# Sharpness of Minima in Deep Matrix Factorization

**Anıl Kamber**[1] **Rahul Parhi**[1]

[1]Department of Electrical and Computer Engineering
University of California, San Diego
akamber@ucsd.edu, rahul@ucsd.edu

## Abstract

Understanding the loss-landscape geometry near a minimum is key to explaining the implicit bias of gradient-based methods in non-convex optimization problems such as deep neural network training and deep matrix factorization. A central quantity for characterizing this geometry is the maximum eigenvalue of the Hessian of the loss. Currently, its precise role has been obfuscated because no exact expressions for this sharpness measure were known in general settings. In this paper, we present the first exact expression for the maximum eigenvalue of the Hessian of the squared-error loss at any minimizer in deep matrix factorization problems, resolving an open question posed by Mulayoff and Michaeli [2020]. This expression reveals a fundamental property of the loss landscape in deep matrix factorization: It is sufficient, but not necessary, for flatness that, across all layers, the product of the spectral norms of the left and right intermediate factors equals a constant determined by the depth and the spectral norm of the target matrix. Most notably, in both depth-2 matrix factorization and deep overparameterized scalar factorization, we show that this condition is both necessary and sufficient for flatness, which implies that flat minima are spectral-norm balanced and not necessarily balanced in depth-2 matrix factorization. To complement our theory, we provide empirical verification, relying on our exact expression, of an escape phenomenon during gradient-based training near a minimizer of a deep matrix factorization problem.

## 1 Introduction

Decades of research in learning theory suggest limiting model complexity to prevent overfitting. However, modern deep learning is heavily overparameterized and has nonetheless achieved unprecedented success in practice over the past decade [Krizhevsky et al., 2012, Vaswani et al., 2017]. Generally, in overparameterized settings, the loss function has infinitely many global minima that achieve zero training error (interpolation regime), yet these models still perform well. This phenomenon has been explored in various settings such as nonparametric regression [Belkin et al., 2019], training two-layer neural networks with logistic loss [Frei et al., 2022], and linear regression [Bartlett et al., 2020].

The observation of this phenomenon naturally raises the question: Why do optimization algorithms find minima that generalize well? In the literature, the propensity of neural network training dynamics to converge to *good minima* is generally attributed to the ability of gradient-based optimization algorithms to avoid *bad minima* [Neyshabur et al., 2017, Zhang et al., 2017]. This is related to the *implicit bias* of gradient descent (GD) [Neyshabur et al., 2014], and a large body of work has focused on its understanding [Gunasekar et al., 2017, 2018, Soudry et al., 2018, Arora et al., 2019, Ji and Telgarsky, 2020, Yun et al., 2021].

It has been observed that *dynamical stability* of GD near a minimum is a key factor in characterizing its implicit bias toward particular solutions [Wu et al., 2018, Nar and Sastry, 2018]. Conceptually, dynamical stability refers to the ability of GD to *stably converge* to a minimum, and it is closely related to the sharpness of the loss landscape in its vicinity [Mulayoff et al., 2021]. This topic has

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Closed-form expressions for the maximum Hessian eigenvalue in the literature. $\Omega$ denotes the set of *all* global minimizers, $\Omega_F \subseteq \Omega$ denotes the set of *flat* global minimizers, and $\Omega_B \subseteq \Omega$ denotes the set of *balanced* global minimizers (see Definition 2.1). The *Layers* column specifies the form of the layers (factors) considered in each work: scalar-valued ($\mathbb{R}$), vector-valued ($\mathbb{R}^a$), or matrix-valued ($\mathbb{R}^{a\times b}$).

| Related Work | Depth | Input Dim. | Output Dim. | $\lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{w}))$ | Layers |
|---|---|---|---|---|---|
| Mulayoff and Michaeli [2020, Theorem 1] | $L$ | $d_0$ | $d_L$ | $\mathbf{w}\in\Omega_F$ | $\mathbb{R}^{a\times b}$ |
| Zhu et al. [2023, Appendix B.1] | 2 | 1 | 1 | $\mathbf{w}\in\mathbb{R}^N$ | $\mathbb{R}$ |
| Singh and Hofmann [2024, Theorem 1] | 2 | 1 | 1 | $\mathbf{w}\in\mathbb{R}^N$ | $\mathbb{R}^a$ |
| Ghosh et al. [2025, Lemma 1] | $L$ | $d_0$ | $d_L$ | $\mathbf{w}\in\Omega_B$ | $\mathbb{R}^{a\times b}$ |
| Theorem 3.1 (This Paper) | $L$ | $d_0$ | $d_L$ | $\mathbf{w}\in\Omega$ | $\mathbb{R}^{a\times b}$ |

been investigated in numerous works [Nar and Sastry, 2018, Wu et al., 2018, Ma and Ying, 2021, Mulayoff et al., 2021, Nacson et al., 2023, Qiao et al., 2024, Liang et al., 2025] within the framework of the classical notion of *linear stability* in dynamical systems [Strogatz, 2024].

Ultimately, this understanding boils down to understanding the *geometry* of the loss landscape near a minimum. The maximum eigenvalue of the Hessian of the loss serves as a key measure to quantify the *sharpness* of the landscape near a minimum. Despite its significance, its precise role is not well-understood, particularly because closed-form expressions are generally unknown, outside a few particular cases. We summarize the current state of understanding as well as the contributions of our paper in Table 1.

Most notably, the seminal work of Mulayoff and Michaeli [2020] derives a closed-form expression for the maximum eigenvalue of the Hessian at *flat* global minima of deep linear networks/deep matrix factorization with squared-error loss. However, obtaining a closed-form expression for *all* global minima in deep linear networks/deep matrix factorization was an open problem. In particular, Mulayoff and Michaeli [2020] claim that finding a closed-form expression for arbitrary global minima is intractable. In this paper, we refute this claim and give a positive answer to the following fundamental question.

*Does a closed-form expression for the maximum eigenvalue of the Hessian exist for overparameterized deep matrix factorization problems?*

In particular, in Theorem 3.1, we provide a closed-form expression for the maximum Hessian eigenvalue at *arbitrary global minima* of depth-$L$ overparameterized deep matrix factorization problems with squared-error loss. We consider a deep matrix factorization setting in which no layer is narrower than both the input and output layers. This guarantees that the factorization is feasible for *every target matrix* in $\mathbb{R}^{d_L\times d_0}$ and hence that the global minima operate in the interpolation regime. Note that the squared-error loss is non-negative, so the existence of an interpolating solution means that the optimal training loss is zero; consequently, every global minimizer is an interpolating solution. Without the above architectural constraint, the global minimizers need not achieve zero training error for an arbitrary target matrix, and our closed-form expression applies only to those target matrices that admit an interpolating solution, a set determined by the dimensions of the factors.

To the best of our knowledge, our analysis provides the first exact expression of the maximum eigenvalue for deep matrix factorization/deep linear neural network training problems. In the case of deep overparameterized scalar factorization (Corollary 3.2) and depth-2 matrix factorization (Corollary 3.3), our closed-form expression simplifies considerably. In Section 4, we present new necessary and sufficient conditions for flat minima in various matrix factorization settings. In Section 5, we demonstrate that different notions of sharpness can lead to misleading conclusions, and we illustrate this with an explicit example.

With our closed-form expression in hand, we then empirically observe in Section 6 the *escape phenomenon*, which was introduced by Wu et al. [2018]. This phenomenon has been observed in numerous works; however, it has been verified empirically only in extremely simplified settings. For instance, Lewkowycz et al. [2020, see Fig. 1] observed this phenomenon for a one-hidden layer linear network trained with squared-error loss on a single scalar data point; Cohen et al. [2021] analyzed it for a convex quadratic loss function; most recently, Ghosh et al. [2025, see Fig. 4] characterized it empirically for the two-layer scalar factorization loss. Although their analysis is not directly related

to the escape phenomenon, Liang and Montúfar [2026] recently studied the convergence regions of GD in deep matrix factorization problems.

In contrast, we provide the first empirical verification of the escape phenomenon during gradient-based training near a minimizer of a deep matrix factorization problem. We find that the escape phenomenon also occurs for overparameterized deep matrix factorization problems. We verify this phenomenon through the lens of *dynamical stability* introduced by Wu et al. [2018].

Consider a twice continuously differentiable loss function $\mathcal{L} : \mathbb{R}^N \to \mathbb{R}$, and let $\mathbf{w}^*$ be a minimizer of $\mathcal{L}$. For $\delta > 0$, denote the open ball $B_\delta(\mathbf{w}^*) := \left\{\mathbf{w} \in \mathbb{R}^N : \|\mathbf{w} - \mathbf{w}^*\|_2 < \delta\right\}$. There exists a $\delta > 0$ such that, for all $\mathbf{w} \in B_\delta(\mathbf{w}^*)$,

$$\mathcal{L}(\mathbf{w}) = \mathcal{L}(\mathbf{w}^*) + (\mathbf{w} - \mathbf{w}^*)^\top \nabla \mathcal{L}(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^\top \nabla^2 \mathcal{L}(\mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*) + o(\|\mathbf{w} - \mathbf{w}^*\|_2^2). \quad (1)$$

**Definition 1.1.** Let $\mathcal{L} : \mathbb{R}^N \to \mathbb{R}$ be a twice continuously differentiable loss function, and let $\mathbf{w}^*$ be a minimizer of $\mathcal{L}$. The linearized GD update rule for $\mathcal{L}$ is given by

$$\tilde{\mathbf{w}}_{t+1} = \tilde{\mathbf{w}}_t - \eta \nabla^2 \mathcal{L}(\mathbf{w}^*)(\tilde{\mathbf{w}}_t - \mathbf{w}^*), \quad (2)$$

where $\eta > 0$ is the step size. Given $\delta > 0$ such that (1) holds, if there exists an initial point $\mathbf{w}_0 \in B_\delta(\mathbf{w}^*)$ such that the residuals $\tilde{\boldsymbol{\epsilon}}_t = \tilde{\mathbf{w}}_t - \mathbf{w}^*$ diverge, i.e.,

$$\lim_{t \to \infty} \|\tilde{\boldsymbol{\epsilon}}_t\|_2 = \infty, \quad (3)$$

then $\mathbf{w}^*$ is said to be *dynamically unstable*[1].

A necessary and sufficient condition for a minimum to be dynamically unstable is that $\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{w}^*)) > 2/\eta$ (cf. Wu et al. 2018, Mulayoff et al. 2021, Chemnitz and Engel 2025). Thus, we use our exact expression for the maximum eigenvalue to verify the escape phenomenon empirically in Section 6.

### 1.1 Contributions

In this paper, we present the first exact expression for the maximum eigenvalue of the Hessian of the squared-error loss at any minimizer in general overparameterized deep matrix factorization problems. Our results lead to the following remarkable observations:

- A minimizer of the deep overparameterized scalar factorization loss is flat **if and only if**, across all layers, the product of the spectral norms of the left and right intermediate factors equals a constant determined by the depth and the magnitude of the target (Corollary 4.1).
- Flat minima are spectral-norm balanced in depth-2 matrix factorization (Corollary 4.2) and not necessarily balanced, contrary to claims made in [Ghosh et al., 2025]. We further discuss this in Section 5.
- A minimizer of the deep matrix factorization loss is flat **if**, across all layers, the product of spectral norms of the left and right intermediate factors equals a constant determined by the depth and the spectral norm of the target matrix (Corollary 4.4). We further show, via an explicit counterexample, that the converse does not hold in general (Remark 4.5).

We also highlight a recent work of Singh et al. [2026] that provides empirical evidence, in two-layer linear networks, that the largest eigenvalue of the Hessian of the loss is correlated with the spectral norm of the weight matrices. Furthermore, for deep linear networks, they observe that this largest Hessian eigenvalue correlates with the spectral norm of the products of the weight matrices. Thus, our results in Theorem 3.1, Corollary 3.2, and Corollary 3.3 provide the first theoretical explanation for their empirical observations.

### 1.2 Related Work

**Flat Minima.** Mulayoff and Michaeli [2020] derived a closed-form expression for the maximum eigenvalue of the Hessian at flat minima for deep linear neural networks. They also showed that when

[1] Note that Definition 1.1 describes the linearized dynamics in a region where the loss is well approximated by its quadratic approximation. Once GD leaves this region, the linearized iterates $\{\tilde{\mathbf{w}}_t\}_{t \geq 0}$ no longer capture the true GD iterates $\{\mathbf{w}_t\}_{t \geq 0}$.

the input and output dimensions are the same, the Hessian at a global minimum is rank-deficient by at least the order of $1 - 1/L$, where $L$ is the depth of the network. Moreover, they showed that the sharpness of the flat minima increases approximately linearly with $L$ if $L \gg 1$. Singh and Hofmann [2024] provided a full characterization of the Hessian spectrum at a point in parameter space for linear and ReLU networks in the *scalar regression* case. They observed that the eigenvalues scale in proportion to the input variance in scalar linear networks with one hidden layer. More recently, Josz [2025] has shown that locally flat minima are globally flat in depth-2 matrix factorization problems.

**Balanced Minima.** Ghosh et al. [2025] provided a full characterization of the Hessian spectrum at balanced minima in deep matrix factorization problems. Furthermore, they showed that the maximum eigenvalue of the Hessian at flat minima is equal to that of the balanced minima. Ding et al. [2024] showed that *norm-minimal*, *balanced* and *flat* solutions coincide in depth-2 matrix factorization problems, where the sharpness is measured by the *scaled trace* of the Hessian matrix of the loss function. We further discuss the results of Ghosh et al. [2025] and Ding et al. [2024] in Section 5. Finally, Wang et al. [2022] showed that large step size GD training induces a *Frobenius-norm balancing effect* between factors in depth-2 matrix factorization problems.

**Dynamical Stability.** In dynamical systems theory, the local stability of a critical point plays a central role in characterizing the convergence behavior of nearby trajectories [Strogatz, 2024]. In the seminal work of Wu et al. [2018] on the dynamical stability analysis of GD training, it was shown that a global minimum is *dynamically stable* for GD if and only if the step size does not exceed $2/\lambda_{\max}$, where $\lambda_{\max}$ denotes the maximum eigenvalue of the Hessian of the loss at the minimum. Mulayoff et al. [2021] investigated this mechanism in the space of learned functions for two-layer overparameterized univariate ReLU networks in the interpolation regime. This was then extended to multivariate ReLU networks by Nacson et al. [2023]. The interpolation assumption was then removed by Qiao et al. [2024] and Liang et al. [2025].

**Edge-of-Stability.** Cohen et al. [2021] observed that neural networks trained with GD typically operate in a regime called *edge of stability*, in which the maximum eigenvalue of the Hessian of the loss hovers just above the value $2/\eta$, where $\eta$ is the step size, and argued that classical optimization theory fails to explain this phenomenon. Recently, Liang et al. [2025] empirically observed that explicit regularization seems to break the edge-of-stability phenomenon.

**Sharpness and Generalization.** It is widely recognized in the literature that flat minima are associated with better generalization [Hochreiter and Schmidhuber, 1997, Keskar et al., 2017]. In a large-scale empirical investigation, Jiang et al. [2020] examined different complexity measures for deep networks and found that a sharpness-based measure exhibited the strongest correlation with generalization. There is also theoretical evidence for this phenomenon in low-rank matrix recovery [Ding et al., 2024]. On the other hand, Dinh et al. [2017] showed that *good minima* can be arbitrarily sharp in deep neural networks.

## 2 Notation and Problem Setup

We denote the *Kronecker product* by $\otimes$, the *Frobenius inner product* by $\langle \cdot, \cdot \rangle$, the *spectral norm* by $\sigma_{\max}(\cdot)$, and the *Frobenius norm* by $\|\cdot\|_F$. We denote by $[L]$ the set of natural numbers up to $L$, that is, $[L] \coloneqq \{1, 2, \ldots, L\}$, and by $\mathbf{I}_{n \times n}$ the $n \times n$ identity matrix. We denote a sequence by $\{x_i\}_{i \geq 0}$.

To simplify the notation for subsequent derivations, we define

$$\prod_{j=n}^{m} \mathbf{W}_j \coloneqq \begin{cases} \mathbf{W}_m \mathbf{W}_{m-1} \ldots \mathbf{W}_n & 1 \leq n \leq m \leq L, \\ \mathbf{I}_{d_m \times d_m} & 0 \leq m \leq L \text{ and } n = m + 1, \end{cases} \tag{4}$$

where $\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$ for any $i \in [L]$.

In this paper, we study the sharpness of the loss landscape of deep matrix factorization problems near any global minimizer. We consider the following optimization problem

$$\min_{\mathbf{w} \in \mathbb{R}^N} \mathcal{L}(\mathbf{w}) \coloneqq \|\mathbf{M} - \mathbf{W}_L \mathbf{W}_{L-1} \cdots \mathbf{W}_1\|_F^2, \tag{5}$$

where $\mathbf{w} = [\mathrm{vec}(\mathbf{W}_1)^\top \ \mathrm{vec}(\mathbf{W}_2)^\top \ \ldots \ \mathrm{vec}(\mathbf{W}_L)^\top]^\top$ denotes the collection of all parameters, and $N$ is the total number of parameters in the model. $\mathbf{M} \in \mathbb{R}^{d_L \times d_0} \setminus \{\mathbf{0}\}$ denotes the target matrix to

be factorized, $L \geq 2$ denotes the depth of factorization and $\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$ is the $i$th factor (layer). This objective is analogous to that of deep linear neural networks. To guarantee the feasibility of factorization for all points in $\mathbb{R}^{d_L \times d_0} \setminus \{\mathbf{0}\}$, we require

$$d_i \geq \min\{d_0, d_L\} \quad \forall i \in [L-1], \tag{6}$$

which follows directly from the fact that

$$\operatorname{rank}(\mathbf{W}_L \cdots \mathbf{W}_1) \leq \min\{\operatorname{rank}(\mathbf{W}_1), \ldots, \operatorname{rank}(\mathbf{W}_L)\}. \tag{7}$$

Linear networks satisfying (6) are called *filling architectures*, and those that do not are called *non-filling architectures* [Trager et al., 2019]. We adopt the same terminology in deep matrix factorization.

We define the set of global minima of $\mathcal{L}(\mathbf{w})$ as

$$\Omega \coloneqq \arg\min_{\mathbf{w} \in \mathbb{R}^N} \mathcal{L}(\mathbf{w}) = \left\{ \mathbf{w} \in \mathbb{R}^N : \prod_{i=1}^{L} \mathbf{W}_i = \mathbf{M} \right\} \tag{8}$$

and define the set of flat minima as

$$\Omega_F \coloneqq \arg\min_{\mathbf{w}^* \in \Omega} \lambda_{\max}\left(\nabla^2 \mathcal{L}(\mathbf{w}^*)\right), \tag{9}$$

where sharpness is measured by the maximum Hessian eigenvalue. This sharpness measure is also known as *worst-case sharpness*.

We further define *balanced* minima that have been investigated in several works as follows.

**Definition 2.1.** A minimizer $\mathbf{w}^* \in \Omega$ is balanced if $(\mathbf{W}_{k+1}^*)^\top \mathbf{W}_{k+1}^* = \mathbf{W}_k^* (\mathbf{W}_k^*)^\top$ for all $k \in [L-1]$. Therefore, we define the set of balanced minimizers as

$$\Omega_B \coloneqq \{\mathbf{w}^* \in \Omega : (\mathbf{W}_{k+1}^*)^\top \mathbf{W}_{k+1}^* = \mathbf{W}_k^* (\mathbf{W}_k^*)^\top \ \ \forall k \in [L-1]\} \tag{10}$$

Additionally, we define the notion of *norm-balanced minima* as follows.

**Definition 2.2.** A minimizer $\mathbf{w}^* \in \Omega$ is $\mathcal{S}^p$-norm balanced if

$$\|\mathbf{W}_1^*\|_{\mathcal{S}^p} = \|\mathbf{W}_2^*\|_{\mathcal{S}^p} = \cdots = \|\mathbf{W}_L^*\|_{\mathcal{S}^p}, \tag{11}$$

where $\|\cdot\|_{\mathcal{S}^p}$ is the Schatten-$p$ norm for $p \in [1, \infty]$. When $p = 2$, the minimizer is *Frobenius-norm balanced*, and when $p = \infty$, it is *spectral-norm balanced*.

*Remark* 2.3. A balanced minimizer $\mathbf{w}^* \in \Omega_B$ is Frobenius-norm balanced. Indeed, this is a direct consequence of the identity $\operatorname{tr}(\mathbf{X}^\top \mathbf{X}) = \operatorname{tr}(\mathbf{X}\mathbf{X}^\top) = \|\mathbf{X}\|_F^2$.

### 2.1 Benign Landscape

Laurent and Brecht [2018] showed that all local minima of deep linear networks with convex and differentiable loss are global if the network is filling. In particular, they proved a more general theorem concerning real-valued functions that take as input a product of matrices. Thus, a corollary of Laurent and Brecht [2018, Theorem 1] is that the overparameterized deep matrix factorization problem with squared-error loss *does not have spurious minima*. Thus, in this paper, we focus on global minima, without loss of generality. Furthermore, under the squared-error loss, deep linear networks have no spurious minima even when the architecture is non-filling [Baldi and Hornik, 1989, Kawaguchi, 2016, Trager et al., 2019].

### 2.2 Deep Linear Neural Network Training

For a training dataset $\mathcal{D} \coloneqq \{(\mathbf{x}_k, \mathbf{y}_k)\}_{k=1}^n$, the deep linear neural network training problem considered by Mulayoff and Michaeli [2020] is as follows:

$$\min_{\mathbf{w} \in \mathbb{R}^N} \mathcal{L}_{\text{DLN}}(\mathbf{w}) \coloneqq \frac{1}{n} \sum_{k=1}^{n} \left\| \mathbf{y}_k - \prod_{i=1}^{L} \mathbf{W}_i \mathbf{x}_k \right\|_2^2 = \frac{1}{n} \left\| \mathbf{Y} - \prod_{i=1}^{L} \mathbf{W}_i \mathbf{X} \right\|_F^2, \tag{12}$$

where $\mathbf{X} \coloneqq [\mathbf{x}_1 \ \ \mathbf{x}_2 \ \ \cdots \ \ \mathbf{x}_n] \in \mathbb{R}^{d_0 \times n}$ and $\mathbf{Y} \coloneqq [\mathbf{y}_1 \ \ \mathbf{y}_2 \ \ \cdots \ \ \mathbf{y}_n] \in \mathbb{R}^{d_L \times n}$. Whenever $\widehat{\mathbf{\Sigma}}_x \coloneqq \frac{1}{n}\mathbf{X}\mathbf{X}^\top$ has full rank and $d_i \geq \min\{d_0, d_L\}$ for all $i \in [L]$, we can construct a deep

matrix factorization problem whose global minimizers coincide with those of (12) [Chou et al., 2024, Equation 4]. For instance, the deep matrix factorization problem considered in (5) with $\mathbf{M} = \widehat{\mathbf{\Sigma}}_{yx}\widehat{\mathbf{\Sigma}}_x^{-1}$, where $\widehat{\mathbf{\Sigma}}_{yx} \coloneqq \frac{1}{n}\mathbf{Y}\mathbf{X}^\top$, has the same set of global minimizers as (12). Furthermore, when $\widehat{\mathbf{\Sigma}}_x = \mathbf{I}_{d_0 \times d_0}$, the *Hessian matrices* of the two squared-error losses coincide at every point of this common set of global minimizers [Mulayoff and Michaeli, 2020, Equation 21]. We therefore emphasize that overparameterized deep matrix factorization and the deep linear neural network setting of Mulayoff and Michaeli [2020] are equivalent for investigating the worst-case sharpness around a minimizer *whenever* $\widehat{\mathbf{\Sigma}}_x = \mathbf{I}_{d_0 \times d_0}$ *and* $d_i \geq \min\{d_0, d_L\}$ *for all* $i \in [L]$.

Our analysis relies on matrix calculus and the formulation of second directional derivatives. We defer the relevant technical preliminaries to Appendix A.

## 3 Overparameterized Deep Matrix Factorization

We now consider the general deep matrix factorization problem. In this section, we prove our main result, which is a closed-form expression for the maximum eigenvalue of the Hessian for any global minimum of the objective (5).

**Theorem 3.1.** *Consider the problem setup in Section 2. If* $\mathbf{w}^* \in \Omega$ *then*

$$\lambda_{\max}\left(\nabla^2 \mathcal{L}\left(\mathbf{w}^*\right)\right) = 2\sigma_{\max}\left(\sum_{i=1}^{L} \mathbf{B}_i^\top \mathbf{B}_i \otimes \mathbf{A}_i \mathbf{A}_i^\top\right), \tag{13}$$

*where* $\mathbf{A}_k = \prod_{i=k+1}^{L} \mathbf{W}_i^*$ *and* $\mathbf{B}_k = \prod_{i=1}^{k-1} \mathbf{W}_i^*$.

The proof relies on an explicit expression for the second directional derivative of $\mathcal{L}$ that simplifies at an arbitrary global minimizer, and the full proof appears in Appendix C.1. Note that the deep overparameterized scalar factorization is a special case of deep matrix factorization where both $\mathbf{B}_i^\top \mathbf{B}_i$ and $\mathbf{A}_i \mathbf{A}_i^\top$ reduce to scalars. In that special case, we have Corollary 3.2 below. Another corollary of Theorem 3.1 is the maximum Hessian eigenvalue for the classical (depth-2) matrix factorization problem. This result may be of independent interest as the expression simplifies considerably.

**Corollary 3.2.** *Consider the following objective function*

$$\mathcal{L}(\mathbf{w}) \coloneqq (m - \mathbf{w}_L^\top \mathbf{W}_{L-1} \cdots \mathbf{W}_2 \mathbf{w}_1)^2, \tag{14}$$

*where* $m \in \mathbb{R} \setminus \{0\}$, $d_0 = d_L = 1$, $\mathbf{w}_L \in \mathbb{R}^{d_{L-1}}$ *and* $\mathbf{w}_1 \in \mathbb{R}^{d_1}$. *For hidden factors (layers), i.e., for all* $i \in \{2, 3, \cdots, L-1\}$, *we have* $\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$. *Then, for all* $\mathbf{w}^* \in \Omega$,

$$\lambda_{\max}\left(\nabla^2 \mathcal{L}\left(\mathbf{w}^*\right)\right) = 2\sum_{i=1}^{L} \sigma_{\max}\left(\prod_{j=i+1}^{L} \mathbf{W}_j^*\right)^2 \sigma_{\max}\left(\prod_{j=1}^{i-1} \mathbf{W}_j^*\right)^2. \tag{15}$$

We also provide a self-contained proof of this corollary in Appendix B.

**Corollary 3.3.** *Consider the following depth-2 matrix factorization objective*

$$\mathcal{L}(\mathbf{L}, \mathbf{R}) = \left\|\mathbf{M} - \mathbf{L}\mathbf{R}^\top\right\|_F^2, \tag{16}$$

*where* $\mathbf{M} \in \mathbb{R}^{d_L \times d_0} \setminus \{\mathbf{0}\}$ *is the target matrix,* $\mathbf{L} \in \mathbb{R}^{d_L \times k}$, *and* $\mathbf{R} \in \mathbb{R}^{d_0 \times k}$. *To ensure the feasibility of the factorization for every point in* $\mathbb{R}^{d_L \times d_0}$, *we choose* $k \geq \min\{d_0, d_L\}$. *We define the set of minimizers as follows:*

$$\Omega \coloneqq \arg\min_{\mathbf{L},\mathbf{R}} \mathcal{L}\left(\mathbf{L}, \mathbf{R}\right) = \left\{(\mathbf{L}, \mathbf{R}) : \mathbf{M} = \mathbf{L}\mathbf{R}^\top\right\}. \tag{17}$$

*If* $(\mathbf{L}^*, \mathbf{R}^*) \in \Omega$, *then*

$$\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{L}^*, \mathbf{R}^*)) = 2(\sigma_{\max}(\mathbf{L}^*)^2 + \sigma_{\max}(\mathbf{R}^*)^2). \tag{18}$$

*Proof.* We have from Theorem 3.1 that $\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{L}, \mathbf{R}))$

$$= 2\sigma_{\max}\left(\mathbf{I} \otimes \mathbf{L}\mathbf{L}^\top + \mathbf{R}\mathbf{R}^\top \otimes \mathbf{I}\right) \tag{19}$$

$$= 2\left(\sigma_{\max}\left(\mathbf{I} \otimes \mathbf{L}\mathbf{L}^\top\right) + \sigma_{\max}\left(\mathbf{R}\mathbf{R}^\top \otimes \mathbf{I}\right)\right) \tag{20}$$

$$= 2\left(\sigma_{\max}\left(\mathbf{L}\right)^2 + \sigma_{\max}\left(\mathbf{R}\right)^2\right) \tag{21}$$

Using the fact from Horn and Johnson [1994, Theorem 4.4.5], we can obtain (20) from (19). Note that for any matrices $\mathbf{A}$ and $\mathbf{B}$, $\sigma_{\max}(\mathbf{A} \otimes \mathbf{B}) = \sigma_{\max}(\mathbf{A})\sigma_{\max}(\mathbf{B})$, and $\sigma_{\max}(\mathbf{A}\mathbf{A}^\top) = \sigma_{\max}(\mathbf{A}^\top\mathbf{A}) = \sigma_{\max}(\mathbf{A})^2$. Hence, we rewrite (20) as (21). □

We also provide an alternative, self-contained proof of this corollary in Appendix C.2. This result was also independently observed by Josz [2025].

## 4 Flatness

In this section, we reveal remarkable aspects of the loss-landscape geometry of general deep matrix factorization problems. First, we show that a minimizer of the deep overparameterized scalar factorization problem is flat if and only if, across all layers, the product of spectral norms of left and right intermediate networks equals $|m|^{1-\frac{1}{L}}$. This is a direct consequence of Corollary 3.2.

**Corollary 4.1.** *Consider the following objective function*

$$\mathcal{L}\left(\mathbf{w}\right) \coloneqq (m - \mathbf{w}_L^\top \mathbf{W}_{L-1} \cdots \mathbf{W}_2 \mathbf{w}_1)^2, \tag{22}$$

*where $m \in \mathbb{R} \setminus \{0\}$, $d_0 = d_L = 1$, $\mathbf{w}_L \in \mathbb{R}^{d_{L-1}}$ and $\mathbf{w}_1 \in \mathbb{R}^{d_1}$. For hidden factors (layers), i.e., for all $i \in \{2, 3, \cdots, L-1\}$, we have $\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$. Then, for any $\mathbf{w}^* \in \Omega$, we have $\mathbf{w}^* \in \Omega_F$ if and only if*

$$\sigma_{\max}(\mathbf{A}_k)\sigma_{\max}(\mathbf{B}_k) = |m|^{1-\frac{1}{L}} \quad \forall k \in [L], \tag{23}$$

*where $\mathbf{A}_k = \prod_{i=k+1}^{L} \mathbf{W}_i^*$ and $\mathbf{B}_k = \prod_{i=1}^{k-1} \mathbf{W}_i^*$.*

*Proof.* For a minimizer $\mathbf{w}^*$, let us assume that $\sigma_{\max}(\mathbf{A}_k)\sigma_{\max}(\mathbf{B}_k) = |m|^{1-\frac{1}{L}}$ for all $k \in [L]$. Then, by Corollary 3.2, $\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{w}^*)) = 2L \times |m|^{2(1-1/L)}$. This implies that $\mathbf{w}^*$ is flat [Mulayoff and Michaeli, 2020, also see Theorem 1]. Now, assume that $\mathbf{w}^*$ is flat. By using the fact from Mulayoff and Michaeli [2020, Theorem 2], for any flat minimum $\mathbf{w}^*$ and $k \in [L]$, we have

$$\sigma_{\max}(\mathbf{A}_k) = |m|^{1-\frac{k}{L}} \text{ and } \sigma_{\max}(\mathbf{B}_k) = |m|^{\frac{k-1}{L}}. \tag{24}$$

Hence, $\sigma_{\max}(\mathbf{A}_k)\sigma_{\max}(\mathbf{B}_k) = |m|^{1-\frac{1}{L}}$ for all $k \in [L]$. □

Second, we show that in depth-2 matrix factorization, a minimizer is flat if and only if the spectral norms of its factors equal $\sqrt{\sigma_{\max}(\mathbf{M})}$. This result is a direct consequence of Corollary 3.3 and the AM-GM inequality [see also Josz, 2025, Lemma 5].

**Corollary 4.2.** *Consider the following depth-2 matrix factorization objective*

$$\mathcal{L}(\mathbf{L}, \mathbf{R}) = \left\|\mathbf{M} - \mathbf{L}\mathbf{R}^\top\right\|_F^2, \tag{25}$$

*where $\mathbf{M} \in \mathbb{R}^{d_L \times d_0} \setminus \{\mathbf{0}\}$ is the target matrix, $\mathbf{L} \in \mathbb{R}^{d_L \times k}$, and $\mathbf{R} \in \mathbb{R}^{d_0 \times k}$ such that $k \geq \min\{d_0, d_L\}$. Then, for any $(\mathbf{L}^*, \mathbf{R}^*) \in \Omega$, we have $(\mathbf{L}^*, \mathbf{R}^*) \in \Omega_F$ if and only if*

$$\sigma_{\max}(\mathbf{L}^*) = \sigma_{\max}(\mathbf{R}^*) = \sqrt{\sigma_{\max}(\mathbf{M})}. \tag{26}$$

*Remark* 4.3. We discuss further implications of this equivalence in Section 5.

Now, we show that a minimizer of deep matrix factorization loss is flat if, across all layers, the product of spectral norms of the left and right intermediate factors equals $\sigma_{\max}(\mathbf{M})^{1-\frac{1}{L}}$.

**Corollary 4.4.** *Consider the optimization objective (5). If $\mathbf{w}^* \in \Omega$, then*

$$\lambda_{\max}\left(\nabla^2 \mathcal{L}(\mathbf{w}^*)\right) \leq 2 \sum_{k=1}^{L} \sigma_{\max}(\mathbf{A}_k)^2 \sigma_{\max}(\mathbf{B}_k)^2, \tag{27}$$

*where $\mathbf{A}_k = \prod_{i=k+1}^{L} \mathbf{W}_i^*$ and $\mathbf{B}_k = \prod_{i=1}^{k-1} \mathbf{W}_i^*$ (see Appendix B). This implies that if $\mathbf{w}^*$ satisfies $\sigma_{\max}(\mathbf{A}_k)\sigma_{\max}(\mathbf{B}_k) = \sigma_{\max}(\mathbf{M})^{1-\frac{1}{L}}$ for all $k \in [L]$, then $\mathbf{w}^*$ is flat.*

*Proof.* Mulayoff and Michaeli [2020] showed that for any $\mathbf{w}^* \in \Omega$,

$$\lambda_{\max}\left(\nabla^2 \mathcal{L}(\mathbf{w}^*)\right) \geq 2L \sigma_{\max}(\mathbf{M})^{2(1-1/L)}. \tag{28}$$

Now, assume that $\sigma_{\max}(\mathbf{A}_k)\sigma_{\max}(\mathbf{B}_k) = \sigma_{\max}(\mathbf{M})^{1-\frac{1}{L}}$ for all $k \in [L]$. Then by (27),

$$\lambda_{\max}\left(\nabla^2 \mathcal{L}(\mathbf{w}^*)\right) = 2L \sigma_{\max}(\mathbf{M})^{2(1-1/L)}. \tag{29}$$

This implies that $\mathbf{w}^*$ is flat [Mulayoff and Michaeli, 2020, Theorem 1]. □

*Remark* 4.5. The constant product of the spectral norms of the left and right intermediate factors across layers is not necessary for flatness. To see this, consider the depth-3 matrix factorization objective

$$\mathcal{L}(\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3) = \|\mathbf{M} - \mathbf{W}_3\mathbf{W}_2\mathbf{W}_1\|_F^2, \qquad \mathbf{M} = \operatorname{diag}\left(1, \tfrac{1}{10}\right), \tag{30}$$

where all factors are $2 \times 2$ matrices. Let $\mathbf{w}^*$ be the point given by

$$\mathbf{W}_3^* = \operatorname{diag}\left(1, \tfrac{1}{2}\right), \quad \mathbf{W}_2^* = \operatorname{diag}\left(1, \tfrac{2}{11}\right), \quad \mathbf{W}_1^* = \operatorname{diag}\left(1, \tfrac{11}{10}\right). \tag{31}$$

Since $\mathbf{W}_3^*\mathbf{W}_2^*\mathbf{W}_1^* = \mathbf{M}$, the point $\mathbf{w}^*$ is a global minimizer. The corresponding left and right intermediate factors are

$$\begin{aligned} &\mathbf{B}_1 = \mathbf{I}_{d_0 \times d_0}, && \mathbf{B}_2 = \mathbf{W}_1^* = \operatorname{diag}\left(1, \tfrac{11}{10}\right), && \mathbf{B}_3 = \mathbf{W}_2^*\mathbf{W}_1^* = \operatorname{diag}\left(1, \tfrac{1}{5}\right), \\ &\mathbf{A}_1 = \mathbf{W}_3^*\mathbf{W}_2^* = \operatorname{diag}\left(1, \tfrac{1}{11}\right), && \mathbf{A}_2 = \mathbf{W}_3^* = \operatorname{diag}\left(1, \tfrac{1}{2}\right), && \mathbf{A}_3 = \mathbf{I}_{d_L \times d_L}. \end{aligned} \tag{32}$$

By Theorem 3.1, $\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{w}^*)) = 6$. Mulayoff and Michaeli [2020] showed that for every $\mathbf{w} \in \Omega_F$, $\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{w})) = 2L\,\sigma_{\max}(\mathbf{M})^{2(1-1/L)}$, which equals 6 for $\mathbf{M} = \operatorname{diag}\left(1, \frac{1}{10}\right)$ and $L = 3$. Since $\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{w}^*))$ attains this value, $\mathbf{w}^*$ is flat. However,

$$\sigma_{\max}(\mathbf{A}_1)\,\sigma_{\max}(\mathbf{B}_1) = 1, \quad \sigma_{\max}(\mathbf{A}_2)\,\sigma_{\max}(\mathbf{B}_2) = \frac{11}{10}, \quad \sigma_{\max}(\mathbf{A}_3)\,\sigma_{\max}(\mathbf{B}_3) = 1. \tag{33}$$

Therefore, flatness *does not imply* that $\sigma_{\max}(\mathbf{A}_k)\,\sigma_{\max}(\mathbf{B}_k)$ is constant across layers.

## 5 Discussion

In this section, we examine the claims made by Ding et al. [2024] and Ghosh et al. [2025] that flat minima are balanced in depth-2 matrix factorization (see Definition 2.1). We then discuss how sharpness measures can be misleading in loss landscape analysis. We conclude by addressing the limitations of our analysis.

**Flat Minima Need Not Be Balanced.** First, we examine the latter work, in which Ghosh et al. [2025] studied the learning dynamics of deep matrix factorization under the squared-error loss beyond the edge of stability. They claim that *flat minima correspond to balanced minima* in deep matrix factorization problems [Ghosh et al., 2025, Proof of Lemma 4]. Globally, however, we can find minimizers that are **flat but not Frobenius-norm balanced**. As we showed in Remark 2.3, Frobenius-norm balancedness is necessary for being balanced; therefore, *flat minima need not be balanced*. To see this, consider a depth-2 matrix factorization problem, i.e.,

$$\min_{\mathbf{L} \in \mathbb{R}^{m \times r} \mathbf{R} \in \mathbb{R}^{n \times r}} \left\{ \mathcal{L}(\mathbf{L}, \mathbf{R}) \coloneqq \left\| \mathbf{M} - \mathbf{L}\mathbf{R}^\top \right\|_F^2 \right\}. \tag{34}$$

Suppose $m = n = r = 3$ and $\mathbf{M} = \operatorname{diag}(11, 2, 1)$. We know that for any flat minimum $(\mathbf{L}', \mathbf{R}') \in \Omega_F$, $\lambda_{\max}(\nabla^2(\mathcal{L}(\mathbf{L}', \mathbf{R}'))) = 4 \times \sigma_{\max}(\mathbf{M}) = 44$. By definition, any minimum whose sharpness is equal to 44 is a flat minimum. Let's investigate a specific minimizer $(\mathbf{L}^*, \mathbf{R}^*)$ such that $\mathbf{L}^* = \operatorname{diag}(\sqrt{11}, \frac{2}{3}, 1)$ and $\mathbf{R}^* = \operatorname{diag}(\sqrt{11}, 3, 1)$. Using Corollary 3.3, $\lambda_{\max}(\nabla^2(\mathcal{L}(\mathbf{L}^*, \mathbf{R}^*))) = 44$, which means that $(\mathbf{L}^*, \mathbf{R}^*)$ is a flat minimum. On the other hand, by definition, $(\mathbf{L}^*, \mathbf{R}^*)$ is not Frobenius-norm balanced. Therefore, a flat minimum is not necessarily balanced. Moreover, we provide a counterexample in Appendix D to show that the singular vector stationary set characterization of Ghosh et al. [2025, Proposition 1] does not hold in general.

**Sharpness Measures Are Delusive.** Ding et al. [2024] showed that flat minima recover the ground truth matrix in low-rank matrix recovery. Furthermore, they showed that *norm-minimal*, *balanced* and *flat* solutions coincide in depth-2 matrix factorization [Ding et al., 2024, Lemma 2.2]. They measured the sharpness of a minimum using the *scaled trace*.[2] Under this measure, flat minima coincide with balanced minima; however, when worst-case sharpness is used as the sharpness measure, flat minima need not even satisfy $\|\mathbf{L}^*\|_F = \|\mathbf{R}^*\|_F$. Hence, under the worst-case sharpness measure, *flat minima need not be balanced*. In fact, we showed that flat minima are spectral-norm balanced when the maximum Hessian eigenvalue is used as the sharpness measure (Corollary 4.2). This means that a loss landscape analysis in non-convex optimization problems using a specific sharpness measure might not lead to the same inferences as those made by using a different metric.

**The Necessity of an Absolute Sharpness Measure.** There is still no consensus in the literature regarding the definition of flatness. For instance, while several works define flat minima as global minimizers that minimize the maximum eigenvalue of the Hessian of the loss [Mulayoff and Michaeli, 2020, Liu et al., 2021, Marion and Chizat, 2024], others define them as global minimizers that minimize the trace of the Hessian [Dinh et al., 2017, Gatmiry et al., 2023]. Even a scaled version of the Hessian trace is designed to define flat minima in depth-2 matrix factorization problems [Ding et al., 2024], as we mentioned before. As we have shown in Section 4 and discussed above, different sharpness measures can lead to contradictory interpretations of balanced minima. Thus, the necessity of a robust sharpness measure that consolidates the interpretations from existing ones is a matter of urgency. Addressing this concern, Josz [2025] showed how minimizing the maximum Hessian eigenvalue over the solution set can disregard the higher-order variation near minimizers and result in misleading interpretations of flat minima. Josz [2025] defined a sharpness measure based on the maximal variation around a point. They showed that, for a smooth loss $\mathcal{L}$, a minimizer $\mathbf{w}^*$ is flat only if it is a local minimum of $\lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{w}))$ subject to $\mathcal{L}(\mathbf{w}) = \mathcal{L}(\mathbf{w}^*)$, and that, in depth-2 matrix factorization with squared-error loss, flat minima are exactly the global minima that minimize the worst-case sharpness measure.

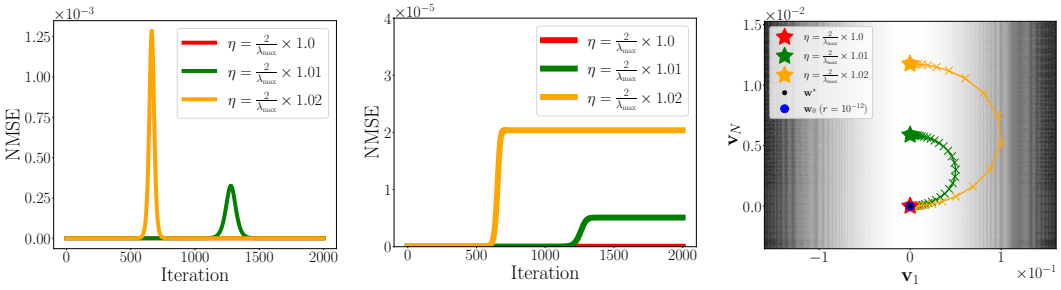
**Limitations.** The dynamical stability analysis relies on how accurate the quadratic approximation of the loss function is in the $\delta$-neighborhood of $\mathbf{w}^*$ [Wu et al., 2018] (also see Definition 1.1). For loss functions whose Hessian is zero at a minimizer, this analysis fails [Wu et al., 2018]. For instance, Jules et al. [2023] observed that for a cross-entropy loss, the quadratic approximation of the training loss near a minimizer is fundamentally inadequate due to the exponential nature of the decision boundary. In contrast, for a nonzero target matrix $\mathbf{M}$, the squared-error loss ensures that the Hessian at any minimizer is nonzero (also see Mulayoff and Michaeli [2020, Lemma 3]). Therefore, our analysis cannot be extended to other loss functions whose Hessians vanish at their global minima. Moreover, the condition required to guarantee the feasibility of factorization for any target matrix $\mathbf{M}$, i.e., Equation (6), limits the applicability of our work in more general settings. Without this condition, our analysis holds only for target matrices $\mathbf{M}$ that admit an interpolating factorization. Finally, when $\widehat{\mathbf{\Sigma}}_x$ is full rank, the deep matrix factorization/deep linear network training problem is not well suited for analyzing the relationship between generalization and the geometric properties of the landscape. This is because, in this case, all minimizers have exactly the same generalization ability, which makes such an analysis inapplicable in our setting.

## 6 Experiments

We initialize GD in a $\delta$-neighborhood of $\mathbf{w}^*$ with $\delta$ of order $10^{-15}$ to $10^{-9}$ so that the linearized GD dynamics approximate the true GD dynamics well. To observe the escape phenomenon clearly, we choose the perturbation direction for the initial point to be the eigenvector of $\nabla^2\mathcal{L}(\mathbf{w}^*)$ corresponding to the largest eigenvalue, so as to avoid choosing a direction that is orthogonal to the eigenspace corresponding to the eigenvalues larger than $2/\eta$.

To measure the distance between the iterates and the minimizer, we plot the normalized squared $\ell^2$-norm of $\mathbf{w}_t - \mathbf{w}^*$ at each iteration as shown in Fig. 1b. We say that the escape phenomenon occurred if the final distance from the minimizer, i.e., $\|\mathbf{w}_T - \mathbf{w}^*\|_2$, is significantly larger than the

[2] For the deep matrix factorization problem in (5), the scaled trace of the Hessian matrix of the squared-error loss takes the form $\mathrm{str}(\nabla^2\mathcal{L}(\mathbf{w}^*)) = 2\|\prod_{i=2}^{L}\mathbf{W}_i^*\|_F^2 + 2\left(\sum_{k=2}^{L-1}\|\prod_{i=k+1}^{L}\mathbf{W}_i^*\|_F^2\|\prod_{i=1}^{k-1}\mathbf{W}_i^*\|_F^2\right) + 2\|\prod_{i=1}^{L-1}\mathbf{W}_i^*\|_F^2$ for any $\mathbf{w}^* \in \Omega$.




(a) Normalized training error across iterations, i.e., $\mathcal{L}(\mathbf{w}_t)/\|\mathbf{M}\|_F^2$.

(b) $\|\mathbf{w}_t - \mathbf{w}^*\|_2^2/\|\mathbf{w}^*\|_2^2$ across iterations.

(c) Trajectories of GD on the contour map of the loss landscape.

Figure 1: The setting is a depth-3 matrix factorization of a random $10 \times 20$ matrix with Gaussian factors $\mathbf{W}_3 \in \mathbb{R}^{10\times 20}$ and $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{20\times 20}$. $\mathbf{v}_1$ and $\mathbf{v}_N$ are the eigenvectors of $\nabla^2\mathcal{L}(\mathbf{w}^*)$ that correspond to its maximum and minimum eigenvalues respectively. $\lambda_{\max}$ is computed via the closed-form expression in Theorem 3.1. GD runs with different step sizes are initialized within $10^{-12}$ of the minimizer $\mathbf{w}^*$ along $\mathbf{v}_1$. The contour plot in (c) projects the parameter space onto $\mathrm{span}\{\mathbf{v}_1, \mathbf{v}_N\}$ (Appendix E.1); darker (brighter) regions indicate higher (lower) loss. See Appendix E.2 for details.

initialization radius. In other words, if the training dynamics move the iterates far from the minimizer relative to how close they started, this indicates that the minimizer is unstable and the dynamics have escaped its vicinity. We observe that if $\eta > 2/\lambda_{\max}$, where $\lambda_{\max} := \lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{w}^*))$, GD escapes from the minimum as shown in Fig. 1. On the other hand, if $\eta \leq 2/\lambda_{\max}$, then GD converges. A catapult in the training error shown in Fig. 1a indicates GD's escape from the basin of a minimum, after which it eventually converges to another minimum unless the step size is very large [Marion and Chizat, 2024]. To the best of our knowledge, this is the first empirical verification of the escape phenomenon during gradient-based training near a minimizer of a deep matrix factorization problem.

## 7 Conclusion

In this paper, we derived an exact expression for the maximum eigenvalue of the Hessian of the squared-error loss in overparameterized deep matrix factorization at any minimizer, and showed that it simplifies substantially in the depth-2 matrix factorization and deep overparameterized scalar factorization settings. We further showed that flat minima in depth-2 matrix factorization are spectral-norm balanced, and that a minimizer of the deep overparameterized scalar factorization loss is flat if and only if, across all layers, the product of the spectral norms of the left and right intermediate factors equals a constant determined by the depth and the magnitude of the target. In deep matrix factorization, the same condition is sufficient but not necessary for flatness. To complement our theory, we conducted GD experiments that crucially rely on our exact expression of the sharpness to observe the escape phenomenon during training. Finally, we discussed how sharpness measures can be misleading for loss landscape analysis, highlighting the need for robust sharpness notions that reconcile existing interpretations. However, we did not compare the computational complexity of our closed-form expressions, particularly those for depth-2 matrix and deep overparameterized scalar factorization, with that of the Lanczos algorithm [Lanczos, 1950]. Although it is clear that our closed-form expressions for these settings do not require forming the Hessian matrix to compute its maximum eigenvalue, their performance against iterative methods based on matrix-vector products has not been investigated in this work. We leave this as future work.

## Acknowledgments

The authors acknowledge support from a Hellman Fellowship from the University of California. The counterexamples provided in Remark 4.5 and Appendix D were suggested by Claude Fable 5 and ChatGPT-6 Astra, respectively, and the authors take full responsibility for their correctness. Beyond this, GPT-4o, GPT-5, GPT-6.1 Sol, and Claude Opus 5.5 were used as spell checkers, grammar suggestion tools, editing aids, and code assistants.

## References


S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. *Advances in neural information processing systems*, 32, 2019.

P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

M. Belkin, A. Rakhlin, and A. B. Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd international conference on artificial intelligence and statistics*, pages 1611–1619. PMLR, 2019.

D. Chemnitz and M. Engel. Characterizing dynamical stability of stochastic gradient descent in overparameterized learning. *Journal of Machine Learning Research*, 26(134):1–46, 2025.

H.-H. Chou, C. Gieshoff, J. Maly, and H. Rauhut. Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *Applied and Computational Harmonic Analysis*, 68: 101595, 2024.

J. Cohen, S. Kaur, Y. Li, J. Z. Kolter, and A. Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

L. Ding, D. Drusvyatskiy, M. Fazel, and Z. Harchaoui. Flat minima generalize for low-rank matrix recovery. *Information and Inference: A Journal of the IMA*, 13(2):iaae009, 2024.

L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

S. Frei, N. S. Chatterji, and P. Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pages 2668–2703. PMLR, 2022.

K. Gatmiry, Z. Li, T. Ma, S. Reddi, S. Jegelka, and C.-Y. Chuang. What is the inductive bias of flatness regularization? a study of deep matrix factorization models. *Advances in Neural Information Processing Systems*, 36:28040–28052, 2023.

A. Ghosh, S. M. Kwon, R. Wang, S. Ravishankar, and Q. Qu. Learning dynamics of deep matrix factorization beyond the edge of stability. In *The Thirteenth International Conference on Learning Representations*, 2025.

I. J. Goodfellow, O. Vinyals, and A. M. Saxe. Qualitatively characterizing neural network optimization problems. *arXiv preprint arXiv:1412.6544*, 2014.

S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in neural information processing systems*, 31, 2018.

S. Hochreiter and J. Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Cambridge university press, 1994.

Z. Ji and M. Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, and S. Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020.

C. Josz. On the geometry of flat minima. *arXiv preprint arXiv:2509.11386*, 2025.

T. Jules, G. Brener, T. Kachman, N. Levi, and Y. Bar-Sinai. Charting the topography of the neural network landscape with thermal-like noise. *arXiv preprint arXiv:2304.01335*, 2023.

K. Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of research of the National Bureau of Standards*, 45(4):255–282, 1950.

T. Laurent and J. Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *International conference on machine learning*, pages 2902–2907. PMLR, 2018.

A. Lewkowycz, Y. Bahri, E. Dyer, J. Sohl-Dickstein, and G. Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

S. Liang and G. Montúfar. Gradient descent with large step sizes: Chaos and fractal convergence region. In *International Conference on Learning Representations*, 2026.

T. Liang, D. Qiao, Y.-X. Wang, and R. Parhi. Stable minima of ReLU neural networks suffer from the curse of dimensionality: The neural shattering phenomenon. In *Advances in Neural Information Processing Systems*, 2025.

T. Liu, Y. Li, S. Wei, E. Zhou, and T. Zhao. Noisy gradient descent converges to flat minima for nonconvex matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, 2021.

C. Ma and L. Ying. On linear stability of SGD and input-smoothness of neural networks. *Advances in Neural Information Processing Systems*, 34:16805–16817, 2021.

P. Marion and L. Chizat. Deep linear networks for regression are implicitly regularized towards flat minima. *Advances in Neural Information Processing Systems*, 37:76848–76900, 2024.

R. Mulayoff and T. Michaeli. Unique properties of flat minima in deep networks. In *International conference on machine learning*, pages 7108–7118. PMLR, 2020.

R. Mulayoff, T. Michaeli, and D. Soudry. The implicit bias of minima stability: A view from function space. *Advances in Neural Information Processing Systems*, 34:17749–17761, 2021.

M. S. Nacson, R. Mulayoff, G. Ongie, T. Michaeli, and D. Soudry. The implicit bias of minima stability in multivariate shallow ReLU networks. In *ICLR*, 2023.

K. Nar and S. Sastry. Step size matters in deep learning. *Advances in Neural Information Processing Systems*, 31, 2018.

B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

D. Qiao, K. Zhang, E. Singh, D. Soudry, and Y.-X. Wang. Stable minima cannot overfit in univariate ReLU networks: Generalization by large step sizes. *Advances in Neural Information Processing Systems*, 37:94163–94208, 2024.

S. P. Singh and T. Hofmann. Closed form of the Hessian spectrum for some neural networks. In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024.

S. P. Singh, W. Ormaniec, and T. Hofmann. Cracking the Hessian: Closed-form Hessian spectra for fundamental neural networks, 2026. URL https://openreview.net/forum?id=XfO8npR1fb.

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

S. H. Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Chapman and Hall/CRC, 2024.

M. Trager, K. Kohn, and J. Bruna. Pure and spurious critical points: a geometric study of linear networks. *arXiv preprint arXiv:1910.01671*, 2019.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Y. Wang, M. Chen, T. Zhao, and M. Tao. Large learning rate tames homogeneity: Convergence and balancing effect. In *International Conference on Learning Representations*, 2022.

L. Wu, C. Ma, et al. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, 31, 2018.

C. Yun, S. Krishnan, and H. Mobahi. A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*, 2021.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

X. Zhu, Z. Wang, X. Wang, M. Zhou, and R. Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *The Eleventh International Conference on Learning Representations*, 2023.

# A Matrix Calculus Background

## A.1 Gâteaux Derivatives

Let $f : \mathbb{R}^{K\times L} \to \mathbb{R}$ be a differentiable function with continuous first- and second-order derivatives on $\mathbb{R}^{K\times L}$. Our objective is to derive closed-form expressions for the first- and second-order directional derivatives of $f$ in the direction of $\mathbf{U} \in \mathbb{R}^{K\times L}$, where $\|\mathbf{U}\|_F < \infty$, denoted respectively by $D_{\mathbf{U}} f(\mathbf{X})$ and $D^2_{\mathbf{U}} f(\mathbf{X})$. By the limit definition of the derivative, the first derivative of $f(\mathbf{X})$ with respect to each entry of $\mathbf{X}$ can be expressed as follows:

$$\frac{\partial f(\mathbf{X})}{\partial X_{ij}} = \lim_{\Delta t\to 0} \frac{f\left(\mathbf{X} + \Delta t \mathbf{e}_i \mathbf{e}_j^\top\right) - f(\mathbf{X})}{\Delta t}, \tag{35}$$

where $\mathbf{e}_i$ is the $i$th standard basis vector of $\mathbb{R}^K$ and $\mathbf{e}_j$ is the $j$th standard basis vector of $\mathbb{R}^L$. If the limit in (35) exists, then by substitution of variables

$$\frac{\partial f(\mathbf{X})}{\partial X_{ij}} U_{ij} = \lim_{\Delta t\to 0} \frac{f\left(\mathbf{X} + \Delta t U_{ij} \mathbf{e}_i \mathbf{e}_j^\top\right) - f(\mathbf{X})}{\Delta t}. \tag{36}$$

By definition, the total change in $f(\mathbf{X})$ in the direction of $\mathbf{U}$ is the sum of changes due to each entry of $\mathbf{X}$. Then

$$D_{\mathbf{U}} f(\mathbf{X}) = \sum_{(i,j)\in[K]\times[L]} \frac{\partial f(\mathbf{X})}{\partial X_{ij}} U_{ij} \tag{37}$$

$$= \sum_{(i,j)\in[K]\times[L]} \lim_{\Delta t\to 0} \frac{f(\mathbf{X} + \Delta t U_{ij} \mathbf{e}_i \mathbf{e}_j^\top) - f(\mathbf{X})}{\Delta t} \tag{38}$$

$$= \lim_{\Delta t\to 0} \frac{f(\mathbf{X} + \Delta t \mathbf{U}) - f(\mathbf{X})}{\Delta t}. \tag{39}$$

We can rewrite (39) as follows:

$$\lim_{\Delta t\to 0} \frac{f\left(\mathbf{X} + (\Delta t + t)\,\mathbf{U}\right) - f(\mathbf{X} + t\mathbf{U})}{\Delta t}\bigg|_{t=0}. \tag{40}$$

This expression can be rewritten as

$$\frac{\partial f(\mathbf{X} + t\mathbf{U})}{\partial t}\bigg|_{t=0}. \tag{41}$$

This is known as the *Gâteaux derivative*, which represents the change in $f(\mathbf{X})$ under a perturbation in the direction of $\mathbf{U}$. By the same reasoning, we obtain the following result.

**Lemma A.1.** *The second directional derivative of $f$ at $\mathbf{X}$ in the direction $\mathbf{U} \in \mathbb{R}^{K\times L}$ is given by*

$$D^2_{\mathbf{U}} f(\mathbf{X}) = \frac{\partial^2}{\partial t^2} f(\mathbf{X} + t\mathbf{U})\,\Big|_{t=0}. \tag{42}$$

*Proof.* We can express the derivative of $\frac{\partial f(\mathbf{X})}{\partial X_{ij}}$ with respect to each entry of $\mathbf{X}$, using the limit definition of the derivative, as follows:

$$\frac{\partial^2 f(\mathbf{X})}{\partial X_{kl}\partial X_{ij}} = \frac{\partial}{\partial X_{kl}}\left(\frac{\partial f(\mathbf{X})}{\partial X_{ij}}\right) = \lim_{\Delta t\to 0} \frac{\partial f(\mathbf{X} + \Delta t \mathbf{e}_i \mathbf{e}_j^\top) - \partial f(\mathbf{X})}{\partial X_{kl}\Delta t} \quad \forall (k,l) \in [K]\times[L] \tag{43}$$

which is equal to

$$\lim_{\Delta h, \Delta t\to 0} \frac{f(\mathbf{X} + \Delta t \mathbf{e}_i \mathbf{e}_j^\top + \Delta h \mathbf{e}_k \mathbf{e}_l^\top) - f(\mathbf{X} + \Delta t \mathbf{e}_i \mathbf{e}_j^\top) - f(\mathbf{X} + \Delta h \mathbf{e}_k \mathbf{e}_l^\top) + f(\mathbf{X})}{\Delta h \Delta t}. \tag{44}$$

By using the substitution of variables as in (36),

$$\frac{\partial^2 f(\mathbf{X})}{\partial X_{kl}\partial X_{ij}} U_{ij} U_{kl} = \frac{\partial}{\partial X_{kl}}\left(\frac{\partial f(\mathbf{X})}{\partial X_{ij}} U_{ij}\right) U_{kl} = \lim_{\Delta t\to 0} \frac{\partial f(\mathbf{X} + \Delta t U_{ij} \mathbf{e}_i \mathbf{e}_j^\top) - \partial f(\mathbf{X})}{\partial X_{kl}\Delta t} U_{kl}. \tag{45}$$

Equivalently,

$$\lim_{\Delta h,\Delta t\to 0}\frac{1}{\Delta h\,\Delta t}\Big[f\big(\mathbf{X}+\Delta t U_{ij}\mathbf{e}_i\mathbf{e}_j^\top+\Delta h U_{kl}\mathbf{e}_k\mathbf{e}_l^\top\big) - f\big(\mathbf{X}+\Delta t U_{ij}\mathbf{e}_i\mathbf{e}_j^\top\big)-f\big(\mathbf{X}+\Delta h U_{kl}\mathbf{e}_k\mathbf{e}_l^\top\big)+f(\mathbf{X})\Big], \tag{46}$$

which can be proved by substitution of variables in (45). In turn, the second order differential in the direction $\mathbf{U}\in\mathbb{R}^{K\times L}$ is

$$D^2_{\mathbf{U}}f(\mathbf{X})=\sum_{i,j}\sum_{k,l}\frac{\partial^2 f(\mathbf{X})}{\partial X_{kl}\partial X_{ij}}U_{ij}U_{kl}=\langle\nabla\langle\nabla f(\mathbf{X}),\mathbf{U}\rangle,\mathbf{U}\rangle, \tag{47}$$

where

$$\nabla f(\mathbf{X})=\begin{bmatrix}\frac{\partial f(\mathbf{X})}{\partial X_{11}} & \frac{\partial f(\mathbf{X})}{\partial X_{12}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial X_{1L}}\\ \frac{\partial f(\mathbf{X})}{\partial X_{21}} & \frac{\partial f(\mathbf{X})}{\partial X_{22}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial X_{2L}}\\ \vdots & \ddots & \cdots & \vdots\\ \frac{\partial f(\mathbf{X})}{\partial X_{K1}} & \frac{\partial f(\mathbf{X})}{\partial X_{K2}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial X_{KL}}\end{bmatrix}\in\mathbb{R}^{K\times L}. \tag{48}$$

Equivalently,

$$D^2_{\mathbf{U}}f(\mathbf{X})=\sum_{k,l}\lim_{\Delta t\to 0}\frac{\partial f(\mathbf{X}+\Delta t\mathbf{U})-\partial f(\mathbf{X})}{\partial X_{kl}\Delta t}U_{kl} \tag{49}$$

$$=\lim_{\Delta t\to 0}\frac{f(\mathbf{X}+2\Delta t\mathbf{U})-2f(\mathbf{X}+\Delta t\mathbf{U})+f(\mathbf{X})}{\Delta t^2} \tag{50}$$

$$=\frac{\partial^2}{\partial t^2}f(\mathbf{X}+t\mathbf{U})\Big|_{t=0}. \tag{51}$$

□

### A.2 Second Directional Derivatives and Maximum Eigenvalue

Consider the following objective function for our real-valued matrix-variable function $\mathcal{L}$ such that

$$\mathcal{L}(\mathbf{W}_1,\cdots,\mathbf{W}_L)=\|\mathbf{M}-\mathbf{W}_L\mathbf{W}_{L-1}\cdots\mathbf{W}_1\|_F^2, \tag{52}$$

where $\mathbf{W}_i\in\mathbb{R}^{d_i\times d_{i-1}}$ for all $i\in[L]$, and $\mathbf{M}\in\mathbb{R}^{d_L\times d_0}$. In this setting, we define the maximum eigenvalue of the Hessian $\nabla^2\mathcal{L}(\mathbf{W}_1,\cdots,\mathbf{W}_L)$ at an arbitrary point in the parameter space as follows:

$$\lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{W}_1,\cdots,\mathbf{W}_L))=\max_{\substack{\mathbf{U}_1,\mathbf{U}_2,\cdots,\mathbf{U}_L:\\ \sum_{i=1}^L\|\mathbf{U}_i\|_F^2=1}}\frac{d^2}{dt^2}\mathcal{L}(\mathbf{W}_1+t\mathbf{U}_1,\cdots,\mathbf{W}_L+t\mathbf{U}_L)\Big|_{t=0}. \tag{53}$$

This is the generalization of the Rayleigh quotient to the case where the Hessian is represented as a tensor and its eigenvectors take the form of matrices. This leads to the following lemma.

**Lemma A.2.** *For any $[\mathbf{W}_1^*,\cdots,\mathbf{W}_L^*]$ such that $\mathbf{M}=\prod_{j=1}^L\mathbf{W}_j^*$, the second directional derivative of $\mathcal{L}$ at $[\mathbf{W}_1^*,\cdots,\mathbf{W}_L^*]$ along $[\mathbf{U}_1,\cdots,\mathbf{U}_L]$ is given by*

$$\nabla^2\mathcal{L}(\mathbf{W}_1^*,\cdots,\mathbf{W}_L^*)[\mathbf{U}_1,\cdots,\mathbf{U}_L]=2\left\|\sum_{i=1}^L\left[\left(\prod_{j=i+1}^L\mathbf{W}_j^*\right)\mathbf{U}_i\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)\right]\right\|_F^2. \tag{54}$$

*Proof.* We can rewrite (52) using the definition of the *Frobenius inner product*

$$\mathcal{L}(\mathbf{W}_1,\cdots,\mathbf{W}_L)=\left\langle\mathbf{M}-\prod_{i=1}^L\mathbf{W}_i,\mathbf{M}-\prod_{i=1}^L\mathbf{W}_i\right\rangle. \tag{55}$$

Then

$$\mathcal{L}(\mathbf{W}_1 + t\mathbf{U}_1, \cdots, \mathbf{W}_L + t\mathbf{U}_L) = \left\langle \mathbf{M} - \prod_{i=1}^{L}(\mathbf{W}_i + t\mathbf{U}_i), \mathbf{M} - \prod_{i=1}^{L}(\mathbf{W}_i + t\mathbf{U}_i) \right\rangle. \tag{56}$$

Let's define $g : \mathbb{R} \to \mathbb{R}^{d_L \times d_0}$ such that

$$g(t) = \mathbf{M} - \prod_{i=1}^{L}(\mathbf{W}_i + t\mathbf{U}_i), \quad \mathcal{L}(\mathbf{W}_1 + t\mathbf{U}_1, \cdots, \mathbf{W}_L + t\mathbf{U}_L) = \langle g(t), g(t) \rangle. \tag{57}$$

First, we need to differentiate $\mathcal{L}(\mathbf{W}_1 + t\mathbf{U}_1, \cdots, \mathbf{W}_L + t\mathbf{U}_L)$ with respect to $t$. Using the fact that $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$, which simplifies the differentiation,

$$\frac{d}{dt}\mathcal{L}(\mathbf{W}_1 + t\mathbf{U}_1, \cdots, \mathbf{W}_L + t\mathbf{U}_L) = 2 \langle g(t), g'(t) \rangle. \tag{58}$$

$$\frac{d^2}{dt^2}\mathcal{L}(\mathbf{W}_1 + t\mathbf{U}_1, \cdots, \mathbf{W}_L + t\mathbf{U}_L) = 2 \langle g'(t), g'(t) \rangle + 2 \langle g(t), g''(t) \rangle. \tag{59}$$

Then, the second directional derivative $\nabla^2 \mathcal{L}(\mathbf{W}_1, \cdots, \mathbf{W}_L)[\mathbf{U}_1, \cdots, \mathbf{U}_L]$ equals

$$\frac{d^2}{dt^2}\mathcal{L}(\mathbf{W}_1 + t\mathbf{U}_1, \cdots, \mathbf{W}_L + t\mathbf{U}_L)\Big|_{t=0} = 2 \langle g'(0), g'(0) \rangle + 2 \langle g(0), g''(0) \rangle. \tag{60}$$

Differentiating $g(t)$ gives

$$g(0) = \mathbf{M} - \prod_{i=1}^{L} \mathbf{W}_i, \tag{61}$$

$$g'(0) = -\sum_{i=1}^{L} \left[ \left( \prod_{j=i+1}^{L} \mathbf{W}_j \right) \mathbf{U}_i \left( \prod_{j=1}^{i-1} \mathbf{W}_j \right) \right], \tag{62}$$

$$g''(0) = -2 \sum_{1 \le k < i \le L} \left[ \left( \prod_{j=i+1}^{L} \mathbf{W}_j \right) \mathbf{U}_i \left( \prod_{j=k+1}^{i-1} \mathbf{W}_j \right) \mathbf{U}_k \left( \prod_{j=1}^{k-1} \mathbf{W}_j \right) \right]. \tag{63}$$

Therefore,

$$\nabla^2 \mathcal{L}(\mathbf{W}_1, \cdots, \mathbf{W}_L)[\mathbf{U}_1, \mathbf{U}_2, \cdots, \mathbf{U}_L] = \tag{64}$$

$$2 \left\langle \sum_{i=1}^{L} \left[ \left( \prod_{j=i+1}^{L} \mathbf{W}_j \right) \mathbf{U}_i \left( \prod_{j=1}^{i-1} \mathbf{W}_j \right) \right], \sum_{i=1}^{L} \left[ \left( \prod_{j=i+1}^{L} \mathbf{W}_j \right) \mathbf{U}_i \left( \prod_{j=1}^{i-1} \mathbf{W}_j \right) \right] \right\rangle \tag{65}$$

$$- 4 \left\langle \mathbf{M} - \prod_{i=1}^{L} \mathbf{W}_i, \sum_{1 \le k < i \le L} \left[ \left( \prod_{j=i+1}^{L} \mathbf{W}_j \right) \mathbf{U}_i \left( \prod_{j=k+1}^{i-1} \mathbf{W}_j \right) \mathbf{U}_k \left( \prod_{j=1}^{k-1} \mathbf{W}_j \right) \right] \right\rangle. \tag{66}$$

Note that for any $\mathbf{w}^* = \mathrm{vec}([\mathbf{W}_1^*, \cdots, \mathbf{W}_L^*]) \in \Omega$, $\mathbf{M} = \prod_{j=1}^{L} \mathbf{W}_j^*$. Hence, for any global minimum

$$\nabla^2 \mathcal{L}(\mathbf{W}_1^*, \cdots, \mathbf{W}_L^*)[\mathbf{U}_1, \mathbf{U}_2, \cdots, \mathbf{U}_L] = \tag{67}$$

$$2 \left\langle \sum_{i=1}^{L} \left[ \left( \prod_{j=i+1}^{L} \mathbf{W}_j^* \right) \mathbf{U}_i \left( \prod_{j=1}^{i-1} \mathbf{W}_j^* \right) \right], \sum_{i=1}^{L} \left[ \left( \prod_{j=i+1}^{L} \mathbf{W}_j^* \right) \mathbf{U}_i \left( \prod_{j=1}^{i-1} \mathbf{W}_j^* \right) \right] \right\rangle. \tag{68}$$

□

## B Alternative, Self-Contained Proof of Corollary 3.2

*Proof.* According to Lemma A.2,

$$\lambda_{\max}\left(\nabla^2\mathcal{L}(\mathbf{w}^*)\right) = \max_{\substack{\mathbf{U}_1,\mathbf{U}_2,\ldots,\mathbf{U}_L:\\ \sum_{i=1}^L \|\mathbf{U}_i\|_F^2=1}} 2\left\|\sum_{i=1}^{L}\left[\left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)\mathbf{U}_i\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)\right]\right\|_F^2 \tag{69}$$

$$\leq \max_{\substack{\mathbf{U}_1,\mathbf{U}_2,\ldots,\mathbf{U}_L:\\ \sum_{i=1}^L \|\mathbf{U}_i\|_F^2=1}} 2\left(\sum_{i=1}^{L}\left\|\left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)\mathbf{U}_i\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)\right\|_F\right)^2 \tag{70}$$

$$= \max_{\substack{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_L:\\ \sum_{i=1}^L \|\mathbf{u}_i\|_2^2=1}} 2\left(\sum_{i=1}^{L}\left\|\left[\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)^\top\otimes\left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)\right]\mathbf{u}_i\right\|_2\right)^2 \tag{71}$$

$$\leq \max_{\substack{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_L:\\ \sum_{i=1}^L \|\mathbf{u}_i\|_2^2=1}} 2\left(\sum_{i=1}^{L}\sigma_{\max}\left(\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)^\top\otimes\left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)\right)\|\mathbf{u}_i\|_2\right)^2 \tag{72}$$

$$= \max_{\substack{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_L:\\ \sum_{i=1}^L \|\mathbf{u}_i\|_2^2=1}} 2\left(\sum_{i=1}^{L}\sigma_{\max}\left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)\sigma_{\max}\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)\|\mathbf{u}_i\|_2\right)^2, \tag{73}$$

where $\mathbf{U}_i \in \mathbb{R}^{d_i\times d_{i-1}}$ and $\mathbf{u}_i \in \mathbb{R}^{d_i d_{i-1}}$. We can upper bound the right-hand side of (69) by using the *triangle inequality*. By applying the *vectorization trick* of the Kronecker product, we can rewrite (70). Then, noting the fact that for any matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{A}\mathbf{x}\|_2 \leq \sigma_{\max}(\mathbf{A})\|\mathbf{x}\|_2$, we can upper bound the right-hand side of (71). Note that for any matrices $\mathbf{A}$ and $\mathbf{B}$, $\sigma_{\max}(\mathbf{A}\otimes\mathbf{B}) = \sigma_{\max}(\mathbf{A})\sigma_{\max}(\mathbf{B})$. Hence, we can rewrite (72). Then, by using the Cauchy–Schwarz inequality,

$$\lambda_{\max}\left(\nabla^2\mathcal{L}(\mathbf{w}^*)\right) \leq 2\sum_{i=1}^{L}\sigma_{\max}\left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)^2\sigma_{\max}\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)^2. \tag{74}$$

Then, it suffices to show that there exists a direction $[\mathbf{U}_1^*, \mathbf{U}_2^*, \ldots, \mathbf{U}_L^*]$ along which the bound in (74) is achieved.

Consider the singular value decomposition (SVD) of $\prod_{j=i+1}^{L}\mathbf{W}_j^*$, and denote by $\mathbf{u}_i^L$ and $\mathbf{v}_i^L$ the left and right singular vectors of $\prod_{j=i+1}^{L}\mathbf{W}_j^*$ corresponding to its largest singular value, respectively. Note that since $\mathbf{w}_L^\top$ is a row vector, we can choose $\mathbf{u}_i^L = 1$ for all $i \in [L]$. Moreover, decompose $\prod_{j=1}^{i-1}\mathbf{W}_j^*$ by SVD, and denote by $\mathbf{u}_i^R$ and $\mathbf{v}_i^R$ the left and right singular vectors of $\prod_{j=1}^{i-1}\mathbf{W}_j^*$ corresponding to the largest singular value, respectively. Note that since $\mathbf{w}_1$ is a vector, we can choose $\mathbf{v}_i^R = 1$ for all $i \in [L]$. Determine a particular direction $[\mathbf{U}_1^*, \mathbf{U}_2^*, \ldots, \mathbf{U}_L^*]$ that achieves the upper bound while satisfying the constraint $\sum_{i=1}^{L}\|\mathbf{U}_i^*\|_F^2 = 1$. Suppose

$$\mathbf{U}_i^* = \frac{\sigma_{\max}\left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)\sigma_{\max}\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)}{\sqrt{\sum_{i=1}^{L}\sigma_{\max}\left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)^2\sigma_{\max}\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)^2}}\mathbf{v}_i^L{\mathbf{u}_i^R}^\top. \tag{75}$$

Then,

$$2\left\|\sum_{i=1}^{L}\left[\left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)\mathbf{U}_i^*\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)\right]\right\|_F^2 = 2\sum_{i=1}^{L}\sigma_{\max}\left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)^2\sigma_{\max}\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)^2. \tag{76}$$

Since the upper bound is achieved, it implies

$$\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{w}^*)) = 2\sum_{i=1}^{L} \sigma_{\max}\left(\prod_{j=i+1}^{L} \mathbf{W}_j^*\right)^2 \sigma_{\max}\left(\prod_{j=1}^{i-1} \mathbf{W}_j^*\right)^2. \tag{77}$$

□

# C Proofs from Section 3

## C.1 Proof of Theorem 3.1

*Proof.* From (53) and (54),

$$\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{w}^*)) = \max_{\substack{\mathbf{U}_1,\mathbf{U}_2,\ldots,\mathbf{U}_L:\\ \sum_{i=1}^{L}\|\mathbf{U}_i\|_F^2=1}} 2\left\|\sum_{i=1}^{L}\left[\left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)\mathbf{U}_i\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)\right]\right\|_F^2 \tag{78}$$

$$= \max_{\substack{\mathbf{U}_1,\mathbf{U}_2,\ldots,\mathbf{U}_L:\\ \sum_{i=1}^{L}\|\mathbf{U}_i\|_F^2=1}} 2\left\|\mathrm{vec}\left(\sum_{i=1}^{L}\left[\left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)\mathbf{U}_i\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)\right]\right)\right\|_2^2 \tag{79}$$

$$= \max_{\substack{\mathbf{U}_1,\mathbf{U}_2,\ldots,\mathbf{U}_L:\\ \sum_{i=1}^{L}\|\mathbf{U}_i\|_F^2=1}} 2\left\|\sum_{i=1}^{L}\mathrm{vec}\left(\left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)\mathbf{U}_i\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)\right)\right\|_2^2 \tag{80}$$

$$= \max_{\substack{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_L:\\ \sum_{i=1}^{L}\|\mathbf{u}_i\|_2^2=1}} 2\left\|\sum_{i=1}^{L}\left[\left(\prod_{j=1}^{i-1}\mathbf{W}_j^*\right)^\top \otimes \left(\prod_{j=i+1}^{L}\mathbf{W}_j^*\right)\right]\mathbf{u}_i\right\|_2^2. \tag{81}$$

Note that vec is a linear operator. Therefore, (79) can be rewritten as (80). Then, by using the vectorization trick of the Kronecker product, we can obtain (81). Let's define a block matrix and a vector such that

$$\mathbf{K} = \left[\mathbf{I}\otimes\textstyle\prod_{j=2}^{L}\mathbf{W}_j^* \,\middle|\, \mathbf{W}_1^{*\top}\otimes\left(\prod_{j=3}^{L}\mathbf{W}_j^*\right) \,\middle|\, \ldots \,\middle|\, \left(\prod_{j=1}^{L-1}\mathbf{W}_j^*\right)^\top\otimes\mathbf{I}\right], \tag{82}$$

$$\mathbf{u} = [\mathbf{u}_1^\top \quad \mathbf{u}_2^\top \quad \cdots \quad \mathbf{u}_L^\top]^\top. \tag{83}$$

Then

$$\lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{w}^*)) = \max_{\mathbf{u}:\|\mathbf{u}\|_2=1} 2\|\mathbf{K}\mathbf{u}\|_2^2 \tag{84}$$

$$= 2\sigma_{\max}(\mathbf{K}^\top\mathbf{K}). \tag{85}$$

Note that $\sigma_{\max}(\mathbf{K}^\top\mathbf{K}) = \sigma_{\max}(\mathbf{K}\mathbf{K}^\top)$, and for any two block matrices $\mathbf{A}$ and $\mathbf{B}$ such that

$$\mathbf{A} = [\mathbf{A}_1 \quad \mathbf{A}_2 \quad \ldots \quad \mathbf{A}_L] \in \mathbb{R}^{M_1\times d}, \quad \mathbf{B} = \begin{bmatrix}\mathbf{B}_1\\ \vdots \\ \mathbf{B}_L\end{bmatrix} \in \mathbb{R}^{d\times M_2} \tag{86}$$

$$\mathbf{A}\mathbf{B} = \sum_{i=1}^{L}\mathbf{A}_i\mathbf{B}_i, \quad \mathbf{A}\mathbf{B}\in\mathbb{R}^{M_1\times M_2}. \tag{87}$$

Furthermore, for any matrices $\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D}$ such that the matrix products $\mathbf{AB}$ and $\mathbf{CD}$ are well defined, we have

$$(\mathbf{A}\otimes\mathbf{C})(\mathbf{B}\otimes\mathbf{D}) = \mathbf{A}\mathbf{B}\otimes\mathbf{C}\mathbf{D}. \tag{88}$$

Using the fact that $(\mathbf{A}\otimes\mathbf{C})^\top = \mathbf{A}^\top\otimes\mathbf{C}^\top$ together with the previous property, it follows that

$$\lambda_{\max}\left(\nabla^2\mathcal{L}(\mathbf{w}^*)\right) = 2\sigma_{\max}\left(\sum_{i=1}^{L}\mathbf{B}_i^\top\mathbf{B}_i\otimes\mathbf{A}_i\mathbf{A}_i^\top\right), \tag{89}$$

where $\mathbf{A}_k = \prod_{i=k+1}^{L}\mathbf{W}_i^*$ and $\mathbf{B}_k = \prod_{i=1}^{k-1}\mathbf{W}_i^*$.

□

### C.2 Proof of Corollary 3.3

*Proof.* According to Lemma A.2, for any $(\mathbf{L}^*, \mathbf{R}^*) \in \Omega$,

$$\nabla^2 \mathcal{L}(\mathbf{L}^*, \mathbf{R}^*)[\mathbf{V}, \mathbf{U}] = 2\left\|\mathbf{L}^*\mathbf{U}^\top + \mathbf{V}\mathbf{R}^{*\top}\right\|_F^2. \tag{90}$$

Then

$$\begin{aligned}
\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{L}^*, \mathbf{R}^*)) &= \max_{\substack{\mathbf{U},\mathbf{V} \\ \|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 = 1}} 2\left\|\mathbf{L}^*\mathbf{U}^\top + \mathbf{V}\mathbf{R}^{*\top}\right\|_F^2 && (91)\\
&\leq \max_{\substack{\mathbf{U},\mathbf{V} \\ \|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 = 1}} 2\left(\left\|\mathbf{L}^*\mathbf{U}^\top\right\|_F + \left\|\mathbf{V}\mathbf{R}^{*\top}\right\|_F\right)^2 && (92)\\
&= \max_{\substack{\mathbf{u},\mathbf{v} \\ \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 = 1}} 2\left(\|(\mathbf{I} \otimes \mathbf{L}^*)\mathbf{u}\|_2 + \|(\mathbf{R}^* \otimes \mathbf{I})\mathbf{v}\|_2\right)^2 && (93)\\
&\leq \max_{\substack{\mathbf{u},\mathbf{v} \\ \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 = 1}} 2\left(\sigma_{\max}(\mathbf{I} \otimes \mathbf{L}^*)\|\mathbf{u}\|_2 + \sigma_{\max}(\mathbf{R}^* \otimes \mathbf{I})\|\mathbf{v}\|_2\right)^2. && (94)
\end{aligned}$$

We can upper bound the right-hand side of (91) using the *triangle inequality*. By applying the *vectorization trick* of the Kronecker product again, we can rewrite (92). Then, noting that for any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{A}\mathbf{x}\|_2 \leq \sigma_{\max}(\mathbf{A})\|\mathbf{x}\|_2$, we can upper bound the right-hand side of (93). Note that for any matrices $\mathbf{A}$ and $\mathbf{B}$, $\sigma_{\max}(\mathbf{A} \otimes \mathbf{B}) = \sigma_{\max}(\mathbf{A})\sigma_{\max}(\mathbf{B})$. Hence,

$$\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{L}^*, \mathbf{R}^*)) \leq \max_{\substack{\mathbf{u},\mathbf{v} \\ \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2 = 1}} 2\left(\sigma_{\max}(\mathbf{L}^*)\|\mathbf{u}\|_2 + \sigma_{\max}(\mathbf{R}^*)\|\mathbf{v}\|_2\right)^2. \tag{95}$$

Then, by using the Cauchy-Schwarz inequality,

$$\lambda_{\max}\left(\nabla^2 \mathcal{L}(\mathbf{L}^*, \mathbf{R}^*)\right) \;\leq\; 2\left(\sigma_{\max}(\mathbf{L}^*)^2 + \sigma_{\max}(\mathbf{R}^*)^2\right). \tag{96}$$

Now, we will show that this upper bound is achievable. Let us decompose $\mathbf{L}^*$ as $\mathbf{U}_L\mathbf{\Sigma}_L\mathbf{V}_L^\top$ by SVD, and denote by $\mathbf{u}_L$ and $\mathbf{v}_L$ the left and right singular vectors corresponding to its largest singular value, respectively. Moreover, decompose $\mathbf{R}^*$ as $\mathbf{U}_R\mathbf{\Sigma}_R\mathbf{V}_R^\top$ by SVD, and denote by $\mathbf{u}_R$ and $\mathbf{v}_R$ the left and right singular vectors corresponding to its largest singular value, respectively. We determine a particular $(\mathbf{U}^*, \mathbf{V}^*)$ such that it achieves the upper bound while satisfying the constraint $\|\mathbf{U}^*\|_F^2 + \|\mathbf{V}^*\|_F^2 = 1$. Choose

$$\mathbf{U}^{*\top} = \frac{\sigma_{\max}(\mathbf{L}^*)}{\sqrt{\sigma_{\max}(\mathbf{L}^*)^2 + \sigma_{\max}(\mathbf{R}^*)^2}}\mathbf{v}_L\mathbf{u}_R^\top, \quad \mathbf{V}^* = \frac{\sigma_{\max}(\mathbf{R}^*)}{\sqrt{\sigma_{\max}(\mathbf{L}^*)^2 + \sigma_{\max}(\mathbf{R}^*)^2}}\mathbf{u}_L\mathbf{v}_R^\top. \tag{97}$$

Using the identity $\left\|\mathbf{x}\mathbf{y}^\top\right\|_F^2 = \|\mathbf{x}\|_2^2\|\mathbf{y}\|_2^2$, we obtain

$$2\left\|\frac{(\sigma_{\max}(\mathbf{L}^*)^2 + \sigma_{\max}(\mathbf{R}^*)^2)}{\sqrt{\sigma_{\max}(\mathbf{L}^*)^2 + \sigma_{\max}(\mathbf{R}^*)^2}}\mathbf{u}_L\mathbf{u}_R^\top\right\|_F^2 = 2(\sigma_{\max}(\mathbf{L}^*)^2 + \sigma_{\max}(\mathbf{R}^*)^2). \tag{98}$$

□

## D On the Singular Vector Stationary Set

In this section, we show that the characterization of the singular vector stationary (SVS) set in Ghosh et al. [2025, Proposition 1] does not hold in general. Ghosh et al. [2025] considered the deep matrix factorization objective in (5) with $d_0 = d_1 = \cdots = d_L = d$ and a target matrix $\mathbf{M}$ with distinct singular values.[3]

Consider gradient flow (GF), $\dot{\mathbf{W}}_i(t) = -\nabla_{\mathbf{W}_i}\mathcal{L}(\mathbf{w}(t))$, and let $\mathbf{M} = \mathbf{U}_\mathbf{M}\mathbf{\Sigma}_\mathbf{M}\mathbf{V}_\mathbf{M}^\top$ and $\mathbf{W}_i(t) = \mathbf{U}_{\mathbf{W}_i}(t)\mathbf{\Sigma}_{\mathbf{W}_i}(t)\mathbf{V}_{\mathbf{W}_i}(t)^\top$ be SVDs for all $i \in [L]$. Ghosh et al. [2025] state in their Proposition 1 that, for any $t \geq 0$, if $\dot{\mathbf{U}}_{\mathbf{W}_i}(t) = \dot{\mathbf{V}}_{\mathbf{W}_i}(t) = \mathbf{0}$ for all $i \in [L]$, then

$$\mathbf{U}_{\mathbf{W}_L}(t) = \mathbf{U}_\mathbf{M}, \qquad \mathbf{V}_{\mathbf{W}_1}(t) = \mathbf{V}_\mathbf{M}, \qquad \mathbf{V}_{\mathbf{W}_{i+1}}(t) = \mathbf{U}_{\mathbf{W}_i}(t) \tag{99}$$

[3]Their objective has an additional factor of $1/2$, which does not affect the arguments below.

for every $i \in [L-1]$. The set of points that satisfy (99) is called the SVS set.

Suppose that each $\mathbf{W}_i(t)$ has full rank with distinct singular values. Then, by Ghosh et al. [2025, Equations (22)–(25)], $\dot{\mathbf{U}}_{\mathbf{W}_i}(t) = \dot{\mathbf{V}}_{\mathbf{W}_i}(t) = \mathbf{0}$ for all $i \in [L]$ if and only if

$$\mathbf{U}_{\mathbf{W}_i}(t)^\top \nabla_{\mathbf{W}_i} \mathcal{L}(\mathbf{w}(t)) \mathbf{V}_{\mathbf{W}_i}(t) \tag{100}$$

is diagonal for all $i \in [L]$, where for any $\mathbf{w}$, $\nabla_{\mathbf{W}_i} \mathcal{L}(\mathbf{w}) = -2\mathbf{A}_i^\top \big(\mathbf{M} - \prod_{j=1}^{L} \mathbf{W}_j\big)\mathbf{B}_i^\top$, with $\mathbf{A}_i$ and $\mathbf{B}_i$ as in Theorem 3.1. The same condition governs GD: If $\nabla_{\mathbf{W}_i} \mathcal{L}(\mathbf{w}) = \mathbf{U}_{\mathbf{W}_i} \mathbf{D}_i \mathbf{V}_{\mathbf{W}_i}^\top$ with $\mathbf{D}_i$ diagonal, then

$$\mathbf{W}_i - \eta \nabla_{\mathbf{W}_i} \mathcal{L}(\mathbf{w}) = \mathbf{U}_{\mathbf{W}_i} \left(\mathbf{\Sigma}_{\mathbf{W}_i} - \eta \mathbf{D}_i\right) \mathbf{V}_{\mathbf{W}_i}^\top, \tag{101}$$

so a GD step leaves the singular vectors unchanged, up to sign and permutation ambiguity.

Ghosh et al. [2025, Proposition 1] states that (100) implies (99). We give a counterexample with $L = 2$ and $d_0 = d_1 = d_2 = 2$, where both GF and GD are initialized at $\mathbf{w}(0)$. The counterexample was suggested by GPT-6 Astra, and we take full responsibility for its correctness. In the counterexample, every factor has full rank with distinct singular values; therefore, the singular vectors are unique up to sign and permutation ambiguity.

**Counterexample.** Consider

$$\mathbf{W}_1(0) = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{W}_2(0) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad \mathbf{M} = \begin{pmatrix} 3 & 2 \\ 5/2 & 0 \end{pmatrix}. \tag{102}$$

Then $\mathbf{M} - \mathbf{W}_2(0)\mathbf{W}_1(0) = \begin{pmatrix} -1 & 1 \\ 1/2 & -2 \end{pmatrix} \neq \mathbf{0}$, so $\mathbf{w}(0) \notin \Omega$, and

$$\nabla_{\mathbf{W}_1} \mathcal{L}(\mathbf{w}(0)) = \begin{pmatrix} 3 & 0 \\ 0 & 6 \end{pmatrix}, \qquad \nabla_{\mathbf{W}_2} \mathcal{L}(\mathbf{w}(0)) = \begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix}. \tag{103}$$

Here, $\mathbf{U}_{\mathbf{W}_1}(0) = \mathbf{V}_{\mathbf{W}_1}(0) = \mathbf{I}_{2\times 2}$ with $\mathbf{\Sigma}_{\mathbf{W}_1}(0) = \mathrm{diag}(2, 1)$, and

$$\mathbf{U}_{\mathbf{W}_2}(0) = \mathbf{V}_{\mathbf{W}_2}(0) = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad \mathbf{\Sigma}_{\mathbf{W}_2}(0) = \mathrm{diag}(3, 1). \tag{104}$$

Thus,

$$\mathbf{U}_{\mathbf{W}_1}(0)^\top \nabla_{\mathbf{W}_1} \mathcal{L}(\mathbf{w}(0)) \mathbf{V}_{\mathbf{W}_1}(0) = \mathrm{diag}(3, 6), \tag{105}$$

$$\mathbf{U}_{\mathbf{W}_2}(0)^\top \nabla_{\mathbf{W}_2} \mathcal{L}(\mathbf{w}(0)) \mathbf{V}_{\mathbf{W}_2}(0) = \mathrm{diag}(2, 6), \tag{106}$$

so (100) holds at $t = 0$. However, $\mathbf{V}_{\mathbf{W}_2}(0) \neq \mathbf{U}_{\mathbf{W}_1}(0)$. Moreover,

$$\mathbf{M}^\top \mathbf{M} = \begin{pmatrix} 61/4 & 6 \\ 6 & 4 \end{pmatrix} \tag{107}$$

is not diagonal, so $\mathbf{V}_{\mathbf{W}_1}(0) \neq \mathbf{V}_{\mathbf{M}}$. Therefore, $\mathbf{w}(0)$ is not in the SVS set.

We note that the converse implication does hold. Under (99), $\mathbf{A}_i$, $\mathbf{B}_i$, and $\mathbf{M} - \prod_{j=1}^{L} \mathbf{W}_j$ share the aligned singular vectors, so $\nabla_{\mathbf{W}_i} \mathcal{L}(\mathbf{w}) = \mathbf{U}_{\mathbf{W}_i} \mathbf{D}_i \mathbf{V}_{\mathbf{W}_i}^\top$ for some diagonal $\mathbf{D}_i$. Hence, the family obtained by fixing these singular-vector bases and allowing arbitrary diagonal coefficients is invariant under GD, which is what Ghosh et al. [2025, Propositions 3 and 4] use. However, it is not a characterization of the points at which the singular vectors are stationary.

# E Additional Experimental Results

## E.1 Visualization of the Contour Map of the Loss Landscape

To study the dynamics of deep matrix factorization, we analyze the trajectories of GD. Previous works have visualized neural network loss landscapes to explore their highly non-convex and non-Euclidean structure [Goodfellow et al., 2014, Li et al., 2018]. However, the high dimensionality prevents full visualization. As a result, only 1-D (line) or 2-D (surface) visualizations are available. In this paper, we focus on contour maps of the loss landscape in the vicinity of a global minimum.

**Contour Plots with Random Projections.** We want to visualize the loss landscape around a global minimum $\mathbf{w}^* \in \mathbb{R}^N$. We select two random vectors, $\zeta$ and $\gamma$, from $\mathbb{R}^N$. Then, for any $K \subset \mathbb{R}^2$, we can define the function $p : K \to \mathbb{R}$ :

$$p(x, y) = \mathcal{L}(\mathbf{w}^* + x\zeta + y\gamma) \quad \forall (x, y) \in K \tag{108}$$

and plot $p$ with a desired resolution.

**Scale Invariance and Manifolds**. Note that our loss function is *scale-invariant*, which means that for any nonzero scalar $c \in \mathbb{R}$, multiplying one layer by $c$ and the next layer by $1/c$, or vice versa, yields the same end-to-end function. This phenomenon forms a manifold for global minimizers in the loss landscape [Dinh et al., 2017]. Furthermore, we know that when $d_0 = d_L$, $\nabla^2\mathcal{L}(\mathbf{w}^*)$ is rank-deficient by at least the order of $1 - 1/L$; that is, at least $1 - 1/L$ of the eigenvalues of $\nabla^2\mathcal{L}(\mathbf{w}^*)$ are zero [Mulayoff and Michaeli, 2020].

**Projection onto the Hessian Eigenvectors.** If we use random projections in visualizations, plots might not be informative for tracking the optimization dynamics of GD due to the phenomenon caused by the scale invariance. To make contour maps as informative as possible, we choose $\zeta$ and $\gamma$ to be $\mathbf{v}_1$ and $\mathbf{v}_N$, respectively — the eigenvectors of the largest and smallest eigenvalues of $\nabla^2\mathcal{L}(\mathbf{w}^*)$.

### E.2 Experiment Details and Additional Experiments

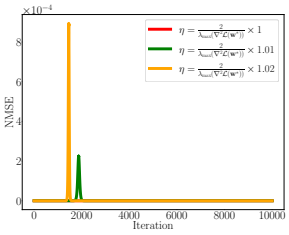
In our experimental setting, we first generate a minimizer $\mathbf{w}^*$ by randomly selecting the layer dimensions and then sampling Gaussian factor matrices of those dimensions. Then we construct $\mathbf{M}$ or $m$ by simply taking the product of the generated Gaussian matrices. To observe the escape phenomenon, we need to chart the loss landscape around the minimizer $\mathbf{w}^*$; however, the parameter space is high-dimensional, and this prevents a full visualization. Therefore, the common practice is to project the loss landscape onto two vectors to chart a 2-D contour plot as we mentioned in Appendix E.1. We select these vectors as eigenvectors of $\nabla^2\mathcal{L}(\mathbf{w}^*)$ that correspond to the maximum and minimum eigenvalues of $\nabla^2\mathcal{L}(\mathbf{w}^*)$. Note that we could have also considered a random perturbation direction sampled uniformly from $\mathbb{S}^{N-1}$. For instance, Mulayoff and Michaeli [2020] showed that for the deep matrix factorization problem where $d_0 = d_L$, which has the same Hessian structure as that of deep linear networks with a full-rank data matrix, the Hessian is rank-deficient at all minima by at least the order of $1 - (1/L)$. This means that at least $1 - (1/L)$ of the eigenvalues are zero at a minimum. Thus, the probability of choosing a direction that is orthogonal to the eigenspace corresponding to eigenvalues larger than $2/\eta$ is 0. Therefore, random perturbations would lead to the escape phenomenon with probability 1. On the other hand, for stable minima, it is important to choose the direction as the eigenvector of $\nabla^2\mathcal{L}(\mathbf{w}^*)$ that corresponds to the largest eigenvalue. The reasoning is the same: if you choose a direction that is not orthogonal to the eigenspace corresponding to the eigenvalues that are zero, then the linearized GD dynamics cannot converge to $\mathbf{w}^*$. This means that if we choose the perturbation direction randomly, then with probability 1, we choose a direction that is not orthogonal to the eigenspace corresponding to the eigenvalues that are zero. Therefore, for the experiment, it is convenient to choose the perturbation direction as the eigenvector of $\nabla^2\mathcal{L}(\mathbf{w}^*)$ corresponding to the largest eigenvalue. Note that the convergence we refer to here is the convergence to the specific global minimizer, not to an arbitrary global minimizer.

To track the training dynamics in the projected space, we project the iterates onto the subspace spanned by these two eigenvectors and indicate each iterate on the 2-D contour plot. We then repeat the experiments while varying the depth, dimensions, and initialization distance $r$ as shown in Figs. 1–5. We note that oscillations occur along the eigenvector corresponding to the maximum eigenvalue of the Hessian.

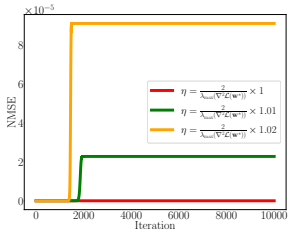
For the scalar factorizations of Figures 3 and 4, the hidden widths are drawn uniformly at random from $[24]$ before the factors are sampled. This yields $(d_0, d_1, d_2, d_3, d_4) = (1, 8, 4, 9, 1)$ for the four-layer experiment, i.e., $N = 85$ parameters, and $d_0 = d_{20} = 1$ with

$$(d_1, \ldots, d_{19}) = (12, 5, 16, 17, 18, 3, 15, 3, 3, 22, 3, 21, 11, 17, 16, 24, 8, 4, 9)$$

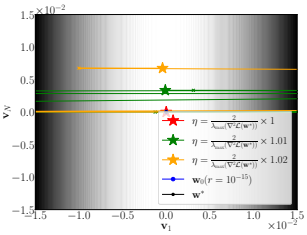
for the twenty-layer experiment, i.e., $N = 2421$ parameters. In all experiments, the entries of the factors $\mathbf{W}_1^*, \mathbf{W}_2^*, \ldots, \mathbf{W}_L^*$ are drawn independently from the standard Gaussian distribution $\mathcal{N}(0, 1)$ without any rescaling. The random instances are generated with NumPy, using `numpy.random.default_rng(12)` for Figure 1 and `numpy.random.RandomState(100)` for Figures 2–5. All computations are performed in IEEE 754 double precision.



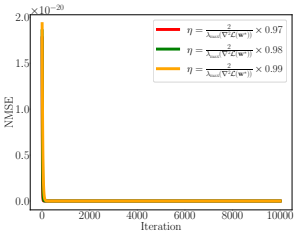

(a) Normalized training error across iterations, i.e., $\mathcal{L}(\mathbf{w}_t)/\|\mathbf{M}\|_F^2$.



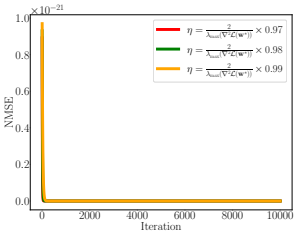

(b) Normalized squared $\ell^2$ distance of $\mathbf{w}_t$ from the minimum $\mathbf{w}^*$.



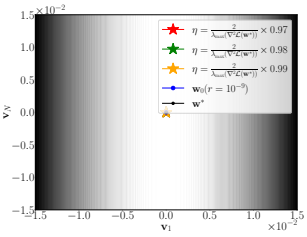

(c) Trajectories of GD on the contour map of the loss landscape.




(d) Normalized training error across iterations, i.e., $\mathcal{L}(\mathbf{w}_t)/\|\mathbf{M}\|_F^2$.

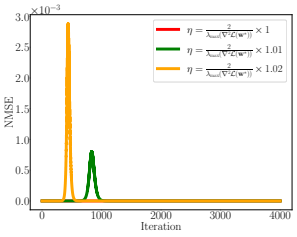



(e) Normalized squared $\ell^2$ distance of $\mathbf{w}_t$ from the minimum $\mathbf{w}^*$.

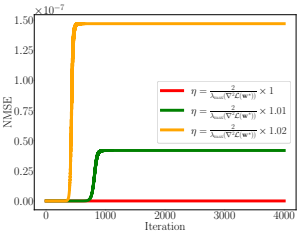



(f) Trajectories of GD on the contour map of the loss landscape.

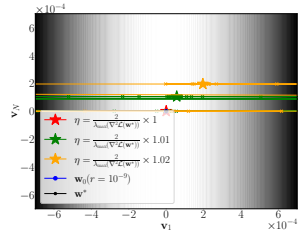
Figure 2: The setting is a depth-2 matrix factorization of a random $10 \times 20$ target matrix, where $\mathbf{L} \in \mathbb{R}^{10\times 30}$ and $\mathbf{R} \in \mathbb{R}^{20\times 30}$ are Gaussian factors. GD dynamics with different step sizes, indicated by different colors, are initialized within a radius of $10^{-15}$ in (a)–(c) and $10^{-9}$ in (d)–(f) from the minimizer $\mathbf{w}^*$. The initialization direction is along the eigenvector of $\nabla^2\mathcal{L}(\mathbf{w}^*)$ that corresponds to its maximum eigenvalue. The value of $\lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{w}^*))$ is computed using the closed-form expression derived in Corollary 3.3.



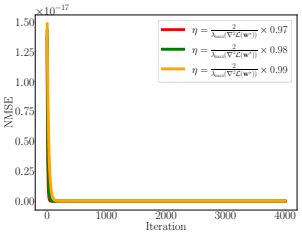

(a) Normalized training error across iterations, i.e., $\mathcal{L}(\mathbf{w}_t)/m^2$.



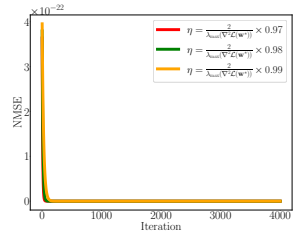

(b) Normalized squared $\ell^2$ distance of $\mathbf{w}_t$ from the minimum $\mathbf{w}^*$.



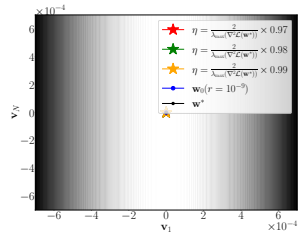

(c) Trajectories of GD on the contour map of the loss landscape.




(d) Normalized training error across iterations, i.e., $\mathcal{L}(\mathbf{w}_t)/m^2$.

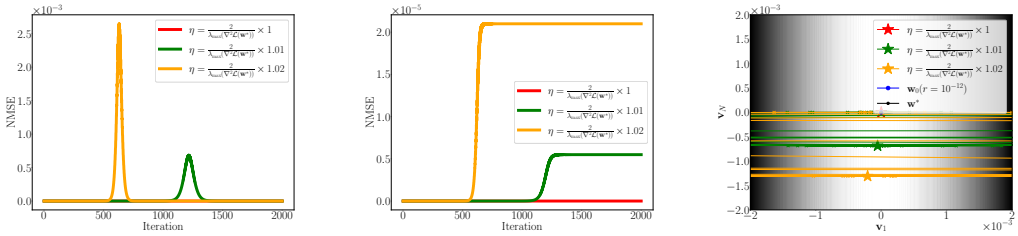



(e) Normalized squared $\ell^2$ distance of $\mathbf{w}_t$ from the minimum $\mathbf{w}^*$.




(f) Trajectories of GD on the contour map of the loss landscape.

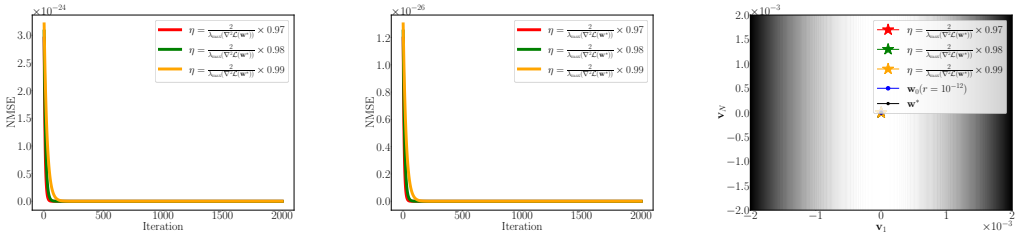
Figure 3: The setting is a 20-layer overparameterized scalar factorization of a random scalar. GD dynamics with different step sizes, indicated by different colors, are initialized within a radius of $10^{-9}$ from the minimizer $\mathbf{w}^*$. The initialization direction is along the eigenvector of $\nabla^2\mathcal{L}(\mathbf{w}^*)$ that corresponds to its maximum eigenvalue. The value of $\lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{w}^*))$ is computed using the closed-form expression derived in Corollary 3.2.




(a) Normalized training error across iterations, i.e., $\mathcal{L}(\mathbf{w}_t)/m^2$.

(b) Normalized squared $\ell^2$ distance of $\mathbf{w}_t$ from the minimum $\mathbf{w}^*$.

(c) Trajectories of GD on the contour map of the loss landscape.

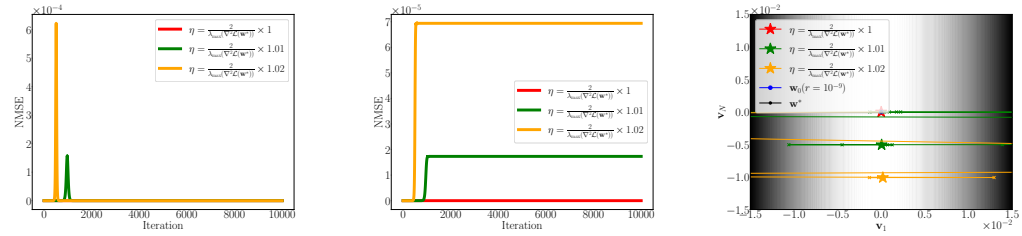



(d) Normalized training error across iterations, i.e., $\mathcal{L}(\mathbf{w}_t)/m^2$.

(e) Normalized squared $\ell^2$ distance of $\mathbf{w}_t$ from the minimum $\mathbf{w}^*$.

(f) Trajectories of GD on the contour map of the loss landscape.

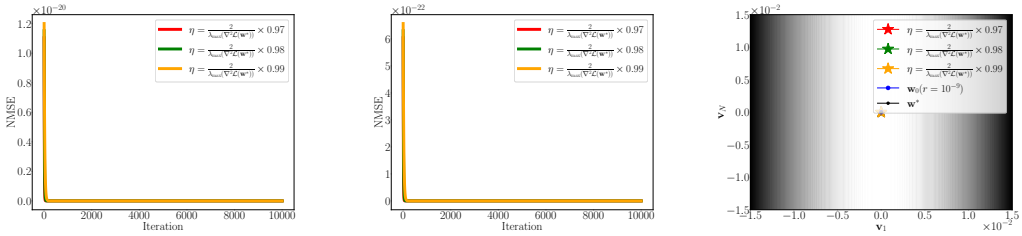
Figure 4: The setting is a 4-layer overparameterized scalar factorization of a random scalar. GD dynamics with different step sizes, indicated by different colors, are initialized within a radius of $10^{-12}$ from the minimizer $\mathbf{w}^*$. The initialization direction is along the eigenvector of $\nabla^2\mathcal{L}(\mathbf{w}^*)$ that corresponds to its maximum eigenvalue. The value of $\lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{w}^*))$ is computed using the closed-form expression derived in Corollary 3.2.




(a) Normalized training error across iterations, i.e., $\mathcal{L}(\mathbf{w}_t)/\|\mathbf{M}\|_F^2$.

(b) Normalized squared $\ell^2$ distance of $\mathbf{w}_t$ from the minimum $\mathbf{w}^*$.

(c) Trajectories of GD on the contour map of the loss landscape.




(d) Normalized training error across iterations, i.e., $\mathcal{L}(\mathbf{w}_t)/\|\mathbf{M}\|_F^2$.

(e) Normalized squared $\ell^2$ distance of $\mathbf{w}_t$ from the minimum $\mathbf{w}^*$.

(f) Trajectories of GD on the contour map of the loss landscape.

Figure 5: The setting is a depth-2 matrix factorization of a random $25 \times 20$ target matrix, where $\mathbf{L} \in \mathbb{R}^{25\times 30}$ and $\mathbf{R} \in \mathbb{R}^{20\times 30}$ are Gaussian factors. GD dynamics with different step sizes, indicated by different colors, are initialized within a radius of $10^{-9}$ from the minimizer $\mathbf{w}^*$. The initialization direction is along the eigenvector of $\nabla^2\mathcal{L}(\mathbf{w}^*)$ that corresponds to its maximum eigenvalue. The value of $\lambda_{\max}(\nabla^2\mathcal{L}(\mathbf{w}^*))$ is computed using the closed-form expression derived in Corollary 3.3.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the contributions and scope of the paper, summarizing the main theoretical results, the problem setting, and the significance of the findings.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide all assumptions for the optimization objectives considered in this paper. To the best of our knowledge, all proofs provided for the theorems are complete and correct.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The main contributions of this paper are theoretical. All information needed to reproduce the experiments, including parameter settings and evaluation metrics, is provided in Section 6 and Appendix E.2.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code and data are not provided, as the main contributions of this paper are theoretical. The experiments presented in Section 6 and Appendix E.2 are described in sufficient detail to be straightforward to set up and run, requiring no specialized hardware or external datasets.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide every detail to reproduce and understand the experiments in Section 6 and Appendix E.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The experiments illustrate deterministic GD trajectories for sampled problem instances. We do not report error bars or statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The experiments presented in Section 6 and Appendix E.2 are computationally lightweight, requiring no specialized hardware or high-performance computing resources, and can be reproduced on a standard consumer-grade CPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

Justification: This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted in the main paper.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: Our work does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our work does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The counterexamples provided in Remark 4.5 and Appendix D were suggested by Claude Fable 5 and ChatGPT-6 Astra, respectively, and we take full responsibility for their correctness. Beyond this, we used LLMs as spell checkers, grammar suggestion tools, editing aids, and basic code assistants.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.