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

Zhiyuan Li, Yuping Luo
Princeton University
{zhiyuanli, yupingl}@cs.princeton.edu

Kaifeng Lyu
Tsinghua University
vfleaking@gmail.com

ABSTRACT

Matrix factorization is a simple and natural test-bed to investigate the implicit regularization of gradient descent. Gunasekar et al. (2017) 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. (2017). We also extend the results to the case where depth $\geq 3$, 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 of 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 $W^* \in \mathbb{R}^{d \times d}$ of rank $r^* \ll d$, one needs to find a low-rank solution $W$ 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 $W^*$ is symmetric and positive semidefinite, one way to solve all these problems is to parameterize $W$ as $W = UU^\top$ for $U \in \mathbb{R}^{d \times r}$ and optimize $\mathcal{L}(U) := \frac{1}{2} f(UU^\top)$, 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 $W^*$. 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.

*Alphabet ordering.
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/attrition effect encourages low-rank, and it is further intensified with depth in deep matrix factorization (i.e., GD optimizes \( f(U_1 \cdots U_L) \) 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 rank-constrained optimization and relaxes the rank constraint by 1 whenever it fails to reach a global minimizer of \( f(\cdot) \) with the current rank constraint. As a by-product, we refute Conjecture 1.1 by demonstrating a counterexample (Example 5.9).

We also extend our results to deep matrix factorization Section 6, where we prove that the trajectory of GF with infinitesimal identity initialization converges to a deep version of GLRL, at least in the early stage of the optimization. We also use this result to confirm 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 E.

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 (Ji 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 by 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 by 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 \( A, B \), we define \( \langle A, B \rangle := \text{Tr}(AB^\top) \) as their inner product. We use \( \|A\|_F, \|A\|_1 \) and \( \|A\|_2 \) to denote the Frobenius norm, nuclear norm and the largest singular value of \( A \) respectively. For a matrix \( A \in \mathbb{R}^{d \times d} \), we use \( \lambda_1(A), \ldots, \lambda_d(A) \) to denote the eigenvalues of \( A \) in decreasing order (if they are all reals). We define \( \mathbb{S}_d \) as the set of symmetric \( d \times d \) matrices and \( \mathbb{S}_d^+ \subseteq \mathbb{S}_d \) as the set of positive semidefinite (PSD) matrices. We write \( A \succeq B \) or \( B \preceq A \) if \( A - B \) is PSD. We use \( \mathbb{S}_{d,r}^+, \mathbb{S}_{d,\leq r}^+ \) to denote the set of \( d \times d \) PSD matrices with rank = \( r \), \( \leq r \) respectively.
Matrix Factorization. Matrix factorization problem asks one to optimize $\mathcal{L}(U, V) := \frac{1}{2} f(UV^\top)$ among $U, V \in \mathbb{R}^{d \times r}$, where $f : \mathbb{R}^{d \times d} \to \mathbb{R}$ is a convex function and in this paper we assume $f$ is $C^3$-smooth. A notable example is matrix sensing. There is an unknown rank-$r^*$ matrix $W^* \in \mathbb{R}^{d \times d}$ with $r^* \ll d$. Given $m$ measurements $X_1, \ldots, X_m \in \mathbb{R}^{d \times d}$, one can observe $y_i := (X_i, W^*)$ through each measurement. The goal of matrix sensing is to reconstruct $W^*$ via minimizing $f(W) := \frac{1}{2} \sum_{i=1}^m \langle (W, X_i) - y_i \rangle^2$. Matrix completion is a notable special case of matrix sensing in which every measurement has the form $X_i = e_{p_i}e_{q_i}^\top$, where $\{e_1, \cdots, e_d\}$ stands for the standard basis (i.e., exactly one entry is observed through each measurement).

Note that matrix factorization in the general case can be reduced to this symmetric case: let $U' = [U']_i \in \mathbb{R}^{2d \times r}$, $f'(\begin{bmatrix} A & B \\ C & D \end{bmatrix}) = \frac{1}{2} f(\mathbf{B}) + \frac{1}{2} f(\mathbf{C})$, then $f(UU^\top) = f'(UU^\top)$. So in this paper we focus on the symmetric case as in previous works (Gunasekar et al., 2017), i.e., finding a low-rank solution for the convex optimization problem: $\min_{W \succeq 0} f(W)$. For this, we parameterize $W$ as $W = UU^\top$ for $U \in \mathbb{R}^{d \times r}$ and optimize $\mathcal{L}(U) := \frac{1}{2} f(UU^\top)$. We assume WLOG throughout this paper that $f(W) = f(UU^\top)$; otherwise, we can set $f'(W) = \frac{1}{2} (f(W) + f(W^*))$ so that $f(W) = f(UU^\top)$ while $\mathcal{L}(U) = \frac{1}{2} f(UU^\top)$ is unaffected. This assumption makes $\nabla f(W)$ symmetric for every symmetric $W$.

Gradient Flow. In this paper, we analyze Gradient Flow (GF) for symmetric matrix factorization, defined as the solution of the following ODE for $U(t) \in \mathbb{R}^{d \times r}$:

$$
\frac{dU}{dt} = -\nabla_\mathcal{L}(U) = -\nabla f(UU^\top)U.
$$

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

$$
\frac{dW}{dt} = -W\nabla f(W) - \nabla f(W)W =: g(W).
$$

We use $\phi(W_0, t)$ to denote the matrix $W(t)$ in (2) when $W(0) = W_0 \succeq 0$. Throughout this paper, we assume $\phi(W_0, t)$ exists for all $t \in \mathbb{R}$, $W_0 \succeq 0$. It is easy to prove that $U$ is a stationary point of $\mathcal{L}(\cdot)$ (i.e., $\nabla \mathcal{L}(U) = 0$) iff $W = UU^\top$ is a critical point of (2) (i.e., $g(W) = 0$); see Lemma C.1 for a proof. If $W$ is a minimizer of $f(\cdot)$ in $S^r_+$, then $W$ is a critical point of (2), but the reverse may not be true, e.g., $g(0) = 0$, but 0 is not necessarily a minimizer.

In this paper, we particularly focus on the overparameterized case, where $r = d$, to understand the implicit regularization of GF when there is no rank constraint for the matrix $W$.

4 Warmup Examples

First, we illustrate how GD performs greedy learning using two warmup examples.

Linearization Around the Origin. In general, for a loss function $\mathcal{L}(U) = \frac{1}{2} f(UU^\top)$, we can always apply Taylor expansion $f(W) \approx f(0) + \langle W, \nabla f(0) \rangle$ around the origin to approximate it with a linear function. This motivates us to study the linear case: $f(W) := f_0 - \langle W, Q \rangle$ for some symmetric matrix $Q$. In this case, the matrix $U$ follows the ODE, $\frac{dU}{dt} = QU$, which can be understood as a continuous version of the classical power iteration method for solving the top eigenvector. Let $Q := \sum_{i=1}^d \mu_i v_i v_i^\top$ be the eigendecomposition of $Q$, where $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_d$ and $v_1, \ldots, v_d$ are orthogonal to each other. Then we can write the solution as:

$$
U(t) = e^{tQ}U(0) = \left( \sum_{i=1}^d e^{\mu_i t}v_i v_i^\top \right)U(0).
$$

When $\mu_1 > \mu_2$, the ratio between $e^{\mu_1 t}$ and $e^{\mu_2 t}$ for $i \neq 1$ increases exponentially fast. As $t \to +\infty$, $U(t)$ and $W(t)$ become approximately rank-1 as long as $v_1^\top U(0) \neq 0$, i.e.,

$$
\lim_{t \to +\infty} e^{-\mu_1 t}U(t) = v_1^\top U(0), \quad \lim_{t \to +\infty} e^{2\mu_1 t}W(t) = (v_1^\top W(0)v_1)v_1 v_1^\top.
$$

The analysis for the simple linear case reveals that GD encourages low-rank through a process similar to power iteration. However, $f(W)$ is non-linear in general, and the linear approximation is close to $f(W)$ only if $W$ 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 $W(t)$ grows to be so large that the linear approximation is far from the actual $f(W)$?

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): $\mathcal{L}(U) = \frac{1}{2} f(UU^\top)$, $f(W) = \frac{1}{2} \| W - W^* \|_F^2$ for some unknown PSD matrix $W^*$. GF in this case can be written as:

Published as a conference paper at ICLR 2021
Through the first warmup example, we may understand why GD learns a rank-1 matrix in the early phase. This means for every $\lambda_{q} > 0$, guaranteed that $\epsilon > 0$ away from the origin. $W_{1} = V_{1}U_{1}$ is a diagonal matrix, since the time derivatives of non-diagonal coordinates stay 0 during training. Let $W(t) = \text{diag}(\sigma_{1}(t), \sigma_{2}(t), \ldots, \sigma_{d}(t))$, then $\sigma_{i}(t)$ satisfies the dynamical equation $\frac{d\sigma_{i}(t)}{dt} = 2\sigma_{i}(t)(\mu_{i} - \sigma_{i}(t))$, and thus $\sigma_{i}(t) = \frac{\alpha_{i}}{\alpha_{i} + (\mu_{i} - \sigma_{i}(t))}$ for large $t$. This shows that every $\sigma_{i}(t)$ increases from $\alpha$ to $\mu_{i}$ over time. As $\alpha \to 0$, every $\sigma_{i}(t)$ has a sharp transition from near 0 to near $\mu_{i}$ at time roughly $\left(\frac{1}{2\mu_{i}} + o(1)\right) \log \frac{1}{\alpha}$, which can be seen from the following limit:

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

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 $i = d$), $\lim_{\alpha \to 0} W(q \log(1/\alpha)) = \text{diag}(\mu_{1}, \mu_{2}, \ldots, \mu_{i}, 0, 0, \ldots, 0)$. Therefore, when the initialization is sufficiently small, GF learns each component of $W^{*}$ 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 (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, … 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). See Appendix A for a comparison with existing greedy algorithms for rank-constrained optimization.

The GLRL algorithm consists of several phases, numbered from 1. In phase $r$, GLRL increases the rank constraint to $r$ and optimizes $L(U_{r}) := \frac{1}{2}f(U_{r}U_{r}^{T})$ among $U_{r} \in \mathbb{R}^{d \times r}$ via GD until it reaches a stationary point $U_{r}(\infty)$, i.e., $\nabla L(U_{r}(\infty)) = 0$. At convergence, $W_{r} := U_{r}(\infty)U_{r}(\infty)^{T}$ is a critical point of (2), and we call it the $r$-th critical point of GLRL. If $W_{r}$ is further a minimizer of $f(\cdot)$ in $S_{d}$, or equivalently, $\lambda_{1}(\nabla f(W_{r})) \leq 0$ (see Lemma C.2), GLRL returns $W_{r}$; otherwise GLRL enters phase $r + 1$.

To set the initial point of GD in phase $r$, GLRL appends a small column vector $\delta_{r} \in \mathbb{R}^{d}$ to the resulting stationary point $U_{r-1}(\infty)$ from the last phase, i.e., $U_{r}(0) = [U_{r-1}(\infty) \delta_{r}] \in \mathbb{R}^{d \times r}$ (in the case of $r = 1$, $U_{1}(0) = [\delta_{1}] \in \mathbb{R}^{d \times 1}$). In this way, $U_{r}(0)U_{r}(0)^{T} = W_{r-1} + \delta_{r}\delta_{r}^{T}$ is perturbed away from the $(r-1)$-th critical point. In GLRL, we set $\delta_{r} = \sqrt{\epsilon}u_{r}$, where $u_{r}$ is the top eigenvector of $-\nabla f(W_{r-1})$ with unit norm $||u_{r}||_{2} = 1$, and $\epsilon > 0$ is a parameter controlling the magnitude of perturbation (preferably very small). Note that it is guaranteed that $\lambda_{1}(\nabla f(W_{r-1})) > 0$; otherwise $W_{r-1}$ is a minimizer of the convex function $f(\cdot)$ in $S_{d}$ and GLRL exits before phase $r$.

**Algorithm 1: Greedy Low-Rank Learning**

**Parameter**: step size $\eta > 0$; small $\epsilon > 0$

$r \leftarrow 0, W_{0} \leftarrow 0 \in \mathbb{R}^{d \times d}$, and $U_{0}(\infty) \in \mathbb{R}^{d \times d}$

while $\lambda_{1}(\nabla f(W_{r})) > 0$ do

$r \leftarrow r + 1$

$u_{r} \leftarrow$ unit top eigenvector of $-\nabla f(W_{r-1})$

$U_{r}(0) \leftarrow \sqrt{\epsilon}u_{r} \in \mathbb{R}^{d \times d}$

for $t = 0, 1, \ldots$ do

$U_{r}(t+1) \leftarrow U_{r}(t) - \eta\nabla L(U_{r}(t))$

end for

$W_{r} \leftarrow U_{r}(\infty)U_{r}(\infty)^{T}$

return $W_{r}$

*In practice, we approximate the infinite time limit by running sufficiently many steps.*
Trajectory of GLRL. We define the (limiting) trajectory of GLRL by taking the learning rate \( \eta \to 0 \). The goal is to show that the trajectory of GLRL is close to that of GF with infinitesimal initialization. Recall that \( \phi(W_0, t) \) stands for the solution \( W(t) \) in (2) when \( W(0) = W_0 \).

**Definition 5.1 (Trajectory of GLRL).** Let \( \mathbf{W}_{0, \epsilon} := 0 \) be the 0th critical point of GLRL. For every \( r \geq 1 \), if the \((r-1)\)th critical point \( \mathbf{W}_{r-1, \epsilon} \) exists and is not a minimizer of \( f(\cdot) \) in \( \mathbb{S}^d_+ \), we define \( \mathbf{W}^{\epsilon}_{r, \epsilon}(t) := \phi(\mathbf{W}_{r-1, \epsilon} + \epsilon \mathbf{u}_{r, \epsilon} \mathbf{u}^\top_{r, \epsilon}, t) \), where \( \mathbf{u}_{r, \epsilon} \) is a top eigenvector of \( \nabla f(\mathbf{W}_{r-1, \epsilon}) \) with unit norm, \( \| \mathbf{u}_{r, \epsilon} \|_2 = 1 \). We define \( \mathbf{W}^{\epsilon}_{r, \epsilon} := \lim_{t \to +\infty} \mathbf{W}^{\epsilon}_{r, \epsilon}(t) \) to be the \( r \)-th critical point of GLRL if the limit exists.

Throughout this paper, we always focus on the case where the top eigenvalue of every \( \nabla f(\mathbf{W}_{r-1, \epsilon}) \) is unique. In this case, the trajectory of GLRL is unique for every \( \epsilon > 0 \), since the normalized top eigenvectors can only be \( \pm \mathbf{u}_{r, \epsilon} \), and both of them lead to the same \( \mathbf{W}^{\epsilon}_{r, \epsilon} \).

5.1 The Limiting Trajectory: A General Theorem for Dynamical System

To prove the equivalence between GF and GLRL, we first introduce our high-level idea by analyzing the behavior of a more general dynamical system around its critical point, say \( 0 \). A specific example is (2) if we set \( \theta \) to be the vectorization of \( W \).

\[
\frac{d\theta}{dt} = g(\theta), \quad \text{where} \quad g(0) = 0. \tag{6}
\]

We use \( \phi(\theta_0, t) \) to denote the value of \( \theta(t) \) in the case of \( \theta(0) = \theta_0 \). We assume that \( g(\theta) \) is \( C^2 \)-smooth with \( J(\theta) \) being the Jacobian matrix and \( \phi(\theta_0, t) \) exists for all \( \theta_0 \) and \( t \). For ease of presentation, in the main text we assume \( J(0) \) is diagonalizable over \( \mathbb{R} \) and defer the same result for the general case into Appendix G.3. Let \( J(0) = \tilde{V} \tilde{D} \tilde{V}^{-1} \) be the eigendecomposition, where \( \tilde{V} \) is an invertible matrix and \( \tilde{D} = \text{diag}(\tilde{\mu}_1, \ldots, \tilde{\mu}_d) \) is the diagonal matrix consisting of the eigenvalues \( \tilde{\mu}_1 \geq \tilde{\mu}_2 \geq \cdots \geq \tilde{\mu}_d \). Let \( \tilde{V} = (\tilde{v}_1, \ldots, \tilde{v}_d) \) and \( \tilde{V}^{-1} = (\tilde{u}_1, \ldots, \tilde{u}_d) \), then \( \tilde{u}_i, \tilde{v}_i \) are left and right eigenvectors associated with \( \tilde{\mu}_i \) and \( \tilde{\mu}^i \). We can rewrite the eigendecomposition as \( J(0) = \sum_{i=1}^d \tilde{\mu}_i \tilde{v}_i \tilde{u}_i^\top \). We also assume the top eigenvalue \( \tilde{\mu}_1 \) is positive and unique. Note \( \tilde{\mu}_1 > 0 \) means the critical point \( \theta = 0 \) is unstable, and in matrix factorization it means \( 0 \) is a strict saddle point of \( L(\cdot) \). 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.** For any \( \theta_0 \in \mathbb{R}^d \) and \( u \in \mathbb{R}^d \), we say that \( \{ \theta_{\delta} \}_{\alpha \in (0, 1)} \) converges to \( \theta_0 \) with positive alignment with \( u \) if \( \lim_{\alpha \to 0} \theta_{\delta} = \theta_0 \) and \( \liminf_{\alpha \to 0} \langle \frac{\theta_{\delta} - \theta_{\delta}}{\| \theta_{\delta} - \theta_{\delta} \|_2}, u \rangle > 0 \).

A special case is the direction of \( \theta_{\delta} \to \theta_0 \) converges, i.e., \( \tilde{\theta} := \lim_{\alpha \to 0} \frac{\theta_{\delta} - \theta_{\delta}}{\| \theta_{\delta} - \theta_{\delta} \|_2} \) exists. In this case, \( \{ \theta_{\delta} \} \) has positive alignment with either \( u \) or \( -u \) except for a zero-measure subset of \( \tilde{\theta} \). This means any convergent sequence generically falls into either of these two categories.

The following theorem shows that if the initial point \( \theta_{\delta} \) converges to \( 0 \) with positive alignment with \( \tilde{u}_1 \) as \( \alpha \to 0 \), the trajectory starting with \( \theta_{\delta} \) converges to a unique trajectory \( z(t) := \phi(\alpha \tilde{v}_1, t + \frac{1}{\mu_1} \log \frac{1}{\alpha}) \). By symmetry, there is another unique trajectory for sequences \( \{ \theta_{\delta} \} \) with positive alignment to \( -\tilde{u}_1 \), which is \( z'(t) := \phi(-\alpha \tilde{v}_1, t + \frac{1}{\mu_1} \log \frac{1}{\alpha}) \). This is somewhat surprising: different initial points should lead to very different trajectories, but our analysis shows that generically there are only two limiting trajectories for infinitesimal initialization. We will soon see how this theorem helps in our analysis for matrix factorization in Sections 5.2 and 5.3.

**Theorem 5.3.** Let \( z_{\alpha}(t) := \phi(\alpha \tilde{v}_1, t + \frac{1}{\mu_1} \log \frac{1}{\alpha}) \), for every \( \alpha > 0 \), then \( z(t) := \lim_{\alpha \to 0} z_{\alpha}(t) \) exists and is also a solution of (6), i.e., \( z(t) = \phi(z(0), t) \). If \( \delta_{\alpha} \) converges to \( 0 \) with positive alignment with \( u_1 \) as \( \alpha \to 0 \), then \( \forall t \in \mathbb{R} \), there is a constant \( C > 0 \) such that

\[
\| \phi(\delta_{\alpha}, t + \frac{1}{\mu_1} \log \frac{1}{\alpha}) - z(t) \|_2 \leq C \cdot \| \delta_{\alpha} \|_2^{\frac{\hat{\gamma}}{\mu_1 + 5}}, \tag{7}
\]

for every sufficiently small \( \alpha \), where \( \hat{\gamma} := \mu_1 - \mu_2 > 0 \) 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 \( \theta \), by Taylor expansion of \( g(\theta) \), the dynamics is approximately \( \frac{d\theta}{dt} \approx J(0) \theta \), which can be understood as a continuous version of power iteration. If the linear approximation is exact, then \( \theta(t) = e^{tJ(0)} \theta(0) \). For large enough \( t_0 \),
\[ e^{\theta_0 J(0)} = \sum_{i=1}^{d} e^{\mu_i \theta_0} \bar{v}_i \bar{u}_i^\top = e^{\mu_1 \theta_0} \bar{v}_1 \bar{u}_1^\top + O(e^{\mu_2 \theta_0}). \]

Therefore, as long as the initial point \( \theta(0) \) has a positive inner product with \( \bar{u}_1 \), \( \bar{J}(t) \) should be very close to \( \epsilon \bar{v}_1 \), for some \( \epsilon > 0 \), and the rest of the trajectory after \( t_0 \) should be close to the trajectory starting from \( \epsilon \bar{v}_1 \). However, here is a tradeoff: we should choose \( t_0 \) to be large enough so that the power iteration takes effect; but if \( t_0 \) is so large that the norm of \( \bar{J}(t_0) \) 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 \( \bar{J}(t_0) \) is well approximated by \( \epsilon \bar{v}_1 \) and the difference between \( \bar{J}(t) \) and \( f(\epsilon \bar{v}_1, t) \) is bounded as well. We defer the details to Appendix G.

5.2 Equivalence Between GD and GLRL: Rank-One Case

Now we establish the equivalence between GD 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 \( g(W) := -W \nabla f(W) - \nabla f(W)W \) and \( J(W) \) be its Jacobian. Then \( J(0) \) is symmetric and thus diagonalizable. Let \( -\nabla f(0) = \sum_{i=1}^{d} \mu_i u_{1[i]} u_{1[i]}^\top \) be the eigendecomposition of the symmetric matrix \( -\nabla f(0) \), where \( \mu_1 \geq \mu_2 \geq \cdots \geq \mu_d \). Then \( J(0) \) has the form:

\[ J(0)[\Delta] = \sum_{i,j=1}^{d} \left( \mu_i \mu_j \right) \left( \Delta, u_{1[i]} u_{1[j]}^\top \right) u_{1[i]} u_{1[j]}^\top, \]

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

We simplify the notation by letting \( u_1 := u_{1[1]} \). A direct corollary of Lemma 5.4 is that \( u_1 u_1^\top \) is the top eigenvector of \( J(0) \). According to Theorem 5.3, now there are only two types of trajectories, which correspond to infinitesimal initialization \( W_\alpha \to 0 \) with positive alignment with \( u_1 u_1^\top \) or \( -u_1 u_1^\top \). As the initialization must be PSD, \( W_\alpha \to 0 \) cannot have positive alignment with \( -u_1 u_1^\top \). For the former case, Theorem 5.6 below states that, for every fixed time \( t \), the GF solution \( \phi(W_\alpha, T(W_\alpha) + t) \) after shifting by a time offset \( T(W_\alpha) := \frac{1}{\mu_1} \log((W_\alpha, u_1 u_1^\top)^{-1}) \) converges to the GLRL solution \( W_1^G(t) \) as \( W_\alpha \to 0 \). The only assumption for this result is that \( 0 \) is not a minimizer of \( f(\cdot) \) in \( S^+_{d\times d} \) (which is equivalent to \( \lambda_1(-\nabla f(0)) > 0 \) and \( -\nabla f(0) \) has an eigenvalue gap. In the full observation case, this assumption is satisfied easily if the ground-truth matrix has a unique top eigenvalue. The proof for Theorem 5.6 is deferred to Appendix I.1.

**Assumption 5.5.** \( \mu_1 > \max\{\mu_2, \mu_0\} \), where \( \mu_1 := \lambda_1(-\nabla f(0)) \), \( \mu_2 := \lambda_2(-\nabla f(0)) \).

**Theorem 5.6.** Under Assumption 5.5, the following limit \( W_1^G(t) \) exists and is a solution of (2).

\[
W_1^G(t) := \lim_{\epsilon \to 0} W_1^G(\epsilon) \left( \frac{1}{2\mu_1} \log \frac{1}{\epsilon} + t \right) = \lim_{\epsilon \to 0} \phi \left( \epsilon u_1 u_1^\top, \frac{1}{2\mu_1} \log \frac{1}{\epsilon} + t \right). 
\]

Let \( \{W_\alpha\} \subseteq S^+_{d\times d} \) be PSD matrices converging to \( 0 \) with positive alignment with \( u_1 u_1^\top \) as \( \alpha \to 0 \), that is, \( \lim_{\alpha \to 0} W_\alpha = 0 \) and \( \exists \alpha_0, q > 0 \) such that \( \langle W_\alpha, u_1 u_1^\top \rangle \geq q \|W_\alpha\|_F \) for all \( \alpha < \alpha_0 \). Then for every sufficiently small \( \alpha \), there is a constant \( C > 0 \) such that

\[
\left\| \phi \left( W_\alpha, \frac{1}{2\mu_1} \log \frac{1}{\langle W_\alpha, u_1 u_1^\top \rangle} + t \right) - W_1^G(t) \right\|_F \leq C \|W_{\alpha}\|_F^{\frac{1}{2\mu_1+\gamma}}
\]

for every sufficiently small \( \alpha \), where \( \gamma := 2\mu_1 - (\mu_1 + \mu_2) = \mu_1 - \mu_2 \).

It is worth to note that \( W_1^G(t) \) has rank \( \leq 1 \) for any \( t \in \mathbb{R} \), since every \( W_1^G(t) \) has rank \( \leq 1 \) and the set \( S^+_{d\times 1} \) is closed. This matches with the first warmup example: GD does start learning with rank-1 solutions. Interestingly, in the case where the limit \( \bar{W}_1 := \lim_{t \to +\infty} W_1^G(t) \) happens to be a minimizer of \( f(\cdot) \) in \( S^+_{d\times d} \), GLRL should exit with the rank-1 solution \( \bar{W}_1 \) after the first phase, and the following theorem shows that this is also the solution found by GF.

**Assumption 5.7.** \( f(W) \) is locally analytic at each point.

**Theorem 5.8.** Under Assumptions 5.5 and 5.7, if \( \|W_1^G(t)\|_F \) is bounded for all \( t \geq 0 \), then the limit \( \bar{W}_1 := \lim_{t \to +\infty} W_1^G(t) \) exists. Further, if \( \bar{W}_1 \) is a minimizer of \( f(\cdot) \) in \( S^+_{d\times d} \), then for PSD matrices \( \{W_\alpha\} \subseteq S^+_{d\times d} \) converging to \( 0 \) with positive alignment with \( u_1 u_1^\top \) as \( \alpha \to 0 \), it holds that \( \lim_{\alpha \to 0} \lim_{t \to +\infty} \phi(W_\alpha, t) = \bar{W}_1 \).
Assumption 5.7 is a natural assumption, since \( f(\cdot) \) in most cases of matrix factorization is a quadratic or polynomial function (e.g., matrix sensing, matrix completion). In general, it is unlikely for a gradient-based optimization process to get stuck at 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 matrices. This means at least for this subclass of problems, the implicit regularization of GD is rather unrelated to 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 loss \( \mathcal{L} : \mathbb{R}^{d \times d} \to \mathbb{R} \) where every rank-1 stationary point of \( \mathcal{L}(U) \) attains the global minimum but none of them is minimizing the nuclear norm. Below we give a concrete matrix completion problem that meets the above requirement. Let \( M \) be a partially observed matrix to be recovered, where the entries in \( \Omega \) are observed and the others (marked with "?") are unobserved. The optimization problem is defined formally by \( \mathcal{L}(U) = \frac{1}{2} f(UU^\top), f(W) = \frac{1}{2} \sum_{(i,j) \in \Omega} (W_{ij} - M_{ij})^2 \).

\[
M = \begin{bmatrix}
? & ? & 1 & R \\
? & ? & R & ? \\
1 & R & ? & ? \\
R & ? & ? & ?
\end{bmatrix}, \quad M_{\text{norm}} = \begin{bmatrix}
R & 1 & 1 & R \\
1 & R & R & 1 \\
1 & R & R & R \\
R & 1 & R & R
\end{bmatrix}, \quad M_{\text{rank}} = \begin{bmatrix}
1 & R & 1 & R \\
R & R^2 & R & R^2 \\
R & 1 & R & R^2 \\
R & R^2 & R & R^2
\end{bmatrix}.
\]

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

### 5.3 Equivalence Between GD and GLRL: General Case

Theorem 5.6 shows that for any fixed time \( t \), the trajectory of GLRL in the first phase approximates GF with infinitesimal initialization, i.e., \( W^G_1(t) = \lim_{\alpha \to 0} \hat{W}_\alpha(t) \), where \( \hat{W}_\alpha(t) := \phi(W_\alpha, \frac{1}{\alpha} \log((W_\alpha, u_i u_i^\top)^{-1}) + t) \). However, \( W^G_1(\infty) \neq \lim_{\alpha \to 0} \hat{W}_\alpha(\infty) \) does not hold in general, unless the prerequisite in Theorem 5.8 is satisfied, i.e., unless \( \hat{W}_1 = W^G_1(\infty) \) is a minimizer of \( f(\cdot) \) in \( S^d_+ \). This is because of the well-known result that GD converges to local minimizers (Lee et al., 2016; 2017). We adapt Theorem 2 of Lee et al. (2017) to the setting of GF (Theorem I.5) and obtain the following result (Theorem 5.10); see Appendix I.4 for the proof.

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

Therefore, for convex \( f(\cdot) \) such as matrix sensing and completion, suppose \( f(\cdot) \) has no rank-1 PSD minimizer, then no matter how small \( \alpha \) is, \( \hat{W}_\alpha(\infty) \) (if exists) is a minimizer of \( f(\cdot) \) with a higher rank and thus away from the rank-1 matrix \( \hat{W}_1 \). In other words, \( W^G_1(t) \) only describes the limiting trajectory of GF in the first phase, i.e., when GF goes from near 0 to near \( \hat{W}_1 \). After a sufficiently long time (depending on \( \alpha \)), GF escapes the critical point \( \hat{W}_1 \), but this is not described by \( W^G_1(t) \).

To understand how GF escapes \( \hat{W}_1 \), a priori, we need to know how GF approaches \( \hat{W}_1 \). Using a similar argument for Theorem 5.3, Theorem 5.11 shows that generically GF only escapes in the direction of \( v_1 v_1^\top \), where \( v_1 \) is the (unique) top eigenvector of \( -\nabla f(\hat{W}_1) \), and thus the limiting trajectory exactly matches with that of GLRL in the second phase until GF gets close to another critical point \( \hat{W}_2 \in S^d_{+, \leq 2} \). If \( \hat{W}_2 \) is still not a minimizer of \( f(\cdot) \) in \( S^d_+ \) (but it is a local minimizer in \( S^d_{+, \leq 2} \) generically), then GF escapes \( \hat{W}_2 \) and the above process repeats until \( \hat{W}_K \) is a minimizer in \( S^d_{+, \leq K} \) for some \( K \). Here by “generically” we hide some technical assumptions and we elaborate on them in Appendix J. 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 I.2, informal). Let \( \hat{W} \) be a critical point of (2) satisfying that \( \hat{W} \) is a local minimizer of \( f(\cdot) \) in \( S^d_{+, \leq r} \) for some \( r \geq 1 \) but not a minimizer in \( S^d_+ \). Let \( -\nabla f(\hat{W}) = \)}
The trajectory of depth-2 GD, $W_{GLRL}(t)$, converges to the trajectory of GLRL, $W_{GLRL}(t)$, as the initialization scale goes to 0. We plot $\delta(t) = \min_{t' \in T} ||W_{GLRL}(t) - W_{GLRL}(t')||_F$ for different initialization scale $||W(0)||_F$, where $T$ is a discrete subset of $\mathbb{R}$ that covers the entire trajectory of GLRL: $\max_{t' \in T} \min_{t \in T} ||W_{GLRL}(t) - W_{GLRL}(t')||_F \leq \delta$ for $\delta \approx 0.00042$. For each $||W(0)||_F$, we run 20 random seeds and plot them separately. The ground truth $W^* \in \mathbb{R}^{20 \times 20}$ is a randomly generated rank-3 matrix with $||W^*||_F = 20$. 30% entries are observed. See more in Appendix E.1.

\[ \sum_{i=1}^{d} \mu_i v_i v_i^\top \text{ be the eigendecomposition of } -\nabla f(W). \] If $\mu_1 > \mu_2$ and if there exists time $T_\alpha \in \mathbb{R}$ for every $\alpha$ so that $\phi(W_{\alpha}, T_\alpha)$ converges to $W$ with positive alignment with the top principal component $v_1 v_1^\top$ as $\alpha \to 0$, then for every fixed $t$, $\lim_{\alpha \to 0} \phi(W_{\alpha}, T_\alpha + 1 + \frac{1}{2\mu_1}, \log \left( \frac{1}{\phi(W_{\alpha}, T_\alpha, v_i v_i^\top)} + t \right)$ exists and is equal to $W^G(t) := \lim_{t \to 0} \phi(W + \epsilon v_1 v_1^\top, \frac{1}{2\mu_1}, \log \frac{1}{\epsilon} + t)$.

Characterization of the trajectory of GF. Generally, the trajectory of GF with small initialization can be split into $K$ phases by $K + 1$ critical points of (2), $\{W_r\}_{r=0}^K$, where in phase $r$ GF escapes from $W_{r-1}$ in the direction of the top principal component of $-\nabla f(W_{r-1})$ and gets close to $W_r$. Each $W_r$ is a local minimizer of $f(\cdot)$ in $S_{d_r}$, but none of them is a minimizer of $f(\cdot)$ in $S_d$. The smaller the initialization is, the longer GF stays around each $W_r$. Moreover, $\{W_r\}_{r=0}^K$ corresponds to $\{W_{r,r}\}_{r=0}^K$ in Definition 5.1 with infinitesimal $\epsilon > 0$.

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 $W = U_1 U_2 \cdots U_L$ on the implicit bias — how does depth encourage GF to find low rank solutions? We take the standard assumption in existing work for the end-to-end dynamics that the weight matrices have a balanced initialization, i.e., $U_{i+1}(0) U_i(0) = U_{i+1}(0) U_i(1)(0)$, $\forall 1 \leq i \leq L - 1$. Arora et al. (2018) showed that if $\{U_i\}_{i=1}^L$ is balanced at initialization, then we have the following end-to-end dynamics. Similar to the depth-2 case, we use $\phi(W(0), t)$ to denote $W(t)$, where

\[ \frac{dW}{dt} = -\sum_{i=0}^{L-1} (W W^\top) \frac{\partial}{\partial W} f(W) (W^\top W)^{-1} \frac{\partial}{\partial W}. \] (11)

The lemma below is the foundation of our analysis for the deep case, which greatly simplifies (11). Due to the space limit, we defer its derivations and applications into Appendix K.

Lemma 6.1. For $M(t) := W(t)^{2/L}$, we have $\frac{dM}{dt} = -\nabla f(M^{L/2}) M^{L/2} - M^{L/2} \nabla f(M^{L/2})$.

Our main result, Theorem 6.2, gives a characterization of the limiting trajectory for deep matrix factorization with infinitesimal identity initialization. Here $W^G(t) := \lim_{\alpha \to 0} W_{\alpha}^G(t)$ is the trajectory of deep GLRL, where $W_{\alpha}^G(t) := \phi(\alpha e_1 e_1^\top, \frac{\alpha^{-(1-1/P)}}{2\mu_1(P-1)} + t)$ (see Algorithm 2). The dynamics for general initialization is more complicated. Please see discussions in Appendix L.

Theorem 6.2. Let $P = \frac{2}{d}$, $L \geq 3$. Suppose $\|\nabla f(0)\|_2 = \lambda_1(-\nabla f(0)) > \max\{\lambda_2(-\nabla f(0)), 0\}$,\footnote{We believe assumption $\|\nabla f(0)\|_2 = \lambda_1(-\nabla f(0))$ could be removed with a more refined analysis.}

for every fixed $t \in \mathbb{R}$,

\[ \left\| \phi(\alpha I, \frac{\alpha^{-(1-1/P)}}{2\mu_1(P-1)} + t) - W(t) \right\|_F = O(\alpha^{\frac{1}{2} + \frac{1}{2P}}), \] (12)

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

\[ \lambda_k \left( \phi(\alpha I, \frac{\alpha^{-(1-1/P)}}{2\mu_1(P-1)} + t) \right) = O(\alpha). \] (13)

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

Definition 6.3 ($r$-low-rankness). For matrix $M \in \mathbb{R}^{d \times d}$, we define the $r$-low-rankness of $M$ as $\sqrt{\sum_{i=r+1}^{d} \sigma_i^2(M)}$, where $\sigma_i(M)$ is the $i$-th largest singular value of $M$. 
Figure 2: GD passes by the same set of critical points as GLRL when the initialization scale is small, and gets much closer to the critical points when \( L \geq 3 \). Depth-2 GD requires a much smaller initialization scale to maintain small low-rankness. Here the ground truth matrix \( W^* \in \mathbb{R}^{20 \times 20} \) is of rank 3 as stated in Appendix E.1. In this case, GLRL has 3 phases and 4 critical points \( \{W_i\}_{i=0}^3 \), where \( W_0 = O \) and \( W_3 = W^* \). For each depth \( L \) and initialization scale \( \|W(0)\|_F \), we plot the distance between the current step of GD and the closest critical point of GLRL, \( \|W_{\text{GD}}(t) - W_i\|_F \), the norm of full gradient, \( \|\nabla_{U_t} L(U_t, I)\|_F \) and the \((r + 1)\)-low-rankness of \( W_{\text{GD}}(t) \) with \( r := \arg \min_{0 \leq i \leq 3} \|W_{\text{GD}}(t) - W_i\|_F \).

Suppose \( f(\cdot) \) admits a unique minimizer \( W_0 \) in \( S^+_{d,1} \), and we run GF from \( \alpha I \) for both depth-2 and depth-\( L \) cases. Intuitively, the 1-low-rankness of the depth-2 solution is \( \Omega(\alpha^{1-\mu}\mu) \), 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 \( O(\alpha) \), as shown in Theorem 6.4. The key idea is to show that there is a basin in the manifold of rank-1 matrices around \( W_0 \) such that any GF starting within the basin converges to \( W_0 \). Based on this, we can prove that starting from any matrix \( O(\alpha) \)-close to the basin, GF converges to a solution \( O(\alpha) \)-close to \( W_0 \). See Appendix M for more details.

**Theorem 6.4.** In the same settings as Theorem 6.2, if \( W(\infty) \) exists and is a minimizer of \( f(\cdot) \) in \( S^+_{d,\leq 1} \), under regularity assumption M.1, we have \( \inf_{t \in [0, 1]} \|\phi(\alpha I, t) - W(\infty)\|_F = O(\alpha) \).

**Interpretation for the advantage of depth with multiple phases.** For depth-2 GLRL, the low-rankness is raised to some power less than 1 per phase (depending on the eigengap). For deep GLRL, we show the low-rankness is only multiplied by some constant for the first phase 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 5) suggest that while being deep is good for generalization, being much deeper may not be much better: once \( L \geq 3 \), increasing the depth does not improve the order of low-rankness significantly. While this theoretical result is only for identity initialization, Theorem F.1 and Corollary F.2 further show that the dynamics of GF (11) with any initialization pointwise converges as \( L \to \infty \), under a suitable time rescaling. See Figure 6 for experimental verification.

7 Conclusion and Future Directions

In this work, we connect gradient descent to Greedy Low-Rank Learning (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.

Acknowledgments

The authors thank Sanjeev Arora and Jason D. Lee for helpful discussions. The authors also thank Runzhe Wang for useful suggestions on writing. ZL and YL acknowledge support from NSF, ONR, Simons Foundation, Schmidt Foundation, Mozilla Research, Amazon Research, DARPA and SRC. ZL is also supported by Microsoft PhD Fellowship.
REFERENCES


A 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 $u_r u_r \top$ as defined in Algorithm 1 into its basis and solve $\min_{u_r} f(\sum_{i=1}^r \alpha_i u_i u_i \top)$ 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 $\alpha_i$, while the entire $U_r$ 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 $\eta, \epsilon$.

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 $-\nabla f(W_r)$, while it solves $\min_{u_r} f(\sum_{i=1}^r \beta_i u_i u_i \top)$ for the rank-$r$ estimation. Khanna et al. (2017) provided convergence guarantee for GECO assuming strong convexity. Haefele and Vidal (2019) proposed a local-descent meta algorithm, of which GLRL can be viewed as a specific realization.

B Deep GLRL Algorithm

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

| parameter: step size $\eta > 0$; small $\epsilon > 0$ |
| $\epsilon' \leftarrow \epsilon^{1/L}, \mathcal{L}(U_1, \ldots, U_L) := f(W_1 \cdots W_L)$. |
| $W_0 \leftarrow 0 \in \mathbb{R}^{d \times d}$ and $U_0, U_{0,1}(\infty), \ldots, U_{0,L}(\infty) \in \mathbb{R}^{d \times 0}$ are empty matrices |
| while $\lambda_1(-\nabla f(W_r)) > 0$ do |
| $r \leftarrow r + 1$ |
| let $u_r$ be a top (unit) eigenvector of $-\nabla f(W_{r-1})$ |
| $U_{r-1,1}(\infty) \leftarrow U_{r-1,1}(\infty) \epsilon' u_r \top \in \mathbb{R}^{d \times r}$ |
| $U_{r,k}(0) \leftarrow \begin{bmatrix} U_{r-1,k}(\infty) & 0 \\ 0 & \epsilon' \end{bmatrix} \in \mathbb{R}^{r \times r}$ for all $2 \leq k \leq L - 1$ |
| $U_{r,L}(0) \leftarrow \begin{bmatrix} U_{r-1,L}(\infty) \\ \epsilon' u_r \top \end{bmatrix} \in \mathbb{R}^{r \times d}$ |
| for $t = 0, 1, \ldots$ do |
| $U_{r,i}(t + 1) \leftarrow U_{r,i}(t) - \eta \nabla U_i \mathcal{L}(U_{r,1}(t), \ldots, U_{r,L}(t)), \forall 1 \leq i \leq L.$ |
| $W_r \leftarrow U_{r,1}(\infty) \cdots U_{r,L}(\infty)$ |
| return $W_r$ |

C Preliminary Lemmas

Lemma C.1. For $U_0 \in \mathbb{R}^{d \times r}$ and $W_0 := U_0 U_0 \top$, the following statements are equivalent:

1. $U_0$ is a stationary point of $\mathcal{L}(U) = \frac{1}{2} f(UU \top)$;
2. $\nabla f(W_0) W_0 = 0$;
3. $W_0 := U_0 U_0 \top$ is a critical point of (2).

Proof. (2) $\Rightarrow$ (3) is trivial. We only prove (1) $\Rightarrow$ (2), (3) $\Rightarrow$ (1).

Proof for (1) $\Rightarrow$ (2). If $U_0$ is a stationary point, then $0 = \nabla \mathcal{L}(U_0) = \nabla f(W_0) U_0$. So $\nabla f(W_0) W_0 = (\nabla f(W_0) U_0) U_0 \top = 0$.

Proof for (3) $\Rightarrow$ (1). If $W_0$ is a critical point, then $0 = \langle g(W_0), \nabla f(W_0) \rangle = -2 \text{Tr}(\nabla f(W_0) W_0 \nabla f(W_0)) = -2\|\nabla f(W_0) U_0\|_F^2$, which implies $\nabla \mathcal{L}(U_0) = 0$. □
Lemma C.2. For a stationary point $U_0 \in \mathbb{R}^{d \times r}$ of $\mathcal{L}(U) = \frac{1}{2} f(UU^\top)$ where $f(\cdot)$ is convex, $W_0 := U_0 U_0^\top$ attains the global minimum of $f(\cdot)$ in $S_+^d := \{ W : W \succeq 0 \}$ iff $\nabla f(W_0) \succeq 0$.

Proof. Since $f(W)$ is a convex function and $S_+^d$ is convex, we know that $W_0$ is a global minimizer of $f(W)$ in $S_+^d$ iff

$$\langle \nabla f(W_0), W - W_0 \rangle \geq 0, \quad \forall W \succeq 0. \quad (14)$$

Note that $\langle \nabla f(W_0), W_0 \rangle = \text{Tr}(\nabla f(W_0) W_0)$. By Lemma C.1, $\langle \nabla f(W_0), W_0 \rangle = 0$. Combining this with (14), we know that $W_0$ is a global minimizer iff

$$\langle \nabla f(W_0), W \rangle \succeq 0, \quad \forall W \succeq 0. \quad (15)$$

It is easy to check that this condition is equivalent to $\nabla f(W_0) \succeq 0$.

D PROOFS FOR COUNTER-EXAMPLE

Conjecture D.1 (Formal Statement, Gunasekar et al. 2017). Suppose $f : \mathbb{R}^{d \times d} \to \mathbb{R}$ is a quadratic function and $\min_{W \succeq 0} f(W) = 0$. Then for any $W_{\text{init}} \succeq 0$ if $W_1 = \lim_{\alpha \to 0} \lim_{t \to +\infty} \phi(\alpha W_{\text{init}}, t)$ exists and $f(W_1) = 0$, then $\|W_1\|_* = \min_{W \succeq 0} \|W\|_*$ s.t. $f(W) = 0$.

Proposition D.2 (Formal Statement for Example 5.9). For constant $R > 1$, let

$$M = \begin{bmatrix} \ast & \ast & 1 & R \\ \ast & \ast & R & ? \\ 1 & R & ? & ? \\ R & ? & ? & ? \end{bmatrix}, M_{\text{norm}} = \begin{bmatrix} R & 1 & 1 & R \\ 1 & R & R & 1 \\ 1 & R & R & 1 \\ R & 1 & 1 & R \end{bmatrix}, \text{ and } \begin{bmatrix} R & 1 & 1 & R \\ R & R^2 & R & R^2 \\ R & R^2 & R & R^2 \\ R & R^2 & R & R^2 \end{bmatrix}$$

and

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

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

Then for any $W_{\text{init}} \succeq 0$, s.t. $u_1^\top W_{\text{init}} u_1 > 0$,

$$\lim_{\alpha \to 0} \lim_{t \to +\infty} \phi(\alpha W_{\text{init}}, t) = M_{\text{rank}}.$$ 

Moreover, we have

$$\|M_{\text{rank}}\|_* = 2R^2 + 2 > 4R = \|M_{\text{norm}}\|_* = \min_{W \succeq 0, f(W) = 0} \|W\|_*.$$ 

Proof. We define $W_{1,\epsilon}^G(t), W_1^G(t)$ in the same way as in Definition 5.1, Theorem 5.6.

$$W_{1,\epsilon}^G(t) := \phi(\epsilon u_1 u_1^\top, t), \quad W_1^G(t) := \lim_{\epsilon \to 0} W_{1,\epsilon}^G(\frac{1}{2\epsilon^3} \log \frac{1}{\epsilon} + t).$$

Below we will show

1. Assumption 5.7 and Assumption 5.5 are satisfied.
2. $\|W_1^G(t)\|_p$ bounded for $t \geq 0$;
3. $\lim_{t \to +\infty} W_1^G(t) = M_{\text{rank}}$;
4. $M_{\text{norm}} = \arg \min_{W \succeq 0, f(W) = 0} \|W\|_*$. 

14
Thus Since $M_{\text{rank}}$ is a global minimizer of $f(\cdot)$, applying Theorem 5.8 finishes the proof.

**Proof for Item 1.** Let $M_0 := \nabla f(0)$, then

$$M_0 = \begin{bmatrix} 0 & 0 & 1 & R \\ 0 & 0 & R & 0 \\ 1 & R & 0 & 0 \\ R & 0 & 0 & 0 \end{bmatrix}.$$ 

Let $A := \begin{bmatrix} 1 & 0 \\ R & 0 \end{bmatrix}$, then we have $\lambda_1(A) = \frac{1+\sqrt{1+R^2}}{2}, \lambda_2(A) = \frac{1-\sqrt{1+R^2}}{2}$, thus $\lambda_1(A) > |\lambda_2(A)| > 0$. Let $v_1 \in \mathbb{R}^2$ be the top eigenvector of $A$. We claim that $u_1 = [v_1] \in \mathbb{R}^d$ is the top eigenvector of $\nabla f(0)$. First by definition it is easy to check that $M_0u_1 = \lambda_1(A)u_1$. Further noticing that $M_0 = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$, we know $\lambda_2(M_0) \in \{\lambda_1^2(A), \lambda_2^2(A)\}$ for all eigenvalues $\lambda_i(M_0)$. That is, $\lambda_1(M_0) = \lambda_1(A), \lambda_2(M_0) = -\lambda_2(A), \lambda_3(M_0) = \lambda_2(A),$ and $\lambda_4(M_0) = -\lambda_1(A).$ Thus Asumption 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 $(x_r(t), y_r(t)) \in \mathbb{R}^2$ be the gradient flow of $g(x, y) = \frac{1}{2}(x^2 - 1)^2 + (xy - R)^2$ starting from $(x_r(0), y_r(0)) = \sqrt{\epsilon}v_1$.

\[
\begin{align*}
    \frac{dx_r(t)}{dt} &= (1 - t^2)x_r(t) - 2y_r(t)(x_r(t)y_r(t) - R) \\
    \frac{dy_r(t)}{dt} &= -2x_r(t)(x_r(t)y_r(t) - R)
\end{align*}
\]  

(16)

Let $W_r(t)$ be the following matrix:

$$W_r(t) := \begin{bmatrix} x_r(t) \\ y_r(t) \\ x_r(t) \\ y_r(t) \end{bmatrix} [x_r(t) \ y_r(t) \ x_r(t) \ y_r(t)].$$

Then it is easy to verify that $W_r(0) = W^G_r(0)$ and $W_r(t)$ satisfies (2). Thus by the existence and uniqueness theorem, we have $W_r(t) = W^G_r(t)$ for all $t$. Taking the limit $\epsilon \to 0$, we know that $W^G_1(t)$ can also be written in the following form:

$$W^G_1(t) = \begin{bmatrix} x(t) \\ y(t) \\ x(t) \\ y(t) \end{bmatrix} [x(t) \ y(t) \ x(t) \ y(t)],$$

and $(x_r(t), y_r(t)) \in \mathbb{R}^2$ 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 over time, and $\lim_{t \to -\infty} g(x(-t), y(-t)) = g(x(\infty), y(\infty)) = g(0, 0) = R^2 + 0.5$, we know $|x(t)y(t)| \leq 3R$ for all $t$. So whenever $g_2(t) - x^2(t) \geq 9R^2$, we have $x^2(t) \leq \frac{9R^2}{y^2(t)} \leq \frac{9R^2}{y^2(t)-x^2(t)} \leq 1$. In this case, $\frac{d(x^2(t)-y^2(t))}{dt} = 2x^2(t)(x^2(t) - 1) \leq 0$. Combining this with $y(-\infty)^2 - x(-\infty)^2 = 0 \leq 9R^2$, we have $g(t) - x^2(t) \leq 9R^2$ for all $t$, which also implies that $y(t)$ is bounded. Noticing that $9R^2 \geq g(x(t), y(t)) \geq (x^2(t) - 1)^2$, we know $x^2(t)$ is also bounded. Therefore, $W^G_1(t)$ is bounded.

**Proof for Item 3.** Note that $(x(\infty), y(\infty))$ is a stationary point of $g(x, y)$. It is clear that $g(x, y)$ only has 3 stationary points $(-0, 0)$, $(-1, R)$ and $(0, -R)$. Thus $W_1$ can only be 0 or $M_{\text{rank}}$. However, since for all $t, f(W^G(t)) < f(0), W_1 = \lim_{t \to -\infty} W^G_1(t)$ cannot be 0. So $W_1$ must be $M_{\text{rank}}$.

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

$$\begin{align*}
    (e_1 - e_4)^T M (e_1 - e_4) \geq 0 &\implies m_{11} + m_{44} \geq m_{14} + m_{41} = 2R \\
    (e_2 - e_3)^T M (e_2 - e_3) \geq 0 &\implies m_{22} + m_{33} \geq m_{23} + m_{32} = 2R
\end{align*}$$

Published as a conference paper at ICLR 2021
### Table 1: Choice of hyperparameters for simulating gradient flow.

For $L = 2$, gradient descent escapes saddles in $O(\log \frac{1}{\epsilon})$ time, where $\epsilon$ is the distance between the initialization and the saddle.

<table>
<thead>
<tr>
<th>Depth $(L)$</th>
<th>Simulation method</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Adaptive LR, $\eta = 10^{-3}$ for $10^5$ iterations</td>
</tr>
<tr>
<td>3</td>
<td>Adaptive LR, $\eta = 2 \times 10^{-3}$ and $\epsilon = 10^{-4}$ for $10^5$ iterations</td>
</tr>
<tr>
<td>4</td>
<td>Adaptive LR, $\eta = 3 \times 10^{-4}$ and $\epsilon = 10^{-3}$ for $10^5$ iterations</td>
</tr>
</tbody>
</table>

Thus $4R = \min_{W \succeq 0, f(W) = 0} \|W\|_F$, 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 $M \succeq 0$.

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

$$M \succeq 0 \implies VMV^T \succeq 0 \implies \begin{bmatrix} 0 & m_{13} - m_{23} & m_{14} - m_{24} \\ m_{31} - m_{41} & R & R \\ m_{41} - m_{42} & R & R \end{bmatrix} \succeq 0,$$

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

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

### E Experiments

#### E.1 General Setup

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

The ground-truth matrix $W_*$ is low-rank by construction: we sample a random orthogonal matrix $U$, a diagonal matrix $S$ with Frobenius norm $\|S\|_F = 1$ and set $W_* = USU^T$. Each measurement $X$ in $X_1, \ldots, X_m$ is generated by sampling two one-hot vectors $u$ and $v$ uniformly and setting $X = \frac{1}{2}uv^T + \frac{1}{2}vu^T$.

In Figures 1, 2, 3 to 5 and 7, the ground truth matrix $W_*$ has shape $20 \times 20$ and rank 3, where $\|W_*\|_F = 20$, $\lambda_1(W_*) = 17.41$, $\lambda_2(W_*) = 8.85$, $\lambda_3(W_*) = 4.31$ and $\lambda_4(-\nabla f(0)) = 6.23$, $\lambda_5(-\nabla f(0)) = 5.41$. $p = 0.3$ is used for generating measurements, except $p = 0.25$ in Figure 3, i.e., each pair of entries of $W_{ij}$ and $W_{ji}$ is observed with probability $p$.

**Gradient Descent.** Let $\tilde{\epsilon} > 0$ be the Frobenius norm of the target random initialization. For the depth-2 case, we sample 2 orthogonal matrices $V_1, V_2$ and a diagonal matrix $D$ with Frobenius norm $\tilde{\epsilon}$, and we set $U = V_1D^{1/2}V_2^T$; for the depth-$L$ case with $L \geq 3$, we sample $L$ orthogonal matrices $V_1, \ldots, V_L$ and a diagonal matrix $D$ with Frobenius norm $\tilde{\epsilon}$, and we set $U_i := V_iD^{1/2}V_{i+1}^T \quad (V_{L+1} = V_L)$. In this way, we can guarantee that the end-to-end matrix $W = U_1 \cdots U_L$ is symmetric and the initialization is balanced for $L \geq 3$.

We discretize the time to simulate gradient flow. When $L > 2$, gradient flow stays around saddle points for most of the time, therefore we use full-batch GD with adaptive learning rate $\tilde{\eta}_t$, inspired by
RMSprop (Tieleman and Hinton, 2012), for faster convergence:

\[ v_{t+1} = \alpha v_t + (1 - \alpha) \left\| \nabla \mathcal{L}(\theta_t) \right\|_2^2, \]
\[ \tilde{\eta}_t = \frac{\eta}{\sqrt{\alpha} + (1 - \alpha \epsilon_t) + \epsilon}, \]
\[ \theta_{t+1} = \theta_t - \tilde{\eta}_t \nabla \mathcal{L}(\theta_t), \]

where \( \alpha = 0.99 \), \( \eta \) is the (unadjusted) learning rate. The choices of hyperparameters are summarized in Table 1. The continuous time for \( \theta_t \) is measured as \( \sum_{t=0}^{\epsilon_t} \tilde{\eta}_t \).

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

E.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. In Figure 2, we first run GLRL and obtain the critical points \( \{W_r\}_{r=0}^3 \) passed by GLRL. We also define the distance of a matrix \( W \) to the critical points to be \( \min_{0 \leq r \leq 3} \|W - W_r\|_F \).

E.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 \( f(W) := \frac{1}{d} \|W - W^*\|_F^2 \) and \( f(0) = 1 \). We run 10 random seeds for GD and plot them separately (most of them overlap).](image)

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

We use the general setting in Appendix E.1. In these experiments, we use the constant learning rate \( 10^{-3} \) for \( 4 \times 10^7 \) iterations. The reference matrix \( W_{ref} \) is obtained by running the first stage of GLRL with \( \|W(0)\|_F = 10^{-48} \) and we pick one matrix in the trajectory with \( \|W_{ref}\|_F \) about 0.6.

For every \( \epsilon = 10^\iota, \iota \in \{-1, -2, -3, -4, -5\} \), we run both gradient descent and the first phase of GLRL with \( \|W(0)\|_F = \epsilon \). For gradient descent, we use random initialization so \( \|W(0)\|_F \) is full rank w.p. 1. The distance of a trajectory to \( W_{ref} \) is defined as \( \min_{t \geq 0} \|W(t) - W_{ref}\|_F \). 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 $W_{ref}$.

The result is shown at Figure 4. We observe that GLRL trajectories are closer to the reference matrix $W_{ref}$ 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 4: Using $\epsilon v_1v_1^\top$ (denoted by “rank 1”) as initialization makes GD much closer to GLRL compared to using random initialization (denoted by “rank $d$”), where $v_1$ is the top eigenvector of $-\nabla f(0)$. 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.

Figure 5: 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 solid curves are the mean and the shaded area indicates one standard deviation. We observe that performance of GD is quite robust to its initialization. Note that for $L > 2$, the shaded area with initialization scale $10^{-7}$ is large, as the sudden decrement of loss occurs at quite different continuous times for different random seeds in this case.

E.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 5. We can see that as the initialization becomes smaller, the final solution gets closer to the ground truth. However, a depth-2 model requires exponentially small initialization, while deeper models require polynomial small initialization, though it takes much longer to converge.

F The Marginal Value of Being Deeper

Theorem F.1 shows that the end-to-end dynamics (17) converges point-wise while $L \to \infty$ if the product of learning rate and depth, $\eta L$, is fixed as constant. Interestingly, (17) also allows us to simulate the dynamics of $W(t)$ for all depths $L$ while the computation time is independent of $L$. In Figure 6, we compare the effect of depth while fixing the initialization and $\eta L$. 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.
Towards infinite depth

Figure 6: 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).

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

$$
\frac{dW}{dt} = -L\tilde{U} \left( (\tilde{U}^\top \nabla f(W) \tilde{V}) \circ K^{(L)} \right) \tilde{V}^\top,
$$

(17)

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

**Proof.** We start from (11):

$$
\frac{dW}{dt} = -\sum_{l=0}^{L-1} (WW^\top)^{L-1-l} \nabla f(W)(W^\top W)^{L-1-l},
$$

$$
= -\sum_{l=0}^{L-1} \tilde{U} \Sigma \Sigma \tilde{U}^\top \nabla f(W) \tilde{V} \Sigma^{2(L-1-l)} \tilde{V}^\top
$$

$$
= -L\tilde{U} \left[ \sum_{l=0}^{L-1} \Sigma \Sigma \tilde{U}^\top \nabla f(W) \tilde{V} \Sigma^{2(L-1-l)} \right] \tilde{V}.
$$

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

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

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

$$
\sum_{l=0}^{L-1} \Sigma \tilde{U}^\top \nabla f(W) \tilde{V} \Sigma^{2(L-1-l)} = \sum_{l=0}^{L-1} (\tilde{U}^\top \nabla f(W) \tilde{V}) \circ H^{(l)}
$$

$$
= (\tilde{U}^\top \nabla f(W) \tilde{V}) \circ K^{(L)},
$$

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

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

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

$$
K_{i,i}^{(L)} = L^{-1} \sum_{l=0}^{L-1} \Sigma \tilde{U}^\top \nabla f(W) \tilde{V} \Sigma^{2(L-1-l)} = \left\{ \begin{array}{ll} \sigma_i^{2-2/L}, & i = j, \\ \frac{\sigma_i^2 - \sigma_j^2}{L \sigma_i^2 + L \sigma_j^2}, & i \neq j. \end{array} \right.
$$

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


Experiment details. We follow the general setting in Appendix E.1. The ground truth $W^*$ is different but is generated in the same manner and has the same shape of $20 \times 20$ and $p = 0.3$ is used for observation generation. We directly apply (17), in which we compute $\tilde{V}$ and $\tilde{U}$ through SVD, to simulate the trajectory together with a constant learning rate of $10^{-3}$ for depth $L$. $W(0)$ is sampled from $10^{-3} \times \mathcal{N}(0, I_d)$.

G Proofs for Dynamical System

In this section, we prove Theorem 5.3 in Section 5.1. In Appendix G.1, we show how to reduce Theorem 5.3 to the case where $J(0)$ is exactly a diagonal matrix, then we prove this diagonal case in Appendix G.2. Finally, in Appendix G.3, we discuss how to extend it to the case where $J(0)$ is non-diagonalizable.

G.1 Reduction to the Diagonal Case

Theorem G.1. If $J(0) = \text{diag}(\hat{\mu}_1, \ldots, \hat{\mu}_d)$ 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 G.1. Let $\frac{d\theta}{dt} = g(\theta)$ be the dynamical system in Theorem 5.3. Let $J(0) = \tilde{V} \hat{D} \tilde{V}^{-1}$ be the eigendecomposition, where $\tilde{V}$ is an invertible matrix and $\hat{D} = \text{diag}(\hat{\mu}_1, \ldots, \hat{\mu}_d)$. Now we define the following new dynamics by changing the basis:

$$\tilde{\theta}(t) = \tilde{V}^{-1} \theta(t).$$

Then $\frac{d\tilde{\theta}(t)}{dt} = \tilde{g}(\tilde{\theta})$ for $\tilde{g}(\tilde{\theta}) := \tilde{V}^{-1} g(\tilde{V} \tilde{\theta})$, and the associated Jacobian matrix is $\tilde{J}(\tilde{\theta}) := \tilde{V}^{-1} J(\hat{V} \tilde{\theta}) \tilde{V}$, and thus $\tilde{J}(0) = \text{diag}(\hat{\mu}_1, \ldots, \hat{\mu}_d)$.

Now we apply Theorem G.1 to $\tilde{\theta}(t)$. Then $\tilde{z}_\alpha(t) := \tilde{V}^{-1} z_\alpha(t)$ converges to the limit $\tilde{z}(t) := \lim_{\alpha \to 0} \tilde{z}_\alpha(t)$. This shows that the limit $\tilde{z}(t) = \tilde{V} \tilde{z}(t)$ exists in Theorem 5.3. We can also verify that $\tilde{z}(t)$ is a solution of (6).

Given $\delta_\alpha$ converging to 0 with positive alignment with $\hat{u}_1$ as $\alpha \to 0$, we can define $\delta_\alpha := \tilde{V}^{-1} \delta_\alpha$, then $\delta_\alpha$ converges to 0 with positive alignment with $\hat{e}_1$, where $\hat{e}_1$ is the first vector in the standard basis and is also the top eigenvector of $\hat{J}(0)$. Therefore, for every $t \in (-\infty, +\infty)$, there is a constant $C > 0$ such that

$$\bigg\| \tilde{V}^{-1} \phi \left( \delta_\alpha, t + \frac{1}{\hat{\mu}_1} \log \frac{1}{\delta_\alpha, \hat{u}_1} \right) - \tilde{z}(t) \bigg\|_2 \leq C \cdot \|\delta_\alpha\|_2^{\frac{1}{1 + \gamma}},$$

for every sufficiently small $\alpha$. As $\tilde{V}$ are invertible, this directly implies (7).}

G.2 Proof for the Diagonal Case

Now we only need to prove Theorem G.1. Let $e_1, \ldots, e_d$ be the standard basis. Then $\hat{u}_1 = \hat{v}_1 = e_1$ in this diagonal case. We only use $e_1$ to stand for $\hat{u}_1$ and $\hat{v}_1$ in the rest of our analysis.

Let $R > 0$. Since $g(\theta)$ is $C^2$-smooth, there exists $\beta > 0$ such that

$$\|J(\theta) - J(\theta + h)\|_2 \leq \beta \|h\|_2$$

for all $\|\theta\|_2, \|\theta + h\|_2 \leq R$. Then the following can be proved by integration:

$$g(\theta + h) - g(\theta) = \left( \int_0^1 J(\theta + \xi h) d\xi \right) h,$$

$$\|g(\theta + h) - g(\theta) - J(\theta) h\|_2 \leq \beta \|h\|_2^2.$$  

By (21), we also have

$$\|g(\theta) - J(0) \theta\|_2 = \|g(\theta) - g(0) - J(0) \theta\|_2 \leq \beta \|\theta\|_2^2.$$
Let $\kappa := \frac{\beta}{\hat{\mu}_1}$. We assume WLOG that $R \leq 1/\kappa$. Let $F(x) = \log x - \log(1 + \kappa x)$. It is easy to see 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 $T_\alpha(r) := \frac{1}{\hat{\mu}_1} (F(r) - F(\alpha)) = \frac{1}{\hat{\mu}_1} \left( \log \frac{r}{\alpha} - \log \frac{1+\kappa r}{1+\kappa \alpha} \right)$.

Our proof only relies on the following properties of $J(0)$ (besides that $\hat{\mu}_1, e_1$ are the top eigenvalue and eigenvector of $J(0)$):

**Lemma G.2.** For $J(0) := \text{diag}(\hat{\mu}_1, \ldots, \hat{\mu}_d)$, we have

1. For any $h \in \mathbb{R}^d$, $h^\top J(0) h \leq \hat{\mu}_1 \|h\|_2^2$.
2. For any $t \geq 0$, $\|e^{tJ(0)} - e^{\hat{\mu}_1 t} t^2 \|_2 = e^{\hat{\mu}_1 t}$.

**Proof.** For Item 1, $h^\top J(0) h = \sum_{i=1}^d \hat{\mu}_i h_i^2 \leq \hat{\mu}_1 \|h\|_2^2$. For Item 2, $\|e^{tJ(0)} - e^{\hat{\mu}_1 t} e_1 e_1^\top \|_2 = \|\text{diag}(0, e^{\hat{\mu}_1 t}, \ldots, e^{\hat{\mu}_1 t})\| = e^{\hat{\mu}_1 t}$.

**Lemma G.3.** For $\theta(t) = \phi(\theta_0, t)$ with $\|\theta_0\|_2 \leq \alpha$ and $t \leq T_\alpha(r)$,

$$\|\theta(t)\|_2 \leq \frac{1 + \kappa \alpha}{1 + \kappa_\alpha} t \leq \frac{1 + \kappa r}{1 + \kappa \alpha} \cdot e^{\hat{\mu}_1 t} \leq r.$$ 

**Proof.** By (22) and Lemma G.2, we have

$$\frac{d}{dt} \|\theta(t)\|_2^2 = 2 \langle \theta(t), J(0)^2 \theta(t) \rangle + \beta \|\theta(t)\|_2^4 \leq 2 \hat{\mu}_1 \|\theta(t)\|_2^3 + \beta \|\theta(t)\|_2^3.$$ 

This implies $\frac{d}{dt} \|\theta(t)\|_2^2 \leq \hat{\mu}_1 \|\theta(t)\|_2^2 + \kappa \|\theta(t)\|_2^3$. Since $F'(x) = \frac{1}{x + \kappa x^2}$, we further have

$$\frac{d}{dt} F(\|\theta(t)\|_2) \leq \hat{\mu}_1.$$ 

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

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

Exponenting $F(\alpha)$ proves the lemma.

**Lemma G.4.** For $\theta(t) = \phi(\theta_0, t)$ with $\|\theta_0\|_2 \leq \alpha$ and $t \leq T_\alpha(r)$, we have

$$\theta(t) = e^{tJ(0)} \theta_0 + O(r^2).$$

**Proof.** Let $\hat{\theta}(t) = e^{tJ(0)} \theta_0$. Then we have

$$\frac{d}{dt} \|\theta(t) - \hat{\theta}(t)\|_2^2 \leq \|g(\theta(t)) - J(0) \hat{\theta}(t), \theta(t) - \hat{\theta}(t)\|_2$$

$$= \langle g(\theta(t)) - J(0) \hat{\theta}(t), \theta(t) - \hat{\theta}(t) \rangle + \|\theta(t) - \hat{\theta}(t)\|_2 \cdot \|\theta(t) - \hat{\theta}(t)\|_2 + \hat{\mu}_1 \|\theta(t) - \hat{\theta}(t)\|_2^2,$$

where the last inequality is due to Lemma G.2. By (22) and Lemma G.3, we have

$$\|g(\theta(t)) - J(0) \theta(t)\|_2 \leq \beta \|\theta(t)\|_2 \leq \beta \left( \frac{1 + \kappa r}{1 + \kappa \alpha} \right)^2 \cdot e^{2\hat{\mu}_1 t}.$$ 

So we have $\frac{d}{dt} \|\theta(t) - \hat{\theta}(t)\|_2 \leq \beta \left( \frac{1 + \kappa r}{1 + \kappa \alpha} \right)^2 \cdot e^{2\hat{\mu}_1 t} + \hat{\mu}_1 \|\theta(t) - \hat{\theta}(t)\|_2$. By Grönwall’s inequality,

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

Evaluating the integral gives

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

which proves the lemma.
Lemma G.5. Let $\theta(t) = \phi(\theta_0, t), \tilde{\theta}(t) = \phi(\tilde{\theta}_0, t)$. If $\max\{\|\theta_0\|_2, \|\tilde{\theta}_0\|_2\} \leq \alpha$, then for $t \leq T_\alpha(r)$,
\[ \|\theta(t) - \tilde{\theta}(t)\|_2 \leq e^{\tilde{\mu}_1 t \cdot (t + \kappa \tau)} \|\theta_0 - \tilde{\theta}_0\|_2. \]

Proof. For $t \leq T_\alpha(r)$, by (20),
\[
\frac{1}{2} \frac{d}{dt} \|\theta(t) - \tilde{\theta}(t)\|_2^2 = \langle g(\theta(t)) - g(\tilde{\theta}(t)), \theta(t) - \tilde{\theta}(t) \rangle \\
= (\theta(t) - \tilde{\theta}(t))^\top \left( \int_0^1 \mathbf{J}((\theta(t) + (1 - \xi)\tilde{\theta}(t)) \, d\xi \right) (\theta(t) - \tilde{\theta}(t)),
\]
where $\theta_\xi(t) := \xi \theta(t) + (1 - \xi) \tilde{\theta}(t)$. By Lemma G.3, $\max\{\|\theta(0)\|_2, \|\tilde{\theta}(0)\|_2\} \leq \frac{1 + \kappa \tau}{1 + k \alpha} \cdot e^{\tilde{\mu}_1 t}$ for all $t \leq T_\alpha(r)$. So $\|\theta(t)\|_2 \leq \frac{1 + \kappa \tau}{1 + k \alpha} \cdot e^{\tilde{\mu}_1 t}$. Combining these with (19) and Lemma G.2, we have
\[
h^\top \mathbf{J}(\theta(t)) h = h^\top \mathbf{J}(0) h + h^\top (\mathbf{J}(\theta(t)) - \mathbf{J}(0)) h \leq \left( \tilde{\mu}_1 + \beta \cdot \frac{1 + \kappa \tau}{1 + k \alpha} \cdot e^{\tilde{\mu}_1 t} \right) \|h\|_2^2,
\]
for all $h \in \mathbb{R}^d$. Thus, $\frac{d}{dt} \|\theta(t) - \tilde{\theta}(t)\|_2 \leq \left( \tilde{\mu}_1 + \beta \cdot \frac{1 + \kappa \tau}{1 + k \alpha} \cdot e^{\tilde{\mu}_1 t} \right) \|\theta(t) - \tilde{\theta}(t)\|_2$. This implies
\[
\frac{1}{2} \frac{d}{dt} \|\theta(t) - \tilde{\theta}(t)\|_2 \leq \int_0^t \left( \tilde{\mu}_1 + \beta \cdot \frac{1 + \kappa \tau}{1 + k \alpha} \cdot e^{\tilde{\mu}_1 \tau} \right) \, d\tau \leq \tilde{\mu}_1 t + \beta \cdot \frac{1 + \kappa \tau}{1 + k \alpha} \cdot e^{\tilde{\mu}_1 t} \leq \tilde{\mu}_1 t + \kappa \tau.
\]
Therefore, $\|\theta(t) - \tilde{\theta}(t)\|_2 \leq e^{\tilde{\mu}_1 t + \kappa \tau} \|\theta(0) - \tilde{\theta}(0)\|_2$. \qed

Lemma G.6. For every $t \in (-\infty, +\infty)$, $z(t)$ exists and $z_\alpha(t)$ converges to $z(t)$ in the following rate:
\[
\|z_\alpha(t) - z(t)\|_2 = O(\alpha),
\]
where $O$ hides constants depending on $g(\theta)$ 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 $\frac{\tilde{\alpha}}{1 + \kappa \alpha} = \alpha$ (i.e., $F(\tilde{\alpha}) = \log \alpha$). Let $\alpha'$ be an arbitrary number less than $\alpha$. Let $t_0 := \frac{1}{\mu_1} \log \frac{\alpha}{\alpha'}$. Then $t_0 = \frac{1}{\mu_1} (F(\tilde{\alpha}) - \log \alpha') \leq T_\alpha(\tilde{\alpha})$. By Lemma G.4, we have
\[
\|\phi(\alpha' e_1, t_0) - \alpha e_1\|_2 = \|\phi(\alpha' e_1, t_0) - e^{\alpha J(0)} \alpha' e_1\|_2 = O(\tilde{\alpha}^2).
\]
Let $r := F^{-1}(\tilde{\mu}_1 t) \leq R$. Then $t + \frac{1}{\mu_1} \log \frac{1}{\alpha} = T_\alpha(r)$ if $\tilde{\alpha} < r$.

By Lemma G.3, $\|\phi(\alpha' e_1, t_0)\|_2 \leq \tilde{\alpha}$. Also, $\|\alpha e_1\|_2 = \frac{\tilde{\alpha}}{1 + \kappa \alpha} \leq \tilde{\alpha}$. By Lemma G.5,
\[
\|z_\alpha(t) - z_{\alpha'}(t)\|_2 \leq O(\tilde{\alpha}^2 \cdot e^{\tilde{\mu}_1 (t + \frac{1}{\mu_1} \log \frac{1}{\alpha'}) + \kappa \tau}) \leq O \left( \frac{\tilde{\alpha}^2}{\alpha} \right).
\]
For $\alpha$ small enough, we have $\tilde{\alpha} = O(\alpha)$, so for any $\alpha' \in (0, \alpha)$,
\[
\|z_\alpha(t) - z_{\alpha'}(t)\|_2 = O(\alpha).
\]
This implies that $\{z_\alpha(t)\}$ satisfies Cauchy’s criterion for every $t$, and thus the limit $z(t)$ exists for $t \leq F(R)/\tilde{\mu}_1$. The convergence rate can be deduced by taking limits for $\alpha' \to 0$ on both sides.
Case 2. For $t = F(R)/\bar{\mu}_1 + \tau$ with $\tau > 0$, $\phi(\theta, \tau)$ is locally Lipschitz with respect to $\theta$. So
\[
\|z_\alpha(t) - z_\alpha'(t)\|_2 = \|\phi(\alpha v_1(F(R)/\bar{\mu}_1), \tau) - \phi(z_\alpha(F(R)/\bar{\mu}_1), \tau)\|_2 = O(\|z_\alpha(F(R)/\bar{\mu}_1) - z_\alpha'(F(R)/\bar{\mu}_1)\|_2) = O(\alpha),
\]
which proves the lemma for $t > F(R)/\bar{\mu}_1$.

Proof for Theorem G.1. The existence of $z(t) := \lim_{\alpha \to 0} z_\alpha(t) = \lim_{\alpha \to 0} \phi(\alpha e_1, t + \frac{1}{\mu_1} \log \frac{1}{\alpha})$ has already been proved in Lemma G.6, where we show \(\|z_\alpha(t) - z(t)\|_2 = O(\alpha)\).

By the continuity of $\phi(\cdot, t)$ for every $t \in \mathbb{R}$, we have
\[
z(t) = \lim_{\alpha \to 0} \phi\left(\alpha v_1, t + \frac{1}{\mu_1} \log \frac{1}{\alpha}\right) = \phi\left(\lim_{\alpha \to 0} \phi\left(\alpha v_1, \frac{1}{\mu_1} \log \frac{1}{\alpha}\right), t\right) = \phi(z(0), t).
\]
Now it is only left to prove (7). WLOG we can assume that $\|\delta_\alpha\|_2$ is decreasing and $\frac{\alpha}{2} \leq \|\delta_\alpha\|_2 \leq \alpha$ (otherwise we can do reparameterization). Then our goal becomes proving
\[
\|\theta_\alpha(t) - z(t)\|_2 = O\left(\frac{\alpha}{\mu_1 + \tau}\right).
\]
where $\theta_\alpha(t) := \phi\left(\delta_\alpha, t + \frac{1}{\mu_1} \log \frac{1}{\delta_\alpha, e_1}\right)$. We prove (23) in the cases of $t \in (-\infty, F(R)/\bar{\mu}_1)$ and $t > F(R)/\bar{\mu}_1$ respectively.

Case 1. Fix $t \in (-\infty, (F(R) + \log q)/\bar{\mu}_1]$. Let $\tilde{\alpha}_1 = \frac{\alpha}{\mu_1 + \tau}$. Let $\alpha_1 := e^{\tilde{F}(\tilde{\alpha}_1)} = \frac{\tilde{\alpha}_1}{1 + \tilde{\alpha}_1 \alpha_1}$. Let $t_0 := \frac{\tilde{\alpha}_1}{1 + \tilde{\alpha}_1 \alpha_1} (F(\tilde{\alpha}_1) - \log \alpha) \leq T(\delta_\alpha, \|\delta_\alpha\|_2)(\tilde{\alpha}_1)$. At time $t_0$, by Lemma G.2 we have
\[
\left\|e^{t_0 J(0)} - e^{\tilde{\mu}_1 t_0 e_1 e_1^T}\right\|_2 = e^{\tilde{\mu}_2 t_0} = e^{\tilde{F}(\tilde{\alpha}_1)-\log \alpha} = \left(\frac{\alpha_1}{\alpha}\right)^{\tilde{\alpha}_1}.
\]
Let $q_\alpha := (<\delta_\alpha, e_1>$. By Definition 5.2, there exists $q > 0$ such that $q_\alpha \geq q$ for all sufficiently small $\alpha$. Then we have
\[
\|\phi(\delta_\alpha, t_0) - \alpha_1 q_\alpha e_1\|_2 = \|\phi(\alpha_1 q_\alpha e_1) - e^{\alpha_1 q_\alpha e_1^T} \delta_\alpha\|_2 + \left\|e^{\alpha_1 q_\alpha e_1^T} - e^{\alpha_1 q_\alpha e_1^T} \delta_\alpha\right\|_2
\]
\[
= O(\alpha_1^2) + \left(\frac{\alpha_1}{\alpha}\right)^{\tilde{\alpha}_1} \|\delta_\alpha\|_2
\]
\[
= O(\alpha_1^2 + \frac{\alpha_1^{\tilde{\alpha}_1} \alpha_1^{-\tilde{\alpha}_1}}{\alpha_1})
\]
\[
= O(\alpha_1^2).
\]
Let $r := F^{-1}(\tilde{\mu}_1 t + \log \frac{1}{\alpha_1}) \leq R$. Then $t + \frac{1}{\mu_1} \log \frac{1}{\alpha_1 q_\alpha} = T(\tilde{\alpha}_1)$ if $\tilde{\alpha}_1 < r$. By Lemma G.3, $\|\phi(\delta_\alpha, t_0)\|_2 \leq \tilde{\alpha}_1$. Also, $\|\alpha_1 q_\alpha e_1\|_2 \leq \alpha_1 = \frac{\tilde{\alpha}_1}{1 + \alpha_1 \tilde{\alpha}_1} \leq \tilde{\alpha}_1$. By Lemma G.5,
\[
\|\theta_\alpha(t) - z_\alpha(t)\|_2 \leq \left\|\phi(\phi(\delta_\alpha, t_0), t + \frac{1}{\mu_1} \log \frac{1}{\alpha_1 q_\alpha}) - \phi(\alpha_1 q_\alpha e_1, t + \frac{1}{\mu_1} \log \frac{1}{\alpha_1 q_\alpha})\right\|_2
\]
\[
= O\left(\alpha_1^2, e^{\tilde{\mu}_1 (t + \frac{1}{\mu_1} \log \frac{1}{\alpha_1 q_\alpha}) + \gamma r}\right)
\]
\[
= O(\alpha_1).
\]
Combining this with the convergence rate for $z_\alpha(t)$, we have
\[
\|\theta_\alpha(t) - z(t)\|_2 \leq \|\theta_\alpha(t) - z_\alpha(t)\|_2 + \|z_\alpha(t) - z(t)\|_2 = O(\alpha_1).
\]

Case 2. For $t = (F(R) + \log q)/\bar{\mu}_1 + \tau$ with $\tau > 0$, $\phi(\theta, \tau)$ is locally Lipschitz with respect to $\theta$. So
\[
\|\theta_\alpha(t) - z(t)\|_2 = \|\phi(\phi((F(R) + \log q)/\bar{\mu}_1), \tau) - \phi(z((F(R) + \log q)/\bar{\mu}_1), \tau)\|_2
\]
\[
= O(\|\theta_\alpha((F(R) + \log q)/\bar{\mu}_1) - z((F(R) + \log q)/\bar{\mu}_1))\|_2
\]
\[
= O(\alpha_1),
\]
which proves (23) for $t > (F(R) + \log q)/\bar{\mu}_1$.

\[\square\]
G.3 Extension to Non-Diagonalizable Case

The proof in Appendix G.2 can be generalized to the case where $J(0)$. Now we state the theorem formally and sketch the proof idea. We use the notations $g(\theta)$, $\phi(\theta_0, t)$, $J(\theta)$ as in Section 5.1, but we do not assume that $J(0)$ is diagonalizable. Instead, we use $\mu_1, \mu_2, \ldots, \mu_d \in \mathbb{C}$ to denote the eigenvalues of $J(0)$, repeated according to algebraic multiplicity. We sort the eigenvalues in the descending order of the real part of each eigenvalue, i.e., $\Re(\mu_1) \geq \Re(\mu_2) \geq \cdots \geq \Re(\mu_d)$, where $\Re(z)$ stands for the real part of a complex number $z \in \mathbb{C}$. We call the eigenvalue with the largest real part the top eigenvalue.

**Theorem G.7.** Assume that $\theta = 0$ is a critical point and the following regularity conditions hold:

1. $g(\theta)$ is $C^{2}$-smooth;
2. $\phi(\theta_0, t)$ exists for all $\theta_0$ and $t$;
3. The top eigenvalue of $J(0)$ is unique and is a positive real number, i.e.,

   \[ \tilde{\mu}_1 > \max\{\Re(\mu_2), 0\}. \]

Let $\tilde{v}_1, \hat{u}_1$ be the left and right eigenvectors associated with $\tilde{\mu}_1$, satisfying $\hat{u}_1^\top \tilde{v}_1 = 1$. Let $z_\alpha(t) := \phi_{(\alpha, \mu_1, t + \frac{1}{\mu_1} \log \frac{1}{\alpha})}$ for every $\alpha > 0$, then $v_\alpha(t) := \lim_{\alpha \to 0} z_\alpha(t)$ exists and $z(t) = \phi(z(0), t)$. If $\delta_\alpha$ converges to 0 with positive alignment with $\tilde{u}_1$ as $\alpha \to 0$, then for any $t \in \mathbb{R}$ and for any $\epsilon > 0$, there is a constant $C > 0$ such that for every sufficiently small $\alpha$,

\[ \left\| \phi \left( \delta_\alpha, t + \frac{1}{\mu_1} \log \frac{1}{\alpha} \right) - z(t) \right\|_2 \leq C \cdot \|\delta_\alpha\|_{\gamma}^{q_1 + q_2} \cdot \epsilon, \]

where $\gamma := \tilde{\mu}_1 - \Re(\mu_2)$ is the eigenvalue gap.

**Proof Sketch.** Define the following two types of matrices. For $r \geq 1, a, \delta \in \mathbb{R}$, we define

\[ J_{a, \delta}^{(r)} := \begin{bmatrix} a & \delta \\ a & \delta \\ & \ddots & \ddots \\ & a & \delta \\ a & \delta \end{bmatrix} \in \mathbb{R}^{r \times r}. \]

For $r \geq 1, a, b, \delta \in \mathbb{R}$, we define

\[ J_{a, b, \delta}^{(r)} := \begin{bmatrix} C & \delta I \\ C & \delta I \\ & \ddots & \ddots \\ & C & \delta I \\ C & \delta I \\ C & \delta I \end{bmatrix} \in \mathbb{R}^{2r \times 2r}, \]

where $C = \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \in \mathbb{R}^{2 \times 2}$. 

By linear algebra, the real matrix $J(0)$ can be written in the real Jordan normal form, i.e., $J(0) = \tilde{V} \text{diag}(J_{[1]}, \ldots, J_{[m]}) \tilde{V}^{-1}$, where $\tilde{V} \in \mathbb{R}^{d \times d}$ is an invertible matrix, and each $J_{[j]}$ is a real Jordan block. Recall that there are two types of real Jordan blocks, $J_{a, \delta}^{(r)}$ or $J_{a, b, \delta}^{(r)}$. The former one is associated with a real eigenvalue $a$, and the latter one is associated with a pair of complex eigenvalues $a \pm bi$. The sum of sizes of all Jordan blocks corresponding to a real eigenvalue $a$ is its algebraic multiplicity. The sum of sizes of all Jordan blocks corresponding to a pair of complex eigenvalues $a \pm bi$ is two times the algebraic multiplicity of $a + bi$ or $a - bi$ (note that $a \pm bi$ have the same multiplicity).

It is easy to see that $J_{a, \delta}^{(r)} = D J_{a, \delta}^{(r)} D^{-1}$ for $D = \text{diag}(\delta^r, \delta^{r-1}, \ldots, \delta) \in \mathbb{R}^{r \times r}$ and $J_{a, b, \delta}^{(r)} = D J_{a, b, \delta}^{(r)} D^{-1}$ for $D = \text{diag}(\delta^r, \delta^r, \delta^{r-1}, \delta^{r-1}, \ldots, \delta, \delta) \in \mathbb{R}^{2r \times 2r}$. This means for every $\delta > 0$
there exists $\tilde{V}_\delta$ such that $J(0) = \tilde{V}_\delta J_\delta \tilde{V}_\delta^{-1}$, where $J_\delta := \text{diag}(J_{\delta[1]}, \ldots, J_{\delta[m]})$, $J_{\delta[j]} := J_{a,\delta}^{(r)}$ if $J_{\delta[j]} := J_a^{(r)}$, or $J_{\delta[a,\delta]} := J_{a,\delta}$ if $J_{\delta[a,\delta]} := J_a^{(r)}$. Since the top eigenvalue of $J(0)$ is positive and unique, $\hat{\mu}_1$ corresponds to only one block $[\hat{\mu}_1] \in \mathbb{R}^{1 \times 1}$. WLOG we let $J_1 = [\hat{\mu}_1]$, and thus $J_{\delta[1]} = [\hat{\mu}_1]$.

We only need to select a parameter $\delta > 0$ and prove the theorem in the case of $J(0) = J_\delta$ since we can change the basis in a similar way as we have done in Appendix G.1. By scrutinizing the proof for Theorem G.1, we can find that we only need to reprove Lemma G.2. However, Lemma G.2 may not be correct since $J(0)$ is not diagonal anymore. Instead, we prove the following:

1. If $\delta \in (0, \check{\gamma})$, then $h^T J_\delta h \leq \hat{\mu}_1 \|h\|^2$ for all $h \in \mathbb{R}^d$;

2. For any $\hat{\mu}_2 \in (\Re(\hat{\mu}_2), \hat{\mu}_1)$, if $\delta \in (0, \hat{\mu}_2 - \Re(\hat{\mu}_2))$, then $\|e^{tJ_\delta} - e^{\hat{\mu}_2 t}e_1e_1^T\|_2 \leq e^{\hat{\mu}_2 t}$ for all $t \geq 0$.

**Proof for Item 1.** Let $K$ be the set of pairs $(k_1, k_2)$ such that $k_1 \neq k_2$ and the entry of $J_\delta$ at the $k_1$-th row and the $k_2$-th column is non-zero. Then we have

$$h^T J_\delta h = h^T J_\delta \frac{J_\delta^T}{2} h = \sum_{k=1}^d \Re(\hat{\mu}_k) h_k^2 + \sum_{(k_1,k_2) \in K} h_{k_1} h_{k_2} \delta \leq \sum_{k=1}^d \Re(\hat{\mu}_k) h_k^2 + \sum_{(k_1,k_2) \in K} h_{k_1}^2 + h_{k_2}^2 \delta.$$

Note that $\Re(\hat{\mu}_k) \leq \Re(\hat{\mu}_2)$ for $k \geq 2$. Also note that there is no pair in $K$ has $k_1 = 1$ or $k_2 = 1$, and for every $k \geq 2$ there are at most two pairs in $K$ has $k_1 = k$ or $k_2 = k$. Combining all these together gives

$$h^T J_\delta h \leq \hat{\mu}_1 h_1^2 + (\Re(\hat{\mu}_2) + \delta) \sum_{k=2}^d h_k^2 \leq \hat{\mu}_1 \|h\|^2,$$

which proves Item 1.

**Proof for Item 2.** Since $J_\delta$ is block diagonal, we only need to prove that $\|e^{tJ_{\delta[j]}}\|_2 \leq e^{\hat{\mu}_2 t}$ for every $j \geq 2$. If $J_{\delta[j]} = J_{a,\delta}^{(r)} = aI + \delta N$, where $N$ is the nilpotent matrix, then

$$e^{tJ_{\delta[j]}} = e^{atI} e^{\delta t N} = e^{atI} e^{\delta t N} = e^{at} e^{\delta t N},$$

where the second equality uses the fact that $I$ and $N$ are commutable. So we have

$$\|e^{tJ_{\delta[j]}}\|_2 \leq e^{at} \|e^{\delta t N}\|_2 = e^{at} e^{\delta t \|N\|_2} \leq e^{(a+\delta)t}.$$

If $J_{\delta[j]} = J_{a,\delta}^{(r)} = D + \delta N^2$, where $D = \text{diag}(C, C, \ldots, C)$ and $N$ is the nilpotent matrix, then

$$e^{tJ_{\delta[j]}} = e^{tD + \delta t N^2} = e^{tD} e^{\delta t N^2},$$

where the second equality uses the fact that $D$ and $N^2$ are commutable. Note that $e^{tC} = e^{at} \left[ \frac{\cos(bt)}{\sin(bt)} - \frac{\sin(bt)}{\cos(bt)} \right]$, which implies $\|e^{tD}\|_2 = \|e^{tC}\|_2 = e^{at}$. So we have

$$\|e^{tJ_{\delta[j]}}\|_2 \leq \|e^{tD}\|_2 \cdot \|e^{\delta t N^2}\|_2 = e^{at} e^{\delta t \|N^2\|_2} \leq e^{(a+\delta)t}.$$

Since $\delta \in (0, \check{\mu}_2 - \Re(\check{\mu}_2))$, we know that $a + \delta < \check{\mu}_2$, which completes the proof.

**Proof for a fixed $\delta$.** Since Item 1 continues to hold for $\delta \in (0, \check{\gamma})$, Lemmas G.3 to G.6 also hold. This proves that $z(t)$ exists and satisfies (6).

It remains to prove (25) for any $\epsilon > 0$. Let $\check{\gamma}' \in (0, \check{\gamma})$ be a number such that $\frac{\check{\gamma}'}{\mu_1 + \check{\gamma}} \geq \frac{\check{\mu}_1}{\hat{\mu}_1 + \check{\gamma}} - \epsilon$. Fix $\check{\mu}_2 = \hat{\mu}_1 - \check{\gamma}' \delta = \hat{\mu}_2 - \Re(\hat{\mu}_2)$. By Item 2, we have $\|e^{tJ_{\delta}} - e^{\hat{\mu}_2 t}e_1e_1^T\|_2 \leq e^{\hat{\mu}_2 t}$ for all $t \geq 0$. By scrutinizing the proof for Theorem G.1, we can find that the only place we use Item 2 in Lemma G.2.
We also write \( \| \delta_{\alpha}, t + \frac{1}{\mu_1} \log \frac{1}{(\delta_{\alpha}, u_1)} \| - \delta^{-1}_t z(t) \leq C \cdot \| \delta^{-1}_t \delta_{\alpha} \|^{\frac{\gamma'}{2 + \gamma}} \), for every sufficiently small \( \alpha \). By definition of \( \gamma', \frac{\gamma'}{2 + \gamma} \geq \frac{\gamma}{2 + \gamma} - \epsilon \). Since \( \delta_{\alpha} \to 0 \) as \( \alpha \to 0 \), we have \( \| \delta^{-1}_t \delta_{\alpha} \| < 1 \) for sufficiently small \( \alpha \). Then we have

\[
\| \phi \left( \delta_{\alpha}, t + \frac{1}{\mu_1} \log \frac{1}{(\delta_{\alpha}, u_1)} \right) - z(t) \| \leq \| \delta^{-1}_t \phi \left( \delta_{\alpha}, t + \frac{1}{\mu_1} \log \frac{1}{(\delta_{\alpha}, u_1)} \right) - \delta^{-1}_t z(t) \| \\
\leq \| \delta^{-1}_t \phi \cdot \| \delta^{-1}_t \delta_{\alpha} \|^{\frac{\gamma'}{2 + \gamma}} \\
\leq C \cdot \| \delta^{-1}_t \| \cdot \| \delta^{-1}_t \delta_{\alpha} \|^{\frac{\gamma'}{2 + \gamma}} \cdot \| \delta_{\alpha} \|^{\frac{\gamma}{2 + \gamma} - \epsilon}.
\]

Absorbing \( \| \delta^{-1}_t \| \cdot \| \delta^{-1}_t \delta_{\alpha} \|^{\frac{\gamma'}{2 + \gamma}} \) into \( C \) proves (25). \( \square \)

H. Eigenvalues of Jacobians and Hessians

In this section we analyze the eigenvalues of the Jacobian \( J(W) \) at critical points of (2).

For notation simplicity, we write \( sz(A) := A + A^T \) to denote the symmetric matrix produced by adding up \( A \) and its transpose, and write \( ac\{A, B\} = AB + BA \) to denote the anticommutator of two matrices \( A, B \). Then \( g(W) \) can be written as \( g(W) := -ac\{\nabla f(W), W\} \).

Let \( U_0 \in \mathbb{R}^{d \times r} \) be a stationary point of the function \( \mathcal{L} : \mathbb{R}^{d \times r} \to \mathbb{R}, \mathcal{L}(U) = \frac{1}{2} f(UU^\top) \), i.e., \( \nabla \mathcal{L}(U_0) = \nabla f(U_0U_0^\top)U_0 = 0 \). By Lemma C.1, this implies

\[
\nabla f(W_0)W_0 = 0 \tag{26}
\]

for \( W_0 := U_0U_0^\top \), and thus \( W_0 \) is a critical point of (2).

For a real-valued or vector-valued function \( F(\theta) \), we use \( D F(\theta)[\delta], D^2 F(\theta)[\delta_1, \delta_2] \) to denote the first- and second-order directional derivatives of \( F(\cdot) \) at \( \theta \).

Let \( \mathcal{X} \) be a linear space, which can be \( \mathbb{R}^{d \times d} \) or \( \mathbb{R}^{d \times r} \). For a function \( F : \mathcal{X} \to \mathcal{X} \), we use \( D F(\theta) \) to denote the directional derivative of \( F \) at \( \theta \), represented by the linear operator

\[
DF(\theta)[\Delta] : \mathcal{X} \to \mathcal{X}, \Delta \mapsto DF(\theta)[\Delta] = \lim_{t \to 0} \frac{F(\theta + t\Delta) - F(\theta)}{t}.
\]

We also write \( DF(\theta)[\Delta_1, \Delta_2] := (DF(\theta)[\Delta_1], \Delta_2) \).

For a function \( F : \mathcal{X} \to \mathbb{R} \), we use \( D^2 F(\theta) = D(\nabla F(\theta)) \) to denote the second directional derivative of \( F \) at \( \theta \), i.e., \( D^2 F(\theta)[\Delta] = D(\nabla F(\theta))[\Delta] \), \( D^2 F(\theta)[\Delta_1, \Delta_2] = D(\nabla F(\theta))[\Delta_1, \Delta_2] \).

Define \( J(W) := Dg(W) \). By simple calculus, we can compute the formula for \( J(W_0) \):

\[
J(W_0)[\Delta] = -ac\{\nabla f(W_0), \Delta \} - ac\{D^2 f(W_0)[\Delta, W_0] \},
\]

\[
J(W_0)[\Delta_1, \Delta_2] = -\langle \nabla f(W_0), sz(\Delta_1 \Delta_2^\top) \rangle - D^2 f(W_0)[\Delta_1, sz(W_0 \Delta_2^\top)],
\]

where \( \Delta, \Delta_1, \Delta_2 \in \mathbb{R}^{d \times d} \).

We can also compute the formula for \( D^2 \mathcal{L}(U_0) \):

\[
D^2 \mathcal{L}(U_0)[\Delta] = \nabla f(W_0) \Delta + D^2 f(W_0)[sz(\Delta U_0^\top)]U_0,
\]

\[
D^2 \mathcal{L}(U_0)[\Delta_1, \Delta_2] = \frac{1}{2} \left( \langle \nabla f(W_0), sz(\Delta_1 \Delta_2^\top) \rangle + D^2 f(W_0)[sz(\Delta_1 U_0^\top), sz(\Delta_2 U_0^\top)] \right),
\]

where \( \Delta, \Delta_1, \Delta_2 \in \mathbb{R}^{d \times r} \).
H.1 Eigenvalues at the Origin

The eigenvalues of $J(0)$ is given in Lemma 5.4. Now we provide the proof.

**Proof for Lemma 5.4.** For $W_0 = 0$, we have

$$J(0)[\Delta] = -\nabla f(0)\Delta - \Delta \nabla f(0)$$

$$J(0)[\Delta_1, \Delta_2] = -\langle \nabla f(0), sz(\Delta_1\Delta_2^T) \rangle$$

It is easy to see from the second equation that $J(0)$ is symmetric.

Let $-\nabla f(0) = \sum_{i=1}^d \mu_i u_{1[i]} u_{1[i]}^T$ be the eigendecomposition of the symmetric matrix $-\nabla f(0)$. Then we have

$$J(0)[\Delta] = \sum_{i=1}^d \mu_i \left( u_{1[i]} u_{1[i]}^T \Delta + \Delta u_{1[i]} u_{1[i]}^T \right)$$

$$= \sum_{i=1}^d \sum_{j=1}^d \mu_i \left( u_{1[i]} u_{1[i]}^T \Delta u_{1[j]} u_{1[j]}^T + u_{1[j]} u_{1[j]}^T \Delta u_{1[i]} u_{1[i]}^T \right)$$

$$= \sum_{i=1}^d \sum_{j=1}^d (\mu_i + \mu_j) u_{1[i]} u_{1[i]}^T \Delta u_{1[j]} u_{1[j]}^T$$

$$= \sum_{i=1}^d \sum_{j=1}^d (\mu_i + \mu_j) \langle \Delta, u_{1[i]} u_{1[i]}^T \rangle u_{1[i]} u_{1[i]}^T,$$

which proves (8).

For $\Delta = u_{1[i]} u_{1[i]}^T + u_{1[j]} u_{1[j]}^T$, we have

$$J(0)[\Delta] = (\mu_i + \mu_j) u_{1[i]} u_{1[i]}^T + (\mu_i + \mu_j) u_{1[j]} u_{1[j]}^T = (\mu_i + \mu_j) \Delta.$$  

So $u_{1[i]} u_{1[i]}^T + u_{1[j]} u_{1[j]}^T$ is an eigenvector of $J(0)$ associated with eigenvalue $\mu_i + \mu_j$. Note that \{u_{1[i]} u_{1[i]}^T + u_{1[j]} u_{1[j]}^T : 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 $\Delta$ (i.e., $\Delta = -\Delta^T$), we have

$$J(0)[\Delta] = J(0)[\Delta^T] = J(0)[-\Delta].$$

So $J(0)[\Delta] = 0$ and every antisymmetric matrix is an eigenvector associated with eigenvalue 0.

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

H.2 Eigenvalues at Second-Order Stationary Points

Now we study the eigenvalues of $J(W_0)$ when $U_0$ is a second-order stationary point of $\mathcal{L}(\cdot)$, i.e., $\nabla \mathcal{L}(U_0) = 0$, $D^2 \mathcal{L}(U_0)[\Delta, \Delta] \geq 0$ for all $\Delta \in \mathbb{R}^{d \times r}$. We further assume that $U_0$ is full-rank, i.e., $\text{rank}(U_0) = r$. This condition is met if $W_0 := U_0 U_0^T$ is a local minimizer of $f(\cdot)$ in $\mathbb{R}^{d \times d}$ but not a minimizer in $\mathbb{S}^+_{d \times d}$.

**Lemma H.1.** For $r \leq d$, if $U_0 \in \mathbb{R}^{d \times r}$ is a second-order stationary point of $\mathcal{L}(\cdot)$, then either $\text{rank}(U_0) = r$, or $W_0$ is a minimizer of $f(\cdot)$ in $\mathbb{S}^+_{d \times d}$, where $W_0 = U_0 U_0^T$.

**Proof.** Assume to the contrary that $U_0$ has rank $< r$ and $W_0$ is a minimizer of $f(\cdot)$ in $\mathbb{S}^+_{d \times d}$. The former one implies that there exists a unit vector $q \in \mathbb{R}^r$ such that $U_0 q = 0$, and the latter one implies that there exists $v \in \mathbb{R}^d$ such that $v^T \nabla f(W_0) v < 0$ by Lemma C.2.

27
Let $\Delta = vq^\top$. Then we have
\[
D^2 \mathcal{L}(U_0) [\Delta, \Delta] = \langle \nabla f(W_0), v v^\top \rangle + \frac{1}{2} D^2 f(W_0)[sz(v(U_0q)^\top), sz(v(U_0q)^\top)] \\
= \langle \nabla f(W_0), v v^\top \rangle + \frac{1}{2} D^2 f(W_0)[0, 0] \\
= \langle \nabla f(W_0), v v^\top \rangle.
\]
So $D^2 \mathcal{L}(U_0)[\Delta, \Delta] < 0$, which leads to a contradiction. \hfill \square

By (26), the symmetric matrices $-\nabla f(W_0)$ and $W_0$ commute, so they can be simultaneously diagonalizable. Since (26) also implies that they have different column spans, we can have the following diagonalization:
\[
-\nabla f(W_0) = \sum_{i=1}^{d-r} \mu_i v_i v_i^\top, \quad W_0 = \sum_{i=d-r+1}^{d} \mu_i v_i v_i^\top. \tag{27}
\]
First we prove the following lemma on the eigenvalues and eigenvectors of the linear operator $-D^2 \mathcal{L}(U_0)$:

**Lemma H.2.** For every $\Delta \in \mathbb{R}^{d \times d}$, if
\[
U_0 \Delta^\top + \Delta U_0^\top = 0 \tag{28}
\]
then $\Delta$ is an eigenvector of the linear operator $-D^2 \mathcal{L}(U_0)[\cdot] : \mathbb{R}^{d \times r} \to \mathbb{R}^{d \times r}$ associated with eigenvalue 0. Moreover, the solutions of (28) spans a linear space of dimension $\frac{r(r-1)}{2}$.

**Proof.** Suppose $U_0 \Delta^\top + \Delta U_0^\top = 0$. Then we have $U_0 \Delta^\top = -\Delta U_0^\top$, and thus $\Delta^\top = -U_0^\top \Delta U_0^\top$, where $U_0^\top$ is the pseudoinverse of the full-rank matrix $U_0$. This implies that there is a matrix $R \in \mathbb{R}^{r \times r}$, such that $\Delta = U_0 R$. Then we have
\[
-D^2 \mathcal{L}(U_0)[\Delta] = -\langle \nabla f(W_0) U_0 R - D^2 f(W_0)[U_0 \Delta^\top + \Delta U_0^\top] U_0 \\
= -\langle \nabla f(W_0) U_0 \rangle R - D^2 f(W_0)[0] U_0 \\
= 0.
\]
Replacing $\Delta$ with $U_0 R$ in (28) gives $U_0 (R + R^\top) U_0^\top = 0$, which is equivalent to $R = -R^\top$ since $U_0$ is full-rank. Since the dimension of $r \times r$ antisymmetric matrices is $\frac{r(r-1)}{2}$, the span spanned by the solutions of (28) also has dimension $\frac{r(r-1)}{2}$. \hfill \square

**Definition H.3** (Eigendecomposition of $-D^2 \mathcal{L}(U_0)$). Let
\[
-D^2 \mathcal{L}(U_0)[\Delta] = \sum_{p=1}^{rd} \xi_p \langle E_p, \Delta \rangle E_p
\]
be the eigendecomposition of the symmetric linear operator $-D^2 \mathcal{L}(U_0)[\cdot] : \mathbb{R}^{d \times r} \to \mathbb{R}^{d \times r}$, where $\xi_1, \ldots, \xi_{rd} \in \mathbb{R}$ are eigenvalues, $E_1, \ldots, E_{rd} \in \mathbb{R}^{d \times r}$ are eigenvectors satisfying $\langle E_p, E_q \rangle = \delta_{pq}$. We enforce $\xi_p$ to be 0 and $E_p$ to be a solution of (28) for every $rd - \frac{r(r-1)}{2} < p \leq rd$.

**Lemma H.4.** Let $\mathbf{A} \in \mathbb{R}^{D \times D}$ be a matrix. If $\{\hat{u}_1, \ldots, \hat{u}_K\}$ is a set of linearly independent left eigenvectors associated with eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_K$ and $\{\hat{v}_1, \ldots, \hat{v}_{D-K}\}$ is a set of linearly independent right eigenvectors associated with eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_{D-K}$, and $\langle \hat{u}_i, \hat{v}_j \rangle = 0$ for all $1 \leq i \leq K, 1 \leq j \leq D - K$, then $\hat{\lambda}_1, \ldots, \hat{\lambda}_K, \hat{\lambda}_1, \ldots, \hat{\lambda}_{D-K}$ are all the eigenvalues of $\mathbf{A}$.

**Proof.** Let $\hat{U} := (\hat{u}_1, \ldots, \hat{u}_K)^\top \in \mathbb{R}^{K \times D}$ and $\hat{V} := (\hat{v}_1, \ldots, \hat{v}_{D-K})^\top \in \mathbb{R}^{D \times (D-K)}$. Then both $\hat{U}$ and $\hat{V}$ are full-rank. Let $\hat{U}^+ = \hat{U}^\top (\hat{U} \hat{U}^\top)^{-1}, \hat{V}^+ = (\hat{V}^\top \hat{V})^{-1} \hat{V}^\top$ be the pseudoinverses of $\hat{U}$ and $\hat{V}$.
Now we define
\[
P := \begin{bmatrix} \hat{U} \\ \hat{V}^+ \end{bmatrix}, \quad Q := \begin{bmatrix} \hat{U}^+ & \hat{V} \end{bmatrix}.
\]
Then we have
\[
PQ = \begin{bmatrix} \hat{U}\hat{U}^+ & \hat{U}\hat{V} \\ \hat{V}\hat{U}^+ & \hat{V}\hat{V}^+ \end{bmatrix}.
\]
Note that \(\hat{U}\hat{U}^+ = I_K, \hat{V}\hat{V} = 0, \hat{V}^+\hat{U}^+ = (\hat{V}^T\hat{V})^{-1}(\hat{U}\hat{V})^T(\hat{U}\hat{U}^+)^{-1} = 0, \hat{V}^+\hat{V} = I_{D-K} \).
So \(PQ = I_D, \) or equivalently \(Q = P^{-1}.\) Then we have
\[
P^{-1}AP = \begin{bmatrix} \text{diag}(\lambda_1, \ldots, \lambda_K) & * \\ 0 & \text{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{D-K}) \end{bmatrix},
\]
where * can be any \(K \times (D-K)\) matrix. Since \(P^{-1}AP\) is upper-triangular, we know that \(P^{-1}AP\) has eigenvalues \(\lambda_1, \ldots, \lambda_K, \tilde{\lambda}_1, \ldots, \tilde{\lambda}_{D-K},\) and so does \(A.\)

**Theorem H.5.** The eigenvalues of \(J(W_0)\) can be fully classified into the following 3 types:

1. \(\mu_i + \mu_j\) is an eigenvalue for every \(1 \leq i \leq j \leq d - r,\) and \(\hat{U}_{ij} := v_i^Tv_j^T + v_j^Tv_i^T\) is an associated left eigenvector.
2. \(\xi_p\) is an eigenvalue for every \(1 \leq p \leq rd - \frac{r(r-1)}{2},\) and \(\hat{V}_p := E_pU_0^T + U_0E_p^T\) is an associated right eigenvector.
3. 0 is an eigenvalue, and any antisymmetric matrix is an associated right eigenvector, which spans a linear space of dimension \(\frac{d(d-1)}{2}.\)

**Proof of Theorem H.5.** We first prove each item respectively, and then prove that these are all the eigenvalues of \(J(W_0)\).

**Proof for Item 1.** For \(\hat{U}_{ij} = v_i^Tv_j^T + v_j^Tv_i^T,\) it is easy to check:
\[
ac\{-\nabla f(W_0), \hat{U}_{ij}\} = (\lambda_i + \lambda_j)\hat{U}_{ij}
\]
\[
\hat{U}_{ij}W_0 = 0
\]
\[
W_0\hat{U}_{ij} = 0
\]
So we have
\[
J(W_0)[\Delta, \hat{U}_{ij}] = (\lambda_i + \lambda_j)\left\langle \Delta, \hat{U}_{ij} \right\rangle - D^2f(W_0)[\Delta, 0] = (\lambda_i + \lambda_j)\left\langle \Delta, \hat{U}_{ij} \right\rangle,
\]
which shows that \(\hat{U}_{ij}\) is a left eigenvector associated with eigenvalue \(\lambda_i + \lambda_j.\)

**Proof for Item 2.** By definition of eigenvector, we have \(-D^2\mathcal{L}(U_0)[E_p] = \xi_pE_p,\) so
\[
\xi_pE_p = -\nabla f(W_0)E_p - D^2f(W_0)[U_0E_p^T + E_pU_0^T]U_0.
\]
Right-multiplying both sides by \(U_0^T,\) we get
\[
\xi_pE_pU_0^T = -\nabla f(W_0)E_pU_0^T - D^2f(W_0)[\hat{V}_p]W_0
\]
\[
= -\nabla f(W_0)[E_pU_0^T + U_0E_p^T] - D^2f(W_0)[\hat{V}_p]W_0
\]
\[
= -\nabla f(W_0)[\hat{V}_p]W_0 - D^2f(W_0)[\hat{V}_p]W_0,
\]
where the second equality uses the fact that \(\nabla f(W_0)U_0 = 0\) since \(U_0\) is a critical point. Taking both sides into \(sz(\cdot)\) gives
\[
\xi_p\hat{V}_p = -sz(\nabla f(W_0)\hat{V}_p) - sz(D^2f(W_0)[\hat{V}_p]W_0)
\]
\[
= J(W_0)[\hat{V}_p],
\]
which proves that \(\hat{V}_p\) is a right eigenvector associated with eigenvalue \(\xi_p.\)
Proof for Item 3. Since $\nabla f(W)$ is symmetric, $g(W)$ is also symmetric. For any $\Delta = -\Delta^\top$, 

$$J(W_0)[\Delta] = J(W_0)[\Delta^\top] = J(W_0)[\Delta].$$

So $J(W_0)[\Delta] = 0$ and $\Delta$ is an eigenvector associated with eigenvalue 0.

No other eigenvalues. Let $S_d$ be the space of symmetric matrices and $A_d$ be the space of antisymmetric matrices. It is easy to see that $S_d$ and $A_d$ are orthogonal to each other, and $S_d$ and $A_d$ are invariant subspaces of $J(W_0)[\Delta]$. Let $h : S_d \to S_d, \Delta \mapsto J(W_0)[\Delta]$ be the linear operator $J(W_0)[\Delta]$ restricted on symmetric matrices. We only need to prove that $h$ is diagonalizable.

It is easy to see that $\{\hat{U}_{ij}\}$ are linearly independent to each other and thus spans a subspace of $S_d$ with dimension $\frac{(d-r)(d-r+1)}{2}$. We can also prove that $\{\hat{V}_p\}$ spans a subspace of $S_d$ with dimension $\frac{rd - r(r-1)}{2}$ by contradiction. Assume to the contrary that there exists scalars $\alpha_p$ for $1 \leq p \leq rd - r(r-1)/2$, not all zero, such that $\sum_{p=1}^{rd-r(r-1)/2} \alpha_p \hat{V}_p = 0$. Then $\sum_{p=1}^{rd-r(r-1)/2} \alpha_p E_p$ is a solution of (28). However, this suggests that $\sum_{p=1}^{rd-r(r-1)/2} \alpha_p E_p$ lies in the span of $\{E_p\}_{rd-r(r-1)/2 < p \leq rd}$, which contradicts to the linear independence of $\{E_p\}_{1 \leq p \leq rd}$.

Note that 

$$\left(\frac{(d-r)(d-r+1)}{2} + \frac{rd - r(r-1)}{2}\right) = \frac{(d+1)}{2} = \dim(S_d).$$

Also note that $\left(\hat{U}_{ij}, \hat{V}_p\right) = 2v_i^\top E_p U_0^\top v_j + 2v_j^\top E_p U_0^\top v_i = 0$. By Lemma H.4, Items 1 and 2 give all the eigenvalues of $h$, and thus Items 1, 2, 3 give all the eigenvalues of $J(W_0)$.

I PROOFS FOR THE DEPTH-2 CASE

I.1 PROOF FOR THEOREM 5.6

Proof for Theorem 5.6. Since $W(t)$ is always symmetric, it suffices to study the dynamics of the lower triangle of $W(t)$. For any symmetric matrix $W \in S_d$, let $\text{vec}_{LT}(W) \in \mathbb{R}^{\frac{d(d+1)}{2}}$ be the vector consisting of the $\frac{d(d+1)}{2}$ entries of $W$ in the lower triangle, permuted according to some fixed order.

Let $g(W)$ be the function defined in (2), which always maps symmetric matrices to symmetric matrices. Let $\tilde{g} : \mathbb{R}^{\frac{d(d+1)}{2}} \to \mathbb{R}^{\frac{d(d+1)}{2}}$ be the function such that $\tilde{g}(\text{vec}_{LT}(W)) = \text{vec}_{LT}(g(W))$ for any $W \in S_d$. For $W(t)$ evolving with (2), we view $\text{vec}_{LT}(W(t))$ as a dynamical system.

$$\frac{d}{dt} \text{vec}_{LT}(W(t)) = \tilde{g}(\text{vec}_{LT}(W(t))).$$

By Lemma 5.4, the spaces of symmetric matrices $S_d$ and antisymmetric matrices $A_d$ are invariant subspaces of $J(0)$, and $\{(\mu_i + \mu_j, u_{1[i]} u_{1[j]}^\top + u_{1[j]} u_{1[i]}^\top)\}_{1 \leq i \leq j \leq d}$ is the set of all the eigenvalues and eigenvectors in the invariant subspace $S_d$. Thus, $\mu_1 := 2\mu_1$ and $\tilde{\mu}_2 := \mu_1 + \mu_2$ are the largest and second largest eigenvalues of the Jacobian of $\tilde{g}(\cdot)$ at $\text{vec}_{LT}(W) = 0$, and $\tilde{u}_1 = \tilde{v}_1 = u_{1[i]} u_{1[j]}^\top$ are the corresponding left and right eigenvectors of the top eigenvalue. Then it is easy to translate Theorem 5.3 to Theorem 5.6.

I.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 I.1. If $\hat{\theta}$ is a local minimizer of $L(\theta)$ and for all $\|\theta - \hat{\theta}\|_2 \leq r$, $\theta$ satisfies Łojasiewicz inequality:

$$\|\nabla L(\theta)\|_2 \geq c (L(\theta) - L(\hat{\theta}))^{\mu}$$

for some $\mu \in [1/2, 1)$, then the gradient flow $\theta(t) = \phi(\theta_0, t)$ converges to a point $\theta_\infty$ near $\hat{\theta}$ if $\theta_0$ is close enough to $\hat{\theta}$, and the distance can be bounded by $\|\theta_\infty - \hat{\theta}\|_2 = O(\|\theta_0 - \hat{\theta}\|_2^{2(1-\mu)})$. 

30
Proof. For every \( t \geq 0 \), if \( \| \theta(t) - \bar{\theta} \|_2 \leq r \),
\[
\frac{d}{dt} \left( \mathcal{L}(\theta(t)) - \mathcal{L}(\bar{\theta}) \right)^{1-\mu} = (1-\mu) \left( \mathcal{L}(\theta(t)) - \mathcal{L}(\bar{\theta}) \right)^{1-\mu} \cdot \left( \nabla \mathcal{L}, \frac{d\theta}{dt} \right)
= - (1-\mu) \left( \mathcal{L}(\theta(t)) - \mathcal{L}(\bar{\theta}) \right)^{1-\mu} \cdot \| \nabla \mathcal{L} \|_2 \cdot \left\| \frac{d\theta}{dt} \right\|_2
\leq - (1-\mu)c \left\| \frac{d\theta}{dt} \right\|_2.
\]
Therefore, \( \| \theta(t) - \theta_0 \|_2 \leq \int_0^t \left\| \frac{d\theta}{dt} \right\|_2 \, dt \leq \frac{1}{(1-\mu)c} \mathcal{L}(\theta_0)^{1-\mu} = O(\| \theta_0 - \bar{\theta} \|_2^{2(1-\mu)}) \). If we choose \( c\| \theta(t) - \bar{\theta} \|_2 \) small enough, then \( \| \theta(t) - \bar{\theta} \|_2 \leq \| \theta(t) - \theta_0 \|_2 + \| \theta_0 - \bar{\theta} \|_2 = O(\| \theta_0 - \bar{\theta} \|_2^{2(1-\mu)}) < r \), and thus \( \int_0^\infty \left\| \frac{d\theta}{dt} \right\|_2 \, dt \) is convergent and finite. This implies that \( \theta_\infty := \lim_{t \to +\infty} \theta(t) \) exists and \( \| \theta_\infty - \bar{\theta} \|_2 = O(\| \theta_0 - \bar{\theta} \|_2^{2(1-\mu)}) \). \( \square \)

Proof for Theorem 5.8. Since \( W^G_1(t) \in \mathbb{S}_{d, \leq 1}^+ \) satisfies \eqref{eq:2}, there exists \( u(t) \in \mathbb{R}^d \) such that \( u(t)u(t)^\top = W^G_1(t) \) and \( u(t) \) satisfies \eqref{eq:1}, i.e., \( \frac{d}{dt} u(t) = -\nabla \mathcal{L}(u(t)) \), where \( \mathcal{L} : \mathbb{R}^d \to \mathbb{R} \). If \( W^G_1(t) \) does not diverge to infinity, then so does \( u(t) \). This implies that there is a limit point \( \bar{u} \) of the set \( \{ u(t) : t \geq 0 \} \).

Let \( \mathcal{U} := \{ u : \mathcal{L}(u) \geq \mathcal{L}(\bar{u}) \} \). Since \( \mathcal{L}(u(t)) \) is non-increasing, we have \( u(t) \in \mathcal{U} \) for all \( t \). Note that \( \bar{u} \) is a local minimizer of \( \mathcal{L}(\cdot) \) in \( \mathcal{U} \). By analyticity of \( f(\cdot) \), Łojasiewicz inequality holds for \( \mathcal{L}(\cdot) \) around \( \bar{u} \) \cite{Lojasiewicz1965}. Applying Lemma 1.1 for \( \mathcal{L} \) restricted on \( \mathcal{U} \), we know that if \( u(t_0) \) is sufficiently close to \( \bar{u} \), the remaining length of the trajectory of \( u(t) \) \((t \geq t_0)\) is finite and thus \( \lim_{t \to +\infty} u(t) \) exists. As \( \bar{u} \) is a limit point, this limit can only be \( \bar{u} \). Therefore, \( W_1 := \lim_{t \to +\infty} W^G_1(t) = \bar{u} \bar{u}^\top \) exists.

If \( W_1 \) is a minimizer of \( f(\cdot) \), \( \mathcal{U} = \{ \bar{u}, 0, \cdots, 0 \} \in \mathbb{R}^{d \times d} \) is also a minimizer of \( \mathcal{L} : \mathbb{R}^{d \times d} \to \mathbb{R} \), \( U \mapsto \frac{1}{2} f(UU^\top) \). By analyticity of \( f(\cdot) \), Łojasiewicz inequality holds for \( \mathcal{L}(\cdot) \) around \( \mathcal{U} \). For every \( \epsilon > 0 \), we can always find a time \( t \), such that \( \| u(t) - \bar{u} \|_2 \leq \epsilon/2 \). On the other hand, by Theorem 5.6, there exists a number \( \alpha \), such that for every \( \alpha < \alpha_\epsilon \),
\[
\| \phi(W, T(W) + t_x) - W_1^G(t_x) \|_2 \leq \epsilon/2, \quad \text{where} \quad T(W) := \frac{1}{2\mu_1} \log \frac{1}{\langle W, u_1 u_1^\top \rangle}.
\]
Combining these together we have \( \| \phi(W_\alpha, T(W_\alpha) + t_x) - W_1^G(t_x) \|_2 \leq \epsilon \).

It is easy to construct a factorization \( \phi(W_\alpha, T(W_\alpha) + t_x) := U_{\alpha, \epsilon} U_{\alpha, \epsilon}^\top \) such that \( \| U_{\alpha, \epsilon} - \bar{U} \|_2 = O(\epsilon) \), e.g., we can find an arbitrary factorization and then right-multiply an orthogonal matrix so that the row vector with the largest norm aligns with the direction of \( \bar{u} \). Applying Lemma 1.1, we know that gradient flow starting with \( U_{\alpha, \epsilon} \) converges to a point that is only \( O(\epsilon^{2(1-\mu)}) \) far from \( \bar{u} \). So we have
\[
\lim_{t \to +\infty} \phi(W_\alpha, T(W_\alpha) + t) - W_1^G(t) = O(\epsilon^{2(1-\mu)}).
\]
Taking \( \epsilon \to 0 \) complete the proof. \( \square \)

I.3 Proof for Theorem 5.11

Theorem I.2. Let \( W \) be a critical point of \eqref{eq:2} satisfying that \( W \) is a local minimizer of \( f(\cdot) \) in \( \mathbb{S}_{d, \leq r}^+ \) for some \( r \geq 1 \) but not a minimizer in \( \mathbb{S}_{d, \leq r}^+ \). Let \( -\nabla f(W) = \sum_{i=1}^d \mu_i v_i v_i^\top \) be the eigendecomposition of \( -\nabla f(W) \). If \( \mu_1 > \mu_2 \), the following limit exists and is a solution of \eqref{eq:2}.
\[
W^G(t) := \lim_{\epsilon \to 0} \phi \left( W + \epsilon v_1 v_1^\top, \frac{1}{2\mu_1} \log \frac{1}{\epsilon} + t \right).
\]
For \( \{ W_\alpha \} \subseteq \mathbb{S}_{d, \leq r}^+ \), if there exists time \( T_\alpha \in \mathbb{R} \) for every \( \alpha \) so that \( \phi(W_\alpha, T_\alpha) \) converges to \( \bar{W} \) with positive alignment with the top principal component \( v_1 v_1^\top \) as \( \alpha \to 0 \), then \( \forall t \in \mathbb{R} \),
\[
\lim_{\alpha \to 0} \phi \left( W_\alpha, T_\alpha + \frac{1}{2\mu_1} \log \frac{1}{\langle \phi(W_\alpha, T_\alpha), v_1 v_1^\top \rangle + t} \right) = W^G(t).
\]
Moreover, there exists a constant $C > 0$ such that
\[
\left\| \phi \left( W_\alpha, T_\alpha + \frac{1}{2\mu_1} \log \frac{1}{\langle \phi(W_\alpha, T_\alpha), v_1v_1^\top \rangle} + t \right) - W^G(t) \right\|_F \leq C \left\| \phi(W_\alpha, T_\alpha) \right\|_{F^{\frac{\gamma}{1 + \gamma}}}
\]
for every sufficiently small $\alpha$, where $\gamma := 2\mu_1 - \max\{\mu_1 + \mu_2, 0\}$.

Proof. Following Appendix I.1, we view $\text{vec}_{LT}(W(t))$ as a dynamical system.

\[
d\text{vec}_{LT}(W(t)) = \tilde{g}(\text{vec}_{LT}(W(t))).
\]

Let $\tilde{W} = \tilde{U} \tilde{U}^\top$ be a factorization of $W$, where $\tilde{U} \in \mathbb{R}^{d \times r}$. Since $\tilde{W}$ is a local minimizer of $f(\cdot)$ in $S_d^{r \times r}$, $\tilde{U}$ is also a local minimizer of $L : \mathbb{R}^{d \times r} \to \mathbb{R}, U \mapsto \frac{1}{2} f(UU^\top)$. Since $\tilde{W}$ is not a minimizer of $f(\cdot)$ in $S_r^d$, by Lemma H.1. $\tilde{U}$ is full-rank. By Theorem H.5, $J(\tilde{W})$ has eigenvalues $\mu + \mu_j, \xi_p, 0$. By a similar argument as in Appendix I.1, the Jacobian of $\tilde{g}$ at $\text{vec}_{LT}(W(t))$ has eigenvalues $\mu + \mu_j, \xi_p$.

Since $\tilde{U}$ is a local minimizer, $\xi_p \leq 0$ for all $p$. If $\mu_1 > \mu_2$, then $2\mu_1$ is the unique largest eigenvalue, and Theorem H.5 shows that $\text{vec}_{LT}(v_1v_1^\top)$ is a left eigenvector associated with $2\mu_1$. The eigenvalue gap $\gamma := 2\mu_1 - \max\{\mu_1 + \mu_2, \max\{\xi_p : 1 \leq p \leq rd - \frac{r(r-1)}{2}\}\} \geq 2\mu_1 - \max\{\mu_1 + \mu_2, 0\}$.

Also note that $\langle \phi(W_\alpha, T_\alpha) - W, v_1v_1^\top \rangle = \langle \phi(W_\alpha, T_\alpha), v_1v_1^\top \rangle$ because $\langle W, v_1v_1^\top \rangle = 0$ by (27). If $\phi(W_\alpha, T_\alpha)$ converges to $W$ as $\alpha \to 0$, then it has positive alignment with $v_1v_1^\top$ iff
\[
\liminf_{\alpha \to 0} \frac{\langle \phi(W_\alpha, T_\alpha), v_1v_1^\top \rangle}{\|\phi(W_\alpha, T_\alpha) - W\|_F} > 0.
\]
Then it is easy to translate Theorem 5.3 to Theorem I.2. \qed

1.4 Gradient Flow only finds minimizers (Proof for Theorem 5.10)

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

Theorem I.3 (Theorem 3.1 in Du and Lee 2018). Let $f : \mathbb{R}^{d \times k} \to \mathbb{R}$ be a $C^2$ convex function. Then $L : \mathbb{R}^{d \times k} \to \mathbb{R}, L(U) = f(UU^\top), k \geq d$ satisfies that (1). Every local minimizer of $L$ 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). 2

Theorem I.4 (Theorem 2 in Lee et al. 2017). Let $g$ be a $C^1$ mapping from $X \to X$ and det $(Dg(x)) \neq 0$ for all $x \in X$. Then the set of initial points that converge to an unstable fixed point has measure zero, $\mu \left( \{x_0 : \lim_{k \to \infty} g^k(x_0) \in A_g^* \} \right) = 0$, where $A_g^* = \{x : g(x) = x, \max_j |\lambda_j(Dg(x))| > 1\}$.

Theorem I.5 (GF only finds minimizers, a continuous analog of Theorem I.4). Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a $C^1$-smooth function, and $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ be the solution of the following differential equation,
\[
\frac{d\phi(x, t)}{dt} = f(\phi(x, t)), \quad \phi(x, 0) = x, \quad \forall x \in \mathbb{R}^d, t \in \mathbb{R}.
\]

Then the set of initial points that converge to an unstable critical point has measure zero,
\[
\mu \left( \{x_0 : \lim_{t \to \infty} \phi(x_0, t) \in \mathcal{U}_f^\star \} \right) = 0, \quad \text{where} \quad \mathcal{U}_f^\star = \{x : f(x) = 0, \lambda_1(Df(x)) > 0\}
\]
and $Df$ is the Jacobian matrix of $f$.

Proof of Theorem I.5. By Theorem 1 in Section 2.3, Perko (2013), we know $\phi(\cdot, \cdot)$ is $C^1$-smooth for both $x, t$. We let $g(x) = \phi(x, 1)$, then we know $g^{-1}(x) = \phi(x, -1)$ and both $g, g^{-1}$ are $C^1$-smooth. Note that $Dg^{-1}(x)$ is the inverse matrix of $Dg(x)$. So both of the two matrices are invertible. Thus we can apply Theorem I.4 and we know $\mu \left( \{x_0 : \lim_{k \to \infty} g^k(x_0) \in A_g^* \} \right) = 0$.

Note that if $\lim_{t \to \infty} \phi(x, t)$ exists, then $\lim_{k \to \infty} \phi^k(x) = \lim_{k \to \infty} \phi(x, t)$. It remains to show that $\mathcal{U}_f^\star \subseteq A_g^*$. For $f(x_0) = 0$, we have $\phi(x_0, t) = x_0$ and thus $g(x_0) = x_0$. Now it suffices to prove

2Though the original theorem is proven for convex functions of form $\sum_{i=1}^n \ell(x_iUU^\top x_i^\top, y_i)$, where $\ell(\cdot, \cdot)$ is $C^2$ convex for its first variable. By scrutinizing their proof, we can see the assumption can be relaxed to $f$ is $C^2$ convex.
that \( \lambda_1(Dg(x_0)) > 1 \). For every \( t \in [0, 1] \), by Corollary of Theorem 1 in Section 2.3, Perko (2013), we have \( \frac{\partial}{\partial t} D\phi(x,t) = Df(\phi(x,t))D\phi(x,t), \forall x,t \). Thus,

\[
\frac{\partial}{\partial t} D\phi(x_0,t) = Df(\phi(x_0,t))D\phi(x_0,t) = Df(x_0)D\phi(x_0,t).
\]

Solving this ODE gives \( Dg(x_0) = D\phi(x,1) = e^{Df(x_0)}D\phi(x,0) = e^{Df(x_0)} \), where the last equality is due to \( D\phi(x,0) \equiv I, \forall x \). Combining this with \( \lambda_1(Df(x_0)) > 0 \), we have \( \lambda_1(Dg(x_0)) > 1 \).

Thus we have \( L_t := \{ x_0 : f(x_0) = 0, \lambda_1(Df(x_0)) > 0 \} \subseteq A_g \), which implies that \( \{ x_0 : \lim_{t \to \infty} \phi(x_0,t) \in U^* \} \subseteq \{ x_0 : \lim_{k \to \infty} g^k(x_0) \in A_g \} \)

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

**Proof of Theorem 5.10.** For (1), by Theorem I.3, we immediately know all the stationary points of \( L(\cdot) \) are either global minimizers or strict saddles. (2) is just a direct consequence of Theorem I.5 by setting \( f \) in the above proof to \( -\nabla L \).

### J Equivalence Between GF and GLRL

In this section we elaborate on the theoretical evidence that GF and GLRL are equivalent generically, including the case where GLRL does not end in the first phase. The word “generically” used when we want to assume one of the following regularity conditions:

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 the top eigenvalue \( \lambda_1(\nabla f(W)) \) is unique for a critical point \( W \) of (2) that is not a minimizer of \( f(\cdot) \) in \( S_d^+ \);
3. We want to assume that a convergent sequence of PSD matrices \( W_\alpha \to \overline{W} \) has positive alignment with \( vv^\top \) for some fixed vector \( v \) with \( \langle W, vv^\top \rangle = 0 \), i.e., for a convergent sequence of PSD matrices \( W_\alpha \to \overline{W} \), it holds for sure that \( \liminf_{\alpha \to 0} \left( \begin{array}{c} W_\alpha - \overline{W}v\overline{v^\top} \\ \overline{W}v\overline{v^\top} \end{array} \right) \geq 0 \), and we further assume that the inequality is strict generically.

Theorem I.2 uncovers how GF with infinitesimal initialization generically behaves. Let \( \overline{W}_0 := 0 \). For every \( r \geq 1 \), if \( \overline{W}_{r-1} \) is a local minimizer in \( S_d^+ \), then \( \lambda_1(\nabla f(\overline{W}_{r-1})) > 0 \) by Lemma C.2. Generically, the top eigenvalue \( \lambda_1(\nabla f(\overline{W}_{r-1})) \) should be unique, i.e., \( \lambda_1(\nabla f(\overline{W}_{r-1})) > \lambda_2(\nabla f(\overline{W}_{r-1})) \). This enables us to apply Theorem I.2 and deduce that the limiting trajectory

\[
W_r^G(t) := \lim_{\epsilon \to 0} \phi \left( \overline{W}_{r-1} + \epsilon u_r, u_r^\top, \frac{1}{2\lambda_1(\nabla f(\overline{W}_{r-1}))} \log \frac{1}{\epsilon + t} \right)
\]

exists, where \( u_r \) is the top eigenvector of \( \nabla f(\overline{W}_{r-1}) \). This \( W^G_r(\cdot) \) is exactly the trajectory of GLRL in phase \( r \) as \( \epsilon \to 0 \).

Note that \( W^G_r(\cdot) \) corresponds to a trajectory of GF minimizing \( \mathcal{L}(\cdot) \) in \( \mathbb{R}^{d \times r} \), which should generically converge to a local minimizer of \( \mathcal{L}(\cdot) \) in \( \mathbb{R}^{d \times r} \). This means the limit \( W_r := \lim_{t \to +\infty} W^G_r(t) \) should generically be a local minimizer of \( f(\cdot) \) in \( S_d^+ \). If \( W_r \) is further a minimizer in \( S_d^+ \), then \( \lambda_1(\nabla f(W_r)) \leq 0 \) and GLRL exits with \( W_r \); otherwise GLRL enters phase \( r + 1 \).

If GF aligns well with GLRL in the beginning of phase \( r \) (defined below), then by Theorem I.2, as \( \alpha \to 0 \), the minimum distance from GF to \( W^G_r(t) \) converges to 0 for every \( t \in \mathbb{R} \). Therefore, GF can get arbitrarily close to the \( r \)-th critical point \( W_r \) of GLRL, i.e., there exists a suitable choice \( T^{(r)}_\alpha \) so that \( \lim_{\alpha \to 0} \phi(W_\alpha, T^{(r)}_\alpha) = W_r \). Note that \( \langle W_r, u_r, u_r^\top \rangle = 0 \) by (27) and thus
\[
\liminf_{\alpha \to 0} \left( \frac{\phi(W_\alpha, T^{(r)}_\alpha) - W_r}{\|W_\alpha, T^{(r)}_\alpha - W_r\|_F}, u_r, u_r^T \right) = \liminf_{\alpha \to 0} \left( \frac{\phi(W_\alpha, T^{(r)}_\alpha), u_r, u_r^T}{\|W_\alpha, T^{(r)}_\alpha - W_r\|_F} \right) \geq 0.
\]
Generically, there should exist a suitable choice of \( T^{(r)}_\alpha \) so that \( \phi(W_\alpha, T^{(r)}_\alpha) \) not only converges to \( W_r \) but also has positive alignment with \( u_r, u_r^T \), that is, GF should generically align well with GLRL in the beginning of phase \( r + 1 \).

**Definition J.1.** We say that GF aligns well with GLRL in the beginning of phase \( r \) if there exists \( T^{(r)}_\alpha \) for every \( \alpha > 0 \) such that \( \phi(W_\alpha, T^{(r)}_\alpha) \) converges to \( W_{r-1} \) with positive alignment with \( u_r, u_r^T \) as \( \alpha \to 0 \).

If the initialization satisfies that \( W_\alpha \) converges to 0 with positive alignment with \( u_1, u_1^T \) as \( \alpha \to 0 \), then GF aligns well with GLRL in the beginning of phase 1, which can be seen by taking \( T^{(1)}_\alpha = 0 \). Now assume that GF aligns well with GLRL in the beginning of phase \( r - 1 \), then the above argument shows that GF should generically align well with GLRL in the beginning of phase \( r \) if GLRL does not exit in phase \( r - 1 \). In the other case, we can use a similar argument as in Theorem 5.8 to show that GF converges to a solution near the minimizer \( W_r \) of \( f(\cdot) \) as \( t \to \infty \), and the distance between the solution and \( W_r \) converges to 0 as \( \alpha \to 0 \). By this induction we prove that GF with infinitesimal initialization is equivalent to GLRL generically.

## K Proofs for Deep Matrix Factorization

### K.1 Preliminary Lemmas

**Lemma K.1.** If \( W(0) \succeq 0 \), then \( W(t) \succeq 0 \) and \( \text{rank}(W(t)) = \text{rank}(W(0)) \) for all \( t \).

**Proof.** Note that we can always find a set of balanced \( U_i(t) \), such that \( U_1(t) \ldots U_d(t) = W(t) \), \( d_2 = d_3 = \cdots = d_k = \text{rank}(W(t)) \) and write the dynamics of \( W(t) \) in the space of \( \{U_i\}_{i=1}^k \). Thus it is clear that for all \( t' \), \( \text{rank}(W(t')) \leq \text{rank}(W(t)) \). We can apply the same argument for \( t' \) and we know \( \text{rank}(W(t')) \leq \text{rank}(W(t')) \). Thus \( \text{rank}(W(t)) \) is constant over time, and we denote it by \( k \). Since eigenvalues are continuous matrix functions, and \( \forall t_\alpha, \lambda_i(W(t)) \), \( i \in [k] \neq 0 \). Thus they cannot change their signs and it must hold that \( W(t) \succeq 0 \).

**Lemma K.2.** \( \forall a, b, P \in \mathbb{R}, \text{ if } a > b \geq 0, \text{ then } \frac{a^p - b^p}{a - b} \leq Pa^{p-1} \).

**Proof.** Let \( f(x) = P(1 - x) - (1 - x^P) \). Since \( f'(x) = -P + Px^{P-1} < 0 \) for all \( x \in [0, 1) \), \( f(x) \geq f(0) = 0 \). Then substituting \( x \) by \( \frac{b}{a} \) completes the proof.

Recall we use \( DF(N)[M] \) to denote the directional derivative along \( M \) of \( F \) at \( N \).

**Lemma K.3.** Let \( F : \mathbb{S}_d^+ \to \mathbb{S}_d^+, M \mapsto M^P \), where \( P \geq 1 \) and \( P \in \mathbb{Q} \). Then \( \forall M, N \succeq 0 \),

\[
\|DF(N)[M]\|_F \leq P \|N\|_F^{P-1} \|M\|_F,
\]

where \( DF(N)[M] := \lim_{t \to 0} \frac{F(N + tM) - F(N)}{t} \) is the directional derivative of \( F \) along \( M \).

**Proof.** Let \( N = USU^\top \), where \( UU^\top = I \) and \( \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_d) \). Note that \( F(UMU^\top) = UF(M)U^\top \) for any \( M \in \mathbb{S}_d^+ \). Then we have

\[
\|DF(N)[M]\|_F = \lim_{t \to 0} \frac{\|F(N + tM) - F(N)\|_F}{t} = \lim_{t \to 0} \frac{\|F(\Sigma + tU^\top MU) - F(\Sigma)\|_F}{t} = \|DF(\Sigma)[U^\top MU]\|_F.
\]
Therefore, it suffices to prove the lemma for the case where \( N \) is diagonal, i.e., \( N = \Sigma \).
Assume $P = \frac{q}{p}$, where $p, q \in \mathbb{N}$ and $q \geq p > 0$. Define $G(N) = N^{\frac{1}{p}}$. Then $G(\Sigma)^p = \Sigma$. Taking directional derivative on both sides along direction $\mathbf{M}$, we have
\[
\sum_{i=1}^{p} G(\Sigma)^{i-1}DG(\Sigma)[\mathbf{M}]G(\Sigma)^{p-1} = \mathbf{M},
\]
So we have
\[
|DG(\Sigma)[\mathbf{M}]|_{ij} = \frac{m_{ij}}{\sum_{k=1}^{p} \frac{k-1}{\sigma_i^{p-1}} \sigma_j^{\frac{q-k}{p}}},
\]
Let $H(G) = G^q$. With the same argument, we know
\[
|DH(G(\Sigma))[\mathbf{M}]|_{ij} = m_{ij} \sum_{k=1}^{q} \frac{k-1}{\sigma_i^{\frac{q-k}{p}}} \sigma_j^{\frac{q-k}{p}}.
\]
Note that $H(G(\Sigma)) = F(\Sigma)$. By chain rule, we have
\[
|DF(\Sigma)[\mathbf{M}] = DH(G(\Sigma))[DG(\Sigma)[\mathbf{M}]].
\]
That is,
\[
|DF(\Sigma)[\mathbf{M}]|_{ij} = m_{ij} \sum_{k=1}^{q} \frac{k-1}{\sigma_i^{\frac{q-k}{p}}} \sigma_j^{\frac{q-k}{p}}.
\]
When $\sigma_i = \sigma_j$, clearly $|DF(\Sigma)[\mathbf{M}]|_{ij} = m_{ij} \cdot \frac{q}{p} \cdot \sigma_i^{\frac{q-p}{p}} = Pm_{ij} \sigma_i^{p-1}$. Otherwise, we assume WLOG that $\sigma_i > \sigma_j$, we multiply $\sigma_i - \sigma_j$ to both numerator and denominator and we have
\[
|DF(\Sigma)[\mathbf{M}]|_{ij} = m_{ij} \frac{\sigma_i^{p} - \sigma_j^{p}}{\sigma_i - \sigma_j} \leq |m_{ij}| P \sigma_i^{p-1} \leq |m_{ij}| \|\Sigma\|_2^{p-1}.
\]
where the first inequality is by Lemma K.2. Thus we conclude the proof.

Lemma K.4. For any $\mathbf{A}, \mathbf{B} \geq 0$ and $P \in \mathbb{R}, P \geq 1$,
\[
\|\mathbf{A}^P - \mathbf{B}^P\|_F \leq P \|\mathbf{A} - \mathbf{B}\|_F \max \left\{ \|\mathbf{A}\|_2^{p-1}, \|\mathbf{B}\|_2^{p-1} \right\}.
\]
Proof. Since both sides are continuous in $P$ and $\mathbb{Q}$ is dense in $\mathbb{R}$, it suffices to prove the lemma for $P \in \mathbb{Q}$. Let $\rho := \max \{\|\mathbf{A}\|_2, \|\mathbf{B}\|_2\}$ and $F(\mathbf{M}) = \mathbf{M}^P$. Define $N : [0, 1] \to \mathbb{S}^+_d$, $N(t) = (1 - t)\mathbf{A} + t\mathbf{B}$, we have
1. $\|N(t)\|_2 \leq \rho$, since $\|\cdot\|_2$ is convex.
2. $\|DF(N(t))[\mathbf{B} - \mathbf{A}]\|_F \leq P \|N(t)\|_2^{p-1} ||\mathbf{B} - \mathbf{A}\|_F$ by Lemma K.3.
Therefore,
\[
\|F(N(1)) - F(N(0))\|_F \leq \int_0^1 \|DF(N(t))\|_F dt \leq \int_0^1 ||DF(N(t))[\mathbf{B} - \mathbf{A}]\|_F dt \leq P \|\mathbf{A} - \mathbf{B}\|_F \rho^{p-1},
\]
which completes the proof.

For a locally Lipschitz function $f(\cdot)$, the Clarke subdifferential (Clarke, 1975; 1990; Clarke et al., 2008) of $f$ at any point $x$ is the following convex set
\[
\frac{\partial^c f(x)}{\partial \mathbf{x}} := \text{co}\left\{ \lim_{k \to \infty} \nabla f(x_k) : x_k \to x, f \text{ is differentiable at } x_k \right\},
\]
where co denotes the convex hull.

Clarke subdifferential generalize the standard notion of gradients in the sense that, when $f$ is smooth, $\frac{\partial f(x)}{\partial \mathbf{x}} = \{\nabla f(x)\}$. Clarke subdifferential satisfies the chain rule:

Published as a conference paper at ICLR 2021
Theorem K.5 (Theorem 2.3.10, Clarke 1990). Let $F : \mathbb{R}^k \to \mathbb{R}^d$ be a differentiable function and $g : \mathbb{R}^d \to \mathbb{R}$ Lipschitz around $F(x)$. Then $f = g \circ F$ is Lipschitz around $x$ and one has
\[
\frac{\partial^p f(x)}{\partial x} \subseteq \frac{\partial^p g(F(x))}{\partial F} \circ \frac{dF(x)}{dx}.
\]
Let $\lambda_m : S_d \to \mathbb{R}, M \mapsto \lambda_m(M)$ be the $m$-th largest eigenvalue of a symmetric matrix $M$. The following theorem gives the Clarke’s subdifferentials of the eigenvalue:

Theorem K.6 (Theorem 5.3, Hiriart-Urruty and Lewis 1999). The Clarke subdifferential of the eigenvalue function $\lambda_m$ is given below, where co denotes the convex hull:
\[
\frac{\partial^p \lambda_m(M)}{\partial M} = \text{co}\{v v^\top : Mv = \lambda_m(M)v, \|v\|_2 = 1\}.
\]

K.2 Proof of Lemma 6.1

The equation to be proved is:
\[
\frac{dM}{dt} = -\nabla f(M^{L/2})M^{L/2} - M^{L/2}\nabla f(M^{L/2}). \tag{29}
\]
Since $W(t) \succeq 0$ by Lemma K.1, (11) can be rewritten as the following:
\[
\frac{dW}{dt} = -\sum_{i=0}^{L-1} W^{\frac{2i}{2}} \nabla f(W)W^{2-2\frac{i+2}{2}}. \tag{30}
\]

Proof for Lemma 6.1. Suppose $W(t)$ is a symmetric solution of (11). By Lemma K.1, we know $W(t)$ also satisfies (30). Now we let $R(t)$ be the solution of the following ODE with $R(0) := (W(0))^\frac{1}{2}$. Note we don’t define $R(t)$ by $(W(t))^\frac{1}{2}$.

\[
\frac{dR}{dt} = -\sum_{i=0}^{L-1} (-1)^i R^i \nabla f(R^L)R^{L-1-i}. \tag{31}
\]

The calculation below shows that $R^L(t)$ also satisfies (30).
\[
\frac{dR^L}{dt} = \sum_{j=0}^{L-1} R^j \frac{dR}{dt} R^{L-1-j} = \sum_{j=0}^{L-1} \sum_{i=0}^{L-1} (-1)^i R^i R^j \nabla f(R^L)R^{2L-2-i-j} = \sum_{i=0}^{2L-2} \left( \sum_{j=0}^{i} (-1)^j \right) R^i \nabla f(R^L)R^{2L-2-i} = \sum_{i=0}^{L-1} (R^L)^{\frac{2i}{2}} \nabla f(R^L)^{2-2\frac{i+2}{2}}.
\]

Since $R^L(0) = W(0)$, by existence and uniqueness theorem, $R^L(t) = W(t), \forall t \in \mathbb{R}$. So
\[
\frac{dM}{dt} = R \frac{dR}{dt} + \frac{dR}{dt} R = -\nabla f(M^{L/2})M^{L/2} - M^{L/2}\nabla f(M^{L/2}),
\]
which completes the proof. \hfill \Box

K.3 Proof for Theorem 6.2

Now we turn to prove Theorem 6.2. Let $P = L/2$. Then (29) can be rewritten as
\[
\frac{dM}{dt} = - (\nabla f(M^P)) M^P + M^P \nabla f(M^P). \tag{32}
\]
The following lemma about the growth rate of $\lambda_k(M)$ is used later in the proof.
Lemma K.7. Suppose $M(t)$ satisfies (32), we have for any $T' > T$, and $k \in [d]$,

$$
\lambda_k(M(T')) - \lambda_k(M(T)) \leq \int_T^{T'} 2\lambda_k(M(t))^P \|\nabla f(M^P(t))\|_2 dt. 
$$

(33)

and

$$
\frac{1}{P-1} \left( \lambda_k^1 - \lambda_k^P(M(T')) - \lambda_k^1 - \lambda_k^P(M(T)) \right) \leq \int_T^{T'} 2\|\nabla f(M^P(t))\|_2 dt. 
$$

(34)

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

$$
\lambda_k(M(T')) - \lambda_k(M(T)) = \int_T^{T'} \frac{d\lambda_k(M(t))}{dt} dt.
$$

When $\frac{d\lambda_k(M(t))}{dt}$ exists, we have

$$
\frac{d\lambda_k(M(t))}{dt} \in \left\{ \left< G, \frac{dM(t)}{dt} \right> : G \in \partial^2 \lambda_k(M) \right\}
$$

$$
= \left\{ 2\lambda_k(M^P(t)) \left< G, -\nabla f(M^P(t)) \right> : G \in \partial^2 \lambda_k(M) \right\}
$$

Note that $\|G\|^F \leq \|G\|^1 = 1$. So $\left< G, -\nabla f(M^P(t)) \right> \leq \|\nabla f(M^P(t))\|_2$. We can prove (34) with a similar argument.

To prove Theorem 6.2, it suffices to consider the case that $M(0) = \hat{\alpha}I$ where $\hat{\alpha} := \alpha^{1/P}$. WLOG we can assume $-\nabla f(0) = \text{diag}(\mu_1, \ldots, \mu_d)$ by choosing a suitable standard basis. By assumption in Theorem 6.2, we have $\mu_1 = \max\{\mu_2, 0\}$ and $\mu_1 = \|\nabla f(0)\|_2$. We use $\phi_m(M_0, t)$ to denote the solution of $M(t)$ when $M(0) = M_0$.

Let $R > 0$. Since $f(\cdot)$ is $C^3$-smooth, there exists $\beta > 0$ such that

$$
\|\nabla f(W_1) - \nabla f(W_2)\|_F \leq \beta \|W_1 - W_2\|_2
$$

for all $W_1, W_2$ with $\|W_1\|_2, \|W_2\|_2 \leq R$.

Let $\kappa = \beta/\mu_1$. We assume WLOG that $R \leq \frac{1}{\kappa(P-1)}$. Let $F_{\hat{\alpha}}(x) := \int_{x-(P-1)}^{x} \frac{dz}{1 + \kappa z^{P/(P-1)}}$.

Then $F'_{\hat{\alpha}}(x) = \frac{(P-1)x^{-P}}{1 + \kappa x^{P-1}} = \frac{P-1}{1 + \kappa x^{P-1}}x^P$. We will use this function to bound norm growth. Let $g_{\hat{\alpha}, c}(t) = \frac{\hat{\alpha}^{-(P-1)} - \kappa^{-(P-1)} - \mu_1}{2\mu_1(P-1)}$. Define $T_{\hat{\alpha}}(r) = \frac{\hat{\alpha}^{-(P-1)} - \kappa^{-(P-1)} - \mu_1}{2\mu_1(P-1)}$. It is easy to verify that $g_{\hat{\alpha}, c}(T_{\hat{\alpha}}(r)) = r^{P-1}$.

Lemma K.8. 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,

$$
\hat{\alpha}^{-(P-1)} - x^{-(P-1)} - F_{\hat{\alpha}}(x) = \int_{x-(P-1)}^{\hat{\alpha}^{-(P-1)}} \frac{\kappa}{z^{P/(P-1)} + \kappa} dz \leq \kappa \int_{x-(P-1)}^{\hat{\alpha}^{-(P-1)}} \frac{1}{z^{P/(P-1)}} dz
$$

$$
= \kappa(P-1) \cdot \frac{-1}{\frac{1}{z^{1/(P-1)}}} \bigg|_{x-(P-1)}^{\hat{\alpha}^{-(P-1)}}
$$

$$
\leq \kappa(P-1)x,
$$

which completes the proof.
Lemma K.9. Let $M_0$ be a PSD matrix with $\|M_0\|_2 \leq 1$. For $M(t) := \phi_m(\hat{\alpha} M_0, t)$ and $t \leq T_{\hat{\alpha}}(c)$,
$$\|M(t)\|_2 = \lambda_1(M(t)) \leq g_{\hat{\alpha},c}(t) \frac{1}{\tau^t}.$$

Proof. Since $\|\nabla f(M^P)\|_2 \leq \|\nabla f(0)\|_2 + \beta \|M\|_2^P \leq \mu_1 + \beta(\lambda_1(M))^P$, by Lemma K.7, we have
$$\lambda_1(M(t)) \leq \lambda_1(M(0)) + \int_0^t 2(\mu_1 + \beta(\lambda_1(M(\tau)))^P)(\lambda_1(M(\tau)))^P d\tau$$
$$= \hat{\alpha} + 2\mu_1(P-1) \int_0^t \frac{d\tau}{F_\hat{\alpha}(\lambda_1(M(\tau)))}.$$

So
$$F_\hat{\alpha}(\lambda_1(M(t))) \leq 2\mu_1(P-1)t.$$

If $\|M(t)\|_2 < \hat{\alpha}$, then $\|M(t)\|_2 \leq g_{\hat{\alpha},c}(t) \frac{1}{\tau^t}$. If $\|M(t)\|_2 \geq \hat{\alpha}$, then by Lemma K.8,
$$F_\hat{\alpha}(\|M(t)\|_2) \leq 2\mu_1(P-1)T_{\hat{\alpha}}(c) = \hat{\alpha}^{-(P-1)} - \kappa(P-1)c - c^{-(P-1)} \leq F_\hat{\alpha}(c),$$
so $\|M(t)\|_2 \leq c$ for all $t \leq T_{\hat{\alpha}}(c)$. Applying Lemma K.8 again, we have
$$\hat{\alpha}^{-(P-1)} - \|M(t)\|_2^{-(P-1)} \leq F(\|M(t)\|_2) + \kappa(P-1)c \leq 2\mu_1(P-1)t + \kappa(P-1)c,$$
which implies $\|M(t)\|_2 \leq g_{\hat{\alpha},c}(t) \frac{1}{\tau^t}$ by definition.

Consider the following ODE:
$$\frac{d \hat{M}}{dt} = -\left(\nabla f(0)\hat{M}^P + \hat{M}^P \nabla f(0)\right).$$

We use $\hat{\phi}_m(M_0, t)$ to denote the solution of $\hat{M}(t)$ when $\hat{M}(0) = \hat{M}_0$. For diagonal matrix $\hat{M}_0$, $\hat{M}(t)$ is also diagonal for any $t$, and it is easy to show that
$$e_i^\top \hat{M}(t)e_i = \begin{cases} \frac{1}{(\hat{\alpha}e_i^\top \hat{M}_0e_i)^{-(P-1)} - 2\mu_1(P-1)t} \frac{1}{\tau^t} & e_i^\top \hat{M}_0e_i \neq 0, \\
0 & e_i^\top \hat{M}_0e_i = 0. \end{cases} \quad (35)$$

Remark K.10. Unlike depth-2 case, the closed form solution, $\hat{M}(t)$ is only tractable for diagonal initialization, i.e., (35) (note that the identity matrix is diagonal). And this is the main barrier for extending our two-phase analysis to the case of general initialization when $L \geq 3$. In Appendix L, we give a more detailed discussion on this barrier.

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

Lemma K.11. Let $M_0$ be a diagonal PSD matrix with $\|M_0\|_2 \leq 1$. For $M(t) := \phi_m(\hat{\alpha} M_0, t)$ and $\hat{M}(t) := \hat{\phi}_m(\hat{\alpha} \hat{M}_0, t)$, we have
$$\|M(T_{\hat{\alpha}}(r)) - \hat{M}(T_{\hat{\alpha}}(r))\|_F = O(r^{P+1}).$$

Proof. We bound the difference $D := M - \hat{M}$ between $M$ and $\hat{M}$.
$$\left\|\frac{dD}{dr}\right\|_F^2 = 2 \left\|\nabla f(0) \left(M^P - \hat{M}^P\right) + (\nabla f(M^P) - \nabla f(0)) M^P \right\|_F^2$$
$$\leq 2 \left(\|\nabla f(0)\|_2 \|M^P - \hat{M}^P\|_F + \|\nabla f(M^P) - \nabla f(0)\|_F \|M^P\|_2\right)^2$$
$$\leq 2 \left(\mu_1 P \max\{\|M\|_2^{P-1}, \|\hat{M}\|_2^{P-1}\}\|D\|_F + \beta \|M\|_2^{2P}\right),$$
where the last step is by Lemma K.4. This implies that
$$\|D(t)\|_F \leq \int_{\tau=0}^t \left\|\frac{dD(\tau)}{d\tau}\right\|_F^2 d\tau \leq \int_0^t 2 \left(\mu_1 P g_{\hat{\alpha},r}(\tau) \|D(\tau)\|_F + \beta g_{\hat{\alpha},r}(\tau)^{2P}\right) d\tau.$$
So
\[
\|D(T_{\alpha}(r))\|_F \leq \int_0^{T_{\alpha}(r)} 2\beta g_{\alpha,r}(t) \frac{\partial_p}{\partial t} \exp \left( 2\mu_1 P \int_t^{T_{\alpha}(r)} g_{\alpha,r}(\tau)d\tau \right) dt \\
= \int_0^{T_{\alpha}(r)} 2\beta g_{\alpha,r}(t) \frac{\partial_p}{\partial t} \exp \left( \frac{P}{P-1} \ln \frac{g_{\alpha,r}(T_{\alpha}(r))}{g_{\alpha,r}(t)} \right) dt \\
= \int_0^{T_{\alpha}(r)} 2\beta g_{\alpha,r}(t) \frac{\partial_p}{\partial t} g_{\alpha,r}(T_{\alpha}(r)) \frac{\partial_p}{\partial t} dt \\
= 2\beta \cdot \frac{1}{2\mu_1} g_{\alpha,r}(T_{\alpha}(r)) \frac{\partial_p}{\partial t} \cdot g_{\alpha,r}(T_{\alpha}(r)) \frac{\partial_p}{\partial t} \\
= \kappa g_{\alpha,r}(T_{\alpha}(r)) \frac{\partial_p}{\partial t} \\
= \kappa r^{P+1},
\]
which proves the bound.

\[\square\]

**Lemma K.12.** Let \(M(t) = \phi_m(\tilde{\alpha}M_0,t),\tilde{M}(t) = \phi_m(\tilde{\alpha}\tilde{M}_0,t).\) If \(\max\{\|M_0\|_2,\|\tilde{M}_0\|_2\} \leq 1.\) For \(t \leq T_{\tilde{\alpha}}(r),\) we have
\[
\|M(t) - \tilde{M}(t)\|_F \leq \left(\frac{r}{\alpha}\right)^P e^{2\kappa r^P} \|M(0) - \tilde{M}(0)\|_F.
\]

**Proof.** Define \(D(t) = M(t) - \tilde{M}(t).\) Then we have
\[
\left\| \frac{dD}{dt} \right\|_F = 2 \left\| \left( \nabla f(M^P) (M^P - \tilde{M}^P) + \left( \nabla f(M^P) - \nabla f(\tilde{M}^P) \right) \tilde{M}^P \right) \right\|_F \\
\leq 2 \left( \left\| \nabla f(M^P) \right\|_2 \left\| M^P - \tilde{M}^P \right\|_F + \beta \left\| M^P - \tilde{M}^P \right\|_F \left\| \tilde{M}^P \right\|_2 \right) \left\| \tilde{M}^P \right\|_2 \\
\leq 2 \left( \mu_1 + \beta \left\| \tilde{M} \right\|_2 + \beta \left\| M \right\|_2 \right)^P P \max\{\left\| M \right\|_2^{-1}, \left\| \tilde{M} \right\|_2^{-1}\} \left\| D \right\|_F,
\]
where the last step is by Lemma K.4. So
\[
\|D(T_{\alpha}(r))\|_F \leq \|D(0)\|_F \cdot \exp \left( 2P \mu_1 \int_0^{T_{\alpha}(r)} \left( 1 + 2\kappa g_{\alpha,r}(t) \frac{\partial_p}{\partial t} \right) g_{\alpha,r}(t) dt \right) \\
\leq \|D(0)\|_F \cdot \exp \left( \frac{P}{P-1} \ln \frac{g_{\alpha,r}(T_{\alpha}(r))}{g_{\alpha,r}(0)} + 2\kappa g_{\alpha,r}(T_{\alpha}(r)) \frac{\partial_p}{\partial t} \right) \\
\leq \|D(0)\|_F \left( \frac{r}{\alpha} \right)^P e^{2\kappa r^P},
\]
which proves the bound.

\[\square\]

Let \(M^G_{\alpha}(t) := \phi_m \left( \alpha e_1 e_1^T, \frac{\hat{\alpha}(-P-1)}{2\mu_1 (P-1)} + t \right).\) Let \(\tilde{M}(t) := \lim_{\hat{\alpha} \to 0} M^G_{\hat{\alpha}}(t).\)

**Lemma K.13.** For every \(t \in (-\infty, +\infty),\) \(\tilde{M}(t)\) exists and \(M^G_{\alpha}(t)\) converges to \(\tilde{M}(t)\) in the following rate:
\[
\|M^G_{\alpha}(t) - \tilde{M}(t)\|_F = O(\hat{\alpha}).
\]

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

**Case 1.** Fix \(t \in (-\infty, \hat{T}].\) Then \(\frac{\hat{\alpha}(-P-1)}{2\mu_1 (P-1)} + t \leq T_{\hat{\alpha}}(c).\) Let \(\hat{\alpha}\) be the unique number such that \(\kappa(P-1)\hat{\alpha} + \hat{\alpha}(-P-1) = \hat{\alpha}(-P-1).\) Let \(\hat{\alpha}' < \hat{\alpha}\) be an arbitrarily small number. Let \(t_0 := T_{\hat{\alpha}'}(\hat{\alpha}) = \frac{\hat{\alpha}'(-P-1) - \hat{\alpha}(-P-1)}{2\mu_1 (P-1)}.\) By Lemma K.11 and (35), we have
\[
\|\phi_m(\hat{\alpha}' e_1 e_1^T, t_0) - \hat{\alpha} e_1 e_1^T\|_F \leq \|\phi_m(\hat{\alpha}' e_1 e_1^T, t_0) - \phi_m(\hat{\alpha} e_1 e_1^T, t_0)\|_F \leq O(\hat{\alpha}'^{P+1}).
\]
By Lemma K.9, $\|\phi_m(\hat{\alpha}'e_1e_1^T, t_0)\|_2 \leq \hat{\alpha}$. Then by Lemma K.12, we have

$$\|\phi_m(\hat{\alpha}'e_1e_1^T, t_0) - \phi(\hat{\alpha}'e_1e_1^T, t)\|_F \leq \left( \frac{C}{\hat{\alpha}} \right)^P e^{2\kappa c^p} \cdot O(\hat{\alpha}^{P+1}) = O(\hat{\alpha}) = O(\hat{\alpha}).$$

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

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

$$\|M^G_G(t) - M^G_G(t)\|_F = \|\phi_m(M^G_G(T), \tau) - \phi_m(M^G_G(T), \tau)\|_F$$

$$= O(\|M^G_G(T) - M^G_G(T)\|_F)$$

$$= O(\hat{\alpha}),$$

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

**Theorem K.14.** For every $t \in (-\infty, +\infty)$, as $\alpha \to 0$, we have:

$$\left\|\phi_m \left(\hat{\alpha}I, \frac{\hat{\alpha}^{-1}(P-1)}{2\mu_1(P-1)} + t \right) - \overline{M}(t)\right\|_F = O(\hat{\alpha}^{-\frac{1}{\tau+\tau}}), \quad (36)$$

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

$$\lambda_k \left(\phi_m \left(\hat{\alpha}I, \frac{\hat{\alpha}^{-1}(P-1)}{2\mu_1(P-1)} + t \right)\right) = O(\hat{\alpha}). \quad (37)$$

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

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

$$\left\|\phi_m \left(\hat{\alpha}I, t_0 \right) - \hat{\alpha}_1 e_1e_1^T\right\|_F \leq \left\|\phi_m \left(\hat{\alpha}I, t_0 \right) - \phi_m \left(\hat{\alpha}I, t_0 \right)\right\|_F + \left\|\phi_m \left(\hat{\alpha}I, t_0 \right) - \phi_m \left(\hat{\alpha}I, t_0 \right)\right\|_F$$

$$= O(\hat{\alpha}_1^{P+1} + \hat{\alpha})$$

$$= O(\hat{\alpha}).$$

By Lemma K.9, $\|\phi_m(\hat{\alpha}'I, t_0)\|_2 \leq \hat{\alpha}_1$. Then by Lemma K.12, we have

$$\left\|M_{\hat{\alpha}}(t) - M^G_{\hat{\alpha}}(t)\right\|_F \leq \left\|\phi_m(\hat{\alpha}I, t_0 + t) - \phi_m(\hat{\alpha}_1 e_1 e_1^T, t)\right\|_F$$

$$\leq \left( \frac{c}{\hat{\alpha}_1} \right)^P e^{2\kappa c^p} \cdot O(\hat{\alpha}) = O(\hat{\alpha}^{\frac{1}{\tau+\tau}}).$$

Combining this with the convergence rate for $M^G_{\hat{\alpha}}(t)$ proves the bound (36).

For (37), by Lemma K.7, we have

$$\lambda_k^{-P}(M_{\hat{\alpha}}(T)) - \lambda_k^{L-P}(M_{\hat{\alpha}}(t_0)) \leq \int_{t_0}^{\overline{T}} 2(P-1) \left\|\nabla f(M_{\hat{\alpha}}(t))\right\|_2 dt$$

$$\leq \int_{t_0}^{\overline{T}} 2(P-1)(\mu_1 + \beta \|M_{\hat{\alpha}}(t)\|_F^p) dt$$

$$\leq -2(P-1) \left( \mu_1(t - T_1) + \frac{\kappa}{2} g_{\hat{\alpha}, c}(t)^{\frac{1}{\tau+\tau}} \right). \quad (38)$$

By Lemma K.11, $\lambda_1(M_{\hat{\alpha}}(T)) = \|M_{\hat{\alpha}}(T)\|_2 = c + O(c^{P+1})$. For $k \geq 2$, $\lambda_k(M_{\hat{\alpha}}(T))^{-1} \geq \Omega(\hat{\alpha}^{-1}(P-1)) - 2(P-1) \left( \mu_1(t - T_1) + \frac{\kappa}{2} c \right)$

$$\geq \Omega(\hat{\alpha}^{-1}(P-1)) - \frac{\hat{\alpha}^{-\frac{1}{\tau+\tau}} - c^{-1}(P-1)}{2\mu_1(P-1)} - O(c)$$

$$\geq \Omega(\hat{\alpha}^{-1}(P-1)).$$
Thus $\lambda_k(M_\alpha(\bar{T})) \leq O(\hat{\alpha})$.

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

$$\|M_\alpha(t) - M^G_{\alpha_1}(t)\|_F = \|\phi_m(M_\alpha(\bar{T}), \tau) - \phi_m(M^G_{\alpha_1}(\bar{T}), \tau)\|_F = O(\|M_\alpha(\bar{T}) - M^G_{\alpha_1}(\bar{T})\|_F) = O(\hat{\alpha}^{\frac{1}{\tau+\tau}}),$$

which proves the bound (36).

For (37), again by Lemma K.7, we have

$$\lambda_k^{1-P}(M_\alpha(\bar{T})) - \lambda_k^{1-P}(M_\alpha(\bar{T} + \tau)) \leq \int_{\bar{T}}^{\bar{T} + \tau} 2(P-1) \|\nabla f(M^P_\alpha(t))\|_2 \, dt \leq \int_{\bar{T}}^{\bar{T} + \tau} 2(P-1)(\beta \|M^P_\alpha(t) - (M^G)^P(t)\|_2 + \|\nabla f((M^G)^P(t))\|_2) \, dt \leq \int_{\bar{T}}^{\bar{T} + \tau} 2(P-1)(O(\hat{\alpha}^{\frac{1}{\tau+\tau}}) + \beta \|M^G(t)\|_2^P) \, dt \leq O(1).$$

Thus $\lambda_k^{1-P}(M_\alpha(\bar{T} + \tau)) = O(\hat{\alpha}^{-(P-1)})$, that is, $\lambda_k(M_\alpha(\bar{T} + \tau)) = O(\hat{\alpha})$, $\forall k \geq 2$.

**Proof of Theorem 6.2.** Note that $(\mathbf{M}(t))^P = \mathbf{W}(t)$ and

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

By Theorem K.14, We have

$$\left\|\phi(\alpha I, \frac{\alpha^{-(1-1/P)} I}{2\mu_1(P - 1)} + t) - \mathbf{W}(t)\right\|_F \leq \left\|\phi_m(\hat{\alpha} I, \frac{\hat{\alpha}^{-(P-1)} I}{2\mu_1(P - 1)} + t)^P - (\mathbf{M}(t))^P\right\|_F \leq P \left\|\phi_m(\hat{\alpha} I, \frac{\hat{\alpha}^{-(P-1)} I}{2\mu_1(P - 1)} + t) - \mathbf{M}(t)\right\|_F \max\left(\left\|\phi_m(\hat{\alpha} I, \frac{\hat{\alpha}^{-(P-1)} I}{2\mu_1(P - 1)} + t)\right\|_2, \|\mathbf{M}(t)\|_2\right)^{P-1}.$$

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

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

**L Escaping direction for deep matrix factorization**

For deep matrix factorization, recall that we only prove that GF with infinitesimal identity initialization escapes in the direction of the top eigenvector. 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 (29), i.e., the analytical solution of (39) is difficult to compute, when $L \geq 3$. Intuitively, the direction that the infinitesimal initialization escapes $0$ is exactly $\mathbf{M} := \lim_{t \to \infty} \frac{\mathbf{M}(t)}{\|\mathbf{M}(t)\|_F}$, where $\mathbf{M}(t)$ is the solution of (39). Showing $\mathbf{M} = v_1 v_1^T$ is a critical step in our analysis towards convergence to GLRL.

$$\frac{d\mathbf{M}}{dt} = -\nabla f(0) \mathbf{M}^{L/2} - \mathbf{M}^{L/2} \nabla f(0).$$  (39)
However, unlike the depth-2 case, \( \overline{M} \) can be different from \( v_1 v_1^T \) even if \( v_1^T M(0) v_1 > 0 \). We here give an example for diagonal \( M(0) \) and \( \nabla f(0) \) at Appendix L.2. Nevertheless, we still conjecture that except for a zero measure set of \( M(0) \), \( \overline{M} = v_1 v_1^T \), based on the following theoretical and experimental evidences:

- If \( v_1^T M(0) v_1 > 0 \) and \( \text{rank}(M(0)) = 1 \), we prove that \( \overline{M} = v_1 v_1^T \). (See Theorem L.1)
- For the counter-example, we show experimentally, even with perturbation of only magnitude \( 10^{-5} \), \( \overline{M} = v_1 v_1^T \). The results are shown at Figure 7. The y-axis indicates \( \langle v_1, u_1(t) \rangle \) where \( u_1(t) \) is the top eigenvector of \( M(t) \). As \( \| W(t) \|_F \) becomes larger, \( u_1(t) \) aligns better with \( v_1 \), which means the noise helps \( M \) escaping from \( v_1 \). The larger the noise is, the faster \( u_1(t) \) converges to \( v_1 \).

### L.1 Rank-One Case

**Theorem L.1** (rank-1 initialization escapes along the top eigenvector). *When \( \text{rank}(M(0)) = 1 \), 
\[
\lim_{t \to \infty} \frac{\|M(t)\|_F}{\|M(t)\|_F} = v_1 v_1^T, \quad \text{if } v_1^T M(0) v_1 > 0.
\]

**Proof.** Let \( u(0) \) be the vector such that \( M(0) = u(0)u(0)^T \) and \( u(t) \in \mathbb{R}^d \) be the solution of
\[
\frac{du(t)}{dt} = \|u(t)\|_2^{L-2} \nabla f(0) u(t).
\]
It is easy to check that \( M(t) = u(t)u(t)^T \) is the solution of (39), because
\[
\frac{dM}{dt} = \frac{du}{dt} u^T + u \frac{du}{dt}^T = -\nabla f(0) M(t) \|u(t)\|_2^{L-2} - M(t) \nabla f(0) \|u(t)\|_2^{L-2} = -\nabla f(0) M^{L/2} - M^{L/2} \nabla f(0).
\]
Let \( \tau(t) = \int_0^t \|u(s)\|_2^{L-2} ds \). Then
\[
\frac{du}{d\tau} = \frac{du}{dt} \frac{dt}{d\tau} = -\frac{1}{\frac{du}{dt}} \|u\|_2^{L-2} \nabla f(0) u = -\nabla f(0) u.
\]
That is, under time rescaling \( t \to \tau(t) \), the trajectory of \( u(t) \) still follows the power iteration, regardless of the depth \( L \). \( \square \)

### L.2 Counter-Example for Escaping Direction

Let \( \nabla f(0) = \text{diag}(2, 0.9, 0.8, \ldots, 0.1) \in \mathbb{R}^{10 \times 10} \) be diagonal. Let \( W(0) \) be also diagonal and \( W(0)_{i,i} \sim \text{Unif}[0.9, 1.1] \cdot \alpha \) for \( i \in [10] \setminus \{2\} \), \( W(0)_{2,2} = 16\alpha \), where \( \alpha = 10^{-16} \) is a small constant. Let the depth be 4.

**Lemma L.2.** With \( \nabla f(0) \) and \( W(0) \) constructed above, \( v_1 M(0) v_1^T \) > 0 and \( \overline{M} \neq v_1 v_1^T \).

**Proof.** It is easy to check that \( v_1 = e_1 \), so \( v_1 M(0) v_1^T \) > 0. Now we prove that \( \overline{M}(\infty) \neq v_1 v_1^T \).
As both \( W(0) \) and \( \nabla f(0) \) are diagonal, \( W(t) \) is always diagonal and has dynamics
\[
\frac{dM(t)}{dt} = -2\nabla f(0)_{i,i} M(t)_{i,i} = -2\nabla f(0)_{i,i} M(t)_{i,i} \quad \forall i \in [10],
\]
therefore we have closed form of \( M(t) \):
\[
M(t)_{i,i}^- = M(0)_{i,i}^- - 2\nabla f(0)_{i,i} t, \quad \forall i \in [10].
\]
For \( i \in [10] \), the time for \( M(t)_{i,i} \) going to infinity is \( (2M(0)_{i,i} \nabla f(0)_{i,i})^{-1} \). By simple calculation, \( M(t)_{2,2} \) goes to infinity the fastest, thus \( \overline{M} = e_2 e_2^T \neq v_1 v_1^T \). \( \square \)

We remark that the scales of \( W(0) \) and \( \nabla f(0) \) do not matter as in gradient flow, as scaling \( \nabla f(0) \) is equivalent to scaling time (by Lemma L.3 below). And for this reason, the x-axis is the chosen as \( \frac{\|W(t)\|_F}{\|W(0)\|_F} \), the relative growth rate.
Towards showing the main convergence result in the section, we make the following assumption.

Lemma L.3. Suppose \( g : \mathbb{R}^d \to \mathbb{R}^d \) is a \( P \)-homogeneous function, that is, \( g(\alpha \theta) = \lambda^P g(\alpha \theta) \) for any \( \alpha > 0 \), and \( \frac{d\theta(t)}{dt} = g(\theta'(t)) \). Then \( \alpha \theta'(\alpha^{P-1} t) \) is the solution of

\[
\frac{d\theta(t)}{dt} = g(\theta(t)), \quad \theta(0) = \alpha \theta'(0).
\]

Proof. Simply plug in \( \theta'(t) = \alpha \theta'(\alpha^{P-1} t) \), then we have

\[
\frac{d\theta(t)}{dt} = \frac{d\alpha \theta'(\alpha^{P-1} t)}{dt} = \alpha^P \frac{d\theta'(\alpha^{P-1} t)}{d\alpha^{P-1} t} = \alpha^P g(\theta'(\alpha^{P-1} t)) = g(\alpha \theta'(\alpha^{P-1} t)) = g(\theta'(t)).
\]

\( \square \)

## M PROOF OF LINEAR CONVERGENCE TO MINIMIZER

In this section, we will present the theorems that guarantee the linear convergence to a minimizer \( W_0 \) of \( f(\cdot) \) if the dynamics (41) is initialized sufficiently close to \( W_0 \), i.e., \( \| W(0) - W_0 \|_F \) is sufficiently small. In Appendix M.3, we will apply this result to prove Theorem 6.4.

\[
\frac{dW}{dt} = -\sum_{i=0}^{L-1} W_i^\top \nabla f(W_i) W_i^2 \sum_{i=0}^{2} =: g(W).
\]

Throughout this section, we assume \( \text{rank}(W_0) = k \) and use \( m := \lambda_k(W_0) \) to denote the \( k \)-th smallest non-zero eigenvalue of \( W_0 \). The tangent space of manifold of rank-\( k \) symmetric matrices at \( W_0 \) is \( \mathcal{T} = \{ V W_0^\top + W_0 V^\top : V \in \mathbb{R}^{d \times k} \} \). It can be shown that \( \dim(\mathcal{T}) = k(d-k) + \frac{k(k+1)}{2} = k^2 - \frac{k(k+1)}{2} \).

Let \( J(W) \) be the Jacobian of \( g(W) \) in (41). For depth-2 case, we have shown that \( \mathcal{T} \) is an invariant subspace of \( J(W_0) \) in Theorem H.5, property 2. This can be generalize to the deep case where \( L \geq 3 \). Therefore, we can use \( J(W_0)|_{\mathcal{T}} : \mathcal{T} \to \mathcal{T} \) to denote the linear operator \( J(W_0) \) restricted on \( \mathcal{T} \).

We also define \( \Pi^L(W) \) as the projection of \( W \in \mathbb{R}^{d \times d} \) on \( \mathcal{T} \), and \( \Pi^L_2(W) := W - \Pi^L_2(W) \).

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

**Assumption M.1.** Suppose \( J(W_0)|_{\mathcal{T}} \) diagonalizable and all eigenvalues are negative real numbers.

\( W_0 \) is a minimizer, so it is clear that \( J(W_0)|_{\mathcal{T}} \) has no eigenvalues with positive real parts (otherwise there is a descending direction of \( f(\cdot) \) from \( W_0 \), since the loss \( f(\cdot) \) strictly decreases along
the trajectory of (41)). If further Assumption M.1 holds, then we know \( J(W_0)|_\mathcal{T} : \mathcal{T} \rightarrow \mathcal{T} \) can be diagonalized as \( J(W_0)|_\mathcal{T} \cdot \mathbf{r} = \mathcal{V}(\Sigma \mathcal{V}^{-1}(\mathbf{r})) \), where \( \Sigma \mathcal{V} = \text{diag}(-\mu_1, \ldots, -\mu_{\dim(\mathcal{T})}) \), \( \mathcal{V} : \mathbb{R}^{\dim(\mathcal{T})} \rightarrow \mathcal{T}, \mathcal{V}(\mathbf{x}) = \sum_{i=1}^{\dim(\mathcal{T})} x_i \mathbf{v}_i \), and \( \mathbf{v}_i \) is the eigenvector associated with eigenvalue \( -\mu_i \).

As shown in Theorem M.3 below, this assumption implies that if \( W(0) \) is rank-\( k \) and is sufficiently close to \( W_0 \), then \( \|W(t) - W_0\|_F \leq C e^{-\rho t} \) for some constant \( C \). For depth-2 case, the above assumption is equivalent to that \( \mathcal{L}(U_0) \) is "strongly convex" at \( U_0 \), except those 0 eigenvalues due to symmetry, by property 2 of Theorem H.5). For the case where \( L \geq 3 \), because this dynamics is not gradient flow, in general it does not correspond to a loss function and strongly convexity does not make any sense. Nevertheless, in experiments we do observe linear convergence to \( W_0 \), so this assumption is reasonable.

M.1 Rank-\( k \) Initialization

For convenience, we define for all \( W \in S_d \),
\[
\|W\|_\mathcal{V} := \left\| \mathcal{V}^{-1} \left( \Pi^2(W) \right) \right\|_F, \quad \|W\|_{F,1} := \left\| \Pi^2(W) \right\|_F, \quad \|W\|_{F,2} := \left\| \Pi^2(W) \right\|_F.
\]

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, \( \|W - W'\|_{F,1} \), dominates that in the orthogonal complement of the tangent space, \( \|W - W'\|_{F,2} \), when both \( W, W' \) get very close to the \( W_0 \) (see a more rigorous statement in Lemma M.2). WLOG, we can assume
\[
\frac{\|\cdot\|_{F,1}}{K} \leq \|\cdot\|_{\mathcal{V}} \leq \|\cdot\|_{F,1},
\]
for some constant \( K \), which may depend on \( f \) and \( W_0 \). This also implies that \( \|\cdot\|_{\mathcal{V}} \leq \|\cdot\|_F \).

Below we also assume for sufficiently small \( R \), and any \( W \) such that \( \|W - W_0\|_F \leq R \), we have \( \|\nabla f(W)\|_2 \leq \rho \) and \( \|J(W)\Delta\|_F \leq \beta \|\Delta\|_F \) for any \( \Delta \). In the proof below, we assume such properties hold as long as we can show the boundedness of \( W(t) - W_0 \).

**Lemma M.2.** Let \( \max\{\|W - W_0\|_{F,1}, \|W' - W_0\|_{F,1}\} = r \), when \( r \leq \frac{m}{2} \), we have
\[
\|W - W'\|_{F,2} \leq \frac{5r}{m} \|W - W'\|_{F,1}.
\]

As a special case, we have
\[
\|W - W_0\|_{F,2} \leq \frac{5\|W - W'\|^2_{F,1}}{m}.
\]

**Proof.** WLOG we can assume \( W_0 \) is only non-zero in the first \( k \) dimension, i.e., \( [W_0]_{ij} = 0 \), for all \( i \geq k + 1, j \geq k + 1 \). We further denote \( W \) and \( W' \) by
\[
W = \begin{bmatrix} A & B' \\ B & C \end{bmatrix}, \quad W' = \begin{bmatrix} A' & B'^T \\ B' & C' \end{bmatrix},
\]
where \( A, A' \in \mathbb{R}^{k \times k}, B, B' \in \mathbb{R}^{(d-k) \times k}, C, C' \in \mathbb{R}^{(d-k) \times (d-k)} \). By definition, we have \( \|A - A'\|_F, \|B - B'\|_F \leq \|W - W'\|_{F,1} \) and \( \|W - W'\|_{F,2} = \|C - C'\|_F \). Moreover, we have \( \lambda_{\min}(A) \geq m - \|W - W_0\|_F \geq m - \|W - W_0\|_{F,1} \geq \frac{m}{2} \).

Since \( W, W' \) is rank-\( k \), we have \( C = BA^{-1}B'^T, C' = B'A'^{-1}B'^T \). Thus
\[
\|W - W'\|_{F,2} = \|C - C'\|_F \\
= \|BA^{-1}B'^T - B'A'^{-1}B'^T\|_F \\
\leq \|B - B'\|_F \|A^{-1}B'^T\|_F + \|BA^{-1}\|_F \|A' - A\|_F \|A'^{-1}B'^T\|_F + \|B'A'^{-1}\|_F \|B'^T - B'^T\|_F \\
\leq \|W - W'\|_{F,1} \frac{2r}{m} + \|W - W'\|_{F,1} \left( \frac{2r}{m} \right)^2 + \|W - W'\|_{F,1} \frac{2r}{m} \\
\leq \|W - W'\|_{F,1} \frac{5r}{m}.
\]

\( \square \)
Theorem M.3 (Linear convergence of rank-$k$ matrices). Suppose that \( \text{rank}(W(0)) = \text{rank}(W_0) = k \) and
\[
\|W(0) - W_0\|_V \leq R := \max \left\{ \frac{\mu_1}{2K^2(29\beta + 10\rho/m)} \right\},
\]
we have \( \|W(t) - W_0\|_V \leq Ce^{-\mu_1 t} \|W(0) - W_0\|_V \) for some constant \( C \) depending on \( W_0 \), where \( W(t) \) satisfies (41).

Proof. For convenience, we define \( W_1(t) := \Pi^d_1 (W(t) - W_0) \), \( W_2(t) := \Pi^d_2 (W(t) - W_0) = \Pi^d_2 (W(t)) \). We also use \( \langle \cdot, \cdot \rangle_{V^{-1}} = \langle V^{-1}(\cdot), V^{-1}(\cdot) \rangle \) for short.

\[
\frac{d}{dt} \|W_1(t)\|_V^2 = \frac{d}{dt} \|W(t) - W_0\|_V^2
\]
\[
= 2 \left\langle \Pi^d_1 \left( \frac{dW(t)}{dt} \right), \Pi^d_1 (W(t) - W_0) \right\rangle_{V^{-1}}
\]
\[
= 2 \left\langle \Pi^d_1 (g(W(t))), W_1(t) \right\rangle_{V^{-1}}
\]
\[
\leq 2 \left\langle \Pi^d_1 (J(W_0)[W(t) - W_0]), W_1(t) \right\rangle_{V^{-1}}
\]
\[
+ 2 \|g(W(t) - W_0) - J(W_0)[W(t) - W_0]\|_V \|W(t) - W_0\|_V
\]
\[
= 2 \left\langle \Pi^d_1 (J(W_0)[W_1(t) + W_2(t)]), W_1(t) \right\rangle_{V^{-1}}
\]
\[
+ 2 \|g(W(t) - W_0) - J(W_0)[W(t) - W_0]\|_V \|W_1(t)\|_V
\]
\[
= 2 \left\langle \Pi^d_1 (J(W_0)[W_1(t)]), W_1(t) \right\rangle_{V^{-1}} + 2 \|J(W_0)[W_2(t)]\|_V \|W_1(t)\|_V
\]
\[
+ 2 \|g(W(t) - W_0) - J(W_0)[W(t) - W_0]\|_V \|W_1(t)\|_V.
\]

For the first term \( \left\langle \Pi^d_1 (J(W_0)[W_1(t)]), W_1(t) \right\rangle_{V^{-1}} \), we know \( W_1(t) \in \mathcal{T} \), and \( \mathcal{T} \) is an invariant space of \( J(W_0) \). Recall \( J(W_0)[\mathcal{T}] = \mathcal{V} (\Sigma V^{-1}(\cdot)) \), we have
\[
2 \left\langle \Pi^d_1 (J(W_0)[W_1(t)]), W_1(t) \right\rangle_{V^{-1}} = 2 \langle \Sigma V^{-1}(W_1(t)), V^{-1}(W_1(t)) \rangle \leq -2\mu_1 \|W_1(t)\|_{F,1}.
\]

For the second term \( 2\beta \|J(W_0)[W_2(t)]\|_V \|W_1(t)\|_V \), we have
\[
2 \|J(W_0)[W_2(t)]\|_V \|W_1(t)\|_V \leq 2 \|J(W_0)[W_2(t)]\|_F \|W_1(t)\|_V = 2\rho \|W_2(t)\|_F.
\]

For the third term \( 2 \|g(W(t) - W_0) - J(W_0)[W(t) - W_0]\|_V \|W_1(t)\|_V \), we have
\[
2 \|g(W(t) - W_0) - J(W_0)[W(t) - W_0]\|_V \leq 2\beta \|W(t) - W_0\|_F^2
\]
\[
\leq 4\beta (\|W_1(t)\|_F^2 + \|W_2(t)\|_F^2)
\]
\[
\leq 4\beta (K^2 \|W_1(t)\|_V^2 + \|W_2(t)\|_V^2).
\]

Thus we have shown the following. Note so far we have not used the assumption that \( W \) is rank-$k$.
\[
\frac{d}{dt} \|W_1(t)\|_V^2 \leq -2\mu_1 \|W_1(t)\|_V^2 + 2\|W_1(t)\|_V \left( \rho \|W_2(t)\|_F + 2\beta K^2 \|W_1(t)\|_V + 2\beta \|W_2(t)\|_F^2 \right),
\]
that is,
\[
\frac{d}{dt} \log \|W_1(t)\|_V^2 \leq -2\mu_1 + 4\beta K^2 \|W_1(t)\|_V + \frac{4\beta \|W_2(t)\|_F + 2\rho \|W_2(t)\|_F}{\|W_1(t)\|_V}.
\]

Let \( T := \sup \{ t \geq 0 : \|W_1(t)\|_V \leq \frac{m}{2K^2} \} \). Setting \( W' = W_0 \) in Lemma M.2, we have for \( t < T \),\( r = \|W(t) - W_0\|_{F,1} \leq \|W(t) - W_0\|_F \leq K \|W(t) - W_0\|_V \leq \frac{m}{2K} \). Thus,\( \|W_2(t)\|_F = \|W_2(t)\|_{F,2} \leq \frac{5\|W(t) - W_0\|_F^2}{m} \leq \frac{5K^2 \|W(t) - W_0\|_V^2}{m} \leq \frac{5}{4} m. \)
Thus, from (42) we can derive that
\[
\frac{d}{dt} \| W_1(t) \|_V^2 \leq -2\mu_1 + K^2(29\beta + 10\rho/m)) \| W_1(t) \|_V \leq -\mu_1. \tag{43}
\]

Since \( \mu_1 < 0 \), \( \| W_1(t) \|_V \) decreases for \([0, T)\). Thus \( T \) must be \( \infty \), otherwise \( \| W_1(T) \|_V = \lim_{t\to T^-} \| W_1(t) \|_V < R_1 \). Contradiction.

Therefore, for any \( t \in [0, \infty) \), we have \( \| W_1(t) \|_V \leq \| W_1(0) \|_V e^{-\mu_1 t} \). That is,
\[
\int_0^\infty \| W_1(t) \|_V dt \leq \frac{2}{\mu_1} \| W_1(0) \|_V \leq \frac{2R}{\mu_1},
\]

Thus from (43), we have
\[
\| W(t) \|_V = \| W_1(t) \|_V \leq \| W_1(0) \|_V \exp \left( -\mu_1 t + \frac{K^2}{2}(29\beta + 10\rho/m) \right) \int_0^\infty \| W_1(t) \|_V dt \]
\[
\leq \| W_1(0) \|_V \exp \left( -\mu_1 t + \frac{K^2}{2}(29\beta + 10\rho/m) \right)
\]
\[
=: C \| W(0) \|_V e^{-\mu_1 t},
\]
which completes the proof. \( \square \)

M.2 Almost Rank-k Initialization

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

Below we show that for deep overparametrized matrix factorization, where \( W(t) \) satisfies (41), if the trajectory is initialized at some \( \| W(0) \|_F \in \) a small neighborhood of the \( k \)-th critical point \( W_0 \) of deep GLRL, and \( W(0) \) is approximately rank-\( k \), in the sense that \( N(0) \) is very small, then \( \inf_{t \geq 0} \| W(t) - W_0 \|_F \) is roughly at the same magnitude of \( N(0) \).

**Theorem M.4** (Linear convergence of almost rank-\( k \) matrices, deep case). Suppose \( W_0 \) is a critical point of rank \( k \) and \( W_0 \) satisfies Assumption M.1, there exists constants \( C_0 \) and \( r \), such that if \( C_0 \| N(0) \|_F \leq \| W_1(0) \|_V \leq r \), then there exists a time \( T \) and constants \( C, C' \), such that

1. \( \| W(t) - W_0 \|_V \leq C e^{-\mu_1 t/2} \| W(0) - W_0 \|_V \), for \( t \leq T \).
2. \( \| W(T) - W_0 \|_F \leq C' \| N(0) \|_F \).

**Proof.** When \( \| W(t) - W_0 \|_F \leq \frac{\lambda_{\text{min}}(W_0)}{4} \| N(t) \|_F \leq \frac{\lambda_{\text{min}}(W_0)}{4} \), thus we have
\[
\| M(t) - W_0 \|_{F,1} \leq \| W(t) - W_0 \|_{F,1} + \| N(t) \|_{F,1} \leq \frac{\lambda_{\text{min}}(W_0)}{2},
\]
thus by Lemma M.2, we have
\[
\| W_{2}(t) \|_{F,2} \leq \| M(t) - W_0 \|_{F,2} + \| N(t) \|_{F,2}
\]
\[
\leq 5 \frac{\| M(t) - W_0 \|_{F,1}^2}{\lambda_{\text{min}}(W_0)} + \| N(t) \|_{F,2}
\]
\[
\leq 10 \frac{\| W_1(t) \|_{F,1}^2 + 10 \| N(t) \|_{F}^2}{\lambda_{\text{min}}(W_0)} + \| N(t) \|_{F,2}
\]
\[
\leq 10K^2 \frac{\| W_1(t) \|_{V}^2 + 10 \| N(t) \|_{F}^2}{\lambda_{\text{min}}(W_0)} + \| N(t) \|_{F,2}.
\]

Thus we can pick constant \( C_0 \) large enough and \( r \) small enough, such that for any \( t \geq 0 \), if \( C_0 \| N(t) \|_F \leq \| W_1(t) \|_V \leq r \), then it holds that:
• The “small terms” in the RHS of (42) satisfies that
\[
4\beta K^2 \|W_1(t)\|_V + \frac{4\beta \|W_2(t)\|_F^2 + 2\rho \|W_2(t)\|_F}{\|W_1(t)\|_V} \leq C_1 \|W_1(t)\|_V + C_2 \|N(t)\|_F \leq \mu_1
\]
for some \(C_1\) and \(C_2\) independent of \(t\).

• The spectral norm \(\frac{1}{2} \|\nabla f(W(t))\|_2 \leq \|\nabla f(W_0)\|_2 =: \rho\) for all \(t \geq 0\).

• \(\forall x < r, \kappa x^{\frac{L-2}{2}} > \frac{1}{\mu_1} \ln \frac{2r}{C_0 x}, \) where \(\kappa_L = 1 - 0.5 \frac{L-2}{r}\).

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., \(C_0 \|N(t)\|_F \leq \|W_1(t)\|_V \leq r\) fails. Thus by (42), for \(t \in [0, T_{C_0, r}]\), we have \(\|W(t)\|_V = \|W_1(t)\|_V \leq \|W_1(0)\|_V e^{-\frac{T}{2}} = \|W_0\|_V e^{-\frac{T}{2}}\). Thus (1) holds for any \(T\) smaller than \(T_{C_0, r}\). If \(T_{C_0, r} = \infty\), then clearly we can pick a sufficiently large \(T\), such that (2) holds. Therefore, below it suffices to consider the case where \(T_{C_0, r}\) is finite. And we know the condition that fails must be \(C_0 \|N(t)\|_F \leq \|W_1(t)\|_V\), i.e. \(C_0 \|N(T_{C_0, r})\|_F = \|W_1(T_{C_0, r})\|_V\).

By (34) in Lemma K.7, we have
\[
\left| \|N(0)\|_F^{\frac{L-2}{2}} - \|N(t)\|_F^{\frac{L-2}{2}} \right| \leq (L-2) \rho t.
\]
Define \(T' := \frac{\kappa_L \|N(0)\|_F^{\frac{L-2}{2}}}{(L-2) \rho}\), we know for any \(t < T'\), we have
\[
\left| \|N(0)\|_F^{\frac{L-2}{2}} - \|N(t)\|_F^{\frac{L-2}{2}} \right| \leq \kappa_L \|N(t)\|_F^{\frac{L-2}{2}}.
\]
That is,
\[
\frac{\|N(t)\|_F^{\frac{L-2}{2}}}{\|N(0)\|_F^{\frac{L-2}{2}}} \in \left[1 - \kappa_L, \frac{1}{1 - \kappa_L} \right] = \left[0.5 \frac{L-2}{L-2}, 0.5 - \frac{L-2}{L-2} \right] \implies \frac{\|N(t)\|_F}{\|N(0)\|_F} \in [1/2, 2].
\]
Now we claim it must hold that \(T' \geq T_{C_0, r}\). Otherwise, we have
\[
\frac{C_0}{2} \|N(0)\|_F \leq C_0 \|N(T')\|_F \leq \|W_1(T')\|_V \leq e^{-\mu_1 T'/2} \|W_1(0)\|_V \leq e^{-\mu_1 T'/2}.\]

Therefore, \(\kappa_L \|N(0)\|_F^{\frac{L-2}{2}} = T' \leq \frac{2}{\mu_1} \ln \frac{2r}{C_0 \|N(0)\|_F}\), which contradicts to the definition of \(C_0\) and \(r\).

As a result, we have
\[
2C_0 \sqrt{d} \|N(0)\|_F = 2C_0 \|N(0)\|_F \geq C_0 \|N(T_{C_0, r})\|_F = \|W_1(T_{C_0, r})\|_V \geq \|W_1(0)\|_V e^{-\mu_1 T_{C_0, r}/2},
\]
and therefore,
\[
T_{C_0, r} \leq \frac{2}{\mu_1} \ln \frac{\|W_1(0)\|_V}{2 \sqrt{d} C_0 \|N(0)\|_F}.
\]
Thus by Lemma M.2, we know
\[
\|W(T_{C_0, r}) - W_0\|_F \leq \|W(T_{C_0, r}) - W_0\|_{F, 1} + \|W(T_{C_0, r}) - W_0\|_{F, 2} \leq K \|W_1(T_{C_0, r})\|_V + \|M(T_{C_0, r}) - W_0\|_{F, 2} + \|N(T_{C_0, r})\|_{F, 2} \leq O(\|N(0)\|_F) + O(\|N(0)\|_F^2) + O(\|N(0)\|_F) = O(\|N(0)\|_F).
\]
M.3 Proof for Theorem 6.4

Proof for Theorem 6.4. Let $C_0, r$ be the constants predicted by Theorem M.4 w.r.t. to $\overline{W}(\infty)$. We claim that we can pick large enough constant $T$, and $\alpha_0$ sufficiently small, such that for all $\alpha \leq \alpha_0$, the initial condition in Theorem M.4 holds, i.e. $C_0 \parallel N(0) \parallel_F \leq \parallel W_1(0) \parallel_V \leq r$, where $W(0) := \phi \left( \alpha I, \frac{\alpha^{-(p-1)}}{2^p \mu_1^{-(p-1)}} + T \right)$.

This is because we can first ensure $\parallel W(T) - \overline{W}(\infty) \parallel_2$ is sufficiently small, i.e., smaller than $\frac{r}{2}$. By Theorem 6.2, we know when $\alpha \to 0$, $\parallel W(T) - W(0) \parallel_V \leq K \parallel W(T) - W(0) \parallel_F = o(1)$ and $\parallel N(0) \parallel_F = O(\alpha)$.

By Theorem M.4, we know there is a time $T$ (either $T_{C_0,r}$ or some sufficiently large number when $T_{C_0,r} = \infty$), such that $\parallel W(T) - W_0 \parallel_F = O(\parallel N(0) \parallel_F) = O(\alpha)$. \qed