|\phi_m\left(\hat{\alpha}\boldsymbol{I}, \frac{\hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)}+t\right)\right\|_2, \left\|\overline{\boldsymbol{M}}(t)\right\|_2\right)^{P-1} \\
= &O(\hat{\alpha}^{\frac{1}{P+1}})O(1) = O(\alpha^{\frac{1}{P(P+1)}}),
\end{aligned}$$

and for $2 \leq k \leq d$,

$$\lambda_k\left(\phi\left(\alpha\boldsymbol{I}, \frac{\hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)}+t\right)\right) \leq \lambda_k\left(\phi_m\left(\hat{\alpha}\boldsymbol{I}, \frac{\hat{\alpha}^{-(P-1)}}{2\mu_1(P-1)}+t\right)\right) = O(\hat{\alpha}^P) = O(\alpha)$$

□

H Proof of Linear Convergence to the Stationary Point

In this section, we will present the theorems that guarantees the linear convergence to local minimum $\boldsymbol{W}_0$ if the dynamics Equation (35) is initialized sufficiently close, i.e. $\|\boldsymbol{W}(0) - \boldsymbol{W}_0\|_F$ sufficiently small (but still at a constant scale).

$$\frac{\mathrm{d}\boldsymbol{W}(t)}{\mathrm{d}t} = \boldsymbol{g}(\boldsymbol{W}(t)) = -\sum_{i=0}^{L-1}(\boldsymbol{W}\boldsymbol{W}^\top)^{\frac{i}{L}}\nabla f(\boldsymbol{W})(\boldsymbol{W}^\top\boldsymbol{W})^{1-\frac{i+1}{L}}. \tag{35}$$

Throughout this section, we assume $\mathrm{rank}(\boldsymbol{W}_0) = k$. The tangent space of manifold of rank-k symmetric matrices at $\boldsymbol{W}_0$ is $\mathcal{T} = \{\boldsymbol{V}\boldsymbol{W}_0^\top + \boldsymbol{W}_0\boldsymbol{V}^\top \mid \boldsymbol{V} \in \mathbb{R}^{d\times k}\}$. It can be shown that $\dim(\mathcal{T}) = k(d-k) + \frac{k(k+1)}{2} = \frac{k(2d-k+1)}{2}$. $\boldsymbol{W}_0$ being the stationary (or the fixed point) of Equation (35) means that $\boldsymbol{g}(\boldsymbol{W}_0)\boldsymbol{W}_0 = \boldsymbol{W}_0\boldsymbol{g}(\boldsymbol{W}_0) = \boldsymbol{0}$. For depth-2 case, we show $\mathcal{T}$ is an invariant subspace of $D\boldsymbol{g}(\boldsymbol{W}_0)$ in Theorem D.4, property 2. Indeed this also holds for $L \geq 3$, as a similar analysis can be applied for that case.

Towards showing the main convergence result in the section, we will make the following assumption.

Assumption H.1. Suppose $D\boldsymbol{g}(\boldsymbol{W}_0)|_{\mathcal{T}}$ diagonalizable and all eigenvalues are real and negative.

If Assumption H.1 holds, then we know $D\boldsymbol{g}(\boldsymbol{W}_0)|_{\mathcal{T}}\colon \mathcal{T} \to \mathcal{T}$ can be diagonalized as $D\boldsymbol{g}(\boldsymbol{W}_0)|_{\mathcal{T}}[\cdot] = \mathcal{V}(\Sigma\mathcal{V}^{-1}(\cdot))$, where $\Sigma_i = \mathrm{diag}([\mu_1, \ldots, \mu_{\dim(\mathcal{T})}])$, $\mathcal{V}: \mathbb{R}^{\dim(\mathcal{T})} \to \mathcal{T}$, $\mathcal{V}(\boldsymbol{x}) = \sum_{i=1}^{\dim(\mathcal{T})} x_i\boldsymbol{V}_i$, with $\boldsymbol{V}_i$ being the eigenvector associated with eigenvector μ_i.

As shown in Theorem H.3, this assumption implies that if $\boldsymbol{W}(0)$ is rank-k and is sufficiently close to $\boldsymbol{W}_0$, then $\|\boldsymbol{W}(t) - \boldsymbol{W}_0\|_F \leq Ce^{\mu_1 t}$ for some constant C. (note here $\mu_1 \leq 0$.) For depth-2 case, the above assumption is equivalent to that $\mathcal{L}(\boldsymbol{U}_0)$ is "strongly convex" at $\boldsymbol{U}_0$, except those 0 eigenvalues due to symmetry, by property 2 of Theorem D.4). For the case where $L \geq 3$, because this dynamics is not gradient flow, in general it doesn't correspond to a loss function and strongly convexity wouldn't make any sense. Nevertheless, in experiments we do observe linear convergence to the stationary point, so this assumption is reasonable.

For convenience, we define $\|\cdot\|_{\mathcal{V}} := \left\|\mathcal{V}^{-1}\left(\Pi_1^{d^2}(\cdot)\right)\right\|_F$, $\|\cdot\|_{F,1} := \left\|\Pi_1^{d^2}(\cdot)\right\|_F$ and $\|\cdot\|_{F,2} := \left\|\Pi_2^{d^2}(\cdot)\right\|_F$. The reason for such definition of norms, as we will see later, is that the norm (or the difference) in the tangent space of the manifold of symmetric rank-r matrices $\|\boldsymbol{W}-\boldsymbol{W}'\|_{F,1}$, will dominate that in the orthogonal complement of the tangent space, i.e. $\|\boldsymbol{W}-\boldsymbol{W}'\|_{F,2}$, when both $\boldsymbol{W},\boldsymbol{W}'$ gets very close to the $\boldsymbol{W}_0$. (See a more rigorous statement in Lemma H.2) W.l.o.g., we can assume $\frac{\|\cdot\|_{F,1}}{K} \le \|\cdot\|_{\mathcal{V}} \le \|\cdot\|_{F,1}$, for some constant K, which may depend on f and $\boldsymbol{W}_0$. This also implies that $\|\cdot\|_{\mathcal{V}} \le \|\cdot\|_F$. Below we also assume for sufficiently small R, and any $\boldsymbol{W}$ such that $\|\boldsymbol{W}-\boldsymbol{W}_0\|_F \le R$, we have $\|\nabla f(\boldsymbol{W})\|_2 \le \rho$ and $\|D_{\boldsymbol{g}}(\boldsymbol{W})[\boldsymbol{\Delta}]\|_F \le \beta\|\boldsymbol{\Delta}\|_F$ for any $\boldsymbol{\Delta}$. In the proof below, as long as we can show the boundedness of $\boldsymbol{W}(t)-\boldsymbol{W}_0$, we will assume such properties hold.

Lemma H.2. *Let* $\max(\|\boldsymbol{W}-\boldsymbol{W}_0\|_{F,1}, \|\boldsymbol{W}'-\boldsymbol{W}_0\|_{F,1}) = r$*, when* $r \le \frac{\lambda_{\min}(\boldsymbol{W}_0)}{2}$*, we have*

$$\|\boldsymbol{W}-\boldsymbol{W}'\|_{F,2} \le \frac{5r}{\lambda_{\min}(\boldsymbol{W}_0)}\|\boldsymbol{W}-\boldsymbol{W}'\|_{F,1} \tag{36}$$

Proof of Lemma H.2. W.l.o.g., we can assume $\Pi = \sum_{i=1}^k \boldsymbol{e}_i\boldsymbol{e}_i^\top$, i.e., $[\boldsymbol{W}_0]_{ij} = 0,\ \forall i,j \ge k+1$. We further denote $\boldsymbol{W}$ and $\boldsymbol{W}'$ by $\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}^\top\\ \boldsymbol{B} & \boldsymbol{C}\end{bmatrix}$ and $\begin{bmatrix}\boldsymbol{A}' & \boldsymbol{B}'^\top\\ \boldsymbol{B}' & \boldsymbol{C}'\end{bmatrix}$, where $\boldsymbol{A},\boldsymbol{A}' \in \mathbb{R}^{k\times k}, \boldsymbol{B},\boldsymbol{B}' \in \mathbb{R}^{(d-k)\times k}, \boldsymbol{C},\boldsymbol{C}' \in \mathbb{R}^{(d-k)\times(d-k)}$. By definition, we have $\|\boldsymbol{A}-\boldsymbol{A}'\|_F, \|\boldsymbol{B}-\boldsymbol{B}'\|_F \le \|\boldsymbol{W}-\boldsymbol{W}'\|_{F,1}$ and $\|\boldsymbol{W}-\boldsymbol{W}'\|_{F,2} = \|\boldsymbol{C}-\boldsymbol{C}'\|_F$. Moreover, we have $\lambda_{\min}(\boldsymbol{A}) \ge \lambda_{\min}(\boldsymbol{W}_0) - \|\boldsymbol{A}-\boldsymbol{W}_0\|_F \ge \lambda_{\min}(\boldsymbol{W}_0) - \|\boldsymbol{W}-\boldsymbol{W}_0\|_{F,1} \ge \frac{\lambda_{\min}(\boldsymbol{W}_0)}{2}$.

Since $\boldsymbol{W},\boldsymbol{W}'$ is rank-k, we have $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^\top, \boldsymbol{C}' = \boldsymbol{B}'\boldsymbol{A}'^{-1}\boldsymbol{B}'^\top$. Thus

$$\begin{aligned}
&\|\boldsymbol{W}-\boldsymbol{W}'\|_{F,2} = \|\boldsymbol{C}-\boldsymbol{C}'\|_F\\
=&\left\|\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^\top - \boldsymbol{B}'\boldsymbol{A}'^{-1}\boldsymbol{B}'^\top\right\|_F\\
\le&\|\boldsymbol{B}-\boldsymbol{B}'\|_F\left\|\boldsymbol{A}^{-1}\boldsymbol{B}^\top\right\|_F + \left\|\boldsymbol{B}\boldsymbol{A}^{-1}\right\|_F\|\boldsymbol{A}'-\boldsymbol{A}\|_F\left\|\boldsymbol{A}'^{-1}\boldsymbol{B}'^\top\right\|_F + \left\|\boldsymbol{B}'\boldsymbol{A}'^{-1}\right\|_F\left\|\boldsymbol{B}^\top-\boldsymbol{B}'^\top\right\|_F\\
\le&\|\boldsymbol{W}-\boldsymbol{W}'\|_{F,1}\frac{2r}{\lambda_{\min}(\boldsymbol{W}_0)} + \|\boldsymbol{W}-\boldsymbol{W}'\|_{F,1}\left(\frac{2r}{\lambda_{\min}(\boldsymbol{W}_0)}\right)^2 + \|\boldsymbol{W}-\boldsymbol{W}'\|_{F,1}\frac{2r}{\lambda_{\min}(\boldsymbol{W}_0)}\\
\le&\|\boldsymbol{W}-\boldsymbol{W}'\|_{F,1}\frac{5r}{\lambda_{\min}(\boldsymbol{W}_0)}
\end{aligned}$$

□

Theorem H.3 (Linear Convergence of rank-k matrices)**.** *Suppose* $\|\boldsymbol{W}(0)-\boldsymbol{W}_0\|_{\mathcal{V}} \le R := \max\left(\frac{\lambda_{\min}(\boldsymbol{W}_0)}{2K}, \sqrt{\frac{5\rho\lambda_{\min}(\boldsymbol{W}_0)}{2\beta}}), \frac{-\mu_1}{4K^2}(\beta+\frac{10\rho}{\lambda_{\min}(\boldsymbol{W}_0)})^{-1}\right)$, $\operatorname{rank}(\boldsymbol{W}(0)) = \operatorname{rank}(\boldsymbol{W}_0) = k$*, we have* $\|\boldsymbol{W}(t)-\boldsymbol{W}_0\|_{\mathcal{V}} \le Ce^{\mu_1 t}\|\boldsymbol{W}(0)-\boldsymbol{W}_0\|_{\mathcal{V}}$*, for some constant* C *depending on* $\boldsymbol{W}_0$*.*

Proof of Theorem H.3. For convenience, we define $\boldsymbol{W}_1(t) := \Pi_1^{d^2}(\boldsymbol{W}(t)-\boldsymbol{W}_0), \boldsymbol{W}_2(t) := \Pi_2^{d^2}(\boldsymbol{W}(t)-\boldsymbol{W}_0) = \Pi_2^{d^2}(\boldsymbol{W}(t))$.

$$
\begin{aligned}
&\frac{\mathrm{d}\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}^2}{\mathrm{d}t}=\frac{\mathrm{d}\left\|\boldsymbol{W}(t)-\boldsymbol{W}_0\right\|_{\mathcal{V}}^2}{\mathrm{d}t}\\
=&2\left\langle\mathcal{V}^{-1}\left(\Pi_1^{d^2}\left(\frac{\mathrm{d}\boldsymbol{W}(t)}{\mathrm{d}t}\right)\right),\mathcal{V}^{-1}\left(\Pi_1^{d^2}\left(\boldsymbol{W}(t)-\boldsymbol{W}_0\right)\right)\right\rangle\\
=&2\left\langle\mathcal{V}^{-1}\left(\Pi_1^{d^2}\left(\frac{\mathrm{d}\boldsymbol{W}(t)}{\mathrm{d}t}\right)\right),\mathcal{V}^{-1}\left(\boldsymbol{W}_1(t)\right)\right\rangle\\
=&2\left\langle\mathcal{V}^{-1}\left(\Pi_1^{d^2}\left(\boldsymbol{g}(\boldsymbol{W}(t))\right)\right),\mathcal{V}^{-1}\left(\boldsymbol{W}_1(t)\right)\right\rangle\\
\leq&2\left\langle\mathcal{V}^{-1}\left(\Pi_1^{d^2}\left(D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}(t)-\boldsymbol{W}_0]\right)\right),\mathcal{V}^{-1}\left(\boldsymbol{W}_1(t)\right)\right\rangle\\
&+2\left\|\boldsymbol{g}(\boldsymbol{W}(t)-\boldsymbol{W}_0)-D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}(t)-\boldsymbol{W}_0]\right\|_{\mathcal{V}}\left\|\boldsymbol{W}(t)-\boldsymbol{W}_0\right\|_{\mathcal{V}}\\
=&2\left\langle\mathcal{V}^{-1}\left(\Pi_1^{d^2}\left(D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}_1(t)+\boldsymbol{W}_2(t)]\right)\right),\mathcal{V}^{-1}\left(\boldsymbol{W}_1(t)\right)\right\rangle\\
&+2\left\|\boldsymbol{g}(\boldsymbol{W}(t)-\boldsymbol{W}_0)-D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}(t)-\boldsymbol{W}_0]\right\|_{\mathcal{V}}\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}\\
=&2\left\langle\mathcal{V}^{-1}\left(\Pi_1^{d^2}\left(D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}_1(t)]\right)\right),\mathcal{V}^{-1}\left(\boldsymbol{W}_1(t)\right)\right\rangle\\
&+2\left\|D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}_2(t)]\right\|_{\mathcal{V}}\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}\\
&+2\left\|\boldsymbol{g}(\boldsymbol{W}(t)-\boldsymbol{W}_0)-D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}(t)-\boldsymbol{W}_0]\right\|_{\mathcal{V}}\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}
\end{aligned}
$$

For the first term $\left\langle\mathcal{V}^{-1}\left(\Pi_1^{d^2}\left(D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}_1(t)]\right)\right),\mathcal{V}^{-1}\left(\boldsymbol{W}_1(t)\right)\right\rangle$, we know $\boldsymbol{W}_1(t)\in\mathcal{T}$, and $\mathcal{T}$ is an invariant space of $D\boldsymbol{g}(\boldsymbol{W}_0)$. Recall $D\boldsymbol{g}(\boldsymbol{W}_0)|_{\mathcal{T}}[\cdot]=\mathcal{V}\left(\Sigma\mathcal{V}^{-1}\left(\cdot\right)\right)$, we have

$$
2\left\langle\mathcal{V}^{-1}\left(\Pi_1^{d^2}\left(D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}_1(t)]\right)\right),\mathcal{V}^{-1}\left(\boldsymbol{W}_1(t)\right)\right\rangle=2\left\langle\Sigma\mathcal{V}^{-1}\left(\boldsymbol{W}_1(t)\right),\mathcal{V}^{-1}\left(\boldsymbol{W}_1(t)\right)\right\rangle\leq 2\mu_1\left\|\boldsymbol{W}_1(t)\right\|_{F,1}.
$$

For the second term $2\beta\left\|D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}_2(t)]\right\|_{\mathcal{V}}\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}$, we have

$$
\begin{aligned}
&2\left\|D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}_2(t)]\right\|_{\mathcal{V}}\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}\\
\leq&2\left\|D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}_2(t)]\right\|_{F}\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}\\
\leq&2\left\|D\boldsymbol{g}(\boldsymbol{W}_0)\right\|_{2}\left\|\boldsymbol{W}_2(t)\right\|_{F}\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}\\
=&2\rho\left\|\boldsymbol{W}_2(t)\right\|_{F}\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}.
\end{aligned}
$$

For the third term $2\left\|\boldsymbol{g}(\boldsymbol{W}(t)-\boldsymbol{W}_0)-D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}(t)-\boldsymbol{W}_0]\right\|_{\mathcal{V}}\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}$, we have

$$
\begin{aligned}
&2\left\|\boldsymbol{g}(\boldsymbol{W}(t)-\boldsymbol{W}_0)-D\boldsymbol{g}(\boldsymbol{W}_0)[\boldsymbol{W}(t)-\boldsymbol{W}_0]\right\|_{\mathcal{V}}\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}\\
\leq&2\beta\left\|\boldsymbol{W}(t)-\boldsymbol{W}_0\right\|_{F}^2\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}\\
\leq&4\beta\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}\left(\left\|\boldsymbol{W}_1(t)\right\|_{F}^2+\left\|\boldsymbol{W}_2(t)\right\|_{F}^2\right)\\
\leq&4\beta\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}\left(K^2\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}^2+\left\|\boldsymbol{W}_2(t)\right\|_{F}^2\right)
\end{aligned}
$$

Thus we have shown the following. Note so far we haven't used any assumption that $\boldsymbol{W}$ is rank-k.

$$
\frac{\mathrm{d}\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}^2}{\mathrm{d}t}\leq 2\mu_1\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}^2+2\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}\left(\rho\left\|\boldsymbol{W}_2(t)\right\|_{F}+2\beta K^2\left\|\boldsymbol{W}_1(t)\right\|_{\mathcal{V}}^2+2\beta\left\|\boldsymbol{W}_2(t)\right\|_{F}^2\right),\tag{37}
$$

that is,

$$\frac{\mathrm{d}\log\|\boldsymbol{W}_1(t)\|_{\mathcal{V}}^2}{\mathrm{d}t} \le 2\mu_1 + 4\beta K^2\|\boldsymbol{W}_1(t)\|_{\mathcal{V}} + 2\frac{\beta\|\boldsymbol{W}_2(t)\|_F^2 + 2\rho\|\boldsymbol{W}_2(t)\|_F}{\|\boldsymbol{W}_1(t)\|_{\mathcal{V}}}. \tag{38}$$

Let $T = \sup_{t\ge 0}\|\boldsymbol{W}_1(t)\|_{\mathcal{V}} \le \frac{\lambda_{\min}(\boldsymbol{W}_0)}{2K}$. Since $\boldsymbol{W}(t)$ is continuous and $\|\boldsymbol{W}(0)\|_{\mathcal{V}} \le \frac{\lambda_{\min}(\boldsymbol{W}_0)}{2K}$, $T > 0$. Setting $\boldsymbol{W}' = \boldsymbol{W}_0$ in Lemma H.2, we have for $t < T$, $r = \|\boldsymbol{W}(t) - \boldsymbol{W}_0\|_{F,1} \le \|\boldsymbol{W}(t) - \boldsymbol{W}_0\|_F \le K\|\boldsymbol{W}(t) - \boldsymbol{W}_0\|_{\mathcal{V}} \le \frac{\lambda_{\min}(\boldsymbol{W}_0)}{2}$. Thus $\|\boldsymbol{W}_2(t)\|_{F,2} \le \frac{5\|\boldsymbol{W}(t)-\boldsymbol{W}_0\|_{F,1}^2}{\lambda_{\min}(\boldsymbol{W}_0)} \le \frac{5K^2\|\boldsymbol{W}(t)-\boldsymbol{W}_0\|_{\mathcal{V}}^2}{\lambda_{\min}(\boldsymbol{W}_0)}$. Also by assumption, $\|\boldsymbol{W}_2(t)\|_{F,2} \le \frac{5\|\boldsymbol{W}(t)-\boldsymbol{W}_0\|_{F,1}^2}{\lambda_{\min}(\boldsymbol{W}_0)} \le \frac{5\|\boldsymbol{W}(t)-\boldsymbol{W}_0\|_{F,1}^2}{\lambda_{\min}(\boldsymbol{W}_0)} \le \frac{2\rho}{\beta}$. Thus from Equation (38) we can derive that

$$\frac{\mathrm{d}\log\|\boldsymbol{W}_1(t)\|_{\mathcal{V}}^2}{\mathrm{d}t} \le 2\mu_1 + 4K^2(\beta + \frac{10\rho}{\lambda_{\min}(\boldsymbol{W}_0)})\|\boldsymbol{W}_1(t)\|_{\mathcal{V}} \le \mu_1. \tag{39}$$

Since $\mu_1 < 0$, $\|\boldsymbol{W}_1(t)\|_{F,1}$ decreases for $[0, T)$. Thus T must be ∞, otherwise $\|\boldsymbol{W}_1(T)\|_{F,1} = \lim_{t\to T^-}\|\boldsymbol{W}_1(t)\|_{F,1} < R_1$. Contradiction.

Therefore, for any $t \in [0, \infty)$, we have $\|\boldsymbol{W}_1(t)\|_{F,1} \le \|\boldsymbol{W}_1(0)\|_{F,1} e^{-\frac{\mu_1}{2}t}$. That is

$$\int_{t=0}^{\infty}\|\boldsymbol{W}_1(t)\|_{\mathcal{V}}\,\mathrm{d}t \le \frac{2}{\mu_1}\|\boldsymbol{W}_1(0)\|_{\mathcal{V}}.$$

Thus from Equation (39), we have

$$\begin{aligned}
&\|\boldsymbol{W}_1(t)\|_{\mathcal{V}} \le \|\boldsymbol{W}_1(0)\|_{\mathcal{V}}\exp\left(\mu_1 t + 4K^2(\beta + \frac{10\rho}{\lambda_{\min}(\boldsymbol{W}_0)})\int_{t=0}^{\infty}\|\boldsymbol{W}_1(t)\|_{\mathcal{V}}\,\mathrm{d}t\right)\\
\le &\|\boldsymbol{W}_1(t)\|_{\mathcal{V}} \le \|\boldsymbol{W}_1(0)\|_{\mathcal{V}}\exp\left(\mu_1 t + \frac{8K^2}{\mu_1}(\beta + \frac{10\rho}{\lambda_{\min}(\boldsymbol{W}_0)})\|\boldsymbol{W}_1(0)\|_{\mathcal{V}}\right)\\
:= &C\|\boldsymbol{W}_1(0)\|_{\mathcal{V}} e^{\mu_1 t}
\end{aligned}$$

□

H.1 Linear Convergence to Stationary Point when depth ≥ 3

We will $\boldsymbol{M}(t)$ to denote the top-k components of $\boldsymbol{W}(t)$ in SVD, and $\boldsymbol{N}(t)$ to denote the rest part, i.e. $\boldsymbol{W}(t) - \boldsymbol{M}(t)$. (One can think $\boldsymbol{M}(t)$ as the main part and $\boldsymbol{N}(t)$ as the negligible part.)

Below we will show that for deep over-parametrized matrix completion, where $mW(t)$ satisfies Equation (25), if the trajectory is initialized at some $\boldsymbol{W}(0)$ in a small neighborhood of $\boldsymbol{W}^0$, a rank-k saddle which will be passed by deep GLRL, and $\boldsymbol{W}(0)$ is approximately rank-k, in the sense that $\boldsymbol{N}(0)$ is very small, then $\inf_{t\ge 0}\|\boldsymbol{W}(t) - \boldsymbol{W}_0\|_{\mathcal{V}}$ is roughly at the same magnitude of $\boldsymbol{N}(0)$.

Theorem H.4 (Almost Linear Convergence of almost rank-k matrices, deep case)**.** *Suppose rank-k stationary point $\boldsymbol{W}_0$ satisfies Assumption H.1, there exists constants C_0 and r, such that if $C_0\|\boldsymbol{N}(0)\|_F \le \|\boldsymbol{W}_1(0)\|_{\mathcal{V}} \le r$, then there exists a time T and constants C, C', such that*

- $\|\boldsymbol{W}(t) - \boldsymbol{W}_0\|_{\mathcal{V}} \le Ce^{\mu_1 t/2}\|\boldsymbol{W}(0) - \boldsymbol{W}_0\|_{\mathcal{V}}$, *for* $t \le T$.
- $\|\boldsymbol{W}(T) - \boldsymbol{W}_0\|_F = C'\|\boldsymbol{N}(0)\|_F$.

Proof. When $\|\boldsymbol{W}(t) - \boldsymbol{W}_0\|_F \le \frac{\lambda_{\min}(\boldsymbol{W}_0)}{4}$, $\|\boldsymbol{N}(t)\|_F \le \frac{\lambda_{\min}(\boldsymbol{W}_0)}{4}$, thus we have $\|\boldsymbol{M}(t) - \boldsymbol{W}_0\|_{F,1} \le \|\boldsymbol{W}(t) - \boldsymbol{W}_0\|_{F,1} + \|\boldsymbol{N}(t)\|_{F,1} \le \frac{\lambda_{\min}(\boldsymbol{W}_0)}{2}$, thus by Lemma H.2, we have

$$\|\boldsymbol{W}_2(t)\|_{F,2} \leq \|\boldsymbol{M}(t)-\boldsymbol{W}_0\|_{F,2} + \|\boldsymbol{N}(t)\|_{F,2} \leq \frac{5\|\boldsymbol{M}(t)-\boldsymbol{W}_0\|_{F,1}^2}{\lambda_{\min}(\boldsymbol{W}_0)} + \|\boldsymbol{N}(t)\|_{F,2}$$
$$\leq \frac{10\|\boldsymbol{W}_1(t)\|_{F,1}^2 + 10\|\boldsymbol{N}(t)\|_F^2}{\lambda_{\min}(\boldsymbol{W}_0)} + \|\boldsymbol{N}(t)\| \leq \frac{10K^2\|\boldsymbol{W}_1(t)\|_{\mathcal{V}}^2 + 10\|\boldsymbol{N}(t)\|_F^2}{\lambda_{\min}(\boldsymbol{W}_0)} + \|\boldsymbol{N}(t)\|$$

Thus we can pick constant C_0 large enough and r small enough, such that for any t, if $C_0\|\boldsymbol{N}(t)\|_F \leq \|\boldsymbol{W}_1(t)\|_{\mathcal{V}} \leq r$, then it holds that :

- The "small terms" in the right hand side of Equation (38) satisfies that $4\beta K^2\|\boldsymbol{W}_1(t)\|_{\mathcal{V}} + 2\frac{\beta\|\boldsymbol{W}_2(t)\|_F^2 + 2\rho\|\boldsymbol{W}_2(t)\|_F}{\|\boldsymbol{W}_1(t)\|_{\mathcal{V}}} \leq C_1\|\boldsymbol{W}_1(t)\|_{\mathcal{V}} + C_2\|\boldsymbol{N}(t)\|_F \leq -\mu_1$ for some C_1 and C_2.
- The spectral norm $\frac{1}{2}\|\nabla f(\boldsymbol{W}(t))\|_2 \leq \|\nabla f(\boldsymbol{W}_0)\|_2 := \rho$.
- $\forall x < r$, $C_3 x^{\frac{2}{L}-1} = T' > -\frac{2}{\mu_1}\ln\frac{2r}{C_0 x}$. ($C_3$ defined below)

Note these conditions can always be satisfied by some C_0 and r because we can first find 3 groups (C_0, r) to satisfy each individual condition, and then take the maximal C_0 and minimal r, it's easy to check these conditions are still verified. And we let $T_{C_0,r}$ be the earliest time that such condition, i.e., $C_0\|\boldsymbol{N}(t)\|_F \leq \|\boldsymbol{W}_1(t)\|_{\mathcal{V}} \leq r$ fails. Thus for $t \in [0, T)$, we have $\|\boldsymbol{W}_1(t)\|_{\mathcal{V}} \leq \|\boldsymbol{W}_1(0)\|_{\mathcal{V}} e^{\frac{\mu_1 t}{2}}$. Therefore, if $T_{C_0,r}$ is finite, the condition that fails must be $C_0\|\boldsymbol{N}(t)\|_F \leq \|\boldsymbol{W}_1(t)\|_{\mathcal{V}}$, i.e. $C_0\|\boldsymbol{N}(T_{C_0,r})\|_F = \|\boldsymbol{W}_1(T_{C_0,r})\|_{\mathcal{V}}$.

By Lemma G.3, we have

$$\left(\|\boldsymbol{N}(0)\|_2^{\frac{2}{L}-1} + 2t(L-2)\rho\right)^{-\frac{1}{1-2/L}} \leq \|\boldsymbol{N}(t)\|_2 \leq \left(\|\boldsymbol{N}(0)\|_2^{\frac{2}{L}-1} - 2t(L-2)\rho\right)^{-\frac{1}{1-2/L}}.$$

Thus there exists constant C_3, such that for $t \leq T' := \min\{C_3\|\boldsymbol{N}(0)\|_2^{\frac{2}{L}-1}, T_{C_0,r}\}$, $\frac{1}{2}\|\boldsymbol{N}(0)\|_2 \leq \|\boldsymbol{N}(t)\|_2 \leq 2\|\boldsymbol{N}(0)\|_2$. Now we claim it must hold that $C_3\|\boldsymbol{N}(0)\|_2^{\frac{2}{L}-1} \geq T_{C_0,r}$. Otherwise, we have $\frac{C_0}{2}\|\boldsymbol{N}(0)\|_2 \leq C_0\|\boldsymbol{N}(0)\|_2 \leq C_0\|\boldsymbol{N}(T')\|_F \leq \|\boldsymbol{W}_1(T')\|_{\mathcal{V}} \leq e^{\mu_1 T'/2}\|\boldsymbol{W}_1(0)\|_{\mathcal{V}} \leq e^{\mu_1 T'/2} r$. Therefore, $C_3\|\boldsymbol{N}(0)\|_2^{\frac{2}{L}-1} = T' \leq -\frac{2}{\mu_1}\ln\frac{2r}{C_0\|\boldsymbol{N}(0)\|_2}$. Contradiction to the definition of C_0 and r.

As a result, we have

$$2C_0\sqrt{d}\|\boldsymbol{N}(0)\|_2 \geq 2C_0\|\boldsymbol{N}(0)\|_F \geq C_0\|\boldsymbol{N}(T_{c_0,r})\|_F = \|\boldsymbol{W}_1(T_{C_0,r})\|_{\mathcal{V}} \geq \|\boldsymbol{W}_1(0)\|_{\mathcal{V}} e^{\mu_1 T_{C_0,r}/2},$$

and

$$T_{C_0,r} \leq \frac{-2}{\mu}\ln\frac{\boldsymbol{W}_1(0)}{2\sqrt{d}C_0\|\boldsymbol{N}(0)\|_F}.$$

Thus by Lemma H.2, we know

$$\|\boldsymbol{W}(T_{C_0,r}) - \boldsymbol{W}_0\|_F \tag{40}$$
$$\leq \|\boldsymbol{W}(T_{C_0,r}) - \boldsymbol{W}_0\|_{F,1} + \|\boldsymbol{W}(T_{C_0,r}) - \boldsymbol{W}_0\|_{F,2} \tag{41}$$
$$\leq K\|\boldsymbol{W}_1(T_{C_0,r})\|_{\mathcal{V}} + \|\boldsymbol{M}(T_{C_0,r}) - \boldsymbol{W}_0\|_{F,2} + \|\boldsymbol{N}(T_{C_0,r})\|_{F,2} \tag{42}$$
$$\leq O(\|\boldsymbol{N}(0)\|_F) + O(\|\boldsymbol{N}(0)\|_F^2) + O(\|\boldsymbol{N}(0)\|_F) \tag{43}$$
$$= O(\|\boldsymbol{N}(0)\|_F). \tag{44}$$

□

Proof of Theorem 6.4. Let C_0, r be the constants predicted by Theorem H.4 w.r.t. to $\overline{\boldsymbol{W}}(\infty)$. We claim that we can pick large enough constant T, and α_0 sufficiently small, such that for all $\alpha \leq \alpha_0$, the initial condition in Theorem H.4 holds, i.e. $C_0 \left\|\boldsymbol{N}(0)\right\|_F \leq \left\|\boldsymbol{W}_1(0)\right\|_{\mathcal{V}} \leq r$, where $\boldsymbol{W}(0) := \phi\left(\alpha \boldsymbol{I}, 0.5\mu_1^{-1}(P-1))^{-1}\alpha^{-(P-1)} + T\right)$.

This is because we can first ensure $\left\|\overline{\boldsymbol{W}}(T) - \overline{\boldsymbol{W}}(\infty)\right\|_2$ is sufficiently small, i.e., smaller than $\frac{r}{2}$. By Theorem 6.2, we know when $\alpha \to 0$, $\left\|\overline{\boldsymbol{W}}(T) - \boldsymbol{W}(0)\right\|_{\mathcal{V}} \leq K \left\|\overline{\boldsymbol{W}}(T) - \boldsymbol{W}(0)\right\|_F = o(1)$ and $\left\|\boldsymbol{N}(0)\right\|_F = O(\alpha)$.

By Theorem H.4, we know there's a time $T_{c_0,r}$, such that $\left\|\boldsymbol{W}(T_{C_0,r} - \boldsymbol{W}_0\right\|_F = O(\left\|\boldsymbol{N}(0)\right\|_F) = O(\alpha)$. □

I PROOFS FOR COUNTER-EXAMPLE

Conjecture I.1 (Formal Statement, Gunasekar et al. (2017))**.** Suppose $f : \mathbb{R}^{d\times d} \to \mathbb{R}$ is a quadratic function and $\min_{\boldsymbol{W}\succeq \boldsymbol{0}} f(\boldsymbol{W}) = 0$. Then for any $\boldsymbol{W}_{\text{init}} \succ \boldsymbol{0}$ if $\overline{\boldsymbol{W}}_1 = \lim_{\alpha\to 0}\lim_{t\to+\infty} \phi(\alpha \boldsymbol{W}_{\text{init}}, t)$ exists and $f(\overline{\boldsymbol{W}}_1) = 0$, then $\left\|\overline{\boldsymbol{W}}_1\right\|_* = \min_{\boldsymbol{W}\succeq \boldsymbol{0}} \left\|\boldsymbol{W}\right\|_* \text{ s.t. } f(\boldsymbol{W}) = 0$.

Propsition I.2 (Formal Statement for Example 5.9)**.** For constant $R > 1$, let

$$\boldsymbol{M} = \begin{bmatrix} ? & ? & 1 & R \\ ? & ? & R & ? \\ 1 & R & ? & ? \\ R & ? & ? & ? \end{bmatrix}, \boldsymbol{M}_{\text{norm}} = \begin{bmatrix} R & 1 & 1 & R \\ 1 & R & R & 1 \\ 1 & R & R & 1 \\ R & 1 & 1 & R \end{bmatrix}, \text{ and } \boldsymbol{M}_{\text{rank}} = \begin{bmatrix} 1 & R & 1 & R \\ R & R^2 & R & R^2 \\ 1 & R & 1 & R \\ R & R^2 & R & R^2 \end{bmatrix}.$$

and

$$\mathcal{L}(\boldsymbol{U}) = \frac{1}{2} f(\boldsymbol{U}\boldsymbol{U}^\top), \quad f(\boldsymbol{W}) = \frac{1}{2}\sum_{(i,j)\in\Omega} (W_{ij} - M_{ij})^2$$

where $\Omega = \{(1,3),(1,4),(2,3),(3,1),(3,2),(4,1)\}$.

Then for any $\boldsymbol{W}_{\text{init}} \succeq \boldsymbol{0}$, s.t. $\boldsymbol{u}_1^\top \boldsymbol{W}_{\text{init}} \boldsymbol{u}_1 > 0$,

$$\lim_{\alpha\to 0}\lim_{t\to+\infty} \phi(\alpha \boldsymbol{W}_{\text{init}}, t) = \boldsymbol{M}_{\text{rank}}.$$

Moreover, we have

$$\left\|\boldsymbol{M}_{\text{rank}}\right\|_* = 2R^2 + 2 > 4R = \left\|\boldsymbol{M}_{\text{norm}}\right\|_* = \min_{\boldsymbol{W}\succeq \boldsymbol{0}} \left\|\boldsymbol{W}\right\|_* \text{ s.t. } f(\boldsymbol{W}) = 0.$$

Proof for Propsition I.2. We define $\boldsymbol{W}_{1,\epsilon}^{\text{G}}(t), \boldsymbol{W}_1^{\text{G}}(t)$ in the same way as in Definition 5.1, Theorem 5.6.

$$\boldsymbol{W}_{1,\epsilon}^{\text{G}}(t) := \phi\left(\epsilon \boldsymbol{u}_1\boldsymbol{u}_1^\top, t\right),$$
$$\boldsymbol{W}_1^{\text{G}}(t) := \lim_{\epsilon\to 0} \boldsymbol{W}_{1,\epsilon}^{\text{G}}(\tfrac{1}{2\mu_1}\log\tfrac{1}{\epsilon} + t).$$

Below we will show

1. Assumption 5.7 and Assumption 5.5 are satisfied.
2. $\left\|\boldsymbol{W}_1^{\text{G}}(t)\right\|_{\text{F}}$ bounded for $t \geq 0$;
3. $\lim_{t\to+\infty} \boldsymbol{W}_1^{\text{G}}(t) = \boldsymbol{M}_{\text{rank}}$;
4. $\boldsymbol{M}_{\text{norm}} = \arg\min_{\boldsymbol{W}\succeq \boldsymbol{0}} \left\|\boldsymbol{W}\right\|_* \text{ s.t. } f(\boldsymbol{W}) = 0$.

Thus Since $\boldsymbol{M}_{\text{rank}}$ is a global minimizer of $f(\,\cdot\,)$, applying Theorem 5.8 finishes the proof.

Proof for Item 1. Let $\boldsymbol{M}_0 := \nabla f(\mathbf{0})$, then

$$\boldsymbol{M}_0 = \begin{bmatrix} 0 & 0 & 1 & R \\ 0 & 0 & R & 0 \\ 1 & R & 0 & 0 \\ R & 0 & 0 & 0 \end{bmatrix}.$$

Let $\boldsymbol{A} := \left[\begin{smallmatrix} 1 & R \\ R & 0 \end{smallmatrix}\right]$, then we have $\lambda_1(\boldsymbol{A}) = \frac{1+\sqrt{1+R^2}}{2}$, $\lambda_2(\boldsymbol{A}) = \frac{1-\sqrt{1+R^2}}{2}$, thus $\lambda_1(\boldsymbol{A}) > |\lambda_2(\boldsymbol{A})| > 0 > \lambda_2(\boldsymbol{A})$. As a result, $\lambda_1(\boldsymbol{A}) = \|\boldsymbol{A}\|_2$. Let $\boldsymbol{v}_1 \in \mathbb{R}^2$ be the top eigenvector of $\boldsymbol{A}$. We claim that $\boldsymbol{u}_1 = \left[\begin{smallmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_1 \end{smallmatrix}\right] \in \mathbb{R}^4$ is the top eigenvector of $\nabla f(\mathbf{0})$. First by definition it is easy to check that $\boldsymbol{M}_0 \boldsymbol{u}_1 = \lambda_1(\boldsymbol{A})\boldsymbol{u}_1$. Further noticing that $\boldsymbol{M}_0^2 = \left[\begin{smallmatrix} \boldsymbol{A}^2 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{A}^2 \end{smallmatrix}\right]$, we know $\lambda_i^2(\boldsymbol{M}_0) \in \{\lambda_1^2(\boldsymbol{A}), \lambda_2^2(\boldsymbol{A})\}$, $\forall i \in \{1,2,3,4\}$. That is, $\lambda_1(\boldsymbol{M}_0) = \lambda_1(\boldsymbol{A})$, $\lambda_2(\boldsymbol{M}_0) = -\lambda_2(\boldsymbol{A})$, $\lambda_3(\boldsymbol{M}_0) = \lambda_2(\boldsymbol{A})$, and $\lambda_4(\boldsymbol{M}_0) = -\lambda_1(\boldsymbol{A})$. Thus Assumption 5.5 is satisfied. Also note that f is quadratic, thus analytic, i,e., Assumption 5.7 is also satisfied.

Proof for Item 2. Let $(x_\epsilon(t), y_\epsilon(t)) \in \mathbb{R}^2$ be the gradient flow of $g(x,y) = \frac{1}{2}(x^2-1)^2 + (xy-R)^2$ starting from $(x_\epsilon(0), y_\epsilon(0)) = \sqrt{\epsilon}\boldsymbol{v}_1$.

$$\begin{aligned} \frac{\mathrm{d}x(t)}{\mathrm{d}t} &= (1-x(t)^2)x(t) - 2y(t)(x(t)y(t)-R) \\ \frac{\mathrm{d}y(t)}{\mathrm{d}t} &= -2x(t)(x(t)y(t)-R) \end{aligned} \tag{45}$$

Let $\boldsymbol{W}_\epsilon(t)$ be the following matrix:

$$\boldsymbol{W}_\epsilon(t) := \begin{bmatrix} x_\epsilon(t) \\ y_\epsilon(t) \\ x_\epsilon(t) \\ y_\epsilon(t) \end{bmatrix} \begin{bmatrix} x_\epsilon(t) & y_\epsilon(t) & x_\epsilon(t) & y_\epsilon(t) \end{bmatrix}.$$

Then it is easy to verify that $\boldsymbol{W}_\epsilon(0) = \boldsymbol{W}_{1,\epsilon}^{\mathrm{G}}(0)$ and $\boldsymbol{W}_\epsilon(t)$ satisfies (2). Thus by the existence and uniqueness theorem, we have $\boldsymbol{W}_\epsilon(t) = \boldsymbol{W}_{1,\epsilon}^{\mathrm{G}}(t)$ for all t. Taking the limit $\epsilon \to 0$, we know that $\boldsymbol{W}_1^{\mathrm{G}}(t)$ can also be written in the following form:

$$\boldsymbol{W}_1^{\mathrm{G}}(t) = \begin{bmatrix} x(t) \\ y(t) \\ x(t) \\ y(t) \end{bmatrix} \begin{bmatrix} x(t) & y(t) & x(t) & y(t) \end{bmatrix},$$

and $(x_\epsilon(t), y_\epsilon(t)) \in \mathbb{R}^2$ is a gradient flow of $g(x,y) = \frac{1}{2}(x^2-1)^2 + (xy-R)^2$.

Since $g(x(t), y(t))$ is non-increasing overtime, and $\lim_{t\to-\infty} g(x(-t), y(-t)) = g(x(-\infty), y(-\infty)) = g(0,0) = R^2 + 0.5$, we know $|x(t)y(t)| \le 3R$ for all t. So whenever $y^2(t) - x^2(t) \ge 9R^2$, we have $x^2(t) \le \frac{9R^2}{y^2(t)} \le \frac{9R^2}{y^2(t)-x^2(t)} \le 1$. In this case, $\frac{\mathrm{d}(y^2(t)-x^2(t))}{\mathrm{d}t} = 2x^2(t)(x^2(t)-1) \le 0$. Combining this with $y(-\infty)^2 - x(-\infty)^2 = 0 \le 9R^2$, we have $y^2(t) - x^2(t) \le 9R^2$ for all t, which also implies that $y(t)$ is bounded. Noticing that $9R^2 \ge g(x(t), y(t)) \ge (x^2(t)-1)^2$, we know $x^2(t)$ is also bounded. Therefore, $\boldsymbol{W}_1^G(t)$ is bounded.

Proof for Item 3. Note that $(x(\infty), y(\infty))$ is a stationary point of $g(x,y)$. It is clear $g(x,y)$ only has 3 stationary points — $(0,0)$, $(1,R)$ and $(-1,-R)$. Thus $\overline{\boldsymbol{W}}_1$ can only be $\mathbf{0}$ or $\boldsymbol{M}_{\mathrm{rank}}$. However, since for all t, $f(\boldsymbol{W}_1^{\mathrm{G}}(t)) < f(\mathbf{0})$, $\overline{\boldsymbol{W}}_1 = \lim_{t\to\infty} \boldsymbol{W}_1^{\mathrm{G}}(t)$ cannot be $\mathbf{0}$. So $\overline{\boldsymbol{W}}_1$ must be $\boldsymbol{M}_{\mathrm{rank}}$.

Proof for Item 4. Let m_{ij} be (i,j)th element of $\boldsymbol{M}$. Suppose $\boldsymbol{M} \succeq \mathbf{0}$, we have

$$(\boldsymbol{e}_1 - \boldsymbol{e}_4)^\top \boldsymbol{M}(\boldsymbol{e}_1 - \boldsymbol{e}_4) \ge 0 \Longrightarrow m_{11} + m_{44} \ge m_{14} + m_{41} = 2R$$
$$(\boldsymbol{e}_2 - \boldsymbol{e}_3)^\top \boldsymbol{M}(\boldsymbol{e}_2 - \boldsymbol{e}_3) \ge 0 \Longrightarrow m_{22} + m_{33} \ge m_{23} + m_{32} = 2R$$

Thus $4R = \min_{\boldsymbol{W}\succeq\mathbf{0}} \|\boldsymbol{W}\|_* \text{ s.t. } f(\boldsymbol{W}) = 0$, where the equality is only attained at $m_{ii} = R, i = 1,2,3,4$. Otherwise, either $\begin{bmatrix} m_{11} & m_{14} \\ m_{41} & m_{44} \end{bmatrix}$ or $\begin{bmatrix} m_{22} & m_{23} \\ m_{32} & m_{33} \end{bmatrix}$ will have negative eigenvalues. Contradiction to that $\boldsymbol{M} \succeq \mathbf{0}$.

Below we will show the rest unknown off-diagonal entries must be 1. Let $\boldsymbol{V} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, we have

$$\boldsymbol{M} \succeq \mathbf{0} \Longrightarrow \boldsymbol{V}\boldsymbol{M}\boldsymbol{V}^\top \succeq \mathbf{0} \Longrightarrow \begin{bmatrix} 0 & m_{13} - m_{23} & m_{14} - m_{24} \\ m_{31} - m_{32} & R & R \\ m_{41} - m_{42} & R & R \end{bmatrix} \succeq \mathbf{0},$$

which implies $m_{13} = m_{23}$, $m_{14} = m_{24}$.

With the same argument for $\boldsymbol{V} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$, we have $m_{13} = m_{14}$, $m_{23} = m_{24}$. Also note $\boldsymbol{M}$ is symmetric and $m_{13} = 1$, thus $m_{ij} = m_{ji} = 1,\ \forall i = 1, 2, j = 3, 4$. Thus $\boldsymbol{M}_{\text{norm}} = \arg\min_{\boldsymbol{W} \succeq \mathbf{0}} \|\boldsymbol{W}\|_* \ \text{ s.t. } \ f(\boldsymbol{W}) = 0$, which is unique. $\square$