Learning Dynamics of Deep Matrix Factorization Beyond the Edge of Stability

Anonymous CPAL submission

Deep neural networks trained using gradient descent with a fixed learning rate η often operate in the regime of "edge of stability" (EOS), where the largest eigenvalue of the Hessian equilibrates about the stability threshold $2/\eta$. In this work, we present a fine-grained analysis of the learning dynamics of (deep) linear networks (DLN) within the deep matrix factorization loss beyond EOS. For DLNs, loss oscillations within EOS follow a period-doubling route to chaos. We theoretically analyze the regime of the 2-period orbit and show that the loss oscillations occur within a small subspace, with the dimension of the subspace precisely characterized by the learning rate. Our analysis contributes to explaining two key phenomena in deep networks: (i) shallow models and simple tasks do not always exhibit EOS [1]; and (ii) oscillations occur within low-dimensional subspaces [2]. We present experiments to support our theory, along with examples demonstrating how these phenomena occur in nonlinear networks and how they differ from those in DLNs.

1. Introduction

Understanding generalization in deep neural networks requires an understanding of the optimization process in gradient descent (GD). In the literature, it has been empirically observed that the learning rate η plays a key role in driving generalization [3, 4]. The "descent lemma" from classical optimization theory says that for a β-smooth loss $\mathcal{L}(\mathbf{\Theta})$ parameterized by $\mathbf{\Theta}$, GD iterates satisfy

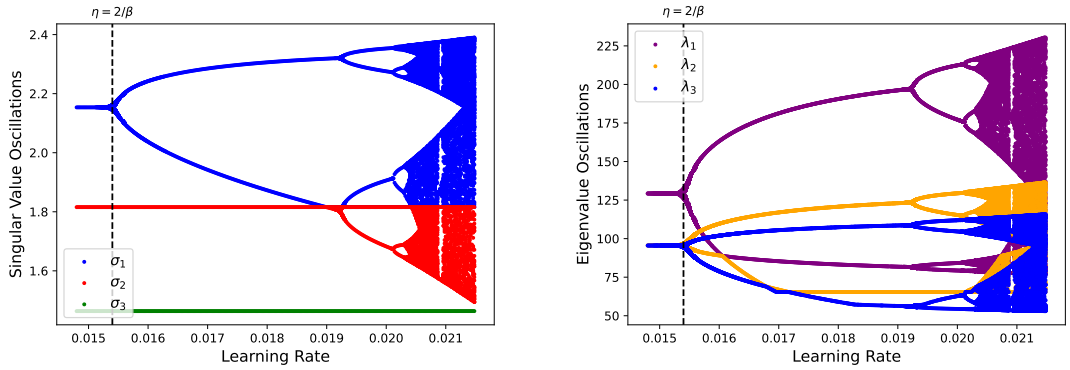
$$\mathcal{L}(\mathbf{\Theta}(t+1)) \leq \mathcal{L}(\mathbf{\Theta}(t)) - \frac{\eta(2-\eta\beta)}{2}\|\nabla\mathcal{L}(\mathbf{\Theta}(t))\|_2^2,$$

and so the learning rate should be chosen as $\eta < 2/\beta$ to monotonically decrease the loss. However, many recent works have shown that the training loss decreases even for $\eta > 2/\beta$, albeit non-monotonically. Surprisingly, it has been observed that choosing such a learning rate often provides better generalization over smaller ones that lie within the stability threshold. This observation has led to a series of works analyzing the behavior of GD within a regime dubbed "the edge of stability" (EOS). By letting $\mathbf{\Theta}$ parameterize a deep network, we formally define EOS as follows:

Definition 1 (Edge of Stability [1])**.** *During training, the sharpness of the loss, defined as* $S(\mathbf{\Theta}) := \|\nabla^2\mathcal{L}(\mathbf{\Theta})\|_2$*, continues to grow until it reaches* $2/\eta$ *(progressive sharpening), after which it stabilizes around* $2/\eta$*. During this process, the training loss behaves non-monotonically over short timescales but consistently decreases over long timescales.*

Using a large learning rate to operate within the EOS regime is hypothesized to give better generalization performance by inducing "catapults" in the training loss [2]. Intuitively, whenever the sharpness $S(\mathbf{\Theta})$ exceeds the local stability limit $2/\eta$, the GD iterates momentarily diverge (or catapults) out of a sharp region and self-stabilizes [5] to settle for a flatter region where the sharpness is below $2/\eta$, which has shown to correlate with better generalization [6–10]. Of course, the dynamics within EOS differ based on the loss landscape. When the loss landscape is highly non-convex with many local valleys, catapults may occur, whereas sustained oscillations may exist for other landscapes. It is of great interest to understand these behaviors within different network architectures to further our understanding of EOS.

Submitted to Second Conference on Parsimony and Learning (CPAL 2025). Do not distribute.



Singular Values of Weights Eigenvalues of Hessian

Figure 1: Bifurcation plot of the oscillations in the singular values (left) and the eigenvalues of the Hessian (right) of a 3-layer end-to-end DLN. The bifurcation plots indicate the existence of a period-doubling route to chaos in DLNs, which we analyze by examining the two-period orbit. Here, $\eta > 2/\beta$ corresponds to the EOS regime, where $\beta = L\sigma_{\star,1}^{2-2/L}$ is the sharpness at the minima, L is the depth of the network and $\sigma_{\star,1}$ is the first singular value of the target matrix $\mathbf{M}_\star$.

From a theoretical perspective, there have been many recent efforts to understand EOS. These works generally focus on analyzing "simple" functions, examples including scalar losses [11–13], quadratic regression models [14], diagonal linear networks [15] and two-layer matrix factorization [16]. However, the simplicity of these functions cannot fully capture the behaviors of deep neural networks within the EOS regime. Specifically, the following observations remain unexplained by existing analyses: (i) mild (or no) sharpening occurs when either networks are shallow or "simple" datasets are used for training (Caveat 2 from [1]); and (ii) the oscillations and catapults in the training loss occur in the span of the top eigenvectors of the NTK [2].

In this work, we present a fine-grained analysis of the learning dynamics of deep linear networks (DLNs) within the EOS regime, demonstrating that these phenomena can be replicated and effectively explained using DLNs. Generally, there are two lines of work for DLNs: (i) those that analyze the effects of depth and initialization scale, and how they implicitly bias the trajectory of gradient flow towards low-rank solutions when the learning rate is chosen to be stable [17–24], and (ii) those that analyze the similarities in behavior between linear and nonlinear networks [25–27]. Our analysis builds upon these works to show that DLNs exhibit intricate and interesting behaviors outside the stability regime and to demonstrate how factors such as depth and initialization scale contribute to the EOS regime. Our main results can be summarized as follows:

- **Oscillations in Low-Dimensional Subspaces.** We show that there exist periodic oscillations within r-dimensional subspaces in DLNs, where r is precisely characterized by the learning rate. For DLNs, a period-doubling route to chaos [28] exists in both the singular values of the DLN and the eigenvalues of the Hessian, as shown in Figure 1. We rigorously characterize the case of the two-period orbit, aiming to contribute to explaining the empirical observations by [2] and [1]. We also prove that the learning rate needed to enter EOS is a function of the network depth, further revealing its role in deep networks.

- **EOS Dynamics for DLNs and Diagonal Linear Networks.** We prove that while DLNs and diagonal linear networks have different curvatures about the global minimum, they exhibit similar behaviors within the EOS regime. We show that the additional eigendirections present in DLNs are not explored during training, making the behavior of the two networks synonymous.

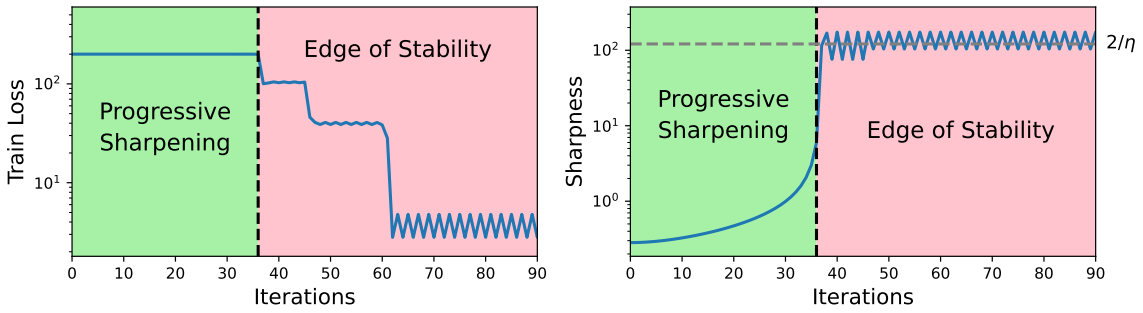


Figure 2: Depiction of the two phases of learning in the deep matrix factorization problem for a network of depth 3. It appears that upon escaping the first saddle point, the GD iterates enter the EOS regime, where the sharpness hovers just above $2/\eta$. Figure (left) shows the training loss undergoes saddle to saddle jump followed by periodic oscillation. Figure (right) plots the corresponding sharpness.

2. Notation and Problem Setup

Notation. We denote vectors with bold lower-case letters (e.g., $\mathbf{x}$) and matrices with bold upper-case letters (e.g., $\mathbf{X}$). We use $\mathbf{I}_n$ to denote an identity matrix of size $n \in \mathbb{N}$. We use $[L]$ to denote the set $\{1, 2, \ldots, L\}$. We use the notation $\sigma_i(\mathbf{A})$ to denote the i-th singular value of the matrix $\mathbf{A}$. We also use the notation $\sigma_{\ell,i}$ to denote the i-th singular value of the matrix $\mathbf{W}_\ell$.

Deep Matrix Factorization Loss. The objective in deep matrix factorization is to model a low-rank matrix $\mathbf{M}_\star \in \mathbb{R}^{d\times d}$ with $\mathrm{rank}(\mathbf{M}_\star) = r$ via a DLN parameterized by a set of parameters $\mathbf{\Theta} = (\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_L)$, which can be estimated by solving

$$\underset{\mathbf{\Theta}}{\operatorname{argmin}}\, f(\mathbf{\Theta}) := \frac{1}{2}\| \underbrace{\mathbf{W}_L \cdot \ldots \cdot \mathbf{W}_1}_{=:\mathbf{W}_{L:1}} - \mathbf{M}_\star \|_{\mathsf{F}}^2, \tag{1}$$

where we adopt the abbreviation $\mathbf{W}_{j:i} = \mathbf{W}_j \cdot \ldots \cdot \mathbf{W}_i$ to denote the end-to-end DLN and is identity when $j < i$. We assume that each weight matrix has dimensions $\mathbf{W}_\ell \in \mathbb{R}^{d\times d}$ to observe the effects of overparameterization. We also assume that the singular values of $\mathbf{M}_\star$ are distinct.

Optimization. We update each weight matrix $\mathbf{W}_\ell \in \mathbb{R}^{d\times d}$ using GD with iterations given by

$$\mathbf{W}_\ell(t) = \mathbf{W}_\ell(t-1) - \eta \cdot \nabla_{\mathbf{W}_\ell} f(\mathbf{\Theta}(t-1)), \quad \forall \ell \in [L], \tag{2}$$

where $\eta > 0$ is the learning rate and $\nabla_{\mathbf{W}_\ell} f(\mathbf{\Theta}(t))$ is the gradient of $f(\mathbf{\Theta})$ with respect to the ℓ-th weight matrix at the t-th GD iterate.

Initialization. In this work, we consider both balanced and unbalanced initializations, respectively:

$$\mathbf{W}_\ell(0) = \alpha \mathbf{I}_d, \quad \forall \ell \in [L], \qquad \text{(Balanced Initialization)}$$
$$\mathbf{W}_L(0) = \mathbf{0}, \quad \mathbf{W}_\ell(0) = \alpha \mathbf{I}_d, \quad \forall \ell \in [L-1], \qquad \text{(Unbalanced Initialization)}$$

where $\alpha > 0$ is a small constant. We assume α is chosen small enough such that $\alpha \in (0, \sigma_{\star,r})$, where $\sigma_{\star,r}$ is the r-th singular value of $\mathbf{M}_\star$. Generally, many existing works on both shallow and deep linear networks assume a zero-balanced initialization (i.e., $\mathbf{W}_i^\top(0)\mathbf{W}_i(0) = \mathbf{W}_j(0)\mathbf{W}_j^\top(0)$ for $i \neq j$). This introduces the invariant $\mathbf{W}_i^\top(t)\mathbf{W}_i(t) = \mathbf{W}_j(t)\mathbf{W}_j^\top(t)$ for all $t > 0$, ensuring two (degenerate) conditions throughout the training trajectory: (i) the singular vectors of each of the layers remain aligned and (ii) the singular values stay balanced. For the unbalanced initialization, the zero weight layer can be viewed as the limiting case of initializing the weights with a (very) small constant $\alpha' \ll \alpha$, and has been similarly explored by [29, 30], albeit for two-layer networks. The zero weight layer relieves the balancing condition of the singular values. Rather than staying balanced, we show that the singular values become increasingly balanced (see Lemma 2). This allows us to jointly analyze the singular values of the weights for either case.

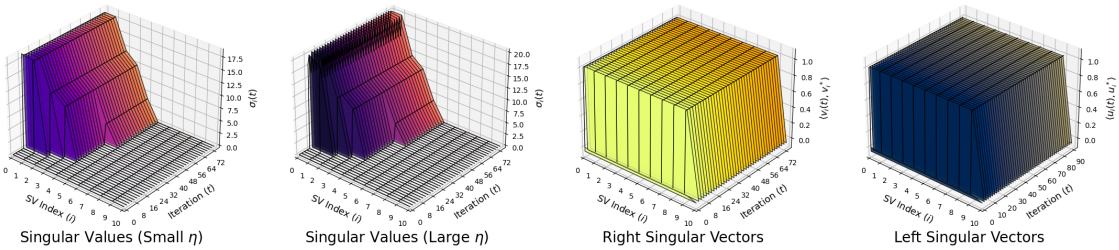


Figure 3: Illustrations of the singular vector and value evolution of the end-to-end DLN. The singular vectors of the network remain static across all iterations, as suggested by the singular vector stationary set, regardless of the learning rate. The angle between the true singular vectors and those of the network remains aligned throughout. The first singular values undergo oscillations in the large η regime, whereas they remain constant in the small η regime.

Nevertheless, we also show that our analysis is not limited to either initialization but applies to *any initialization* that converges to a set we call the singular vector stationary set (see Proposition 1). To the best of our knowledge, it is common to assume that the singular vectors remain aligned, as many existing works make the same assumption [17, 19, 29, 31–33].

3. Deep Matrix Factorization Beyond the Edge of Stability

When using a large learning rate, the learning dynamics can typically be separated into two distinct stages: (i) progressive sharpening and (ii) the edge of stability. Within the progressive sharpening stage, the sharpness lies below $2/\eta$ and tends to continually rise. Our goal is to analyze the EOS stage under the deep matrix factorization formulation. Here, we observe that the training loss fluctuates due to layerwise singular value oscillations, as illustrated in Figure 2.

3.1. Main Results

Before we present our main results, we provide a definition of what we refer to as a strict balanced state of the singular values of the weight matrices. The parameters are said to be in a strict balanced state if the singular values of each weight matrix take the same values across all layers.

Definition 2 (Strict Balanced State)**.** *The parameters $\mathbf{\Theta}$ of the DLN from Equation (1) are said to be in a strict balanced state if for some $t \geq 0$*

$$\sigma_i(\mathbf{W}_\ell(t)) = \sigma_i(\mathbf{W}_k(t)), \quad \forall i \in [r], \quad \forall \ell, k \in [L],$$

where $\sigma_i(\mathbf{W}_\ell)$ denotes the i-th singular value of the ℓ-th layer and r is the rank of the matrix $\mathbf{M}_\star$.

It is straightforward to show that the parameters are in a strictly balanced state for all $t \geq 0$ if we initialize the singular values to be the same across all weight matrices $\mathbf{W}_\ell$. Hence, it immediately holds that the balanced initialization is in a strictly balanced state. However, the one-zero initialization (i.e., the unbalanced initialization) sets the singular values of $\mathbf{W}_L$ to zero, meaning the parameters are not initially in a strictly balanced state. However, in Lemma 2 of Section 3.2, we prove that the balancing increasingly occurs throughout GD iterations within the EOS regime. Consequently, we assume a strictly balanced state for the remainder of this paper and analyze the EOS regime in relation to the balanced minimum. Next, we derive the eigenvalues of the Hessian at convergence, such that we can identify the learning rate needed to enter the EOS regime for DLNs.

Lemma 1 (Eigenvalues of Hessian at the Balanced Minimum)**.** *Consider running GD on the deep matrix factorization loss $f(\mathbf{\Theta})$ defined in Equation (1). The set of all non-zero eigenvalues of the training*

loss Hessian at the balanced minimum is given by

$$\lambda_{\mathbf{\Theta}} = \underbrace{\left\{L\sigma_{\star,i}^{2-\frac{2}{L}}, \sigma_{\star,i}^{2-\frac{2}{L}}\right\}_{i=1}^{r}}_{\textit{self-interaction}} \bigcup \underbrace{\left\{\sum_{\ell=0}^{L-1}\left(\sigma_{\star,i}^{1-\frac{1}{L}-\frac{1}{L}\ell}\cdot\sigma_{\star,j}^{\frac{1}{L}\ell}\right)^2\right\}_{i\neq j}^{r}}_{\textit{interaction with other singular values}} \bigcup \underbrace{\left\{\sum_{\ell=0}^{L-1}\left(\sigma_{\star,k}^{1-\frac{1}{L}-\frac{1}{L}\ell}\cdot\alpha^{\ell}\right)^2\right\}_{k=1}^{r}}_{\textit{interaction with initialization}}$$

where $\sigma_{\star,i}$ is the i-th singular value of the target matrix $\mathbf{M}_\star \in \mathbb{R}^{d\times d}$, $\alpha \in \mathbb{R}$ is the initialization scale, L is the depth of the network, and the second element of the set under "self-interaction" has a multiplicity of $d-r$.

We defer all of the proofs to Appendix C. Let λ_i denote the i-th largest eigenvalue of the Hessian. By Lemma 1, we observe that the sharpness is equal to $\lambda_1 = \|\nabla^2 f(\mathbf{\Theta})\|_2 = L\sigma_{\star,1}^{2-\frac{2}{L}}$ at the balanced minimum. In Lemma 8, we show that the sharpness at the balanced minimum is the flattest, and hence $L\sigma_{\star,1}^{2-\frac{2}{L}}$ represents the smallest sharpness value among all global minima. Thus, if η is set such that $\eta > 2/\lambda_1$, oscillations in the loss will occur, as the step size is large enough to induce oscillations even in the flattest region. Interestingly, notice that all non-zero eigenvalues are a function of network depth. For a deeper network, the sharpness will be larger, implying that a smaller learning rate can be used to drive the DLN into EOS. This provides a unique perspective on how the learning rate should be chosen as networks become deeper and explains the observation made by [1], who observed that sharpness scales with the depth of the network. Equipped with the eigenvalues, we show in the following result that oscillations occur in a two-period orbit along the i-th eigenvector direction of the balanced minimum, given that the learning rate is set to be greater than $2/\lambda_i$.

Theorem 1 (Stable Subspace Oscillations)**.** *Let $\alpha' \coloneqq \left(\ln\left(\frac{2\sqrt{2}}{\eta\lambda_1}\right)\cdot\frac{\sigma_{\star,1}^{4/L}}{L^2\cdot 2^{\frac{2L-3}{L}}}\right)^{1/4}$. Consider running GD on the loss in (1) with initialization scale $0<\alpha<\alpha'$. If $\eta=\frac{2}{K}$ with $\lambda_i \leq K < \lambda_{i+1}$, then 2-period orbit oscillation occurs in the direction of Δ_i, where λ_i and Δ_i denote the i-th largest eigenvalue and eigenvector of the Hessian at the balanced minimum, respectively.*

The complete proof is provided in Appendix C.1, where we derive all eigenvectors at the balanced minimum and demonstrate that the necessary conditions from Lemma 12 (restated from [16]) are satisfied for a two-period orbit. The condition on the initialization scale is chosen to ensure balanced behavior, as demonstrated in Lemma 2. To understand Theorem 1 more clearly, consider the first eigenvector of the Hessian, which we derived to be

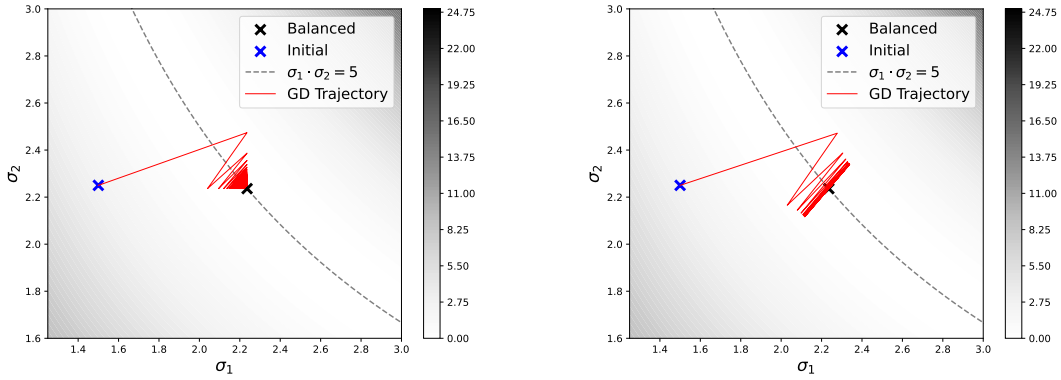
$$\Delta_1 = \frac{1}{\sqrt{L}}\cdot\operatorname{vec}\left(\mathbf{u}_{\star,1}\mathbf{v}_{\star,1}^\top, \mathbf{v}_{\star,1}\mathbf{v}_{\star,1}^\top, \ldots, \mathbf{v}_{\star,1}\mathbf{v}_{\star,1}^\top\right),$$

where $\mathbf{u}_{\star,1}, \mathbf{v}_{\star,1} \in \mathbb{R}^d$ are the first left and right singular vectors of $\mathbf{M}_\star$, respectively. By Proposition 2, we know that the each weight layer takes the form $\mathbf{W}_\ell = \mathbf{V}_\star\mathbf{\Sigma}_\ell\mathbf{V}_\star^\top$ for all $\ell \in [L-1]$ and $\mathbf{W}_L = \mathbf{U}_\star\mathbf{\Sigma}_L\mathbf{V}_\star^\top$ at convergence, starting from the unbalanced initialization. By stacking and flattening these weights, consider the direction

$$\widetilde{\Delta} \coloneqq \sum_{i=1}^{d}\operatorname{vec}\left(\sigma_{L,i}\cdot\mathbf{u}_{\star,i}\mathbf{v}_{\star,i}^\top, \sigma_{L-1,i}\cdot\mathbf{v}_{\star,i}\mathbf{v}_{\star,i}^\top, \ldots, \sigma_{1,i}\cdot\mathbf{v}_{\star,i}\mathbf{v}_{\star,i}^\top\right).$$

Since $\widetilde{\Delta}^\top\Delta_1$ is only non-zero in the rank-1 components of $\widetilde{\Delta}$, this implies that if $\lambda_1 \leq K < \lambda_2$ from Theorem 1, the oscillations only occur in the rank-1 components of the weights. The following result substantiates this claim by demonstrating that, with an appropriately chosen learning rate, oscillations occur in the singular values in the top-p directions given that $p \leq r$.

Theorem 2 (Rank-p Periodic Subspace Oscillations)**.** *Let $\mathbf{M}_\star = \mathbf{U}_\star\mathbf{\Sigma}_\star\mathbf{V}_\star^\top$ denote the SVD of the target matrix and define $S_p \coloneqq L\sigma_{\star,p}^{2-\frac{2}{L}}$ and $K_p' \coloneqq \max\left\{S_{p+1}, \frac{S_p}{2\sqrt{2}}\right\}$. If we run GD on the deep matrix factorization loss with initialization scale $\alpha<\alpha'$ from Theorem 1 and learning rate $\eta=\frac{2}{K}$, where $K_p' < K < S_p$, then under strict balancing, the top-p singular values of the end-to-end DLN oscillates in a 2-period orbit*



$\eta < 0.2$ $\eta > 0.2$

Figure 4: Illustration of the GD trajectories for two different learning rates for minimizing the function $f(\sigma_1, \sigma_2) = \frac{1}{2}(\sigma_2 \cdot \sigma_1 - 5)^2$, starting from an unbalanced initial point. The sharpness at the balanced minimum is 10, and so the learning rate needed to enter EOS is $\eta = 0.2$. These plots show that while GD initially arrives at an unbalanced solution in both cases, the iterates oscillate about the other minima to arrive at the balanced solution within the EOS regime.

$(j \in \{1, 2\})$ *around the balanced minimum and admits the following decomposition:*

$$\mathbf{W}_{L:1} = \underbrace{\sum_{i=1}^{p} \rho_{i,j}^{L} \cdot \mathbf{u}_{\star,i} \mathbf{v}_{\star,i}^{\top}}_{\textit{oscillation subspace}} + \underbrace{\sum_{k=p+1}^{d} \sigma_{\star,k} \cdot \mathbf{u}_{\star,k} \mathbf{v}_{\star,k}^{\top}}_{\textit{stationary subspace}}, \quad j \in \{1, 2\}, \quad \forall \ell \in [L-1], \tag{3}$$

where $\rho_{i,1} \in \left(0, \sigma_{\star,i}^{1/L}\right)$ *and* $\rho_{i,2} \in \left(\sigma_{\star,i}^{1/L}, (2\sigma_{\star,i})^{1/L}\right)$ *are the two real roots of the polynomial* $g(\rho_i) = 0$ *and*

$$g(\rho_i) = \rho_i^L \cdot \frac{1 + \left(1 + \eta L(\sigma_{\star,i} - \rho_i^L) \cdot \rho_i^{L-2}\right)^{2L-1}}{1 + \left(1 + \eta L(\sigma_{\star,i} - \rho_i^L) \cdot \rho_i^{L-2}\right)^{L-1}} - \sigma_{\star,i}.$$

In contrast to Theorem 1, Theorem 2 explicitly identifies the subspaces that exhibit a two-period orbit based on the range of the learning rate. It also provides a rough characterization of the oscillation amplitude, which is determined by $\rho_{i,1}$ and $\rho_{i,2}$—values below and above the balanced minimum, respectively. Since there is no closed-form solution for an arbitrary higher-order polynomial, $\rho_{i,1}$ and $\rho_{i,2}$ are defined as solutions to the polynomial $g(\rho_i)$. Overall, this aims to theoretically explain why (i) oscillations occur primarily within the top subspaces of the network, as observed by [2], and (ii) oscillations are more pronounced in the directions of stronger features, as measured by the magnitudes of their singular values.

Notice that Theorem 2 demonstrates the existence of a two-period orbit only around the balanced minimum. While this assumption was initially made for ease of analysis, we empirically observe that the two-period orbit actually *only* occurs around the balanced solution. To illustrate this claim, in Figure 4, we provide a plot of the GD trajectory for minimizing a two-layer scalar loss $f(\sigma_1, \sigma_2) = \frac{1}{2}(\sigma_2 \cdot \sigma_1 - 5)^2$, starting from an unbalanced initial point $(\sigma_1, \sigma_2) = (1.5, 2.25)$. Notice that, by Lemma 1, the sharpness around the balanced minimum is $L\sigma_{\star,1}^{2-2/L} = 10$, and thus the necessary learning rate to enter the EOS regime is $\eta = 0.2$. We plot the GD trajectory under two cases for the learning rate: (i) slightly below the EOS learning rate, $\eta = 0.1999$, and (ii) slightly above it, $\eta = 0.2010$. When $\eta = 0.2010$, GD first arrives at an unbalanced solution, then oscillates until it reaches the balanced minimum, where it sustains a two-period orbit around the balanced solution. The other minima are too narrow to sustain the oscillations induced by the large learning rate, causing the GD iterates to bounce out of those minima and settle at the flattest, balanced solution. When the learning rate is slightly below $\eta = 0.2$, it remains large enough to induce oscillations around

other unbalanced minima, but ultimately GD converges to the balanced solution, where it exhibits no oscillations, as predicted by Theorem 2. This empirical observation suggests that two-period oscillations and balancing occur simultaneously, which we leave for future investigation.

Finally, we conclude by remarking that our result also aims to generalize the recent theoretical findings of [16], where they proved the existence of a certain class of scalar functions $f(x)$ for which GD does not diverge even when operating beyond the stability threshold. They demonstrated that there exists a range in which the loss oscillates around the local minima with a certain periodicity. These oscillations gradually progress into higher periodic orbits (e.g., 2, 4, 8 periods), transition into chaotic behavior, and ultimately result in divergence. In our work, we prove that this oscillatory behavior beyond the stability threshold also occurs in DLNs.

3.2. Assumptions and Analytical Tools

This section presents the two main tools used in our analyses: the singular vector stationary set and singular value balancedness. First, we present the singular vector stationary set, which allows us to encompass a wider range of initialization schemes. This set defines a broad class of initialization for which singular vector alignment occurs, simplifying the dynamics to only singular values.

Proposition 1 (Singular Vector Stationary Set)**.** *Consider the deep matrix factorization loss in Equation (1). Let $\mathbf{M}_\star = \mathbf{U}_\star \mathbf{\Sigma}_\star \mathbf{V}_\star^\top$ and $\mathbf{W}_\ell(t) = \mathbf{U}_\ell(t)\mathbf{\Sigma}_\ell(t)\mathbf{V}_\ell^\top(t)$ denote the compact SVD for the target matrix and the ℓ-th layer weight matrix at time t, respectively. For any time $t \geq 0$, if $\dot{\mathbf{U}}_\ell(t) = \dot{\mathbf{V}}_\ell(t) = 0$ for all $\ell \in [L]$, then the singular vector stationary points for each weight matrix are given by*

$$\mathrm{SVS}(f(\mathbf{\Theta})) = \begin{cases} (\mathbf{U}_L, \mathbf{V}_L) & = (\mathbf{U}_\star, \mathbf{Q}_L), \\ (\mathbf{U}_\ell, \mathbf{V}_\ell) & = (\mathbf{Q}_{\ell+1}, \mathbf{Q}_\ell), \quad \forall \ell \in [2, L-1], \\ (\mathbf{U}_1, \mathbf{V}_1) & = (\mathbf{Q}_2, \mathbf{V}_\star), \end{cases}$$

where $\{\mathbf{Q}_\ell\}_{\ell=2}^L$ are any set of orthogonal matrices.

The singular vector stationary set states that for any set of weights where the gradients with respect to the singular vectors become zero, the singular vectors become fixed points for subsequent iterations. Once the singular vectors become stationary, running GD further isolates the dynamics on the singular values. Hence, throughout our analysis, we re-write and consider the loss

$$\frac{1}{2}\left\|\mathbf{W}_{L:1}(t) - \mathbf{M}^\star\right\|_{\mathsf{F}}^2 = \frac{1}{2}\|\mathbf{\Sigma}_{L:1} - \mathbf{\Sigma}^\star\|_{\mathsf{F}}^2 = \frac{1}{2}\sum_{i=1}^{r}\left(\sigma_i(\mathbf{\Sigma}_{L:1}(t)) - \sigma_{\star,i}\right)^2, \tag{4}$$

where $\mathbf{\Sigma}_{L:1}$ are the singular values of $\mathbf{W}_{L:1}$. This allows us to decouple the dynamics of the singular vectors and singular values, focusing on the periodicity that occurs in the singular values within the EOS regime. In Propositions 2 and 3, we prove that both the unbalanced and balanced initializations belong to this set respectively, with an illustration in Figure 3. Specifically, we show that the singular vectors belongs to the singular vector stationary set after GD iteration $t = 1$ (far before entering the EOS regime) with $\mathbf{Q}_\ell = \mathbf{V}_\star$, allowing us to consider the loss in Equation (4). Next, we present a result to validate our use of the strictly balanced assumption on the unbalanced initialization case by showing that the singular values become increasingly balanced throughout the GD iterations.

Lemma 2 (Balancing)**.** *Let $\sigma_{\star,i}$ and $\sigma_{\ell,i}(t)$ denote the i-th singular value of $\mathbf{M}_\star \in \mathbb{R}^{d\times d}$ and $\mathbf{W}_\ell(t)$, respectively and define $S_i \coloneqq L\sigma_{\star,i}^{2-\frac{2}{L}}$. Consider GD on the i-th index of the simplified loss in (4) with the unbalanced initialization and learning rate $\frac{2}{S_i} < \eta < \frac{2\sqrt{2}}{S_i}$. If the initialization scale α satisfies $0 < \alpha < \left(\ln\left(\frac{2\sqrt{2}}{\eta S_i}\right) \cdot \frac{\sigma_{\star,i}^{4/L}}{L^2 \cdot 2^{\frac{2L-3}{L}}}\right)^{1/4}$, then there exists a constant $c \in (0,1]$ such that for all $\ell \in [L-1]$, we have $\left|\sigma_{L,i}^2(t+1) - \sigma_{\ell,i}^2(t+1)\right| < c\left|\sigma_{L,i}^2(t) - \sigma_{\ell,i}^2(t)\right|$.*

This result has been shown to hold similarly for two-layer matrix factorization [16, 34, 35], and our analysis extends it to the deep case. Lemma 2 considers the scalar loss for a single singular value

index and states that, as long as α is chosen below a threshold dependent on $\sigma_{\star,i}$, the i-th singular value will become increasingly balanced. To ensure that balancing occurs for all singular values of the loss in (4), we can select the learning rate to induce oscillations in all singular values (assuming they remain below the divergence limit) and choose α with $\sigma_{\star,1}$ such that it is the smallest α that satisfies the condition for all singular values $\sigma_{\star,i}$.

If the constant were $c < 1$, note that the balancing gap would approach zero infinitesimally. However, our analysis currently only shows that $c < 1$ when the product of singular values across all layers $\sigma_i(\mathbf{\Sigma}_{L:1}) < \sigma_{\star,i}$ and but $c = 1$ when $\sigma_i(\mathbf{\Sigma}_{L:1}) > \sigma_{\star,i}$. Since we start from a small initialization scale, we generally mostly operate within the regime $\sigma_i(\mathbf{\Sigma}_{L:1}) < \sigma_{\star,i}$, and only transition to the latter regime when oscillations occur. Note that $\sigma_i(\mathbf{\Sigma}_{L:1}) = \sigma_{\star,i}$ cannot occur since the learning rate is chosen to be within the EOS regime – equality can only arise in the stable regime, where balancing does not occur. In Figure 5, we plot the balancing gap between the top-3 singular values of a weight matrix initialized to zero and those initialized to α for a rank-3 matrix. This plot shows that the gap decreases and goes to zero empirically, and this is consistently the case across all of our experiments, with additional results provided in Appendix B. This provides empirical evidence that our analysis can be further improved such that $c < 1$ for both cases. To this end, we use this insights to assume that strict balancing holds for both the considered initialization schemes. This allows us to write the loss of the singular values into the form $\sigma_i(\mathbf{\Sigma}_{L:1}(t)) = \sigma_i^L(t)$, which allows us to focus on the dynamics in the singular values.

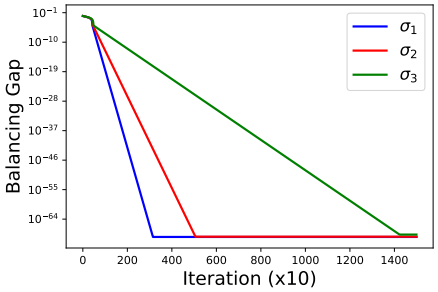


Figure 5: Plot of $|\sigma_{L,i}^2(t) - \sigma_{\ell,i}^2(t)|$ showing strict balancing.

3.3. Relation to Diagonal Linear Networks

In this section, we derive the necessary conditions and characteristics of two-period oscillations in diagonal linear networks to establish their similarities with DLNs.

Theorem 3 (Subspace Oscillation for Diagonal Linear Networks)**.** *Consider an L-layer diagonal linear network on the loss*

$$\mathcal{L}\left(\{\mathbf{s}_\ell\}_{\ell=1}^L\right) := \frac{1}{2}\|\mathbf{s}_1 \odot \ldots \odot \mathbf{s}_L - \mathbf{s}_\star\|_2^2, \tag{5}$$

where $\mathbf{s}_\star \in \mathbb{R}^d$ be an r-sparse vector with ordered coordinates such that $s_{\star,1} > \ldots > s_{\star,d}$ and define $S_p := Ls_{\star,p}^{2-\frac{2}{L}}$ and $\alpha' := \left(\ln\left(\frac{2\sqrt{2}}{\eta L s_{\star,1}^{2-\frac{2}{L}}}\right) \cdot \frac{s_{\star,1}^{\frac{4}{L}}}{L^2 \cdot 2^{\frac{2L-3}{L}}}\right)^{\frac{1}{4}}$. For any $p < r-1$ and $\alpha < \alpha'$, suppose we run GD on Equation (5) with learning rate $\eta = \frac{2}{K}$, where $S_p \geq K > S_{p+1}$ with initialization $\mathbf{s}_\ell = \alpha \mathbf{1}_d$ for all $\ell \in [L-1]$ and $\mathbf{s}_L = \mathbf{0}_d$. Then, under strict balancing, the top-p coordinates of $\mathbf{s}_\ell$ oscillate within a 2-period fixed orbit around the minima in the form

$$s_{\ell,i}(t) = \rho_{i,j}(t), \quad \forall i < p, \forall \ell \in [L],$$

where $\rho_{i,j}(t) \in \{\rho_{i,1}, \rho_{i,2}\}$, $\rho_{i,1} \in \left(0, s_{\star,i}^{1/L}\right)$ and $\rho_{i,2} \in \left(s_{\star,i}^{1/L}, (2s_{\star,i})^{1/L}\right)$ are two real roots of the polynomial $h(\rho) = 0$:

$$h(\rho) = \rho^L \cdot \frac{1 + \left(1 + \eta L(s_{\star,i} - \rho^L) \cdot \rho^{L-2}\right)^{2L-1}}{1 + \left(1 + \eta L(s_{\star,i} - \rho^L) \cdot \rho^{L-2}\right)^{L-1}} - s_{\star,i}.$$

From Lemma 1, we observe that DLNs exhibit additional dominant curvature directions that are not present in diagonal linear networks—specifically, the eigenvalues corresponding to the "interaction with other singular values" are absent in diagonal linear networks. These eigenvalues arise due to

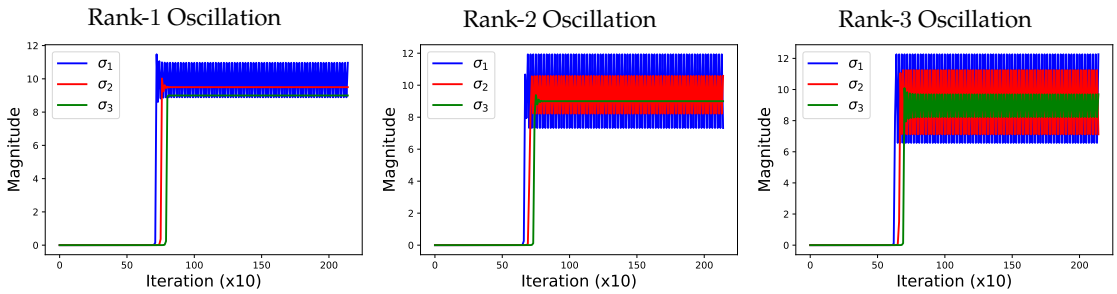


Figure 6: Evolution of the singular values of the end-to-end 3-layer network for fitting a rank-3 target matrix with singular values 10, 9.5, and 9. We use a learning rate of $\eta = 2/S_i$ with $S_i := L\sigma_{\star,i}^{2-2/L}$. The oscillations occur exactly with learning rate ranges specified in Theorem 2.

the zero off-diagonal elements of the singular value diagonal matrix in (4). However, despite these extra directions, Theorem 3 demonstrates that the behavior of diagonal linear networks and DLNs is essentially the same. This can be understood using the argument from Theorem 1: the eigenvectors corresponding to these directions are always orthogonal to the flattened weights $\widetilde{\Delta}$, thereby making the behaviors of the two network types synonymous, even in the EOS regime.

4. Experimental Results

4.1. Subspace Oscillations in Deep Networks

Firstly, we provide experimental results corroborating Theorem 2. We let the target matrix be $\mathbf{M}_\star \in \mathbb{R}^{50\times 50}$ with rank 3, with dominant singular values $\sigma_\star = 10, 9.5, 9$. For the DLN, we consider a 3-layer network, with each layer as $\mathbf{W}_\ell \in \mathbb{R}^{50\times 50}$ and use an initialization scale of $\alpha = 0.01$. In Figure 6, we present the behaviors of the singular values of the end-to-end network under different learning rate regimes. Recall that by Theorem 2, the i-th singular value undergoes periodic oscillations when K is set to be $S_i < K < S_{i+1}$, where $S_i = L\sigma_{\star,i}^{2-2/L}$. Figure 6 illustrates this clearly – we only observe oscillations in the i-th coordinate depending on the learning rate. Interestingly, notice that σ_2 also begins to oscillate in the rank-1 oscillation region before settling at a minimum. This occurs because, while the learning rate is large enough to catapult out of a minimum, it is not sufficiently large to induce periodic oscillations.

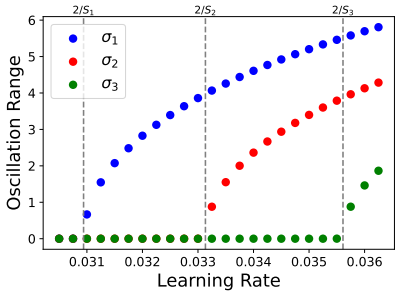


Figure 7: Oscillation range as a function of the learning rate.

In Figure 7, we present an experiment demonstrating the relationship between the oscillation range and the learning rate by plotting the amplitude of the singular value oscillations in the end-to-end network. The oscillations begin to occur starting from each region $\eta = 2/S_i$, and the oscillation range (or amplitude) increases as the learning rate increases. This can also be observed in Figure 6; the amplitude of σ_1 increases as we move from the rank-1 to the rank-3 oscillation region.

4.2. Similarities and Differences Between Linear and Nonlinear Nets at EOS

Mild Sharpening. "Mild" sharpening refers to the sharpness not rising to $2/\eta$ throughout learning, and generally occurs in tasks with low complexity as discussed in Caveat 2 of [1]. We il-

lustrate mild sharpening in Figure 10, where we plot sharpness in two settings: (i) regression with simple images and (ii) classification with an MLP using a subset of the CIFAR-10 dataset.

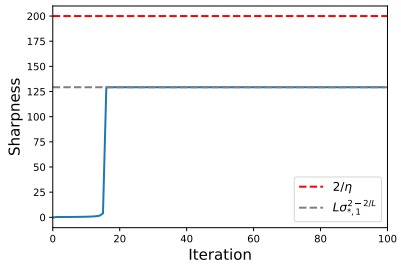


Figure 8: DLNs do not enter EOS regime if $L\sigma_1^{2-\frac{2}{L}} < 2/\eta$.

For the regression task, we minimize the loss $\mathcal{L}(\Theta) = \|G(\Theta) - \mathbf{y}_{\text{image}}\|_2^2$, where $G(\Theta)$ is a UNet parameterized by Θ, and $\mathbf{y}_{\text{image}}$ denotes one of the images in Figure 10b. We observe that when $\mathbf{y}_{\text{image}}$ is a smooth, low-frequency image, the sharpness of the loss generally remains low. However, when $\mathbf{y}_{\text{image}}$ has higher frequency content, the sharpness increases and enters the EOS regime (Figure 10a). Similarly, for the classification task, we train a 2-layer fully connected neural network on N labeled training images from the CIFAR-10 dataset using MSE loss and plot the sharpness in Figure 10c. The sharpness links to N, the number of data points used for training. For small N values, such as 100 or 200, the network learns only a limited set of latent features, resulting in mild sharpening, and it does not reach the EOS threshold. However, when N exceeds 1000, the sharpness increases and reaches the EOS threshold. The intrinsic dimension update in neural networks for such low complexity tasks is usually smaller [36] which could cause the sharpness to be small. Similar observations can also be seen in DLNs. In Figure 8, we show that the sharpness reaches $L\sigma_{\star,1}^{2-\frac{2}{L}}$, where $\sigma_{\star,1}$ is the singular value of the target matrix. Whenever $L\sigma_{\star,1}^{2-\frac{2}{L}} < 2/\eta$, the network will not enter the EOS regime. This can be viewed as low-complexity learning, as $\sigma_{\star,1}$ corresponds to the magnitude of the strongest feature of the target matrix. Hence, when $\sigma_{\star,1}$ is not large enough, the sharpness will not rise to $2/\eta$. While these observations do not fully explain mild sharpening, our experiments demonstrate that interpreting sharpness as a measure of complexity, combined with our findings from DLNs, marks an important first step toward fully understanding this phenomenon.

Difference in Oscillation Behaviors. Here, we discuss the differences in oscillations that arise in DLNs compared to catapults that occur in practical deep nonlinear networks. The main difference lies in the loss landscape—at convergence, the Hessian for DLNs is positive semi-definite, as shown in Lemma 1, meaning there are only directions of positive curvature and flat directions (in the null space of the Hessian). In this landscape, oscillations occur because the basin walls bounce off, without the direction of escape. However, in deep nonlinear networks, it has been frequently observed that the Hessian at the minima has negative eigenvalues [37, 38]. This enables an escape direction along the negative curvature, preventing sustained oscillations. In Figure 9, we demonstrate these two differences by visualizing the loss landscapes and the iterates throughout GD marked in red. The Holder table function Figure 9 (left) exhibits numerous local minima, causing the loss to exhibit a sharp "catapult" when a large learning rate is used. In contrast, for DLNs (shown in the right) the loss oscillates in a periodic orbit around the minima since there are no spurious local minima [39–43].

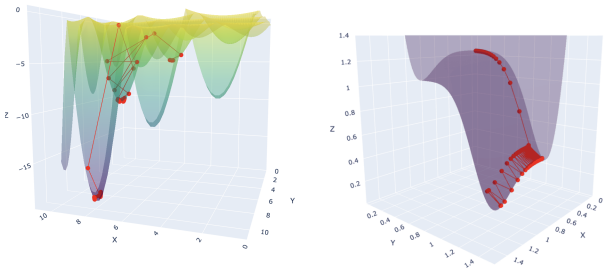
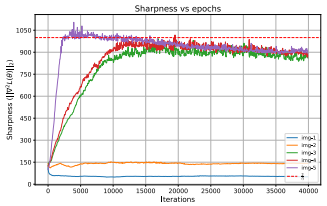
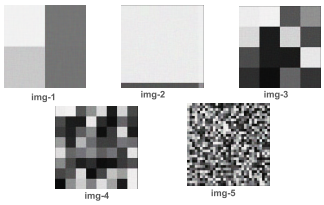


Figure 9: Loss landscape of the Holder table function and DLNs, respectively (left–right). The Holder table function is non-convex which allows catapulting to other minima, whereas DLNs do not have spurious local minima.

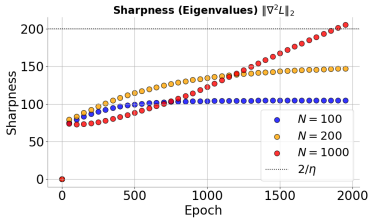
Lastly, [5] studies self-stabilization, where sharpness decreases below $2/\eta$ after initially exceeding $2/\eta$. Their analysis requires assumptions such as $\nabla L(\theta) \cdot u(\theta) = 0$ and $\nabla S(\theta)$ lies in the null space of the Hessian, where $S(\theta)$ and $u(\theta)$ denotes the sharpness and its corresponding



(a) Sharpness plots for training image generator networks using SGD with learning rate $\eta = 2 \times 10^{-4}$.



(b) Target images (denoted as $\mathbf{y}_{\text{image}}$) with different frequencies used for training.



(c) 2-layer FC network trained with small number N of CIFAR-10 dataset with $\eta = 10^{-2}$

Figure 10: Illustration of Caveat 2 by [1] on how mild sharpening occurs on simple datasets and network. (a) Regression task showing the evolution of the sharpness when an UNet (with fixed initialization) is trained to fit a single image shown in (b). (c) Evolution of the minimal progressive sharpening on a classification task of a 2-layer MLP trained on a subset of CIFAR-10.

eigenvector respectively. These assumptions do not hold exactly in DLNs. Rather, the sharpness oscillates about $2/\eta$ as shown in Figure 2 as the condition for stable oscillation holds along each eigenvector of the Hessian.

5. Conclusion, Limitations and Future Work

In this paper, we presented a fine-grained analysis of the learning dynamics of deep matrix factorization with the aim of understanding unexplained phenomena in deep nonlinear networks within the EOS regime. Our analysis revealed that within EOS, DLNs exhibit periodic oscillations in small subspaces, where the subspace dimension is exactly characterized by the learning rate. There are two limitations to our work: we require (i) the dynamics converge to the singular vector stationary set, and (ii) strict balancing of the singular values. However, we provide thorough empirical evidence validating the use of these assumptions, along with more results in Appendix B. For the balancing assumption, we leave for future work on alleviating the assumption of strict balancing, and rigorously show that this holds before entering the EOS regime.

References

[1] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

[2] Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Catapults in SGD: spikes in the training loss and their impact on generalization through feature learning. *arXiv preprint arXiv:2306.04815*, 2023.

[3] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=NEv8YqBROO.

[4] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

[5] Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nhKHA59gXz.

[6] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

[7] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2019. URL https://arxiv.org/abs/1803.05407.

[8] Henning Petzka, Michael Kamp, Linara Adilova, Cristian Sminchisescu, and Mario Boley. Relative flatness and generalization. In *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=sygvo7ctb_.

[9] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

[10] Khashayar Gatmiry, Zhiyuan Li, Tengyu Ma, Sashank J. Reddi, Stefanie Jegelka, and Ching-Yao Chuang. What is the inductive bias of flatness regularization? A study of deep matrix factorization models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=2hQ7MBQApp.

[11] Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=p7EagBsMAEO.

[12] Yuqing Wang, Zhenghao Xu, Tuo Zhao, and Molei Tao. Good regularity creates large learning rate implicit biases: edge of stability, balancing, and catapult. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023. URL https://openreview.net/forum?id=6O15A3h2yl.

[13] Itai Kreisler, Mor Shpigel Nacson, Daniel Soudry, and Yair Carmon. Gradient descent monotonically decreases the sharpness of gradient flow solutions in scalar networks and beyond. In *International Conference on Machine Learning*, pages 17684–17744. PMLR, 2023.

[14] Atish Agarwala, Fabian Pedregosa, and Jeffrey Pennington. Second-order regression models exhibit progressive sharpening to the edge of stability. *arXiv preprint arXiv:2210.04860*, 2022.

[15] Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (S)GD over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates, 2023.

[17] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *2nd International Conference on Learning Representations, ICLR*, 2014. URL http://arxiv.org/abs/1312.6120.

[18] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International conference on machine learning*, pages 244–253. PMLR, 2018.

[19] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/c0c783b5fc0d7d808f1d14a6e9c8280d-Paper.pdf.

[20] Chong You, Zhihui Zhu, Qing Qu, and Yi Ma. Robust recovery via implicit bias of discrepant learning rates for double over-parameterization. *Advances in Neural Information Processing Systems*, 33:17733–17744, 2020.

[21] Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. In *International Conference on Machine Learning*, pages 14153–14172. PMLR, 2022.

[22] Xitong Zhang, Ismail R Alkhouri, and Rongrong Wang. Structure-preserving network compression via low-rank induced training through linear layers composition. *arXiv preprint arXiv:2405.03089*, 2024.

[23] Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. *Advances in Neural Information Processing Systems*, 36:7475–7505, 2023.

[24] Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2022.

[25] Peng Wang, Xiao Li, Can Yaras, Zhihui Zhu, Laura Balzano, Wei Hu, and Qing Qu. Understanding deep representation learning via layerwise feature compression and discrimination. *arXiv preprint arXiv:2311.02960*, 2024. URL https://arxiv.org/abs/2311.02960.

[26] Yedi Zhang, Andrew M Saxe, and Peter E. Latham. When are bias-free reLU networks like linear networks? In *High-dimensional Learning Dynamics 2024: The Emergence of Structure and Reasoning*, 2024. URL https://openreview.net/forum?id=LdYBMeWOG3.

[27] Can Yaras, Peng Wang, Wei Hu, Zhihui Zhu, Laura Balzano, and Qing Qu. The law of parsimony in gradient descent for learning deep linear networks. *arXiv preprint arXiv:2306.01154*, 2023.

[28] Edward Ott. *Chaos in Dynamical Systems*. Cambridge University Press, 2 edition, 2002.

[29] Aditya Vardhan Varre, Maria-Luiza Vladarean, Loucas PILLAUD-VIVIEN, and Nicolas Flammarion. On the spectral bias of two-layer linear networks. In *Advances in Neural Information Processing Systems*, volume 36, pages 64380–64414, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/cad2fd66cf88226d868f90a7cbaa4a53-Paper-Conference.pdf.

[30] Zhenghao Xu, Yuqing Wang, Tuo Zhao, Rachel Ward, and Molei Tao. Provable acceleration of nesterov's accelerated gradient for rectangular matrix factorization and linear neural networks. *arXiv preprint arXiv:2410.09640*, 2024.

[31] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/f39ae9ff3a81f499230c4126e01f421b-Paper.pdf`.

[32] Hung-Hsu Chou, Carsten Gieshoff, Johannes Maly, and Holger Rauhut. Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *Applied and Computational Harmonic Analysis*, 68:101595, 2024.

[33] Soo Min Kwon, Zekai Zhang, Dogyoon Song, Laura Balzano, and Qing Qu. Efficient low-dimensional compression of overparameterized models. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 1009–1017. PMLR, 02–04 May 2024. URL `https://proceedings.mlr.press/v238/min-kwon24a.html`.

[34] Yuqing Wang, Minshuo Chen, Tuo Zhao, and Molei Tao. Large learning rate tames homogeneity: Convergence and balancing effect. *arXiv preprint arXiv:2110.03677*, 2021.

[35] Tian Ye and Simon S Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. *Advances in Neural Information Processing Systems*, 34:1429–1439, 2021.

[36] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

[37] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019.

[38] Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

[39] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in neural information processing systems*, 29, 2016.

[40] Kenji Kawaguchi. Deep learning without poor local minima. *Advances in neural information processing systems*, 29, 2016.

[41] Haihao Lu and Kenji Kawaguchi. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.

[42] Li Zhang. Depth creates no more spurious local minima. *arXiv preprint arXiv:1901.09827*, 2019.

[43] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small nonlinearities in activation functions create bad local minima in neural networks. *arXiv preprint arXiv:1802.03487*, 2018.

[44] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. *arXiv preprint arXiv:2002.09572*, 2020.

[45] Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the relation between the sharpest directions of dnn loss and the sgd step length. *arXiv preprint arXiv:1807.05031*, 2018.

[46] Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pages 948–1024. PMLR, 2022.

[47] Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. *Advances in Neural Information Processing Systems*, 35:34689–34708, 2022.

[48] Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. *Advances in Neural Information Processing Systems*, 36, 2024.

[49] Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability. *Advances in Neural Information Processing Systems*, 35:9983–9994, 2022.

[50] Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*, 36, 2024.

[51] Minhak Song and Chulhee Yun. Trajectory alignment: understanding the edge of stability phenomenon via bifurcation theory. *arXiv preprint arXiv:2307.04204*, 2023.

[52] Dayal Singh Kalra, Tianyu He, and Maissam Barkeshli. Universal sharpness dynamics in neural network training: Fixed point analysis, edge of stability, and route to chaos. *arXiv preprint arXiv:2311.02076*, 2023.

[53] Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Quadratic models for understanding neural network dynamics. *arXiv preprint arXiv:2205.11787*, 2022.

[54] Xuxing Chen, Krishnakumar Balasubramanian, Promit Ghosal, and Bhavya Agrawalla. From stability to chaos: Analyzing gradient descent dynamics in quadratic regression. *arXiv preprint arXiv:2310.01687*, 2023.

[55] Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1lj0nNFwB.

[56] Hung-Hsu Chou, Carsten Gieshoff, Johannes Maly, and Holger Rauhut. Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *Applied and Computational Harmonic Analysis*, 68:101595, 2024. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2023.101595. URL https://www.sciencedirect.com/science/article/pii/S1063520323000829.

[57] Can Yaras, Peng Wang, Laura Balzano, and Qing Qu. Compressible dynamics in deep overparameterized low-rank learning & adaptation. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=uDkXoZMzBv.

[58] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[59] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

[60] Leon Mirsky. A trace inequality of John von Neumann. *Monatshefte für mathematik*, 79(4): 303–306, 1975.

[61] Pierre Marion and Lénaïc Chizat. Deep linear networks for regression are implicitly regularized towards flat minima. *arXiv preprint arXiv:2405.13456*, 2024.

Appendix

Contents

A. Discussion on Related Work

Implicit Bias of Edge of Stability. Edge of stability was first coined by [1], where they showed that the Hessian of the training loss plateaus around $2/\eta$ when deep models were trained using GD. However, [44, 45] previously demonstrated that the step size influences the sharpness along the optimization trajectory. Due to the important practical implications of the edge of stability, there has been an explosion of research dedicated to understanding this phenomenon and its implicit regularization properties. Here, we survey a few of these works. [5] explained edge of stability through

a mechanism called "self-stabilization", where they showed that during the momentary divergence of the iterates along the sharpest eigenvector direction of the Hessian, the iterates also move along the negative direction of the gradient of the curvature, which leads to stabilizing the sharpness to $2/\eta$. [14] proved that second-order regression models (the simplest class of models after the linearized NTK model) demonstrate progressive sharpening of the NTK eigenvalue towards a slightly different value than $2/\eta$. [46] mathematically analyzed the edge of stability, where they showed that the GD updates evolve along some deterministic flow on the manifold of the minima. [47] showed that the normalization layers had an important role in the edge of stability – they showed that these layers encouraged GD to reduce the sharpness of the loss surface and enter the EOS regime. [48] established the phenomenon in two-layer networks and find phase transitions for step-sizes in which networks fail to learn "threshold" neurons. [49] also analyze a two-layer network, but provide a theoretical proof for the change in sharpness across four different phases. [15] analyzed the edge of stability in diagonal linear networks and found that oscillations occur on the sparse support of the vectors. Lastly, [50] analyzed the convergence at the edge of stability for constant step size GD for logistic regression on linearly separable data.

Edge of Stability in Toy Functions. To analyze the edge of stability in slightly simpler settings, many works have constructed scalar functions to analyze the prevalence of this phenomenon. For example, [16] studied a certain class of scalar functions and identified conditions in which the function enters the edge of stability through a two-step convergence analysis. [12] showed that the edge of stability occurs in specific scalar functions, which satisfies certain regularity conditions and developed a global convergence theory for a family of non-convex functions without globally Lipschitz continuous gradients. [11] analyzed local oscillatory behaviors for 4-layer scalar networks with balanced initialization. [51, 52] provide analyses of learning dynamics at the EOS in simplified settings such as two-layer networks. [53, 54] study GD dynamics for quadratic models in large learning rate regimes. Overall, all of these works showed that the necessary condition for the edge of stability to occur is that the second derivative of the loss function is non-zero, even though they assumed simple scalar functions. Our work takes one step further to analyze the prevalence of the edge of stability in DLNs. Although our loss simplifies to a loss in terms of the singular values, they precisely characterize the dynamics of the DLNs for the deep matrix factorization problem.

Deep Linear Networks. Over the past decade, many existing works have analyzed the learning dynamics of DLNs as a surrogate for deep nonlinear networks to study the effects of depth and implicit regularization [17–19, 22]. Generally, these works focus on unveiling the dynamics of a phenomenon called "incremental learning", where small initialization scales induce a greedy singular value learning approach [17, 33, 55], analyzing the learning dynamics via gradient flow [17, 19, 56], or showing that the DLN is biased towards low-rank solution [19, 33, 57], amongst others. However, these works do not consider the occurence of the edge of stability in such networks. On the other hand, while works such as those by [57] and [33] have similar observations in that the weight updates occur within an invariant subspace as shown by Proposition 2, they do not analyze the edge of stability regime.

B. Additional Results

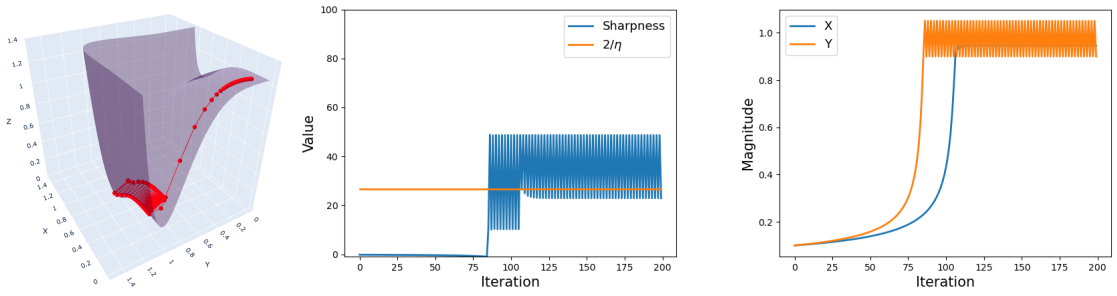
B.1. Experimental Details

In this section, we provide additional details regarding the experiments used to generate the figures in the main text. For Figure 1, we consider a rank-3 target matrix $\mathbf{M}_\star \in \mathbb{R}^{5\times 5}$ with ordered singular values $10, 6, 3$. We use a 3-layer DLN to fit the target matrix. Since $\sigma_{\star,1} = 10$, the network enters the EOS regime at

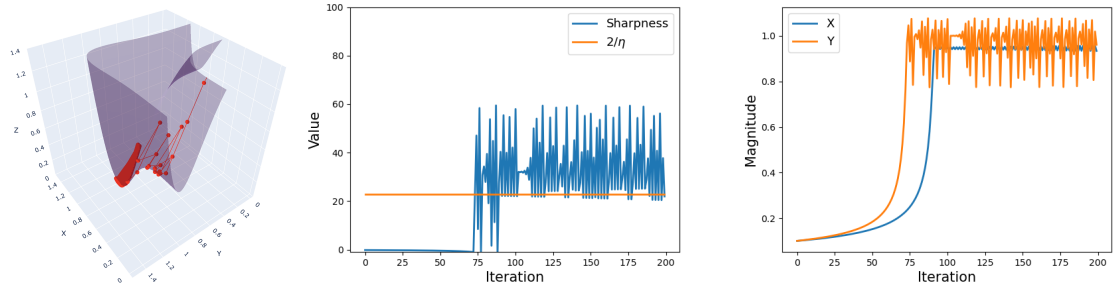
$$\eta = \frac{2}{L\sigma_{\star,1}^{2-2/L}} = 0.0309.$$

We show that there exists a two-period orbit after $0.0309/2 = 0.0154$, as we do not have a scaling of $1/2$ in the objective function for the code used to generate the figures.

Oscillation along Y-axis: $2/\lambda_2 > \eta > 2/\lambda_1$



Oscillation along both X and Y-axis: $\eta > 2/\lambda_2$



Loss Landscape | Sharpness | Oscillatory Components

Figure 11: Demonstration of the EOS dynamics of a 2-dimensional depth-4 scalar network as shown in Equation (6). X, Y axes are the eigenvectors of the Hessian with eigenvalues λ_1 and λ_2 respectively. Top: when $\eta > 2/\lambda_1$, the X component remains fixed, while the Y component oscillates with a periodicity of 2. Bottom: for $\eta > 2/\lambda_2$, the iterates oscillation in both directions.

In Figure 9 and 11, we compared the landscape of DLNs with that of a more complicated non-convex function such as the Holder table function. To mimic the DLN, we considered the loss function

$$z = L(x, y) = (x^4 - 0.8)^2 + (y^4 - 1)^2, \tag{6}$$

which corresponds to a 4-layer network. Here the eigenvector of the Hessian at the global minima coincides with the x, y-axis. We calculate the eigenvalues λ_1 and λ_2 at the minimum $(0.8^{0.25}, 1)$ and plot the dynamics of the iterates for step size range $\frac{2}{\lambda_2} > \eta > \frac{2}{\lambda_1}$ and $\eta > \frac{2}{\lambda_2}$. When $\frac{2}{\lambda_2} > \eta > \frac{2}{\lambda_1}$ the x-coordinate stays fixed at the minima $0.8^{0.25}$ and the y-coordinate oscillates around its minimum at $y = 1$. This is evident in the landscape figure. Similarly, when $\eta > \frac{2}{\lambda_2}$, oscillations occur in both the x and y direction. The loss landscape $z = L(x, y)$ does not have spurious local minima, so sustained oscillations take place in the loss basin.

For the non-convex landscape as shown in Figure 9 and 12, we consider the Holder table function:

$$f(x, y) = -\left|\sin(x)\cos(y)\exp\left(1 - \frac{\sqrt{x^2 + y^2}}{\pi}\right)\right|.$$

By observation, we initialize near a sharp minima and run GD with an increasing learning rate step size as shown in the lefthand side of Figure 12. When the learning rate is fixed, we observe that oscillations take place inside the local valley, but when learning rate is increased, it jumps out of the local valley to find a flatter basin. Similar to the observations by [1], the sharpness of the GD iterates are "regulated" by the threshold $2/\eta$, as it seems to closely follow this value as shown in Figure 12.

Overall, these examples aim to highlight the difference in linear and complex loss landscapes. The former consists of *only* saddles and global minima, and hence (stably) oscillate about the global

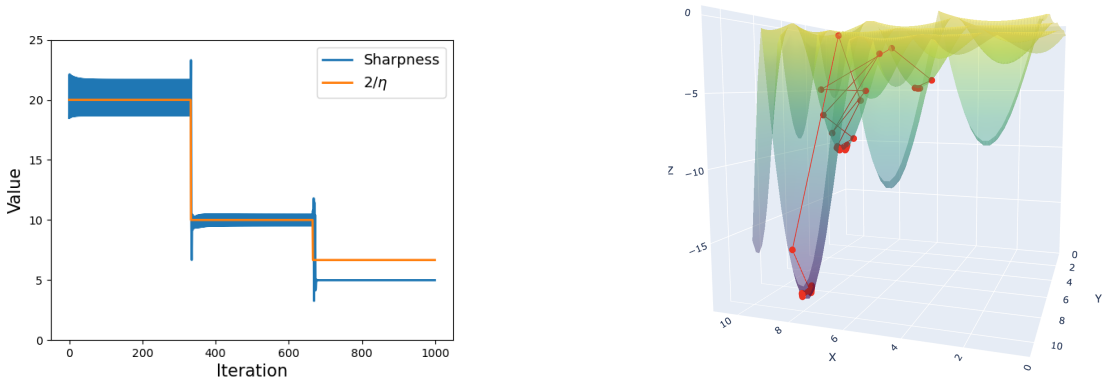


Figure 12: EOS dynamics at various step learning rates from the Holder table function. Left: plot of the learning rate steps and sharpness, showing that sharpness follows the EOS limit $2/\eta$. Right: Plot showing that the iterates catapult out of a local basin when the learning rate is increased and jumps out to a surface where the sharpness is about $2/\eta$.

minimum. However, in more complicated non-convex landscapes, sharpness regularization due to large learning rates enable catapulting to flatter loss basins, where sharpness is smaller than $2/\eta$.

B.2. Initialization Outside Singular Vector Invariant Set

In this section, we present an initialization example that is outside the singular vector stationary set. We consider the following initialization:

$$\mathbf{W}_L(0) = \mathbf{0}, \qquad \mathbf{W}_\ell(0) = \alpha \mathbf{P}_\ell, \quad \forall \ell \in [L-1], \tag{7}$$

where $\mathbf{P}_\ell \in \mathbb{R}^{d\times d}$ is an orthogonal matrix. Note that here for $\ell > 1$, the singular vectors do not align and lies outside the SVS set we defined in Proposition 2. We consider the deep matrix factorization problem with a target matrix $\mathbf{M}_\star \in \mathbb{R}^{d\times d}$, where $d = 100$, $r = 5$, and $\alpha = 0.01$. We empirically obtain that the decomposition after convergence admits the form:

$$\mathbf{W}_L(t) = \mathbf{U}^\star \begin{bmatrix} \mathbf{\Sigma}_L(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \left[\left(\prod_{i=L-1}^{1} \mathbf{P}_i \right) \mathbf{V}^\star \right]^\top, \tag{8}$$

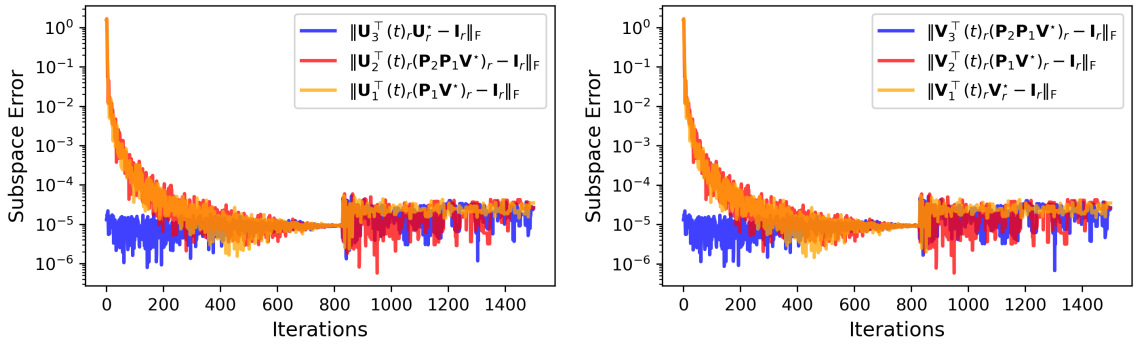
$$\mathbf{W}_\ell(t) = \left[\left(\prod_{i=\ell}^{1} \mathbf{P}_i \right) \mathbf{V}^\star \right] \begin{bmatrix} \mathbf{\Sigma}_\ell(t) & \mathbf{0} \\ \mathbf{0} & \alpha \mathbf{I}_{d-r} \end{bmatrix} \left[\left(\prod_{i=\ell-1}^{1} \mathbf{P}_i \right) \mathbf{V}^\star \right]^\top, \quad \forall \ell \in [2, L-1], \tag{9}$$

$$\mathbf{W}_1(t) = \mathbf{P}_1 \mathbf{V}^\star \begin{bmatrix} \mathbf{\Sigma}_1(t) & \mathbf{0} \\ \mathbf{0} & \alpha \mathbf{I}_{d-r} \end{bmatrix} \mathbf{V}^{\star\top}, \tag{10}$$

where $\mathbf{W}_L(0) = \mathbf{0}$ and $\mathbf{W}_\ell(0) = \alpha \mathbf{P}_l$, $\forall \ell \in [L-1]$. The decomposition after convergence lies in the SVS set as the singular vectors now align with each other. This demonstrates an example where even when the initialization is made outside the SVS set, GD aligns the singular vectors such that after certain iterations it lies in the SVS set.

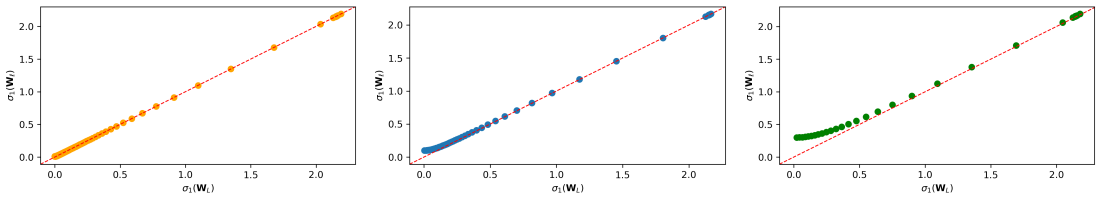
B.3. Balancing of Singular Values

In this section, we present additional experimental results on Lemma 2 and how close the iterates become for different initialization scales. To this end, we consider the same setup from the previous section, where we have a target matrix $\mathbf{M}_\star \in \mathbb{R}^{d\times d}$, where $d = 100$, $r = 5$, and varying initialization α. In Figure 14, we observe that for larger values of α, the balancing quickly occurs, whereas for smaller values of α, the balancing is almost immediate. This is to also highlight that our bound on α in Lemma 2 may be an artifact of our analysis, and can choose larger values of α in practice.



Left Singular Vectors Right Singular Vectors

Figure 13: Empirical verification of the decomposition for initialization with orthogonal matrices (lying outside SVS set) in that after some GD iterations, the singular vectors of the intermediate matrices align to lie within SVS set, displaying singular vector invariance.



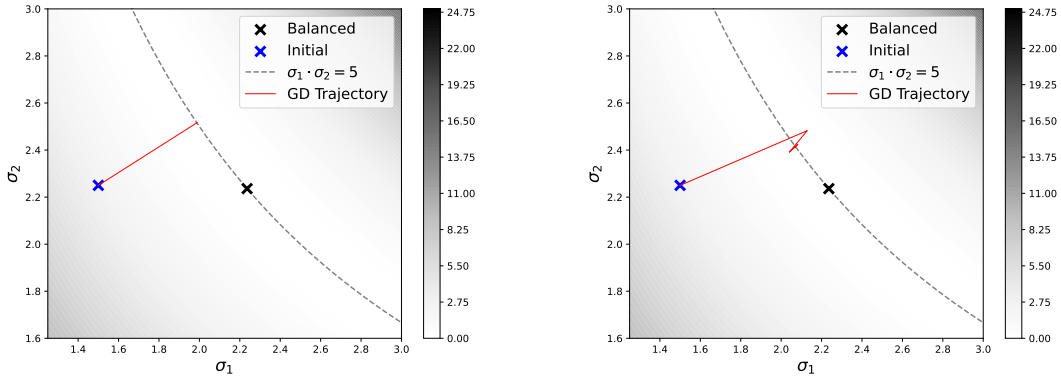
$\alpha = 0.01$ $\alpha = 0.10$ $\alpha = 0.30$

Figure 14: Observing the balancedness between the singular value initialized to 0 and a singular value initialized to α. The scattered points are successive GD iterations (going left to right). The initial gap between the two values is larger for a larger α, but quickly gets closer over more GD iterations.

Balancing in the Stable Regime. Here, we discuss the difference in the balancing of singular values within the stable regime versus the EOS regime. In Figure 4, we noted that strict balancing and the two-period orbit occur simultaneously, which we used as a basis for considering strict balancing in our analyses. However, we observe that strict balancing does not actually occur within the stable regime. To illustrate this, we present additional contour plots in Figure 15, following the setup in Figure 4. When starting from an unbalanced initialization, GD tends to converge to any solution that minimizes the objective function to zero (i.e., $\sigma_1\sigma_2 = 5$), which is not necessarily the balanced solution.

In Figure 16, we plot the balancing gap between the singular value initialized to zero and the one initialized to α for fitting a rank-3 matrix with singular values 10, 9.5, and 9 using a 3-layer DLN. In the stable regime, where oscillations do not occur, the balancing gap plateaus. However, when the learning rate is large enough to induce oscillations in the first singular value of the DLN, the balancing gap for the first singular value strictly goes to zero. Similarly, when oscillations occur in the second singular value, the balancing gap at the second index also goes to zero. This suggests that the balancing gap strictly goes to zero only in the presence of oscillations (i.e., within the EOS regime).

We provide intuition for why balancing does not occur in the stable regime as it does in the EOS regime. In the proof for Lemma 2, we define the balancing dynamics between singular value σ_i and



$\eta = 0.180$ $\eta = 0.195$

Figure 15: Illustration of the GD trajectories for two different learning rates for minimizing the function $f(\sigma_1, \sigma_2) = \frac{1}{2}(\sigma_2 \cdot \sigma_1 - 5)^2$, starting from an unbalanced initial point. The sharpness at the balanced minimum is 10, and so the learning rate needed to enter EOS is $\eta = 0.2$.

σ_j at iteration t as

$$b_{i,j}^{(t+1)} := \left(\sigma_i^{(t+1)}\right)^2 - \left(\sigma_j^{(t+1)}\right)^2$$
$$= b_{i,j}^{(t)} \left(1 - \eta^2 (\pi^{(t)} - \sigma_\star)^2 \frac{\left(\pi^{(t)}\right)^2}{\left(\sigma_i^{(t)}\right)^2 \left(\sigma_j^{(t)}\right)^2}\right),$$

where $\pi^{(t)}$ denotes the product of the singular values across all layers at time t. Then, notice that in the stable regime, as soon as we reach a solution such that $\pi^{(t)} = \sigma_\star$, we obtain $b_{i,j}^{(t+1)} = b_{i,j}^{(t)}$, and so the balancing gap must plateau. However, in the EOS regime, we can never achieve $\pi^{(t)} = \sigma_\star$ due to oscillations, which allows the balancing gap to further decrease.

B.4. Additional Experiments for Balancing, Singular Vector Invariance, and Theory

Our theory relied on two tools and assumptions: balancing of singular values and stationarity of the singular vectors. In this section, we investigate how the dynamics at EOS are affected if these two assumptions do not hold.

Balancing. By Lemma 2, recall that balancing only holds as long as α chosen below a certain threshold. To this end, we consider the dynamics of a 3-layer DLN to fit a target matrix $\mathbf{M}_\star \in \mathbb{R}^{10\times 10}$ of rank-3 with ordered singular values $10, 8, 6$. We use a learning rate of $\eta = 0.0166$, which corresponds to oscillations in the top-2 singular values. In Figure 17, we show the dynamics of when the initialization scale is $\alpha = 0.01$ and $\alpha = 0.5$, where balancing holds theoretically for the former but not for the latter. Clearly, we observe that balancing does not hold for $\alpha = 0.5$. However, examining the middle plots reveals that the oscillations in the singular values still have the same amplitude in both cases and for both singular values. This suggests that balancing is merely a tool for analysis, as the oscillations of interest remain prevalent in both scenarios.

Singular Vector Stationarity. Throughout this paper, we considered two initializations in Equation (**??**), where balancing holds immediately and one where balancing holds for a sufficiently small initialization scale. In this section, we investigate different initializations with aim to observe (i) if they do not converge to the SVS set and (ii) how they affect the oscillations if they do not belong to

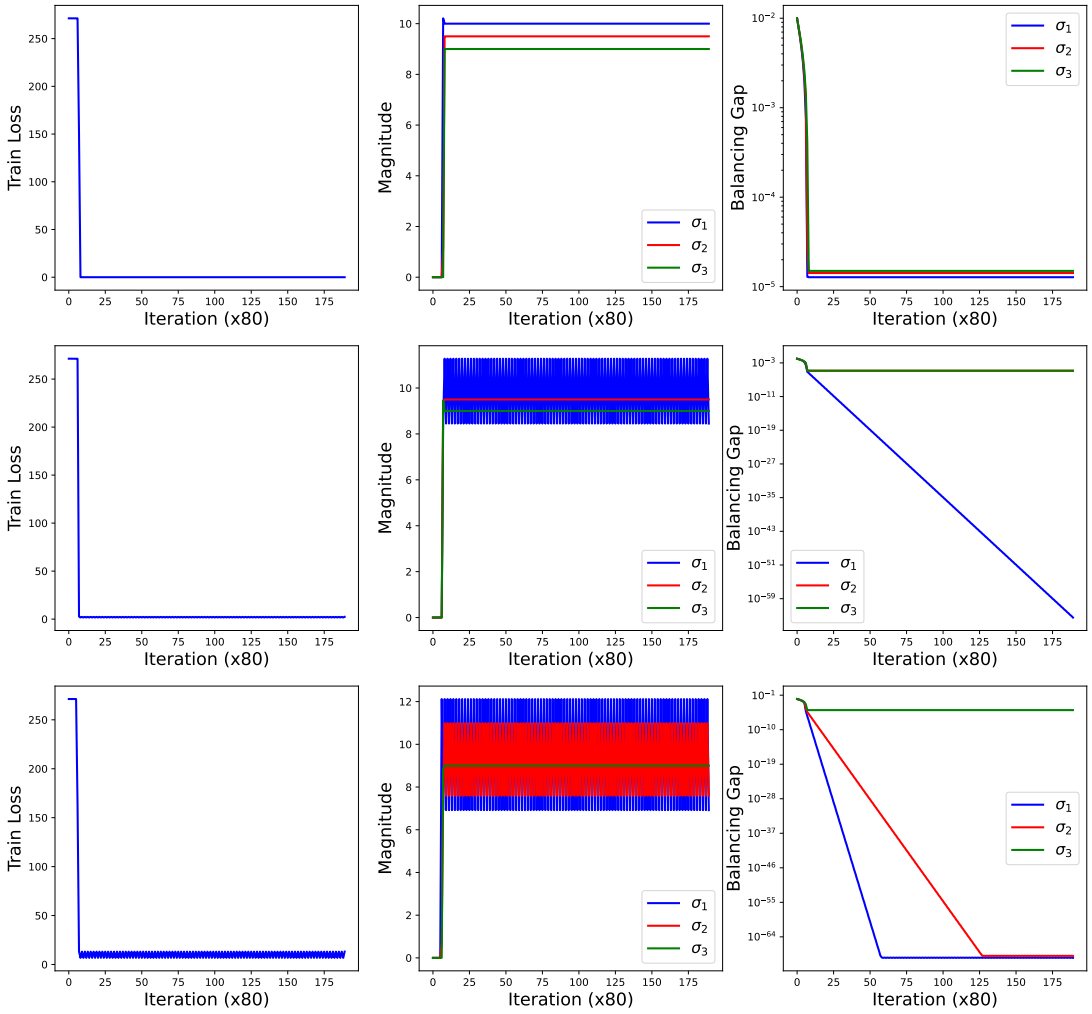


Figure 16: Plots of the training loss, singular value magnitude, and the balancing gap over iterations for different learning rates: $\eta = 0.030, 0.032, 0.034$ (top to bottom). When the learning rate is stable ($\eta < 0.031$ since the top singular value is $\sigma_{\star,1} = 10$), the balancing gap plateaus, whereas the balancing gap goes strictly to zero when the oscillations occur.

the SVS set. To this end, we consider the following:

$$\mathbf{W}_L(0) = \mathbf{0}, \quad \mathbf{W}_\ell(0) = \alpha \mathbf{I}_d, \quad \forall \ell \in [L-1], \qquad \text{(Original)}$$
$$\mathbf{W}_L(0) = \mathbf{0}, \quad \mathbf{W}_\ell(0) = \alpha \mathbf{P}_\ell, \quad \forall \ell \in [L-1], \qquad \text{(Orthogonal)}$$
$$\mathbf{W}_L(0) = \mathbf{0}, \quad \mathbf{W}_\ell(0) = \alpha \mathbf{H}_\ell, \quad \forall \ell \in [L-1], \qquad \text{(Random)}$$

where $\mathbf{P}_\ell$ is an orthogonal matrix and $\mathbf{H}_\ell$ is a random matrix with Gaussian entries. For all of these initialization schemes, we consider the same setup as in the balancing case, with an initialization scale of $\alpha = 0.01$. To observe if singular vector stationarity holds, we consider the subspace distance as follows:

$$\text{Subspace Distance} = \|\mathbf{U}_{\ell-1,r}^\top \mathbf{V}_{\ell,r} - \mathbf{I}_r\|_{\mathsf{F}}, \tag{11}$$

where $\mathbf{U}_{\ell,r}$ and $\mathbf{V}_{\ell,r}$ are the top-r left and right singular vectors of layer $\mathbf{W}_\ell$, respectively. Since Proposition 1 implies that the intermediate singular vectors cancel, the initialization converges to the SVS set if the subspace distance goes to zero. In Figure 18, we plot the dynamics for all of the initializations. Generally, we observe that the subspace distance for all cases go to zero, validating the use of the SVS set for analysis purposes.

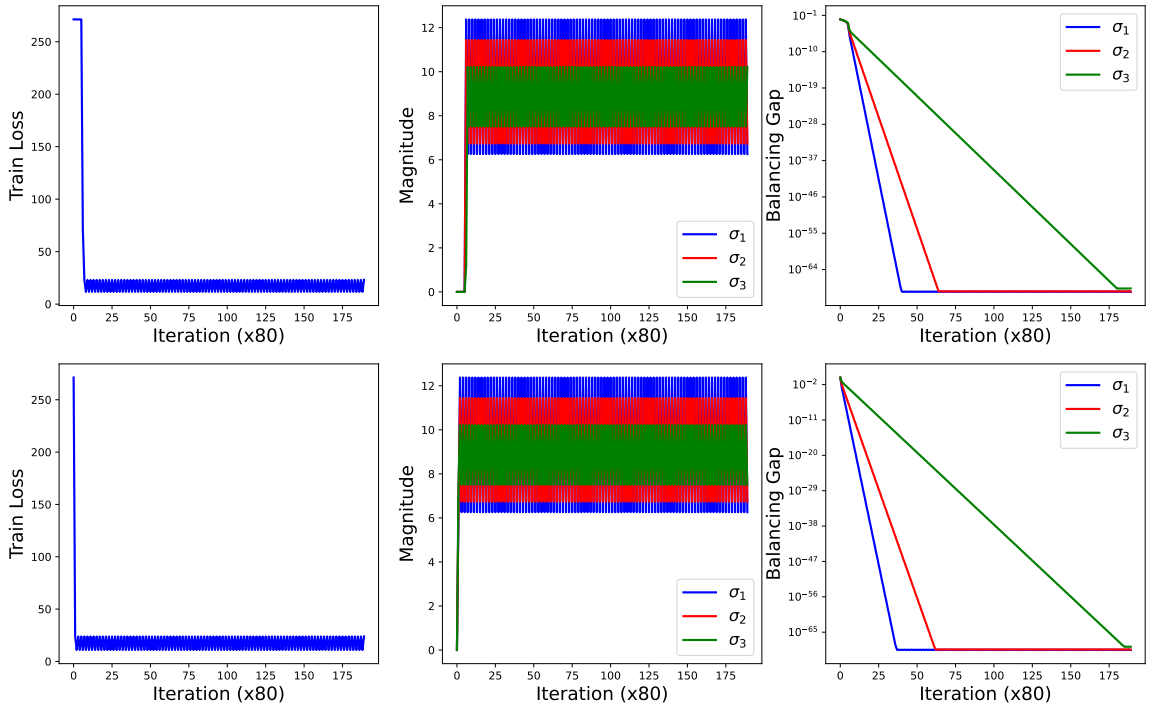


Figure 17: Top: EOS dynamics of a 3-layer DLN with initialization scale $\alpha = 0.01$, where balancing theoretically holds. Bottom: EOS dynamics of the DLN with initialization scale $\alpha = 0.5$. While the balancing does not hold for $\alpha = 0.5$, the oscillations in the singular values are still prevalent, with the same amplitude.

Additional Results. In this section, we provide more experimental results to corroborate our theory. Recall that in Lemma 1, we proved that the learning rate needed to enter the EOS is a function of the depth, and that deeper networks can enter EOS using a smaller learning rate. To verify this claim, we provide an additional experiment where the target matrix is $\mathbf{M}_\star \in \mathbb{R}^{5\times 5}$ with the top singular value set to $\sigma_{\star,1} = 0.5$. We use an initialization scale of $\alpha = 0.01$. In Figure 19, we can clearly see that shallower networks need a larger learning rate, and vice versa to enter EOS. Here, black refers to stable learning and white refers to regions in which oscillations occur (EOS regime).

B.5. Periodic and Free Oscillations

In this section, we present additional experiments on oscillation and catapults in both deep linear and nonlinear networks to supplement the results in the main paper. First, we consider a 3-layer MLP without bias terms for the weights, with each hidden layer consisting of 1000 units. The network is trained using MSE loss with a learning rate of $\eta = 4$, along with random weights scaled by $\alpha = 0.01$ and full-batch gradient descent on a 5K subset of the MNIST dataset, following [1]. The motivation for omitting bias terms comes from the findings of [26], where they provably show that a ReLU network without bias terms behaves similarly to a linear network. With this in mind, we aimed to investigate how oscillations manifest in comparison to deep linear networks (DLNs). In Figure 20, we plot the training loss, top-5 singular values, and sharpness throughout training. Interestingly, despite the non-convexity of the loss landscape, the oscillations appear to be almost periodic across all three plots. It would be of great interest to theoretically study the behavior of EOS for this network architecture and determine whether our analyses extend to this case as well.

Next, we consider the DLN setting to corroborate our result from Theorem 2. We consider modeling rank-3 target matrix with singular values $\sigma_{\star,i} = \{10, 9, 8\}$ with a 3-layer DLN with initialization scale $\alpha = 0.1$. By computing the sharpness under these settings, notice that $2/\lambda_1 = L\sigma_{\star,1}^{2-\frac{2}{L}} \approx 0.01547$

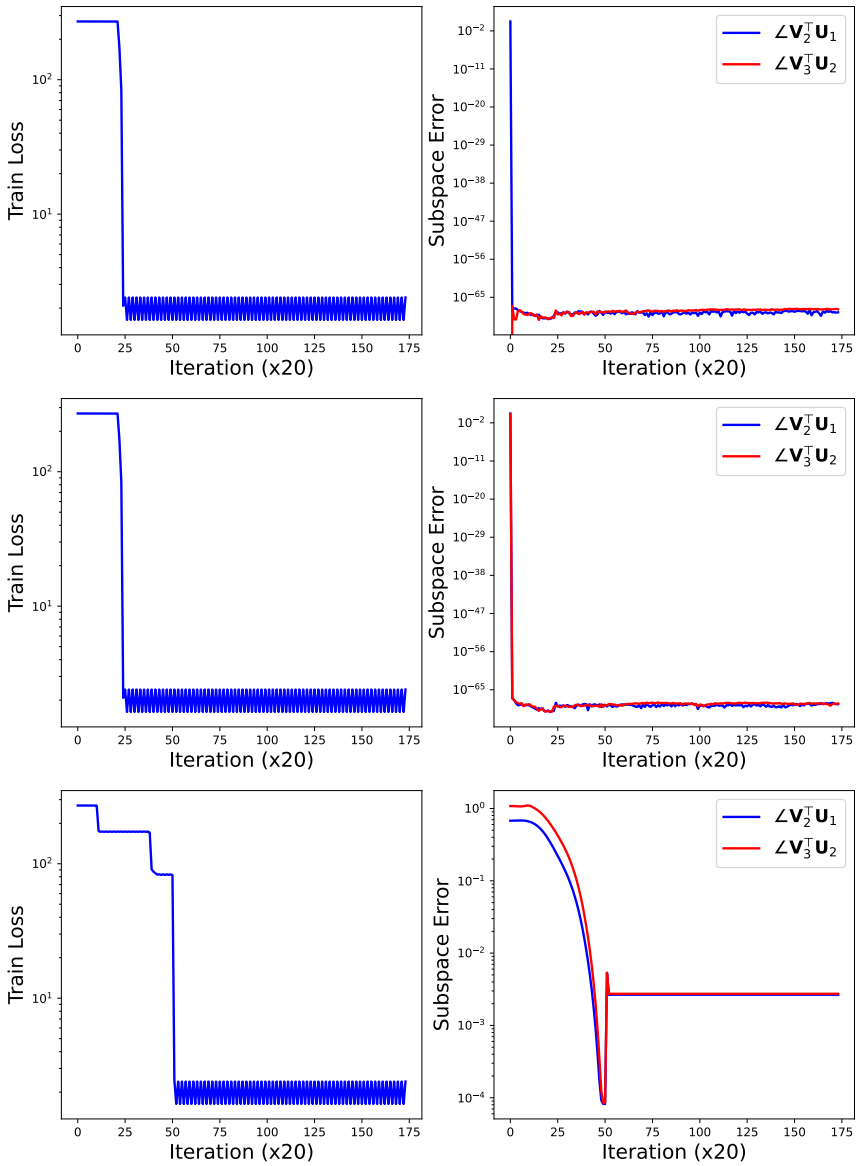


Figure 18: EOS dynamics of a 3-layer DLN for different initializations where it all converges to the SVS set. The subspace distance is defined in Equation (11). Top: Dynamics with the original identity initialization. Middle: Dynamics with orthogonal initialization. Bottom: Dynamics with random initialization.

and $2/\lambda_2 \approx 0.01657$. In Figure 21, we use learning rates near these values, and plot the oscillations in the singular values. Here, we can see that the oscillations follow exactly our theory.

Lastly, we provide additional experiments demonstrating stronger oscillation in feature directions as measured by the singular values. To this end, we consider a 4-layer MLP with ReLU activations with hidden layer size in each unit of 200 for classification on a subsampled 20K set on MNIST and CIFAR-10. In Figure 22, we show that the oscillations in the training loss are artifacts of jumps only in the top singular values, which is also what we observe in the DLN setting.

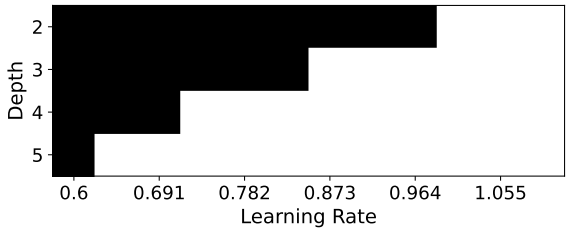


Figure 19: Demonstrating that deeper networks requires a smaller learning rate to enter the EOS regime for DLNs, as implied by Lemma 2, for a target matrix with top singular value $\sigma_{\star,1} = 0.5$ and initialization $\alpha = 0.01$. Black refers to stable learning and white refers to regions in which oscillations in the loss and singular values occur. The EOS limit exactly matches $\eta = \frac{2}{L\sigma_{\star,i}^{2-\frac{2}{L}}}$.

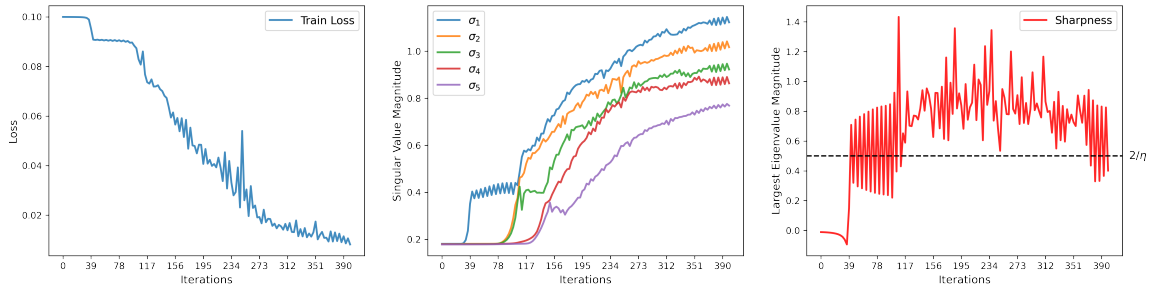


Figure 20: Plot of the training loss, singular values, and sharpness for an MLP network with no bias. Similar to the DLN case, there are oscillations in each of the plots throughout iterations.

B.6. Investigation of Oscillations in Low-Rank Adaptors

Previously, we investigated the differences in oscillations (i.e., oscillations versus catapults) in deep linear and nonlinear networks, and how changes in the landscape present one behavior or the other. Low-rank adaptation (LoRA) [58] has arguably become one of the most popular methods for fine-tuning deep neural networks. By viewing the adaptations as individual low-rank matrix factorization problems, then this formulation closely aligns with our theoretical setup with a depth of 2. Here we pose the question (i) does oscillations may appear in such a setup and (ii) what these oscillations may imply in terms of generalization.

Briefly, the main idea behind LoRA is that rather than training from scratch, we can update two low-rank factor matrices to "append" onto an existing weight matrix. That is, give a pre-trained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d_1 \times d_2}$, LoRA involves updating two low-rank factors commonly referred to as "adaptors":

$$\underbrace{\mathbf{W}_\star}_{\text{new weight}} = \underbrace{\mathbf{W}_0}_{\text{pre-trained weight}} + \underbrace{\mathbf{A}\mathbf{B}^\top}_{\text{adaptors}}.$$

For a sufficiently small rank r, upon training only $\mathbf{A} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{B} \in \mathbb{R}^{d_2 \times r}$, $\mathbf{W}_\star \in \mathbb{R}^{d_1 \times d_2}$ is used for inference.

For the experimental settings, we follow the setup used by [57] and consider a pre-trained BERT [59] base model and apply adaptation on all attention and feedforward weights in the transformer, resulting in 72 adapted layers in total. For initialization, we use random weights and scale them using an initialization scale of $\alpha = 10^{-3}$ for the adaptors and randomly sample 512 examples from the STS-B [59] dataset for fine-tuning. We choose a batch size of 64 with a maximum sequence length of 128 tokens. First, we experiment how large the rank of the adaptors must be to drive the entire network to EOS. Using a learning rate of $\eta = 10^{-4}$, Figure 23 shows oscillatory behavior across all ranks. However, this behavior may also be an artifact of the stochasticity induced by updating with only a batch of samples.

In Figure 24, we present the training loss and Pearson correlation for different learning rates using rank $r = 8$. When $\eta = 10^{-4}$, the training loss catapults from a magnitude of 7 to 4 whereas other learning rates cannot decrease the loss with such a magnitude. Consequently, $\eta = 10^{-4}$ achieves the best Pearson correlation coefficient. This suggests that learning rate plays an important role when the optimization is restricted to a small subspace as used in LoRA. We leave this for future work a careful study of this observation, aiming to accurately select the learning rate to maximize the efficiency of LoRA.

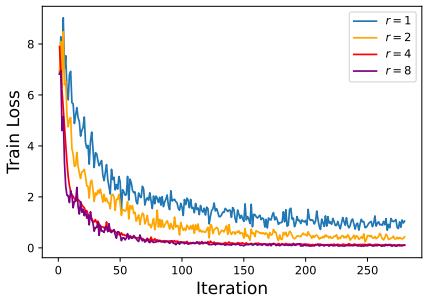
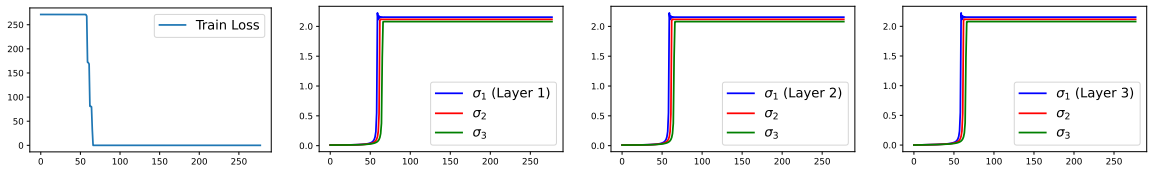
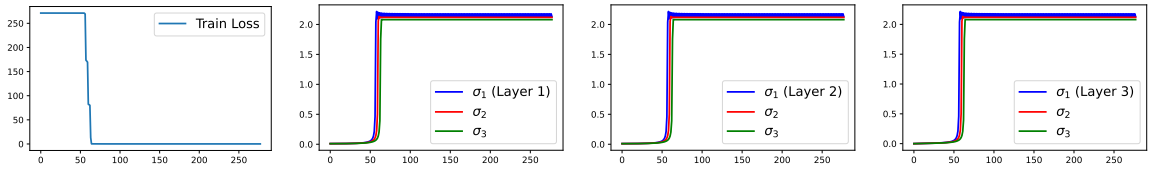


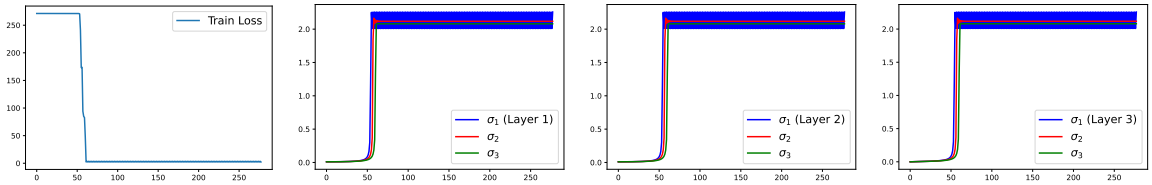
Figure 23: Catapults in the training loss for different ranks for LoRA.



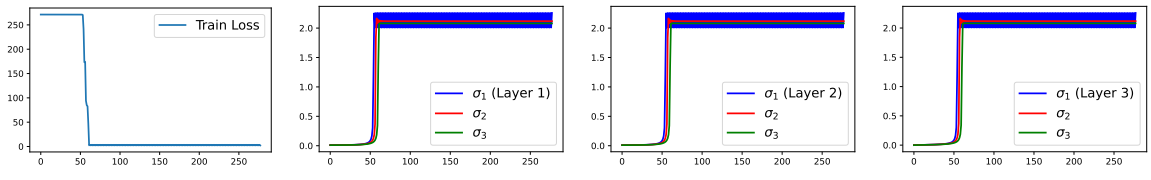
Train Loss ($\eta = 0.0300$) Layer 1 σ_i ($\eta = 0.0300$) Layer 2 σ_i ($\eta = 0.0300$) Layer 3 σ_i ($\eta = 0.0300$)



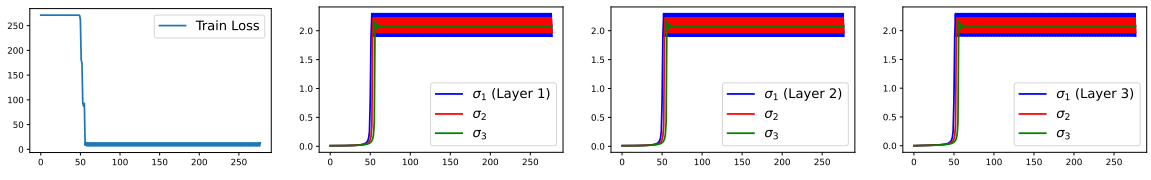
Train Loss ($\eta = 0.031$) Layer 1 σ_i ($\eta = 0.0310$) Layer 2 σ_i ($\eta = 0.0310$) Layer 3 σ_i ($\eta = 0.0310$)



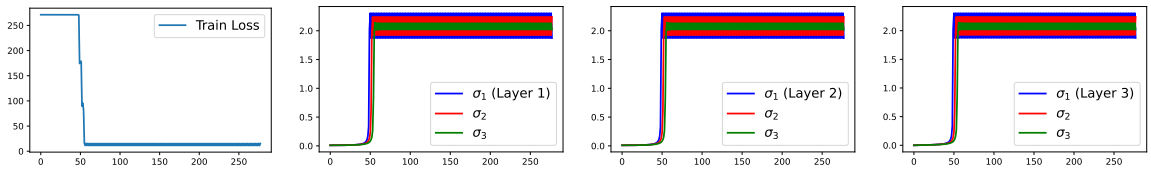
Train Loss ($\eta = 0.0325$) Layer 1 σ_i ($\eta = 0.0325$) Layer 2 σ_i ($\eta = 0.0325$) Layer 3 σ_i ($\eta = 0.0325$)



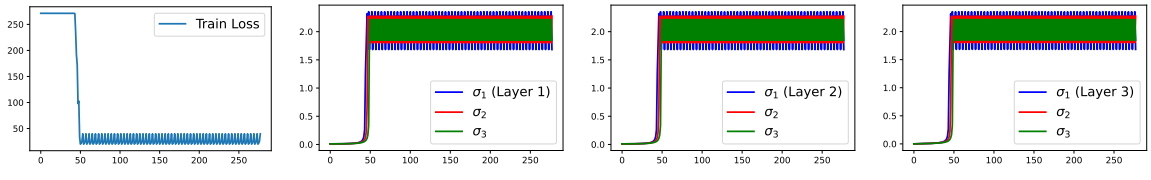
Train Loss ($\eta = 0.0335$) Layer 1 σ_i ($\eta = 0.0335$) Layer 2 σ_i ($\eta = 0.0335$) Layer 3 σ_i ($\eta = 0.0335$)



Train Loss ($\eta = 0.0350$) Layer 1 σ_i ($\eta = 0.0350$) Layer 2 σ_i ($\eta = 0.0350$) Layer 3 σ_i ($\eta = 0.0350$)

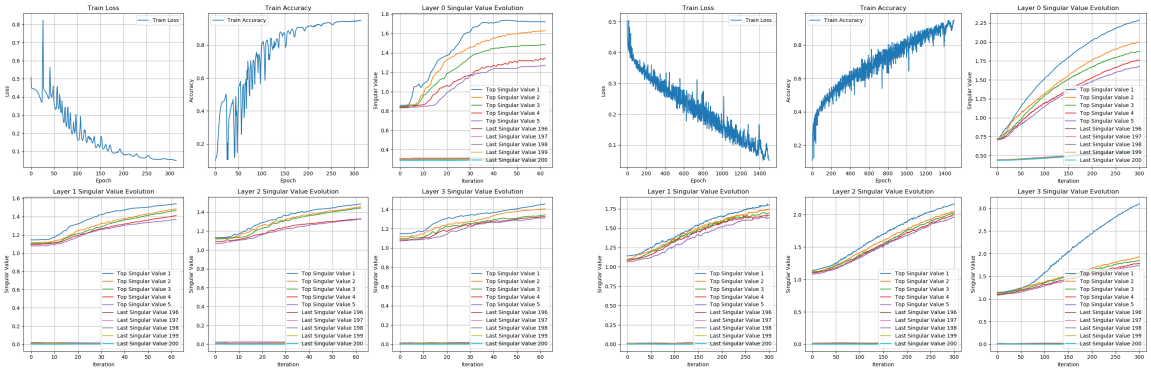


Train Loss ($\eta = 0.0360$) Layer 1 σ_i ($\eta = 0.0360$) Layer 2 σ_i ($\eta = 0.0360$) Layer 3 σ_i ($\eta = 0.0360$)



Train Loss ($\eta = 0.0400$) Layer 1 σ_i ($\eta = 0.0400$) Layer 2 σ_i ($\eta = 0.0040$) Layer 3 σ_i ($\eta = 0.0400$)

Figure 21: Depiction of the training loss and the singular values of each weight matrix for fitting a rank-3 matrix with singular values $10, 9.5, 9$. The weights enter the EOS regime based on the learning rate $\eta > 2/K$, where $K = L\sigma_{\star,i}^{2-2/L}$ and $L = 3$. For a sufficiently large learning rate (e.g., $\eta = 0.04$), the singular values start to enter a period-4 orbit.



MNIST Dataset with 4-Layer MLP CIFAR-10 Dataset with 4-Layer MLP

Figure 22: Prevalence of oscillatory behaviors in top subspaces in 4-layer networks with ReLU activations on two different datasets.

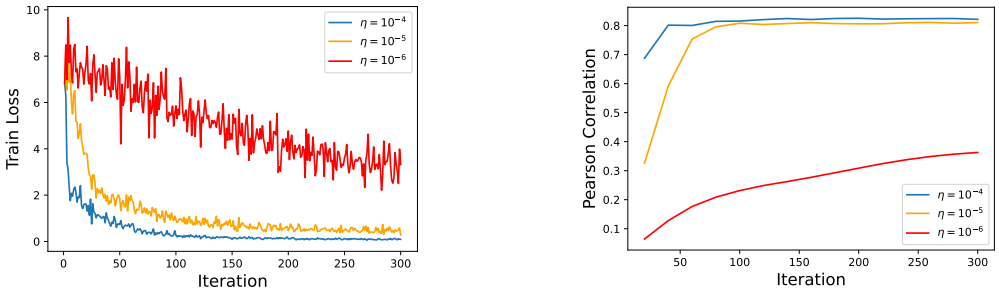


Figure 24: Illustration of different behaviors in the training loss for various learning rates with a fixed rank of $r = 8$ for fine-tuning BERT using LoRA. These plots indicate that larger learning rates lead to higher Pearson correlations. When $\eta = 10^{-4}$, the training loss catapults from a magnitude of 7 to 4 whereas other learning rates do not decrease the loss with such large magnitude.

C. Deferred Proofs

C.1. Deferred Proofs for Oscillations

Proposition 1 (Singular Vector Stationary Set)**.** *Consider the deep matrix factorization loss in Equation (1). Let $\mathbf{M}_\star = \mathbf{U}_\star \mathbf{\Sigma}_\star \mathbf{V}_\star^\top$ and $\mathbf{W}_\ell(t) = \mathbf{U}_\ell(t)\mathbf{\Sigma}_\ell(t)\mathbf{V}_\ell^\top(t)$ denote the compact SVD for the target matrix and the ℓ-th layer weight matrix at time t, respectively. For any time $t \geq 0$, if $\dot{\mathbf{U}}_\ell(t) = \dot{\mathbf{V}}_\ell(t) = 0$ for all $\ell \in [L]$, then the singular vector stationary points for each weight matrix are given by*

$$\mathrm{SVS}(f(\mathbf{\Theta})) = \begin{cases} (\mathbf{U}_L, \mathbf{V}_L) & = (\mathbf{U}_\star, \mathbf{Q}_L), \\ (\mathbf{U}_\ell, \mathbf{V}_\ell) & = (\mathbf{Q}_{\ell+1}, \mathbf{Q}_\ell), \quad \forall \ell \in [2, L-1], \\ (\mathbf{U}_1, \mathbf{V}_1) & = (\mathbf{Q}_2, \mathbf{V}_\star), \end{cases}$$

where $\{\mathbf{Q}_\ell\}_{\ell=2}^L$ can be any orthogonal matrices.

Proof. Let us consider the dynamics of $\mathbf{W}_\ell(t)$ in terms of its SVD with respect to time:

$$\dot{\mathbf{W}}_\ell(t) = \dot{\mathbf{U}}_\ell(t)\mathbf{\Sigma}_\ell(t)\mathbf{V}_\ell^\top(t) + \mathbf{U}_\ell(t)\dot{\mathbf{\Sigma}}_\ell(t)\mathbf{V}_\ell^\top(t) + \mathbf{U}_\ell(t)\mathbf{\Sigma}_\ell(t)\dot{\mathbf{V}}_\ell^\top(t). \tag{12}$$

By left multiplying by $\mathbf{U}_\ell^\top(t)$ and right multiplying by $\mathbf{V}_\ell(t)$, we have

$$\mathbf{U}_\ell^\top(t)\dot{\mathbf{W}}_\ell(t)\mathbf{V}_\ell(t) = \mathbf{U}_\ell^\top(t)\dot{\mathbf{U}}_\ell(t)\mathbf{\Sigma}_\ell(t) + \dot{\mathbf{\Sigma}}_\ell(t) + \mathbf{\Sigma}_\ell(t)\dot{\mathbf{V}}_\ell^\top(t)\mathbf{V}_\ell(t), \tag{13}$$

where we used the fact that $\mathbf{U}_\ell(t)$ and $\mathbf{V}_\ell(t)$ have orthonormal columns. Now, note that we also have

$$\mathbf{U}_\ell^\top(t)\mathbf{U}_\ell(t) = \mathbf{I}_r \implies \dot{\mathbf{U}}_\ell^\top(t)\mathbf{U}_\ell(t) + \mathbf{U}_\ell^\top(t)\dot{\mathbf{U}}_\ell(t) = \mathbf{0},$$

which also holds for $\mathbf{V}_\ell(t)$. This implies that $\dot{\mathbf{U}}_\ell^\top(t)\mathbf{U}_\ell(t)$ is a skew-symmetric matrix, and hence have zero diagonals. Since $\mathbf{\Sigma}_\ell(t)$ is diagonal, $\mathbf{U}_\ell^\top(t)\dot{\mathbf{U}}_\ell(t)\mathbf{\Sigma}_\ell(t)$ and $\mathbf{\Sigma}_\ell(t)\dot{\mathbf{V}}_\ell^\top(t)\mathbf{V}_\ell(t)$ have zero diagonals as well. On the other hand, since $\dot{\mathbf{\Sigma}}_\ell(t)$ is a diagonal matrix, we can write

$$\hat{\mathbf{I}}_r \odot \left(\mathbf{U}_\ell^\top(t)\dot{\mathbf{W}}_\ell(t)\mathbf{V}_\ell(t)\right) = \mathbf{U}_\ell^\top(t)\dot{\mathbf{U}}_\ell(t)\mathbf{\Sigma}_\ell(t) + \mathbf{\Sigma}_\ell(t)\dot{\mathbf{V}}_\ell^\top(t)\mathbf{V}_\ell(t), \tag{14}$$

where $\odot$ stands for the Hadamard product and $\hat{\mathbf{I}}_r$ is a square matrix holding zeros on its diagonal and ones elsewhere. Taking transpose of Equation (14), while recalling that $\mathbf{U}_\ell^\top(t)\dot{\mathbf{U}}_\ell(t)$ and $\mathbf{V}_\ell^\top(t)\dot{\mathbf{V}}_\ell(t)$ are skew-symmetric, we have

$$\hat{\mathbf{I}}_r \odot \left(\mathbf{V}_\ell^\top(t)\dot{\mathbf{W}}_\ell^\top(t)\mathbf{U}_\ell(t)\right) = -\mathbf{\Sigma}_\ell(t)\mathbf{U}_\ell^\top(t)\dot{\mathbf{U}}_\ell(t) - \dot{\mathbf{V}}_\ell^\top(t)\mathbf{V}_\ell(t)\mathbf{\Sigma}_\ell(t). \tag{15}$$

Then, by right multiplying Equation (14) by $\mathbf{\Sigma}_\ell(t)$, left-multiply Equation (15) by $\mathbf{\Sigma}_\ell(t)$, and by adding the two terms, we get

$$\begin{aligned} \hat{\mathbf{I}}_r \odot \Big(\mathbf{U}_\ell^\top(t)\dot{\mathbf{W}}_\ell(t)\mathbf{V}_\ell(t)\mathbf{\Sigma}_\ell(t) + \mathbf{\Sigma}_\ell(t)\mathbf{V}_\ell^\top(t)\dot{\mathbf{W}}_\ell^\top(t)\mathbf{U}_\ell(t)\Big) \\ = \mathbf{U}_\ell^\top(t)\dot{\mathbf{U}}_\ell(t)\mathbf{\Sigma}_\ell^2(t) - \mathbf{\Sigma}_\ell^2(t)\dot{\mathbf{V}}_\ell^\top(t)\mathbf{V}_\ell(t). \end{aligned}$$

Since we assume that the singular values of $\mathbf{M}_\star$ are distinct, the top-r diagonal elements of $\mathbf{\Sigma}_\ell^2(t)$ are also distinct (i.e., $\Sigma_r^2(t) \neq \Sigma_{r'}^2(t)$ for $r \neq r'$). This implies that

$$\mathbf{U}_\ell^\top(t)\dot{\mathbf{U}}_\ell(t) = \mathbf{H}(t) \odot \left[\mathbf{U}_\ell^\top(t)\dot{\mathbf{W}}_\ell(t)\mathbf{V}_\ell(t)\mathbf{\Sigma}_\ell(t) + \mathbf{\Sigma}_\ell(t)\mathbf{V}_\ell^\top(t)\dot{\mathbf{W}}_\ell^\top(t)\mathbf{U}_\ell(t)\right],$$

where the matrix $\mathbf{H}(t) \in \mathbb{R}^{d\times d}$ is defined by:

$$H_{r,r'}(t) := \begin{cases} \left(\Sigma_{r'}^2(t) - \Sigma_r^2(t)\right)^{-1}, & r \neq r', \\ 0, & r = r'. \end{cases} \tag{16}$$

Then, multiplying from the left by $\mathbf{U}_\ell(t)$ yields

$$\mathbf{P}_{\mathbf{U}_\ell(t)}\dot{\mathbf{U}}_\ell(t) = \mathbf{U}_\ell(t)\left(\mathbf{H}(t)\odot\left[\mathbf{U}_\ell^\top(t)\dot{\mathbf{W}}_\ell(t)\mathbf{V}_\ell(t)\mathbf{\Sigma}_\ell(t) + \mathbf{\Sigma}_\ell(t)\mathbf{V}_\ell^\top(t)\dot{\mathbf{W}}_\ell^\top(t)\mathbf{U}_\ell(t)\right]\right), \tag{17}$$

with $\mathbf{P}_{\mathbf{U}_\ell(t)} := \mathbf{U}_\ell(t)\mathbf{U}_\ell^\top(t)$ being the projection onto the subspace spanned by the (orthonormal) columns of $\mathbf{U}_\ell(t)$. Denote by $\mathbf{P}_{\mathbf{U}_{\ell\perp}(t)}$ the projection onto the orthogonal complement (i.e., $\mathbf{P}_{\mathbf{U}_{\ell\perp}(t)} := \mathbf{I}_r - \mathbf{U}_\ell(t)\mathbf{U}_\ell^\top(t)$). Apply $\mathbf{P}_{\mathbf{U}_{\ell\perp}(t)}$ to both sides of Equation (12):

$$\mathbf{P}_{\mathbf{U}_{\ell\perp}(t)}\dot{\mathbf{U}}_\ell(t) = \mathbf{P}_{\mathbf{U}_{\ell\perp}(t)}\dot{\mathbf{U}}_\ell(t)\mathbf{\Sigma}_\ell(t)\mathbf{V}_\ell^\top(t) + \mathbf{P}_{\mathbf{U}_{\ell\perp}(t)}\mathbf{U}_\ell(t)\dot{\mathbf{\Sigma}}_\ell(t)\mathbf{V}_\ell^\top(t) \tag{18}$$

$$+ \mathbf{P}_{\mathbf{U}_{\ell\perp}(t)}\mathbf{U}_\ell(t)\mathbf{\Sigma}_\ell(t)\dot{\mathbf{V}}_\ell^\top(t). \tag{19}$$

Note that $\mathbf{P}_{\mathbf{U}_{\ell\perp}(t)}\mathbf{U}_\ell(t) = 0$, and multiply from the right by $\mathbf{V}_\ell(t)\mathbf{\Sigma}_\ell^{-1}(t)$ (the latter is well-defined since we have the compact SVD and the top-r elements are non-zero):

$$\mathbf{P}_{\mathbf{U}_{\ell\perp}(t)}\dot{\mathbf{U}}_\ell(t) = \mathbf{P}_{\mathbf{U}_{\ell\perp}(t)}\dot{\mathbf{W}}_\ell(t)\mathbf{V}_\ell(t)\mathbf{\Sigma}_\ell^{-1}(t) = (\mathbf{I}_r - \mathbf{U}_\ell(t)\mathbf{U}^\top(t))\dot{\mathbf{W}}(t)\mathbf{V}_\ell(t)\mathbf{\Sigma}_\ell^{-1}(t). \tag{20}$$

Then by adding the two equations above, we obtain an expression for $\dot{\mathbf{U}}(t)$:

$$\begin{aligned}\dot{\mathbf{U}}_\ell(t) &= \mathbf{P}_{\mathbf{U}_\ell(t)}\dot{\mathbf{U}}_\ell(t) + \mathbf{P}_{\mathbf{U}_{\ell\perp}(t)}\dot{\mathbf{U}}_\ell(t)\\ &= \mathbf{U}_\ell(t)\left(\mathbf{H}(t)\odot\left[\mathbf{U}_\ell^\top(t)\dot{\mathbf{W}}_\ell(t)\mathbf{V}_\ell(t)\mathbf{\Sigma}_\ell(t) + \mathbf{\Sigma}_\ell(t)\mathbf{V}_\ell^\top(t)\dot{\mathbf{W}}_\ell^\top(t)\mathbf{U}_\ell(t)\right]\right)\\ &\quad + (\mathbf{I}_r - \mathbf{U}_\ell(t)\mathbf{U}_\ell^\top(t))\dot{\mathbf{W}}(t)\mathbf{V}_\ell(t)\mathbf{\Sigma}_\ell^{-1}(t).\end{aligned} \tag{21}$$

We can similarly derive the dynamics for $\dot{\mathbf{V}}_\ell(t)$ and $\dot{\mathbf{\Sigma}}_\ell(t)$:

$$\dot{\mathbf{V}}_\ell(t) = \mathbf{V}_\ell(t)\left(\mathbf{H}(t)\odot\left[\mathbf{\Sigma}_\ell(t)\mathbf{U}_\ell^\top(t)\dot{\mathbf{W}}_\ell(t)\mathbf{V}_\ell(t) + \mathbf{V}_\ell^\top(t)\dot{\mathbf{W}}_\ell{}^\top(t)\mathbf{U}_\ell(t)\mathbf{\Sigma}_\ell(t)\right]\right) \tag{22}$$

$$+ \left(\mathbf{I}_r - \mathbf{V}_\ell(t)\mathbf{V}_\ell^\top(t)\right)\dot{\mathbf{W}}_\ell{}^\top(t)\mathbf{U}_\ell(t)\mathbf{\Sigma}_\ell^{-1}(t), \tag{23}$$

$$\dot{\mathbf{\Sigma}}_\ell(t) = \mathbf{I}_r\odot\left[\mathbf{U}_\ell^\top(t)\dot{\mathbf{W}}_\ell(t)\mathbf{V}_\ell(t)\right].$$

Now, we will left multiply $\dot{\mathbf{U}}_\ell(t)$ and $\dot{\mathbf{V}}_\ell(t)$ with $\mathbf{U}_\ell^\top(t)$ and $\mathbf{V}_\ell^\top(t)$, respectively, to obtain

$$\begin{aligned}\mathbf{U}_\ell^\top(t)\dot{\mathbf{U}}_\ell(t) &= -\mathbf{H}(t)\odot\left[\mathbf{U}_\ell^\top(t)\nabla_{\mathbf{W}_\ell}f(\mathbf{\Theta})\mathbf{V}_\ell(t)\mathbf{\Sigma}_\ell(t) + \mathbf{\Sigma}_\ell(t)\mathbf{V}_\ell^\top(t)\nabla_{\mathbf{W}_\ell}f(\mathbf{\Theta})\mathbf{U}_\ell(t)\right],\\ \mathbf{V}_\ell^\top(t)\dot{\mathbf{V}}_\ell(t) &= -\mathbf{H}(t)\odot\left[\mathbf{\Sigma}_\ell(t)\mathbf{U}_\ell^\top(t)\nabla_{\mathbf{W}_\ell}f(\mathbf{\Theta})\mathbf{V}_\ell(t) + \mathbf{V}_\ell^\top(t)\nabla_{\mathbf{W}_\ell}f(\mathbf{\Theta})\mathbf{U}_\ell(t)\mathbf{\Sigma}_\ell(t)\right],\end{aligned}$$

where we replaced $\dot{\mathbf{W}}_\ell(t) \coloneqq -\nabla_{\mathbf{W}_\ell}f(\mathbf{\Theta})$, as $\dot{\mathbf{W}}_\ell(t)$ is the gradient of $f(\mathbf{\Theta})$ with respect to $\mathbf{W}_\ell$ by definition. By rearranging and multiplying by $\mathbf{\Sigma}_\ell(t)$, we have

$$\mathbf{U}_\ell^\top(t)\dot{\mathbf{U}}_\ell(t)\mathbf{\Sigma}_\ell(t) - \mathbf{\Sigma}_\ell(t)\mathbf{V}^T(t)\dot{\mathbf{V}}_\ell(t) = -\hat{\mathbf{I}}_r\odot[\mathbf{U}_\ell^\top(t)\nabla_{\mathbf{W}_\ell}f(\mathbf{\Theta})\mathbf{V}_\ell(t)]. \tag{24}$$

Hence, when $\dot{\mathbf{U}}_\ell(t) = 0$ and $\dot{\mathbf{V}}_\ell(t) = 0$, it must be that the left-hand side is zero and so $\mathbf{U}_\ell^\top(t)\nabla_{\mathbf{W}_\ell}f(\mathbf{\Theta})\mathbf{V}_\ell(t)$ is a diagonal matrix.

Now, notice that for the given loss function $f(\mathbf{\Theta})$, we have

$$-\dot{\mathbf{W}}_\ell(t) = \nabla_{\mathbf{W}_\ell}f(\mathbf{\Theta}(t)) = \mathbf{W}_{L:\ell+1}^\top(t)\cdot(\mathbf{W}_{L:1}(t) - \mathbf{M}_\star)\cdot\mathbf{W}_{\ell-1:1}^\top(t).$$

Then, from Equation (24), when the singular vectors are stationary, we have

$$\mathbf{U}_\ell^\top(t)\mathbf{W}_{L:\ell+1}^\top(t)\cdot(\mathbf{W}_{L:1}(t) - \mathbf{M}_\star)\cdot\mathbf{W}_{\ell-1:1}^\top(t)\mathbf{V}_\ell(t)$$

must be a diagonal matrix for all $\ell\in[L]$. The only solution to the above should be (since the intermediate singular vectors need to cancel to satisfy the diagonal condition), is the set

$$\mathrm{SVS}(f(\mathbf{\Theta})) = \begin{cases}(\mathbf{U}_L,\mathbf{V}_L) &= (\mathbf{U}_\star,\mathbf{Q}_L),\\ (\mathbf{U}_\ell,\mathbf{V}_\ell) &= (\mathbf{Q}_{\ell+1},\mathbf{Q}_\ell), \quad \forall\ell\in[2,L-1],\\ (\mathbf{U}_1,\mathbf{V}_1) &= (\mathbf{Q}_2,\mathbf{V}_\star),\end{cases}$$

where $\{\mathbf{Q}_\ell\}_{\ell=2}^{L}$ are any set of orthogonal matrices. Then, notice that when the singular vectors are stationary, the dynamics become isolated on the singular values:

$$\dot{\mathbf{\Sigma}}_\ell(t) = \mathbf{I}_r \odot \left[\mathbf{U}_\ell^\top(t)\dot{\mathbf{W}}_\ell(t)\mathbf{V}_\ell(t)\right],$$

since $\left[\mathbf{U}_\ell^\top(t)\dot{\mathbf{W}}_\ell(t)\mathbf{V}_\ell(t)\right]$ is diagonal. This completes the proof.

□

Theorem 1 (Stable Subspace Oscillations)**.** *Let* $\alpha' \coloneqq \left(\ln\left(\frac{2\sqrt{2}}{\eta\lambda_1}\right)\cdot\frac{\sigma_{\star,1}^{4/L}}{L^2\cdot 2^{\frac{2L-3}{L}}}\right)^{1/4}$*. Consider running GD on the loss in (1) with initialization scale* $0<\alpha<\alpha'$*. If* $\eta=\frac{2}{K}$ *with* $\lambda_i \le K < \lambda_{i+1}$*, then under strict balancing, 2-period orbit oscillation occurs in the direction of* Δ_i*, where* λ_i *and* Δ_i *denote the* i*-th largest eigenvalue and eigenvector of the Hessian at the balanced minimum, respectively.*

Proof. Define f_{Δ_i} as the 1-D function at the cross section of the loss landscape and the line following the direction of Δ_i passing the (balanced) minima. To prove the result, we will invoke Lemma 12, which states that two-period orbit oscillation occurs in the direction of Δ_i if the minima of f_{Δ_i} satisfies $f_{\Delta_i}^{(3)} > 0$ and $3[f_{\Delta_i}^{(3)}]^2 - f_{\Delta_i}^{(2)} f_{\Delta_i}^{(4)} > 0$, for $\eta > \frac{2}{\lambda_i}$. Recall that the initialization condition is an artifact of Lemma 2, which allows us to consider the balanced minimum.

First, we will derive the eigenvectors of the Hessian of the training loss at convergence (i.e., $\mathbf{M}_\star = \mathbf{W}_{L:1}$). To obtain the eigenvectors of the Hessian of parameters $(\mathbf{W}_L, \ldots, \mathbf{W}_2, \mathbf{W}_1)$, consider a small perturbation of the parameters:

$$\mathbf{\Theta} \coloneqq (\Delta\mathbf{W}_\ell + \mathbf{W}_\ell)_{\ell=1}^{L} = (\mathbf{W}_L + \Delta\mathbf{W}_L, \ldots, \mathbf{W}_2 + \Delta\mathbf{W}_2, \mathbf{W}_1 + \Delta\mathbf{W}_1).$$

Given that $\mathbf{W}_{L:1} = \mathbf{M}_\star$, consider and evaluate the loss function at this minima:

$$\mathcal{L}(\mathbf{\Theta}) = \frac{1}{2}\Bigg\| \sum_\ell \mathbf{W}_{L:\ell+1}\Delta\mathbf{W}_\ell\mathbf{W}_{\ell-1:1} \tag{25}$$

$$+ \sum_{\ell<m} \mathbf{W}_{L:\ell+1}\Delta\mathbf{W}_\ell\mathbf{W}_{\ell-1:m+1}\Delta\mathbf{W}_m\mathbf{W}_{m-1:1} + \ldots + \Delta\mathbf{W}_{L:1}\Bigg\|_{\mathsf{F}}^2. \tag{26}$$

By expanding each of the terms and splitting by the orders of $\Delta\mathbf{W}_\ell$ (perturbation), we get that the second-order term is equivalent to

$$\Theta\left(\sum_{\ell=1}^{L}\|\Delta\mathbf{W}_\ell\|^2\right):\ \frac{1}{2}\Bigg\|\sum_\ell \mathbf{W}_{L:\ell+1}\Delta\mathbf{W}_\ell\mathbf{W}_{\ell-1:1}\Bigg\|_{\mathsf{F}}^2$$

$$\Theta\left(\sum_{\ell=1}^{L}\|\Delta\mathbf{W}_\ell\|^3\right):\ \mathrm{tr}\left[\left(\sum_\ell \mathbf{W}_{L:\ell+1}\Delta\mathbf{W}_\ell\mathbf{W}_{\ell-1:1}\right)^\top\left(\sum_{\ell<m} \mathbf{W}_{L:\ell+1}\Delta\mathbf{W}_\ell\mathbf{W}_{\ell-1:m+1}\Delta\mathbf{W}_m\mathbf{W}_{m-1:1}\right)\right]$$

$$\Theta\left(\sum_{\ell=1}^{L}\|\Delta\mathbf{W}_\ell\|^4\right):\ \frac{1}{2}\|\sum_{\ell<m} \mathbf{W}_{L:\ell+1}\Delta\mathbf{W}_\ell\mathbf{W}_{\ell-1:m+1}\Delta\mathbf{W}_m\mathbf{W}_{m-1:1}\|_{\mathsf{F}}^2$$

$$+\mathrm{tr}\left[\sum_l (\mathbf{W}_{L:\ell+1}\Delta\mathbf{W}_\ell\mathbf{W}_{\ell-1:1})^\top\left(\sum_{l<m<p} \mathbf{W}_{L:\ell+1}\Delta\mathbf{W}_\ell\mathbf{W}_{\ell-1:m+1}\Delta\mathbf{W}_m\mathbf{W}_{m-1:p+1}\Delta\mathbf{W}_p\mathbf{W}_{p-1:1}\right)\right]$$

The direction of the steepest change in the loss at the minima correspond to the largest eigenvector direction of the Hessian. Since higher order terms such as $\Theta\left(\sum_{\ell=1}^{L}\|\Delta\mathbf{W}_\ell\|^3\right)$ are insignifcant compared to the second order terms $\Theta\left(\sum_{\ell=1}^{L}\|\Delta\mathbf{W}_\ell\|^2\right)$, finding the direction that maximizes the

second order term leads to finding the eigenvector of the Hessian. Then, the eigenvector corresponding to the maximum eigenvalue of $\nabla^2 \mathcal{L}$ is the solution of

$$\Delta_1 \coloneqq \mathrm{vec}(\Delta \mathbf{W}_L, \dots \Delta \mathbf{W}_1) = \underset{\|\Delta \mathbf{W}_L\|_{\mathsf{F}}^2 + \ldots + \|\Delta \mathbf{W}_1\|_{\mathsf{F}}^2 = 1}{\operatorname{argmax}} f\left(\Delta \mathbf{W}_L, \dots, \Delta \mathbf{W}_1\right), \tag{27}$$

where

$$f(\Delta \mathbf{W}_L, \dots, \Delta \mathbf{W}_1) \coloneqq \frac{1}{2} \|\Delta \mathbf{W}_L \mathbf{W}_{L-1:1} + \ldots + \mathbf{W}_{L:3} \Delta \mathbf{W}_2 \mathbf{W}_1 + \mathbf{W}_{L:2} \Delta \mathbf{W}_1\|_{\mathsf{F}}^2. \tag{28}$$

While the solution of Equation (27) gives the maximum eigenvector direction of the Hessian, Δ_1, the other eigenvectors can be found by solving

$$\Delta_r \coloneqq \underset{\substack{\|\Delta \mathbf{W}_L\|_{\mathsf{F}}^2 + \ldots + \|\Delta \mathbf{W}_1\|_{\mathsf{F}}^2 = 1, \\ \Delta_r \perp \Delta_{r-1}, .., \Delta_r \perp \Delta_1}}{\operatorname{argmax}} f\left(\Delta \mathbf{W}_L, \dots, \Delta \mathbf{W}_1\right). \tag{29}$$

By expanding $f(\cdot)$, we have that

$$\begin{aligned} f(\Delta \mathbf{W}_L, \dots, \Delta \mathbf{W}_1) = \|\Delta \mathbf{W}_L \mathbf{W}_{L-1:1}\|_{\mathsf{F}}^2 + \ldots + \|\mathbf{W}_{L:3} \Delta \mathbf{W}_2 \mathbf{W}_1\|_{\mathsf{F}}^2 + \|\mathbf{W}_{L:2} \Delta \mathbf{W}_1\|_{\mathsf{F}}^2 \\ + \operatorname{tr}\left[\left(\Delta \mathbf{W}_L \mathbf{W}_{L-1:1}\right)^\top \left(\mathbf{W}_{L:3} \Delta \mathbf{W}_2 \mathbf{W}_1 + \ldots + \mathbf{W}_{L:2} \Delta \mathbf{W}_1\right) \right] + \ldots + \\ \operatorname{tr}\left[\left(\mathbf{W}_{L:2} \Delta \mathbf{W}_1\right)^\top \left(\mathbf{W}_{L:3} \Delta \mathbf{W}_2 \mathbf{W}_1 + \ldots + \mathbf{W}_{L:3} \Delta \mathbf{W}_2 \mathbf{W}_1\right) \right]. \end{aligned} \tag{30}$$

We can solve Equation (27) by maximizing each of the terms, which can be done in two steps:

(i) Each Frobenius term in the expansion is maximized when the left singular vector of $\Delta \mathbf{W}_\ell$ aligns with $\mathbf{W}_{L:\ell+1}$ and the right singular vector aligns with $\mathbf{W}_{\ell-1:1}$. This is a result of Von Neumann's trace inequality [60]. Similarly, each term in the trace is maximized when the singular vector of the perturbations align with the products.

(ii) Due to the alignment, Equation (27) can be written in just the singular values. Let $\Delta s_{\ell,i}$ denote the i-th singular value of the perturbation matrix $\Delta \mathbf{W}_\ell$. Recall that all of the singular values of $\mathbf{M}_\star$ are distinct (i.e., $\sigma_{\star,1} > \ldots > \sigma_{\star,r}$). Hence, it is easy to see that Equation (27) is maximized when $\Delta s_{\ell,i} = 0$ (i.e, all the weight goes to $\Delta s_{\ell,1}$). Thus, each perturbation matrix must be rank-1.

Now since each perturbation is rank-1, we can write each perturbation as

$$\Delta \mathbf{W}_\ell = \Delta s_\ell \Delta \mathbf{u}_\ell \Delta \mathbf{v}_\ell^\top, \quad \forall \ell \in [L], \tag{31}$$

for $\Delta s_\ell > 0$ and orthonormal vectors $\Delta \mathbf{u}_\ell \in \mathbb{R}^d$ and $\Delta \mathbf{v}_\ell \in \mathbb{R}^d$ with $\sum_{\ell=1}^L \Delta s_\ell^2 = 1$. Plugging this in each term, we obtain:

$$\|\mathbf{W}_{L:\ell+1} \Delta_1 \mathbf{W}_\ell \mathbf{W}_{\ell-1:1}\|_2^2 = \Delta_1 s_\ell^2 \cdot \left\| \underbrace{\mathbf{V}_\star \sigma_\star^{\frac{L-\ell}{L}} \mathbf{V}_\star^\top \Delta \mathbf{u}_\ell}_{=:\mathbf{a}} \underbrace{\Delta \mathbf{v}_\ell^\top \mathbf{V}_\star \sigma_\star^{\frac{\ell-1}{L}} \mathbf{V}_\star^\top}_{=:\mathbf{b}^\top} \right\|_2^2.$$

Since alignment maximizes this expression as discussed in first point, we have:

$\mathbf{u}_\ell = \mathbf{v}_\ell = \mathbf{v}_{\star,1}$ for all $\ell \in [2, L-1]$, then

$$\mathbf{a} = \sigma_{\star,1}^{\frac{L-\ell}{L}} \mathbf{v}_{\star,1} \quad \text{and} \quad \mathbf{b}^\top = \sigma_{\star,1}^{\frac{\ell-1}{L}} \mathbf{v}_{\star,1}^\top \implies \mathbf{a}\mathbf{b}^\top = \sigma_{\star,1}^{1-\frac{1}{L}} \cdot \mathbf{v}_{\star,1} \mathbf{v}_{\star,1}^\top.$$

The very same argument can be made for the trace terms. Hence, in order to maximize $f(\cdot)$, we must have

$$\begin{aligned} &\mathbf{v}_L = \mathbf{v}_{\star,1}, \quad \text{and} \quad \mathbf{u}_1 = \mathbf{v}_{\star,1}, \\ &\mathbf{u}_\ell = \mathbf{v}_\ell = \mathbf{v}_{\star,1}, \quad \forall \ell \in [2, L-1]. \end{aligned}$$

To determine $\mathbf{u}_L$ and $\mathbf{v}_1$, we can look at one of the trace terms:

$$\operatorname{tr}\left[(\Delta_1 \mathbf{W}_L \mathbf{W}_{L-1:1})^\top (\mathbf{W}_{L:3} \Delta_1 \mathbf{W}_2 \mathbf{W}_1 + \ldots + \mathbf{W}_{L:2} \Delta_1 \mathbf{W}_1)\right] \leq \left(\frac{L-1}{L}\right) \cdot \sigma_{\star,1}^{2-\frac{2}{L}}.$$

To reach the upper bound, we require $\mathbf{u}_L = \mathbf{u}_{\star,1}$ and $\mathbf{v}_1 = \mathbf{v}_{\star,1}$. Finally, as the for each index, the singular values are balanced, we will have $\Delta_1 s_\ell = \frac{1}{\sqrt{L}}$ for all $\ell \in [L]$ to satisfy the constraint. Finally, we get that the leading eigenvector is

$$\Delta_1 \coloneqq \operatorname{vec}\left(\frac{1}{\sqrt{L}} \mathbf{u}_1 \mathbf{v}_1^\top, \frac{1}{\sqrt{L}} \mathbf{v}_1 \mathbf{v}_1^\top, \ldots, \frac{1}{\sqrt{L}} \mathbf{v}_1 \mathbf{v}_1^\top\right).$$

Notice that we can also verify that $f(\Delta_1) = L\sigma_{\star,1}^{2-\frac{2}{L}}$, which is the leading eigenvalue (or sharpness) derived in Lemma 1.

To derive the remaining eigenvectors, we need to find all of the vectors in which $\Delta_i^\top \Delta_j = 0$ for $i \neq j$, where

$$\Delta_i = \operatorname{vec}(\Delta_i \mathbf{W}_L, \ldots \Delta_i \mathbf{W}_1),$$

and $f(\Delta_i) = \lambda_i$, where λ_i is the i-th largest eigenvalue. By repeating the same process as above, we find that the eigenvector-eigenvalue pair as follows:

$$\begin{aligned}
\Delta_1 &= \operatorname{vec}\left(\frac{1}{\sqrt{L}} \mathbf{u}_1 \mathbf{v}_1^\top, \frac{1}{\sqrt{L}} \mathbf{v}_1 \mathbf{v}_1^\top, \ldots, \frac{1}{\sqrt{L}} \mathbf{v}_1 \mathbf{v}_1^\top\right), \quad \lambda_1 = L\sigma_{\star,1}^{2-\frac{2}{L}} \\
\Delta_2 &= \operatorname{vec}\left(\frac{1}{\sqrt{L}} \mathbf{u}_1 \mathbf{v}_2^\top, \frac{1}{\sqrt{L}} \mathbf{v}_1 \mathbf{v}_2^\top, \ldots, \frac{1}{\sqrt{L}} \mathbf{v}_1 \mathbf{v}_2^\top\right), \quad \lambda_2 = \left(\sum_{i=0}^{L-1} \sigma_{\star,1}^{1-\frac{1}{L}-\frac{1}{L}i} \cdot \sigma_{\star,2}^{\frac{1}{L}i}\right) \\
\Delta_3 &= \operatorname{vec}\left(\frac{1}{\sqrt{L}} \mathbf{u}_2 \mathbf{v}_1^\top, \frac{1}{\sqrt{L}} \mathbf{v}_2 \mathbf{v}_1^\top, \ldots, \frac{1}{\sqrt{L}} \mathbf{v}_2 \mathbf{v}_1^\top\right), \quad \lambda_3 = \left(\sum_{i=0}^{L-1} \sigma_{\star,1}^{1-\frac{1}{L}-\frac{1}{L}i} \cdot \sigma_{\star,2}^{\frac{1}{L}i}\right) \\
&\vdots \\
\Delta_d &= \operatorname{vec}\left(\frac{1}{\sqrt{L}} \mathbf{u}_2 \mathbf{v}_2^\top, \frac{1}{\sqrt{L}} \mathbf{v}_2 \mathbf{v}_2^\top, \ldots, \frac{1}{\sqrt{L}} \mathbf{v}_2 \mathbf{v}_2^\top\right), \quad \lambda_d = L\sigma_{\star,2}^{2-\frac{2}{L}} \\
&\vdots \\
\Delta_{dr+r} &= \operatorname{vec}\left(\frac{1}{\sqrt{L}} \mathbf{u}_d \mathbf{v}_r^\top, \frac{1}{\sqrt{L}} \mathbf{v}_d \mathbf{v}_r^\top, \ldots, \frac{1}{\sqrt{L}} \mathbf{v}_d \mathbf{v}_r^\top\right),
\end{aligned}$$

which gives a total of $dr + r$ eigenvectors.

Second, equipped with the eigenvectors, let us consider the 1-D function f_{Δ_i} generated by the cross-section of the loss landscape and each eigenvector Δ_i passing the minima:

$$\begin{aligned}
f_{\Delta_i}(\mu) &= \mathcal{L}(\mathbf{W}_L + \mu \Delta \mathbf{W}_L, \ldots, \mathbf{W}_2 + \mu \Delta \mathbf{W}_2, \mathbf{W}_1 + \mu \Delta \mathbf{W}_1), \\
&= \mu^2 \cdot \frac{1}{2} \|\Delta \mathbf{W}_L \mathbf{W}_{L-1:1} + \ldots + \mathbf{W}_{L:3} \Delta \mathbf{W}_2 \mathbf{W}_1 + \mathbf{W}_{L:2} \Delta \mathbf{W}_1\|_{\mathsf{F}}^2 \\
&+ \mu^3 \cdot \sum_{\ell=1, \ell<m}^{L} \operatorname{tr}\left[(\mathbf{W}_{L:\ell+1} \Delta \mathbf{W}_\ell \mathbf{W}_{\ell-1:1})^\top (\mathbf{W}_{L:\ell+1} \Delta \mathbf{W}_\ell \mathbf{W}_{\ell-1:m+1} \Delta \mathbf{W}_m \mathbf{W}_{m-1:1})\right] \\
&+ \mu^4 \cdot \frac{1}{2} \left\| \left(\sum_{\ell<m} \mathbf{W}_{L:\ell+1} \Delta \mathbf{W}_\ell \mathbf{W}_{\ell-1:m+1} \Delta \mathbf{W}_m \mathbf{W}_{m-1:1} \right) \right\|_{\mathsf{F}}^2 \\
&+ \mu^4 \cdot \sum_{\ell<m<p}^{L} \operatorname{tr}\left[(\mathbf{W}_{L:\ell+1} \Delta \mathbf{W}_\ell \mathbf{W}_{\ell-1:1})^\top (\mathbf{W}_{L:\ell+1} \Delta \mathbf{W}_\ell \mathbf{W}_{\ell-1:m+1} \Delta \mathbf{W}_m \mathbf{W}_{m-1:p+1} \Delta \mathbf{W}_p \mathbf{W}_{p-1:1})\right].
\end{aligned}$$

Then, the several order derivatives of $f_{\Delta_i}(\mu)$ at $\mu = 0$ can be obtained from Taylor expansion as

$$
\begin{aligned}
f^{(2)}_{\Delta_i}(0) &= \|\Delta_i \mathbf{W}_L \mathbf{W}_{L-1:1} + \ldots + \mathbf{W}_{L:3}\Delta_i \mathbf{W}_2 \mathbf{W}_1 + \mathbf{W}_{L:2}\Delta_i \mathbf{W}_1\|_{\mathsf{F}}^2 = \lambda_i^2 \\
f^{(3)}_{\Delta_i}(0) &= 6\sum_{\ell=1}^{L} \mathrm{tr}\left[\left(\mathbf{W}_{L:\ell+1}\Delta_i \mathbf{W}_\ell \mathbf{W}_{\ell-1:1}\right)^\top \left(\mathbf{W}_{L:\ell+2}\Delta_i \mathbf{W}_{\ell+1}\mathbf{W}_\ell \Delta_i \mathbf{W}_{\ell-1}\mathbf{W}_{\ell-2:1}\right)\right] \\
&= 6\left\| \sum_{\ell} \mathbf{W}_{L:\ell+1}\Delta_i \mathbf{W}_\ell \mathbf{W}_{\ell-1:1} \right\|_{\mathsf{F}} \cdot \left\| \left(\sum_{\ell<m} \mathbf{W}_{L:\ell+1}\Delta \mathbf{W}_\ell \mathbf{W}_{\ell-1:m+1}\Delta \mathbf{W}_m \mathbf{W}_{m-1:1} \right) \right\|_{\mathsf{F}} \\
&:= 6\lambda_i \cdot \beta_i \\
f^{(4)}_{\Delta_i}(0) &= 12\|\Delta_i \mathbf{W}_L \Delta_i \mathbf{W}_{L-1}\mathbf{W}_{L-2:1} + \ldots + \mathbf{W}_{L:4}\Delta_i \mathbf{W}_3 \mathbf{W}_2 \Delta_i \mathbf{W}_1 + \mathbf{W}_{L:3}\Delta_i \mathbf{W}_2 \Delta_i \mathbf{W}_1\|_{\mathsf{F}}^2 \\
&\quad + 24\sum_{\ell=1}^{L} \mathrm{tr}\left[\left(\mathbf{W}_{L:\ell+1}\Delta_i \mathbf{W}_\ell \mathbf{W}_{\ell-1:1}\right)^\top \left(\sum_{l<m<p} \mathbf{W}_{L:\ell+1}\Delta \mathbf{W}_\ell \mathbf{W}_{\ell-1:m+1}\Delta \mathbf{W}_m \mathbf{W}_{m-1:p+1}\Delta \mathbf{W}_p \mathbf{W}_{p-1:1} \right)\right] \\
&:= 12\beta_i^2 + 24\lambda_i \cdot \delta_i,
\end{aligned}
$$

where we defined

$$
\begin{aligned}
\lambda_i &= \left\| \sum_{\ell} \mathbf{W}_{L:\ell+1}\Delta_i \mathbf{W}_\ell \mathbf{W}_{\ell-1:1} \right\|_{\mathsf{F}} && \text{(Total } \tbinom{L}{1} \text{ terms)} \\
\beta_i &= \left\| \left(\sum_{\ell<m} \mathbf{W}_{L:\ell+1}\Delta \mathbf{W}_\ell \mathbf{W}_{\ell-1:m+1}\Delta \mathbf{W}_m \mathbf{W}_{m-1:1} \right) \right\|_{\mathsf{F}} && \text{(Total } \tbinom{L}{2} \text{ terms)} \\
\delta_i &= \left\| \left(\sum_{l<m<p} \mathbf{W}_{L:\ell+1}\Delta \mathbf{W}_\ell \mathbf{W}_{\ell-1:m+1}\Delta \mathbf{W}_m \mathbf{W}_{m-1:p+1}\Delta \mathbf{W}_p \mathbf{W}_{p-1:1} \right) \right\|_{\mathsf{F}}, && \text{(Total } \tbinom{L}{3} \text{ terms)}
\end{aligned}
$$

and used the fact that $\mathrm{tr}(\mathbf{A}^\top \mathbf{B}) = \|\mathbf{A}\|_{\mathsf{F}} \cdot \|\mathbf{B}\|_{\mathsf{F}}$ under singular vector alignment.

Then, since β_i has $\binom{L}{2}$ terms inside the sum, when the Frobenium term is expanded, it will have $\frac{\binom{L}{2}(\binom{L}{2}+1)}{2}$ number of terms. Under alignment and balancedness, $\beta_i^2 = \Delta s_\ell^2 \sigma_i^{2-\frac{4}{L}} \times \frac{\binom{L}{2}(\binom{L}{2}+1)}{2}$ and $\lambda_i \delta_i = \Delta s_\ell^2 \sigma_i^{2-\frac{4}{L}} \times \binom{L}{3} L$. Thus, we have the expression

$$
\begin{aligned}
2\beta_i^2 - \lambda_i \delta_i &= \Delta s_\ell^2 \sigma_i^{2-\frac{4}{L}} \left(2\frac{\binom{L}{2}\left(\binom{L}{2}+1\right)}{2} - \binom{L}{3} L \right) \\
&= \Delta s_\ell^2 \sigma_i^{2-\frac{4}{L}} \binom{L}{3} L \times \left(\frac{3\left(\frac{L(L-1)}{2}+1\right)}{L(L-2)} - 1 \right) \\
&= \Delta s_\ell^2 \sigma_i^{2-\frac{4}{L}} \frac{2\binom{L}{3} L}{L(L-2)} \times \left((L-1)^2 + 5\right) > 0,
\end{aligned}
$$

for any depth $L > 2$. Finally, the condition of stable oscillation of 1-D function is

$$
3[f^{(3)}_{\Delta_i}]^2 - f^{(2)}_{\Delta_i} f^{(4)}_{\Delta_i} = 108\lambda_i^2 \beta_i^2 - (\lambda_i^2)(12\beta_i^2 + 24(2\lambda_i)(\delta_i)) = 48\lambda_i^2(2\beta_i^2 - \lambda_i \delta_i) > 0,
$$

which we have proven to be positive for any depth $L > 2$, for all the eigenvector directions corresponding to the non-zero eigenvalues. This completes the proof.

□

Theorem 2 (Rank-p Periodic Subspace Oscillations)**.** *Let $\mathbf{M}_\star = \mathbf{U}_\star \mathbf{\Sigma}_\star \mathbf{V}_\star^\top$ denote the SVD of the target matrix and define $S_p := L\sigma_{\star,p}^{2-\frac{2}{L}}$ and $K'_p := \max\left\{ S_{p+1}, \frac{S_p}{2\sqrt{2}} \right\}$. If we run GD on the deep matrix factorization loss with initialization scale $\alpha < \alpha'$ from Theorem 1 and learning rate $\eta = \frac{2}{K}$, where $K'_p < K < S_p$,*

then under strict balancing, the top-p singular values of the end-to-end DLN oscillates in a 2-period orbit ($j \in \{1,2\}$) around the balanced minimum and admits the following decomposition:

$$\mathbf{W}_{L:1} = \underbrace{\sum_{i=1}^{p} \rho_{i,j}^{L} \cdot \mathbf{u}_{\star,i}\mathbf{v}_{\star,i}^{\top}}_{\textit{oscillation subspace}} + \underbrace{\sum_{k=p+1}^{d} \sigma_{\star,k} \cdot \mathbf{u}_{\star,k}\mathbf{v}_{\star,k}^{\top}}_{\textit{stationary subspace}}, \quad j \in \{1,2\}, \quad \forall \ell \in [L-1], \tag{32}$$

where $\rho_{i,1} \in \left(0, \sigma_{\star,i}^{1/L}\right)$ and $\rho_{i,2} \in \left(\sigma_{\star,i}^{1/L}, (2\sigma_{\star,i})^{1/L}\right)$ are the two real roots of the polynomial $g(\rho_i) = 0$ and

$$g(\rho_i) = \rho_i^L \cdot \frac{1 + \left(1 + \eta L(\sigma_{\star,i} - \rho_i^L) \cdot \rho_i^{L-2}\right)^{2L-1}}{1 + \left(1 + \eta L(\sigma_{\star,i} - \rho_i^L) \cdot \rho_i^{L-2}\right)^{L-1}} - \sigma_{\star,i}.$$

Proof. To prove the result, we will consider the GD step on the i-th singular value and show that the 2-period orbit condition holds given the learning rate $\eta = \frac{2}{K}$. For ease of exposition, let us denote the i-th singular value of each $\mathbf{W}_\ell$ as $\sigma_i \coloneqq \sigma_{\ell,i}$. Under balancing, consider the two-step GD update on the first singular value:

$$\begin{aligned}
\sigma_i(t+1) &= \sigma_i(t) + \eta L \cdot \left(\sigma_{\star,i} - \sigma_i^L(t)\right) \cdot \sigma_i^{L-1}(t) \\
\sigma_i(t) = \sigma_i(t+2) &= \sigma_i(t+1) + \eta L \cdot \left(\sigma_{\star,i} - \sigma_i^L(t+1)\right) \cdot \sigma_i^{L-1}(t+1). \qquad \text{(By 2-period orbit)}
\end{aligned}$$

Define $z \coloneqq \left(1 + \eta L \cdot \left(\sigma_{\star,i} - \sigma_i^L(t)\right) \cdot \sigma_i^{L-2}(t)\right)$ and by plugging in $\sigma_i(t+1)$ for $\sigma_i(t)$, we have

$$\begin{aligned}
&\sigma_i(t) = \sigma_i(t)z + \eta L \cdot \left(\sigma_{\star,i} - \sigma_i^L(t)z^L\right) \cdot \sigma_i^{L-1}(t)z^{L-1} \\
\implies &1 = z + \eta L \cdot \left(\sigma_{\star,i} - \sigma_i^L(t)z^L\right) \cdot \sigma_i^{L-2}(t)z^{L-1} \\
\implies &1 = \left(1 + \eta L \cdot \left(\sigma_{\star,i} - \sigma_i^L(t)\right) \cdot \sigma_i^{L-2}(t)\right) + \eta L \cdot \left(\sigma_{\star,i} - \sigma_i^L(t)z^L\right) \cdot \sigma_i^{L-2}(t)z^{L-1} \\
\implies &0 = \left(\sigma_{\star,i} - \sigma_i^L(t)\right) + \left(\sigma_{\star,i} - \sigma_i^L(t)z^L\right) \cdot z^{L-1}
\end{aligned}$$

Simplifying this expression further, we have

$$\begin{aligned}
&0 = \sigma_{\star,i} - \sigma_i^L(t) + \sigma_{\star,i}z^{L-1} - \sigma_i^L(t)z^{2L-1} \\
\implies &\sigma_i^L(t) + \sigma_i^L(t)z^{2L-1} = \sigma_{\star,i} + \sigma_{\star,i}z^{L-1} \\
\implies &\sigma_i^L(t) \cdot \left(1 + z^{2L-1}\right) = \sigma_{\star,i} \cdot \left(1 + z^{L-1}\right) \\
\implies &\sigma_i^L(t) \frac{\left(1 + z^{2L-1}\right)}{\left(1 + z^{L-1}\right)} = \sigma_{\star,i},
\end{aligned}$$

and by defining $\rho_i \coloneqq \sigma_i(t)$, we obtain the polynomial

$$\sigma_{\star,i} = \rho_i^L \frac{1 + z^{2L-1}}{1 + z^{L-1}}, \quad \text{where } z \coloneqq \left(1 + \eta L(\sigma_{\star,i} - \rho_i^L) \cdot \rho_i^{L-2}\right).$$

Next, we show the existence of (real) roots within the ranges for $\rho_{i,1}$ and $\rho_{i,2}$. We note that these roots only exist within the EOS regime. First, consider $\rho_{i,1} \in \left(0, \sigma_{\star,i}^{1/L}\right)$. We will show that for two values within this range, there is a sign change for all $L \geq 2$. More specifically, we show that there exists $\rho_i \in \left(0, \sigma_{\star,i}^{1/L}\right)$ such that

$$\rho_i^L \frac{1 + z^{2L-1}}{1 + z^{L-1}} - \sigma_{\star,i} > 0 \quad \text{and} \quad \rho_i^L \frac{1 + z^{2L-1}}{1 + z^{L-1}} - \sigma_{\star,i} < 0.$$

For the positive case, consider $\rho_i = (\frac{1}{2}\sigma_{\star,i})^{1/L}$. We need to show that

$$\frac{1 + z^{2L-1}}{1 + z^{L-1}} = \frac{1 + \left(1 + \eta L \cdot \left(\frac{\sigma_{\star,i}}{2}\right) \frac{\sigma_{\star,i}^{1-\frac{2}{L}}}{2^{1-\frac{2}{L}}}\right)^{2L-1}}{1 + \left(1 + \eta L \cdot \left(\frac{\sigma_{\star,i}}{2}\right) \frac{\sigma_{\star,i}^{1-\frac{2}{L}}}{2^{1-\frac{2}{L}}}\right)^{L-1}} > 2.$$

To do this, we will plug in the smallest possible value of $\eta = \frac{2}{L\sigma_{\star,i}^{2-\frac{2}{L}}}$ to show that the fraction is still greater than 2, which gives us

$$u(L) \coloneqq \frac{1 + \left(1 + \frac{1}{2^{1-\frac{2}{L}}}\right)^{2L-1}}{1 + \left(1 + \frac{1}{2^{1-\frac{2}{L}}}\right)^{L-1}}, \tag{33}$$

which is an increasing function of L for all $L \geq 2$. Since $u(2) > 2$, Equation (33) must always be greater than 2. For the negative case, we can simply consider $\rho_i = 0$. Hence, since the polynomial is continuous, by the Intermediate Value Theorem (IVT), there must exist a root within the range $\rho_i \in \left(0, \sigma_{\star,i}^{1/L}\right)$.

Next, consider the range $\rho_{i,2} \in \left(\sigma_{\star,i}^{1/L}, (2\sigma_{\star,i})^{1/L}\right)$. Similarly, we will show sign changes for two values in $\rho_{i,2}$. For the positive case, consider $\rho_i = \left(\frac{3}{2}\sigma_{\star,i}\right)^{1/L}$. For η, we can plug in the smallest possible value within the range to show that this value of ρ_i provides a positive quantity. Specifically, we need to show that

$$\frac{1+z^{2L-1}}{1+z^{L-1}} > \frac{2}{3} \implies \frac{1 + \left(1 + \frac{2}{\sigma_{\star,i}^{2-\frac{2}{L}}} \cdot (\sigma_{\star,i} - \frac{3}{2}\sigma_{\star,i}) \cdot \left(\frac{3}{2}\sigma_{\star,i}\right)^{1-\frac{2}{L}}\right)^{2L-1}}{1 + \left(1 + \frac{2}{\sigma_{\star,i}^{2-\frac{2}{L}}} \cdot (\sigma_{\star,i} - \frac{3}{2}\sigma_{\star,i}) \cdot \left(\frac{3}{2}\sigma_{\star,i}\right)^{1-\frac{2}{L}}\right)^{L-1}} > \frac{2}{3}.$$

We can simplify the fraction as follows:

$$\frac{1 + \left(1 + \frac{2}{\sigma_{\star,1}^{2-\frac{2}{L}}} \cdot (\sigma_{\star,1} - \frac{3}{2}\sigma_{\star,1}) \cdot \left(\frac{3}{2}\sigma_{\star,1}\right)^{1-\frac{2}{L}}\right)^{2L-1}}{1 + \left(1 + \frac{2}{\sigma_{\star,1}^{2-\frac{2}{L}}} \cdot (\sigma_{\star,1} - \frac{3}{2}\sigma_{\star,1}) \cdot \left(\frac{3}{2}\sigma_{\star,1}\right)^{1-\frac{2}{L}}\right)^{L-1}} = \frac{1 + \left(1 - (\frac{3}{2})^{1-\frac{2}{L}}\right)^{2L-1}}{1 + \left(1 - (\frac{3}{2})^{1-\frac{2}{L}}\right)^{L-1}}.$$

Then, since we are subtracting by $(\frac{3}{2})^{1-\frac{2}{L}}$, we can plug in its largest value for $L \geq 2$, which is $3/2$. This gives us

$$\frac{1 + (-0.5)^{2L-1}}{1 + (-0.5)^{L-1}} > \frac{2}{3},$$

as for odd values of L, the function increases to 1 starting from $L = 2$, and decreases to 1 for even L. To check negativity, let us define

$$h(\rho) \coloneqq \frac{f(\rho)}{g(\rho)} \coloneqq \frac{\rho^L\left(1 + z^{2L-1}\right)}{1 + z^{L-1}}.$$

We will show that $h'\left(\sigma_{\star,i}^{1/L}\right) < 0$:

$$\begin{aligned}
h'\left(\sigma_{\star,i}^{1/L}\right) &= \frac{f'\left(\sigma_{\star,i}^{1/L}\right) g\left(\sigma_{\star,i}^{1/L}\right) - f\left(\sigma_{\star,i}^{1/L}\right) g'\left(\sigma_{\star,i}^{1/L}\right)}{g^2\left(\sigma_{\star,i}^{1/L}\right)} \\
&= \frac{f'\left(\sigma_{\star,i}^{1/L}\right) - \sigma_{\star,i} \cdot g'\left(\sigma_{\star,i}^{1/L}\right)}{2} \\
&= \frac{L\sigma_{\star,i}^{1-\frac{1}{L}} - \sigma_{\star,i}(2L-1)\left(\eta L^2 \sigma_{\star,i}^{2-\frac{3}{L}}\right) - \sigma_{\star,i}(L-1)\left(\eta L^2 \sigma_{\star,i}^{2-\frac{3}{L}}\right)}{2} \\
&= \frac{L\sigma_{\star,i}^{1-\frac{1}{L}} - (3L-2)\left(\eta L^2 \sigma_{\star,i}^{3-\frac{3}{L}}\right)}{2} < 0,
\end{aligned}$$

as otherwise we need $\eta \leq \frac{\sigma_{\star,i}^{2/L-2}}{3L^2-2L}$, which is out of the range of interest. Since $h'(\rho) < 0$, it follows that there exists a $\delta > 0$ such that $h(\rho) > h(x)$ for all x such that $\rho < x < \rho + \delta$. Lastly, since $h(\rho) - \sigma_{\star,i} = 0$ for $\rho = \sigma_{\star,i}^{1/L}$, it follows that $h(\rho) - \sigma_{\star,i}$ must be negative at $\rho + \delta$. Similarly, by IVT, there must exist a root within the range $\rho_{i,2} \in \left(\sigma_{\star,i}^{1/L}, (2\sigma_{\star,i})^{1/L}\right)$. This proves that the i-th singular value undergoes a two-period orbit with the roots $\rho_{i,1}$ and $\rho_{i,2}$. Then, notice that if the learning rate is large enough to induce oscillations in the i-th singular value, then it is also large enough to have oscillations in all singular values from 1 to the $(i-1)$-th singular value (assuming that it is not large enough for divergence). Finally, at the (balanced) minimum, we can express the dynamics as

$$\mathbf{W}_{L:1} = \underbrace{\sum_{i=1}^{p} \rho_{i,j}^{L} \cdot \mathbf{u}_{\star,i}\mathbf{v}_{\star,i}^{\top}}_{\text{oscillation subspace}} + \underbrace{\sum_{k=p+1}^{d} \sigma_{\star,k} \cdot \mathbf{u}_{\star,k}\mathbf{v}_{\star,k}^{\top}}_{\text{stationary subspace}}, \quad j \in \{1,2\}, \quad \forall \ell \in [L-1]. \tag{34}$$

This completes the proof.

□

Lemma 1 (Hessian Eigenvalues at Convergence)**.** *Consider running GD on the deep matrix factorization loss $f(\mathbf{\Theta})$ defined in Equation (1). The set of all non-zero eigenvalues of the training loss Hessian at the balanced minimum is given by*

$$\lambda_{\mathbf{\Theta}} = \underbrace{\left\{L\sigma_{\star,i}^{2-\frac{2}{L}}, \sigma_{\star,i}^{2-\frac{2}{L}}\right\}_{i=1}^{r}}_{\textit{self-interaction}} \bigcup \underbrace{\left\{\sum_{\ell=0}^{L-1}\left(\sigma_{\star,i}^{1-\frac{1}{L}-\frac{1}{L}\ell} \cdot \sigma_{\star,j}^{\frac{1}{L}\ell}\right)^2\right\}_{i\neq j}^{r}}_{\textit{interaction with other singular values}} \bigcup \underbrace{\left\{\sum_{\ell=0}^{L-1}\left(\sigma_{\star,k}^{1-\frac{1}{L}-\frac{1}{L}\ell} \cdot \alpha^{\ell}\right)^2\right\}_{k=1}^{r}}_{\textit{interaction with initialization}}$$

where $\sigma_{\star,i}$ is the i-th singular value of the target matrix $\mathbf{M}_\star \in \mathbb{R}^{d\times d}$, $\alpha \in \mathbb{R}$ is the initialization scale, L is the depth of the network, and the second element of the set under "self-interaction" has a multiplicity of $d-r$.

Proof. By Proposition 2, notice that we can re-write the loss in Equation (1) as

$$\frac{1}{2}\left\|\mathbf{W}_{L:1} - \mathbf{M}_\star\right\|_{\mathsf{F}}^2 = \frac{1}{2}\|\mathbf{\Sigma}_{L:1} - \mathbf{\Sigma}_\star\|_{\mathsf{F}}^2,$$

where $\mathbf{\Sigma}_{L:1}$ are the singular values of $\mathbf{W}_{L:1}$. We will first show that the eigenvalues of the Hessian with respect to the weight matrices $\mathbf{W}_\ell$ are equivalent to those of the Hessian taken with respect to its singular values $\mathbf{\Sigma}_\ell$. To this end, consider the vectorized form of the loss:

$$f(\mathbf{\Theta}) \coloneqq \frac{1}{2}\|\mathbf{W}_{L:1} - \mathbf{M}_\star\|_{\mathsf{F}}^2 = \frac{1}{2}\|\mathrm{vec}(\mathbf{W}_{L:1}) - \mathrm{vec}(\mathbf{M}_\star)\|_2^2.$$

Then, each block of the Hessian $\nabla_{\mathbf{\Theta}}^2 f(\mathbf{\Theta}) \in \mathbb{R}^{d^2L\times d^2L}$ with respect to the vectorized parameters is given as

$$\left[\nabla_{\mathbf{\Theta}}^2 f(\mathbf{\Theta})\right]_{m,\ell} = \nabla_{\mathrm{vec}(\mathbf{W}_m)}\nabla_{\mathrm{vec}(\mathbf{W}_\ell)}^{\top} f(\mathbf{\Theta}) \in \mathbb{R}^{d^2\times d^2}.$$

By the vectorization trick, each vectorized layer matrix has an SVD of the form $\mathrm{vec}(\mathbf{W}_\ell) = \mathrm{vec}(\mathbf{U}_\ell\mathbf{\Sigma}_\ell\mathbf{V}_\ell^{\top}) = (\mathbf{V}_\ell \otimes \mathbf{U}_\ell)\cdot\mathrm{vec}(\mathbf{\Sigma}_\ell)$. Then, notice that we have

$$\nabla_{\mathrm{vec}(\mathbf{W}_\ell)} f(\mathbf{\Theta}(t)) = (\mathbf{V}_\ell \otimes \mathbf{U}_\ell)\cdot\nabla_{\mathrm{vec}(\mathbf{\Sigma}_\ell)} f(\mathbf{\Theta}(t)),$$

which gives us that each block of the Hessian is given by

$$\begin{aligned}\left[\nabla_{\mathbf{\Theta}}^2 f(\mathbf{\Theta})\right]_{m,\ell} &= \nabla_{\mathrm{vec}(\mathbf{W}_m)}\nabla_{\mathrm{vec}(\mathbf{W}_\ell)}^{\top} f(\mathbf{\Theta}) \\ &= (\mathbf{V}_m \otimes \mathbf{U}_m)\cdot\underbrace{\nabla_{\mathrm{vec}(\mathbf{\Sigma}_m)}\nabla_{\mathrm{vec}(\mathbf{\Sigma}_\ell)}^{\top} f(\mathbf{\Theta})}_{=:\mathbf{H}_{m,\ell}}\cdot(\mathbf{V}_\ell \otimes \mathbf{U}_\ell)^{\top}.\end{aligned}$$

Then, since the Kronecker product of two orthogonal matrices is also an orthogonal matrix by Lemma 10, we can write the overall Hessian matrix as

$$\widetilde{\mathbf{H}} = \begin{bmatrix} \mathbf{R}_1\mathbf{H}_{1,1}\mathbf{R}_1 & \mathbf{R}_1\mathbf{H}_{1,2}\mathbf{R}_2 & \dots & \mathbf{R}_1\mathbf{H}_{1,L}\mathbf{R}_L \\ \mathbf{R}_2\mathbf{H}_{2,1}\mathbf{R}_1 & \mathbf{R}_2\mathbf{H}_{2,2}\mathbf{R}_2 & \dots & \mathbf{R}_2\mathbf{H}_{2,L}\mathbf{R}_L \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_L\mathbf{H}_{L,1}\mathbf{R}_1 & \mathbf{R}_L\mathbf{H}_{L,2}\mathbf{R}_2 & \dots & \mathbf{R}_L\mathbf{H}_{L,L}\mathbf{R}_L \end{bmatrix},$$

for orthogonal matrices $\{\mathbf{R}_\ell\}_{\ell=1}^L$. Then, by Lemma 9, the eigenvalues of $\widetilde{\mathbf{H}}$ are the same as those of $\mathbf{H}$, where $\mathbf{H} \in \mathbb{R}^{d^2L \times d^2L}$ is the Hessian matrix with respect to the vectorized $\mathbf{\Sigma}_\ell$:

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{1,1} & \mathbf{H}_{1,2} & \dots & \mathbf{H}_{L,1} \\ \mathbf{H}_{2,1} & \mathbf{H}_{2,2} & \dots & \mathbf{H}_{L,2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{1,L} & \mathbf{H}_{2,L} & \dots & \mathbf{H}_{L,L} \end{bmatrix}.$$

Now, we can consider the following vectorized loss:

$$\begin{aligned} f(\mathbf{\Theta}) = \frac{1}{2}\|\mathbf{\Sigma}_{L:1} - \mathbf{\Sigma}_\star\|_{\mathsf{F}}^2 &= \frac{1}{2}\left\|\mathrm{vec}\left(\mathbf{\Sigma}_{L:1} - \mathbf{\Sigma}_\star\right)\right\|_2^2 \\ &= \frac{1}{2}\| \underbrace{\left(\mathbf{\Sigma}_{\ell-1:1}^\top \otimes \mathbf{\Sigma}_{L:\ell+1}\right)}_{=:\mathbf{A}_\ell} \cdot \mathrm{vec}(\mathbf{\Sigma}_\ell) - \mathrm{vec}(\mathbf{\Sigma}_\star)\|_2^2. \end{aligned}$$

Then, the gradient with respect to $\mathrm{vec}(\mathbf{\Sigma}_\ell)$ is given by

$$\nabla_{\mathrm{vec}(\mathbf{\Sigma}_\ell)} f(\mathbf{\Theta}) = \mathbf{A}_\ell^\top \left(\mathbf{A}_\ell \cdot \mathrm{vec}(\mathbf{\Sigma}_\ell) - \mathrm{vec}(\mathbf{\Sigma}_\star)\right).$$

Then, for $m = \ell$, we have

$$\mathbf{H}_{\ell,\ell} = \nabla^2_{\mathrm{vec}(\mathbf{\Sigma}_\ell)} f(\mathbf{\Theta}) = \mathbf{A}_\ell^\top \mathbf{A}_\ell.$$

For $m \neq \ell$, we have

$$\begin{aligned} &\mathbf{H}_{m,\ell} = \nabla_{\mathrm{vec}(\mathbf{\Sigma}_m)} \nabla_{\mathrm{vec}(\mathbf{\Sigma}_\ell)} f(\mathbf{\Theta}) = \nabla_{\mathrm{vec}(\mathbf{\Sigma}_m)} \left[\mathbf{A}_\ell^\top (\mathbf{A}_\ell \mathrm{vec}(\mathbf{\Sigma}_\ell) - \mathrm{vec}(\mathbf{M}^\star))\right] \\ &= \nabla_{\mathrm{vec}(\mathbf{\Sigma}_m)} \mathbf{A}_\ell^\top \cdot \underbrace{(\mathbf{A}_\ell \mathrm{vec}(\mathbf{\Sigma}_\ell) - \mathrm{vec}(\mathbf{M}^\star))}_{=0 \text{ at convergence}} + \mathbf{A}_\ell^\top \cdot \nabla_{\mathrm{vec}(\mathbf{\Sigma}_m)} (\mathbf{A}_\ell \mathrm{vec}(\mathbf{\Sigma}_\ell) - \mathrm{vec}(\mathbf{M}^\star)) \\ &= \mathbf{A}_\ell^\top \mathbf{A}_m, \end{aligned}$$

where we have used the product rule along with the fact that $\mathbf{A}_\ell \mathrm{vec}(\mathbf{\Sigma}_\ell) = \mathbf{A}_m \mathrm{vec}(\mathbf{\Sigma}_m)$.

Overall, the Hessian at convergence for GD is given by

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_1^\top \mathbf{A}_1 & \mathbf{A}_1^\top \mathbf{A}_2 & \dots & \mathbf{A}_1^\top \mathbf{A}_L \\ \mathbf{A}_2^\top \mathbf{A}_1 & \mathbf{A}_2^\top \mathbf{A}_2 & \dots & \mathbf{A}_2^\top \mathbf{A}_L \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_L^\top \mathbf{A}_1 & \mathbf{A}_L^\top \mathbf{A}_2 & \dots & \mathbf{A}_L^\top \mathbf{A}_L \end{bmatrix}$$

Now, we can derive an explicit expression for each $\mathbf{A}_{m,\ell}$ by considering the implicit balancing effect of GD in Lemma 2. Under balancing and Proposition 2, we have that at convergence,

$$\mathbf{\Sigma}_{L:1} = \mathbf{\Sigma}_\star \implies \mathbf{\Sigma}_\ell = \begin{bmatrix} \mathbf{\Sigma}_{\star,r}^{1/L} & \mathbf{0} \\ \mathbf{0} & \alpha \cdot \mathbf{I}_{d-r} \end{bmatrix}, \quad \forall \ell \in [L-1], \quad \text{and } \mathbf{\Sigma}_L = \mathbf{\Sigma}_\star^{1/L}.$$

Thus, we have

$$\mathbf{H}_{m,\ell} = \begin{cases} \mathbf{\Sigma}_\ell^{2(\ell-1)} \otimes \mathbf{\Sigma}_\star^{\frac{2(L-\ell)}{L}} & \text{for } m = \ell, \\ \mathbf{\Sigma}_\ell^{m+\ell-2} \otimes \mathbf{\Sigma}_\star^{2L-m-\ell} & \text{for } m \neq \ell. \end{cases}$$

Now, we are left with computing the eigenvalues of $\mathbf{H} \in \mathbb{R}^{d^2L \times d^2L}$. To do this, let us block diagonalize $\mathbf{H}$ into $\mathbf{H} = \mathbf{PCP}^\top$, where $\mathbf{P}$ is a permutation matrix and

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 & & \\ & \ddots & \\ & & \mathbf{C}_{d^2} \end{bmatrix} \in \mathbb{R}^{d^2L \times d^2L},$$

where each (i,j)-th entry of $\mathbf{C}_k \in \mathbb{R}^{L \times L}$ is the k-th diagonal element of $\mathbf{H}_{i,j}$. Since $\mathbf{C}$ and $\mathbf{H}$ are similar matrices, they have the same eigenvalues. Then, since $\mathbf{C}$ is a block diagonal matrix, its eigenvalues (and hence the eigenvalues of $\mathbf{H}$) are the union of each of the eigenvalues of its blocks.

By observing the structure of $\mathbf{H}_{m,\ell}$, notice that each $\mathbf{C}_k$ is a rank-1 matrix. Hence, when considering the top-r diagonal elements of $\mathbf{H}_{m,\ell}$ corresponding to each Kronecker product to construct $\mathbf{C}_k$, each $\mathbf{C}_k$ can be written as an outer product $\mathbf{u}\mathbf{u}^\top$, where $\mathbf{u} \in \mathbb{R}^L$ is

$$\mathbf{u}^\top = \begin{bmatrix} \sigma_{\star,i}^{1-\frac{1}{L}} \sigma_{\star,j}^0 & \sigma_{\star,i}^{1-\frac{2}{L}} \sigma_{\star,j}^{\frac{1}{L}} & \sigma_{\star,i}^{1-\frac{3}{L}} \sigma_{\star,j}^{\frac{2}{L}} & \dots & \sigma_{\star,i}^0 \sigma_{\star,j}^{1-\frac{1}{L}} \end{bmatrix}^\top. \tag{35}$$

Then, the non-zero eigenvalue of this rank-1 matrix is simply $\|\mathbf{u}\|_2^2$, which simplifies to

$$\|\mathbf{u}\|_2^2 = \sum_{\ell=0}^{L-1} \left(\sigma_{\star,i}^{1-\frac{1}{L}-\frac{1}{L}\ell} \cdot \sigma_{\star,j}^{\frac{1}{L}\ell} \right)^2.$$

Next, we can consider the remaining $d-r$ components of each Kronecker product of $\mathbf{H}_{m,\ell}$. Notice that for $m = \ell = L$, we have

$$\mathbf{H}_{L,L} = \begin{bmatrix} \sigma_{\star,1}^{\frac{2(L-1)}{L}} \cdot \mathbf{I}_d & & & \\ & \ddots & & \\ & & \sigma_{\star,r}^{\frac{2(L-1)}{L}} \cdot \mathbf{I}_d & \\ & & & \alpha^{2(L-1)} \mathbf{I}_{d-r} \otimes \mathbf{I}_d \end{bmatrix}.$$

This amounts to a matrix $\mathbf{C}_k$ with a single element $\sigma_{\star,i}^{\frac{2(L-1)}{L}}$ and 0 elsewhere. This gives an eigenvalue $\sigma_{\star,i}^{\frac{2(L-1)}{L}}$ for all $i \in [r]$, with multiplicity $d-r$.

Lastly, we can consider the diagonal components of $\mathbf{H}_{m,\ell}$ that is a function of the initialization scale α. For this case, each $\mathbf{C}_k$ can be written as an outer product $\mathbf{v}\mathbf{v}^\top$, where

$$\mathbf{v}^\top = \begin{bmatrix} \sigma_{\star,i}^{1-\frac{1}{L}} \alpha^0 & \sigma_{\star,i}^{1-\frac{2}{L}} \alpha & \sigma_{\star,i}^{1-\frac{3}{L}} \alpha^2 & \dots & \sigma_{\star,i}^0 \alpha^{L-1} \end{bmatrix}^\top. \tag{36}$$

Similarly, the non-zero eigenvalue is simply $\|\mathbf{v}\|_2^2$, which corresponds to

$$\|\mathbf{v}\|_2^2 = \sum_{\ell=0}^{L-1} \left(\sigma_{\star,k}^{1-\frac{1}{L}-\frac{1}{L}\ell} \cdot \alpha^\ell \right)^2.$$

This completes the proof. □

Lemma 2 (Balancing)**.** *Let $\sigma_{\star,i}$ and $\sigma_{\ell,i}(t)$ denote the i-th singular value of $\mathbf{M}_\star \in \mathbb{R}^{d\times d}$ and $\mathbf{W}_\ell(t)$, respectively and define $S_i \coloneqq L\sigma_{\star,i}^{2-\frac{2}{L}}$. Consider GD on the i-th index of the simplified loss in (4) with the unbalanced initialization and learning rate $\frac{2}{S_i} < \eta < \frac{2\sqrt{2}}{S_i}$. If the initialization scale α satisfies $0 < \alpha < \left(\ln\left(\frac{2\sqrt{2}}{\eta S_i}\right) \cdot \frac{\sigma_{\star,i}^{4/L}}{L^2 \cdot 2^{\frac{2L-3}{L}}} \right)^{1/4}$, then there exists a constant $c \in (0,1]$ such that for all $\ell \in [L-1]$, we have $\left|\sigma_{L,i}^2(t+1) - \sigma_{\ell,i}^2(t+1)\right| < c\left|\sigma_{L,i}^2(t) - \sigma_{\ell,i}^2(t)\right|$.*

Proof. To prove the result, we will use a quantity defined as the gradient flow solution (GFS) sharpness following Kreisler et al. [13]:

Definition 3 (GFS Sharpness)**.** *The GFS sharpness denoted by $\psi(x)$ is the sharpness achieved by the global minima which lies in the same GF trajectory of x (i.e., $\|\nabla^2 L(z)\|$ such that $L(z) = 0$ and $z = GF(x)$, where $GF(\cdot)$ denotes the gradient flow solution).*

Then, let us consider the i-th index of the simplified loss in (4):

$$\frac{1}{2}\left(\prod_{\ell=1}^{L}\sigma_{\ell,i} - \sigma_{\star,i}\right)^2 =: \frac{1}{2}\left(\prod_{\ell=1}^{L}\sigma_{\ell} - \sigma_{\star}\right)^2,$$

and omit the dependency on i for ease of exposition. Our goal is to show that the L-th singular value σ_L initialized to zero become increasingly balanced to σ_ℓ which are initialized to α. To that end, let us define the balancing dynamics between σ_i and σ_j as $b_{i,j}^{(t+1)} \coloneqq \left(\sigma_i^{(t+1)}\right)^2 - \left(\sigma_j^{(t+1)}\right)^2$ and $\pi^{(t)} \coloneqq \prod_{\ell=1}^{L}\sigma_\ell(t)$ for the product of singular values at iteration t. We can simplify the balancing dynamics as follows:

$$\begin{aligned}
b_{i,j}^{(t+1)} &= \left(\sigma_i^{(t+1)}\right)^2 - \left(\sigma_j^{(t+1)}\right)^2 \\
&= \left(\sigma_i^{(t)} - \eta\left(\pi^{(t)} - \sigma_\star\right)\frac{\pi^{(t)}}{\sigma_i^{(t)}}\right)^2 - \left(\sigma_j^{(t)} - \eta\left(\pi^{(t)} - \sigma_\star\right)\frac{\pi^{(t)}}{\sigma_j^{(t)}}\right)^2 \\
&= \left(\sigma_i^{(t)}\right)^2 - \left(\sigma_j^{(t)}\right)^2 + \eta^2\left(\pi^{(t)} - \sigma_\star\right)^2\left(\frac{\left(\pi^{(t)}\right)^2}{\left(\sigma_i^{(t)}\right)^2} - \frac{\left(\pi^{(t)}\right)^2}{\left(\sigma_j^{(t)}\right)^2}\right) \\
&= \left(\left(\sigma_i^{(t)}\right)^2 - \left(\sigma_j^{(t)}\right)^2\right)\left(1 - \eta^2(\pi^{(t)} - \sigma_\star)^2\frac{\left(\pi^{(t)}\right)^2}{\left(\sigma_i^{(t)}\right)^2\left(\sigma_j^{(t)}\right)^2}\right) \\
&= b_{i,j}^{(t)}\left(1 - \eta^2(\pi^{(t)} - \sigma_\star)^2\frac{\left(\pi^{(t)}\right)^2}{\left(\sigma_i^{(t)}\right)^2\left(\sigma_j^{(t)}\right)^2}\right).
\end{aligned}$$

Then, in order to show that $\left|b_{i,j}^{(t+1)}\right| < c\left|b_{i,j}^{(t)}\right|$ for some $0 < c \leq 1$, we need to prove that

$$\left|1 - \eta^2(\pi^{(t)} - \sigma_\star)^2\frac{\left(\pi^{(t)}\right)^2}{\left(\sigma_i^{(t)}\right)^2\left(\sigma_j^{(t)}\right)^2}\right| < c,$$

for all iterations t. We complete this using the following two steps:

(i) We show that for all scalars σ in the trajectory, if $\psi(\sigma) < \frac{2\sqrt{1+c}}{\eta}$ and $\sigma > 0$, then it holds that $\sum_{i=1}^{\min\{2,L-1\}}\frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-i}^2\sigma_L^2} \leq 1 + c$, where $\pi(\sigma)$ denotes the product given the trajectory of all σ_i. This case is analyzed when $\pi(\sigma) \in [0, \sigma_\star)$ where $0 < c < 1$ and when $\pi(\sigma) > \sigma_\star$ where $c = 1$.

(ii) If $\sum_{i=1}^{\min\{2,L-1\}}\frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-i}^2\sigma_L^2} \leq 1+c$, then iterates become more balanced, i.e, $|b_{i,j}^{(t+1)}| < c|b_{i,j}^{(t)}|$.

We prove (i) in Lemma 3 and (ii) in Lemma 4. Both of the proofs are originally from [13], which we adapted using our notation for ease of the reader. Then, in Lemma 5, we show that for each $\sigma_\star$, as long as the initialization scale satisfies

$$\alpha < \left(\ln\left(\frac{2\sqrt{2}}{\eta L\sigma_\star^{2-\frac{2}{L}}}\right)\cdot\frac{\sigma_\star^{\frac{4}{L}}}{L^2\cdot 2^{\frac{2L-3}{L}}}\right)^{\frac{1}{4}},$$

then it holds that the GFS sharpness satisfies $\psi(\sigma) < \frac{2\sqrt{2}}{\eta}$, which is the necessary condition for balancing. Then, to satisfy this condition for all singular values $\sigma_{\star,i}$ for all $i \in [r]$, we need

$$\alpha < \left(\ln \left(\frac{2\sqrt{2}}{\eta L \sigma_{\star,1}^{2-\frac{2}{L}}} \right) \cdot \frac{\sigma_{\star,1}^{\frac{4}{L}}}{L^2 \cdot 2^{\frac{2L-3}{L}}} \right)^{\frac{1}{4}} \implies \eta < \frac{2\sqrt{2}}{L\sigma_{\star,1}^{2-\frac{2}{L}}}, \tag{37}$$

for the validity of the initialization scale. Thus, as long as the conditions in Equation (37) hold, we will have balancing. This completes the proof. □

Lemma 3. *If the GFS sharpness $\psi(\sigma) \leq \frac{2\sqrt{1+c}}{\eta}$ and $\sigma > 0$, then $\sum_{i=1}^{\min\{2,L-1\}} \frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{[L-i]}^2\sigma_{[D]}^2} \leq (1+c)$ for some $0 < c \leq 1$.*

Proof. We will consider two cases: (i) $\pi(\sigma) \in [0, \sigma_\star)$ and (ii) $\pi(\sigma) > \sigma_\star$ [1] .

Case 1: Let $\sigma \in \mathbb{R}^D$ and consider the case where $\pi(\sigma) \in [0, \sigma_\star)$. Then, we have

$$\sum_{i=1}^{\min\{2,L-1\}} \frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-i}^2\sigma_L^2} \leq \frac{\eta^2\pi^2(\sigma)}{\sigma_{L-i}^2\sigma_L^2}.$$

Our goal is to show that if $\psi(\sigma) \leq \frac{2\sqrt{1+c}}{\eta}$ for some $0 < c < 1$ then,

$$\sum_{i=1}^{\min\{2,L-1\}} \frac{\eta^2(\pi(\sigma)-1)^2\pi^2(\sigma)}{\sigma_{L-i}^2\sigma_L^2} \leq \frac{\eta^2\pi^2(\sigma)}{\sigma_{L-i}^2\sigma_L^2} \leq 1+c.$$

Since the GFS sharpness is constant for all the weights on the gradient flow (GF) trajectory by definition, we can focus on the singular values (or weights) at the global minima. Consider $z = \mathrm{GF}(\sigma)$, the GF solution of σ. In Lemma 6, we proved that GF preserves unbalancedness, such that $\sigma_l^2 - \sigma_m^2 = z_l^2 - z_m^2$ for all layers. Hence, it is sufficient to show that $\sum_{i=1}^{\min\{2,L-1\}} \frac{\eta^2\pi(z)^2}{z_{L-i}^2 z_L^2} \leq 1+c$ in order to ensure $\sum_{i=1}^{\min\{2,L-1\}} \frac{\eta^2\pi^2(\sigma)}{\sigma_{L-i}^2\sigma_L^2} \leq 1+c$. Note that $\pi(z) = \sigma_\star$, since it lies on the global minima. Then,

$$\sum_{i=1}^{\min\{2,L-1\}} \frac{\eta^2\pi^2(z)}{z_{L-i}^2 z_L^2} = \sum_{i=1}^{\min\{2,L-1\}} \frac{\eta^2\sigma_\star^2}{z_{L-i}^2 z_L^2}. \tag{38}$$

From Lemma 7, we know that the sharpness at the global minima is given as

$$\psi(\sigma) = \left\| \nabla^2 L(z) \right\| = \sum_{i=1}^{L} \frac{\sigma_\star^2}{z_i^2}. \tag{39}$$

This immediately implies that $\frac{\sigma_\star^2}{z_L^2} \leq \psi(\sigma)$ and equivalently, $\exists \alpha \in [0,1]$ such that $\frac{\sigma_\star^2}{z_L^2} = \alpha\psi(\sigma)$. Therefore, we have

$$\sum_{i=1}^{\min\{2,L-1\}} \frac{\sigma_\star^2}{z_{L-i}^2} \leq (1-\alpha)\psi(\sigma). \tag{40}$$

Substituting Equations (39) and (40) into the expression we aim to bound, we obtain

$$\sum_{i=1}^{\min\{2,L-1\}} \frac{\eta^2(\pi(\sigma)-\sigma_\star^2)^2\pi^2(\sigma)}{\sigma_{L-i}^2\sigma_L^2} = \sum_{i=1}^{\min\{2,L-1\}} \frac{\eta^2\sigma_\star^2}{z_{L-i}^2 z_L^2} \leq \eta^2\alpha(1-\alpha)\psi^2(\sigma) \leq \frac{\eta^2}{4}\psi^2(\sigma) \leq 1+c,$$

[1] We ignore the case $\pi^{(t)} = \sigma_\star$ when we get $b_{i,j}^{(t+1)} = b_{i,j}^{(t)}$. Since the occurence $\pi^{(t)} = \sigma_\star$ holds with a probability of zero where EOS ceases to exist.

where we used the fact that the maximum of $\alpha(1-\alpha)$ is $\frac{1}{4}$ when $\alpha = \frac{1}{2}$ and $\psi(\sigma) \leq \frac{2\sqrt{1+c}}{\eta}$. Thus, if $\psi(\sigma) \leq \frac{2\sqrt{1+c}}{\eta}$, then for every weight σ lying on its GF trajectory, we have

$$\sum_{i=1}^{\min\{2,L-1\}} \frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-i}^2\sigma_L^2} \leq 1+c.$$

Case 2: Consider the case in which $\pi(\sigma) > \sigma_\star$. We already have that $\sigma > 0$ throughout the trajectory (refer to Lemma 3.11 in [13]) and so $\pi(\sigma) > 0$. So, the GD update from σ_i will also stay positive

$$\sigma_i - \eta(\pi(\sigma)-\sigma_\star)\pi(\sigma)\frac{1}{\sigma_i} > 0.$$

From this, we get

$$2 > \frac{\eta(\pi(\sigma)-\sigma_\star)\pi(\sigma)}{\sigma_i^2} > 0,$$

This implies $\sum_{i=1}^{\min\{2,L-1\}} \frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-i}^2\sigma_L^2} \leq (1+c)$ with $c=1$. This completes the proof. □

Lemma 4. *If* $\sum_{i=1}^{\min\{2,L-1\}} \frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-i}^2\sigma_L^2} \leq 1+c$ *for* $i,j \in [L]$ *for some* $0 < c \leq 1$, *then* $\left|b_{i,j}^{(t+1)}\right| < c\left|b_{i,j}^{(t)}\right|$.

Proof. Recall that the condition for balancing was given by

$$b_{i,j}^{(t+1)} = b_{i,j}^{(t)}\left(1 - \eta^2(\pi^{(t)}-\sigma_\star)^2\frac{\pi^{(t)2}}{\left(\sigma_i^{(t)}\right)^2\left(\sigma_j^{(t)}\right)^2}\right). \tag{41}$$

WLOG, suppose that the σ are sorted such that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_L$. We know that

$$\sum_{i=1}^{\min\{2,L-1\}} \frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-i}^2\sigma_L^2} \leq 1+c,$$

which implies

$$\frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-1}^2\sigma_L^2} < 1+c \quad \text{and} \quad \frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_i^2\sigma_j^2} < \frac{1+c}{2}, \tag{42}$$

for all $i \in [L]$, $j \in [L-2]$ and $i < j$. Notice that the latter inequality comes from the fact that

$$\begin{aligned}\frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-2}^2\sigma_L^2} + \frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-2}^2\sigma_L^2} &< \frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-1}^2\sigma_L^2} + \frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-2}^2\sigma_L^2} \\ &< 1+c,\end{aligned}$$

which implies that

$$2\frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-2}^2\sigma_L^2} < 1+c \implies \frac{\eta^2(\pi(\sigma)-\sigma_\star)^2\pi^2(\sigma)}{\sigma_{L-2}^2\sigma_L^2} < \frac{1+c}{2},$$

and since σ are sorted, it holds for all other σ. Therefore from Equation (41), we have for all $i \in [L-2]$,

$$b_{i,i+1}^{(t+1)} < cb_{i,i+1}^{(t)} \quad \text{and} \quad b_{L-2,L}^{(t+1)} < cb_{L-2,L}^{(t)} \quad \text{and} \quad -cb_{L-1,L}^{(t)} < b_{L-1,L}^{(t+1)} < cb_{L-1,L}^{(t)}. \tag{43}$$

Then, notice that since we initialized all of the singular values σ_ℓ for $\ell \in [L-1]$ to be the same, they follow the same dynamics. Since we already showed that $|b_{L-1,L}^{(t+1)}| < c|b_{L-1,L}^{(t)}|$, it must follow that

$$\left|b_{i,j}^{(t+1)}\right| < c\left|b_{i,j}^{(t)}\right| \quad \text{for } i,j \in [L].$$

This completes the proof.

□

Lemma 5. *Consider running GD with learning rate η in Equation (2) on the scalar loss*

$$\mathcal{L}(\{\sigma_i\}_{i=1}^d) = \frac{1}{2}\left(\prod_{i=1}^{L}\sigma_i - \sigma_\star\right)^2,$$

with initialization $\sigma_L(0) = 0$ and $\sigma_\ell(0) = \alpha$ for all $\ell \in [L-1]$. If $\alpha < \left(\ln\left(\frac{2\sqrt{2}}{\eta L \sigma_\star^{2-\frac{2}{L}}}\right)\cdot\frac{\sigma_\star^{\frac{4}{L}}}{L^2\cdot 2^{\frac{2L-3}{L}}}\right)^{\frac{1}{4}}$, then the GFS sharpness $\psi(\sigma) \le \frac{2\sqrt{1+c}}{\eta}$ for some $0 < c < 1$.

Proof. Since the singular values σ_ℓ for all $\ell \in [L-1]$ are initialized to α, note that they all follow the same dynamics. Then, let us define

$$y \coloneqq \sigma_1 = \ldots = \sigma_{L-1} \quad \text{and} \quad x \coloneqq \sigma_L.$$

The gradient flow (GF) solution is the intersection between

$$xy^{L-1} = \sigma_\star \quad \text{and} \quad x^2 - y^2 = -\alpha^2,$$

where the first condition comes from convergence and the second comes from the conservation flow law of GF which we prove in Lemma 6. Then, if we can find a solution at the intersection such that

$$(\hat{x}(\alpha), \hat{y}(\alpha)) = \begin{cases} xy^{L-1} = \sigma_\star \\ x^2 - y^2 = -\alpha^2, \end{cases} \tag{44}$$

solely in terms of α, then we can plug in $(\hat{x}(\alpha), \hat{y}(\alpha))$ into the GFS[2] from Lemma 7

$$\psi(\hat{x}(\alpha), \hat{y}(\alpha)) = \psi(\sigma) = \sum_{i=1}^{L}\frac{\sigma_\star^2}{\sigma_i^2} = \sigma_\star^2\left(\frac{1}{\hat{x}(\alpha)^2} + \frac{L-1}{\hat{y}(\alpha)^2}\right) < \frac{2\sqrt{2}}{\eta}$$

and solve to find an upper bound on α. The strict inequality ensures that we can find a c in $c \in [0, 1)$ such that $\psi(\alpha) < \frac{2\sqrt{1+c}}{\eta}$. However, the intersection $(\hat{x}(\alpha), \hat{y}(\alpha))$ is a $2L$-th order polynomial in $\hat{y}(\alpha)$ which does not have a straightforward closed-form solution solely in terms of α. Hence, we aim to find the upper bound on α by using a calculus of variations. By plugging in x, the solution $\hat{y}(\alpha)$ satisfies

$$y^{2L} - \alpha^2 y^{2L-2} = \sigma_\star^2.$$

Then, by differentiating the relation with respect to α, we obtain the following variational relation:

$$\begin{aligned} &2Ly^{2L-1}dy - \alpha^2 2(L-1)y^{2L-3}dy - 2\alpha y^{2L-2}d\alpha = 0 \\ &\implies y^{2L-3}(y^2L - \alpha^2(L-1))dy = \alpha y^{2(L-1)}d\alpha \\ &\implies dy = \frac{y\alpha}{(y^2L - \alpha^2(L-1))}d\alpha, \end{aligned} \tag{45}$$

where we used the fact that $y^{2L-2} > 0$ from Lemma 3 in the last line. Then, in order to have $\frac{dy}{d\alpha} > 0$, we need $y > \sqrt{\frac{L-1}{L}}\alpha$, which is always true since $y > \alpha$ from initialization. Then, since $\alpha \to 0$, $\lim_{\alpha\to 0}\hat{y}(\alpha) = \sigma_\star^{\frac{1}{L}}$ and $\lim_{\alpha\to 0}\hat{x}(\alpha) = \sigma_\star^{\frac{1}{L}}$, as it corresponds to exact balancing. Hence, $\frac{dy}{d\alpha} > 0$ implies as α increases from 0, $\hat{y}(\alpha)$ would increase from $\sigma_\star^{\frac{1}{L}}$ and $\hat{y}(\alpha)$ is an increasing function of α. Similarly, the intersection at the global minima would satisfy the following relation for $\hat{x}(\alpha)$:

$$\begin{aligned} &x^{(2+\frac{2}{L-1})} + x^{\frac{2}{(L-1)}}\alpha^2 = \sigma_\star^{\frac{2}{L-1}} \\ &\implies \left(2 + \frac{2}{L-1}\right)x^{1+\frac{2}{L-1}}dx + \left(\frac{2}{L-1}\right)\alpha^2 x^{\frac{2}{L-1}-1}dx + x^{\frac{2}{L-1}}(2\alpha d\alpha) = 0 \\ &\implies dx = \frac{-\alpha}{\left(\frac{L}{L-1}x + \frac{\alpha^2}{L-1}\frac{1}{x}\right)}d\alpha. \end{aligned} \tag{46}$$

[2]Note that throughout the proof $(\hat{x}(\alpha), \hat{y}(\alpha))$ denotes the gradient flow solution as function of α. It does not refer to the GF trajectory.

Note that since $x > 0$, we will always have $\frac{dx}{d\alpha} < 0$. Then, since $\lim_{\alpha \to 0} \hat{x}(\alpha) = \sigma_\star^{\frac{1}{L}}$, $\frac{dx}{d\alpha} < 0$ implies that as α increases, $\hat{x}(\alpha)$ would decrease from $\sigma_\star^{\frac{1}{L}}$. Now, with the variational relations $\frac{d\hat{x}}{d\alpha}$ and $\frac{d\hat{y}}{d\alpha}$ in place, we aim to find $\frac{d\psi}{d\alpha}$:

$$
\begin{aligned}
&\Psi(\alpha) \coloneqq \psi(\hat{x}(\alpha), \hat{y}(\alpha)) = \sigma_\star^2 \left(\frac{1}{\hat{x}(\alpha)^2} + \frac{L-1}{\hat{y}(\alpha)^2} \right) \\
&\implies d\Psi = \sigma_\star^2 \left(-\frac{2}{\hat{x}^3} d\hat{x} - \frac{2(L-1)}{\hat{y}^3} d\hat{y} \right) \\
&\implies d\Psi = \frac{1}{\hat{x}^3} \left[\frac{2\alpha\sigma_\star^2}{\left(\frac{L}{L-1}\right)\hat{x} + \left(\frac{\alpha^2}{L-1}\right)\frac{1}{\hat{x}}} \right] d\alpha - \left[\frac{(L-1)}{\hat{y}^3} \frac{2\alpha\hat{y}\sigma_\star^2}{(\hat{y}^2 L - \alpha^2(L-1))} \right] d\alpha \\
&\implies d\Psi = \left[\frac{1}{\hat{x}^4 + \frac{\alpha^2}{L}\hat{x}^2} - \frac{1}{(\hat{y}^4 - \alpha^2 \hat{y}^2 \frac{(L-1)}{L})} \right] 2\frac{(L-1)\sigma_\star^2}{L} \alpha d\alpha \\
&\implies d\Psi = G(\alpha) d\alpha,
\end{aligned}
$$

where we defined $G(\alpha) \coloneqq \left[\frac{1}{\hat{x}^4 + \frac{\alpha^2}{L}\hat{x}^2} - \frac{1}{(\hat{y}^4 - \alpha^2 \hat{y}^2 \frac{(L-1)}{L})} \right] 2\frac{(L-1)\sigma_\star^2}{L} \alpha$ and used the notation $\hat{x} = \hat{x}(\alpha)$ and $\hat{y} = \hat{y}(\alpha)$ for simplicity.

Next, we will show the three following steps:

(i) Prove that $G(\alpha) > 0$ for all $\alpha > 0$ to show that the sharpness $\Psi(\alpha)$ is an increasing function of α.

(ii) Solve the differential $d\Psi$ to find the relationship between $d\Psi$ and $\Psi(\alpha)$.

(iii) Find an upper bound on a part of $\frac{d\Psi}{\Psi(\alpha)}$ found in Step 2.

These series of steps comes from the fact that the intersection does not have a closed-form solution. The goal is to find a function in which we can upper bound $\frac{d\Psi}{\Psi(\alpha)}$ with a function with a closed-form solution to find a bound on α such that the sharpness $\psi(\alpha) < \frac{2\sqrt{2}}{\eta}$.

Step 1: Prove $G(\alpha) > 0$ to show sharpness $\Psi(\alpha)$ is an increasing function of α.

There have been several lines of work such as those by [13] and [61] which showed that GD would decrease the sharpness of the solution. The more balanced the solution (which corresponds to smaller α), the smaller the sharpness. We prove this again here:

$$
\begin{aligned}
G(\alpha) > 0 &\implies \hat{y}^4 - \alpha^2 \hat{y}^2 \frac{(L-1)}{L} > \hat{x}^4 + \frac{\alpha^2}{L}\hat{x}^2 \\
&\implies (\hat{y}^4 - \hat{x}^4) > \alpha^2 \left(\frac{1}{L}\hat{x}^2 + \hat{y}^2 \frac{(L-1)}{L} \right) \\
&\implies \underbrace{(\hat{y}^2 - \hat{x}^2)}_{=\alpha^2} (\hat{y}^2 + \hat{x}^2) > \alpha^2 \left(\frac{1}{L}\hat{x}^2 + \hat{y}^2 \frac{(L-1)}{L} \right) \\
&\implies \hat{x}^2 (1 - \frac{1}{L}) + \hat{y}^2 \frac{1}{L} > 0,
\end{aligned}
$$

where the last inequality always holds since we have $L > 2$. This proves that Ψ is an increasing function of α since for $d\Psi = G(\alpha) d\alpha$, as it always holds that $G(\alpha) > 0$ for any $L > 2$ and $\alpha > 0$.

Step 2: Solve the differential to establish the relation between $\Psi(\alpha)$ and α.

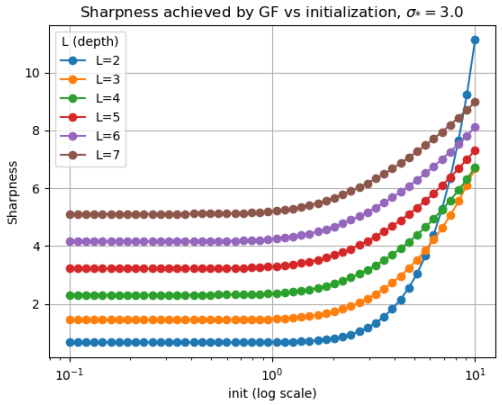


Figure 25: Sharpness $\Psi(\alpha)$ as a function of initialization α. The theoretical approximation bound $\Psi = \Psi_0 \exp\left(\frac{L^2 \cdot 2^{\frac{2(L-1)}{L}}}{2\sigma_\star^{\frac{4}{L}}}\alpha^4\right)$ serves as proxy upper bound to this increasing function.

Rewriting the expression for sharpness and establishing an equation we have

$$\Psi(\alpha) = \sigma_\star^2\left(\frac{1}{\hat{x}(\alpha)^2} + \frac{L-1}{\hat{y}(\alpha)^2}\right) \implies \Psi(\alpha) = \sigma_\star^2\left(\frac{\hat{y}^2 + (L-1)\hat{x}^2}{\hat{x}^2\hat{y}^2}\right) \tag{47}$$

$$\implies \frac{\hat{y}^2}{L} + \left(1 - \frac{1}{L}\right)\hat{x}^2 = \frac{\Psi(\alpha)\hat{x}^2\hat{y}^2}{L\sigma_\star^2}. \tag{48}$$

Now, we revisit the original differential between $\Psi(\alpha)$ and α:

$$d\Psi = \left[\frac{1}{(\hat{x}^4 + \frac{\alpha^2}{L}\hat{x}^2)} - \frac{1}{(\hat{y}^4 - \alpha^2\hat{y}^2\frac{(L-1)}{L})}\right] 2\frac{(L-1)\sigma_\star^2}{L}\alpha d\alpha$$

$$\implies d\Psi = \frac{\hat{y}^4 - \hat{x}^4 - \alpha^2(\frac{\hat{x}^2}{L} + (1 - \frac{1}{L}\hat{y}^2))}{(\hat{x}^4 + \frac{\alpha^2}{L}\hat{x}^2)(\hat{y}^4 - \alpha^2\hat{y}^2\frac{(L-1)}{L})} 2\left(1 - \frac{1}{L}\right)\sigma_\star^2\alpha d\alpha$$

$$\implies d\Psi = \frac{\alpha^2\left(\frac{\hat{y}^2}{L} + (1 - \frac{1}{L})\hat{x}^2\right)}{(\hat{x}^4 + \frac{\alpha^2}{L}\hat{x}^2)(\hat{y}^4 - \alpha^2\hat{y}^2\frac{(L-1)}{L})} 2\left(1 - \frac{1}{L}\right)\sigma_\star^2\alpha d\alpha \tag{49}$$

Using the expression for $\frac{\hat{y}^2}{L} + (1 - \frac{1}{L})\hat{x}^2$ derived in Equation (48) and plugging it into Equation (49), we obtain

$$d\Psi = \frac{\alpha^2\left(\frac{\Psi(\alpha)\hat{x}^2\hat{y}^2}{L\sigma_\star^2}\right)}{(\hat{x}^4 + \frac{\alpha^2}{L}\hat{x}^2)(\hat{y}^4 - \alpha^2\hat{y}^2\frac{(L-1)}{L})} 2\left(1 - \frac{1}{L}\right)\sigma_\star^2\alpha\, d\alpha$$

$$\implies \frac{d\Psi}{\Psi(\alpha)} = \frac{2}{(\hat{x}^2 + \frac{\alpha^2}{L})(\hat{y}^2 - \alpha^2\frac{(L-1)}{L})}\left(\frac{1}{L} - \frac{1}{L^2}\right)\alpha^3 d\alpha$$

$$\implies \frac{d\Psi}{\Psi(\alpha)} = \frac{2}{(\hat{x}^2 + \frac{\alpha^2}{L})^2}\left(\frac{1}{L} - \frac{1}{L^2}\right)\alpha^3 d\alpha \tag{50}$$

$$\implies \frac{d\Psi}{\Psi(\alpha)} = P(\alpha)\alpha^3 d\alpha, \tag{51}$$

$$\tag{52}$$

where we have defined $P(\alpha) \coloneqq \frac{2}{(\hat{x}^2 + \frac{\alpha^2}{L})^2}\left(\frac{1}{L} - \frac{1}{L^2}\right)$.

Solving the differential $\frac{d\Psi}{\Psi(\alpha)} = P(\alpha)\alpha^3 d\alpha$ in exact closed-form is difficult since $\hat{x}$ is also an function of α. However, in Step 1, we proved that $\Psi(\alpha)$ is an increasing function of α, and so instead of

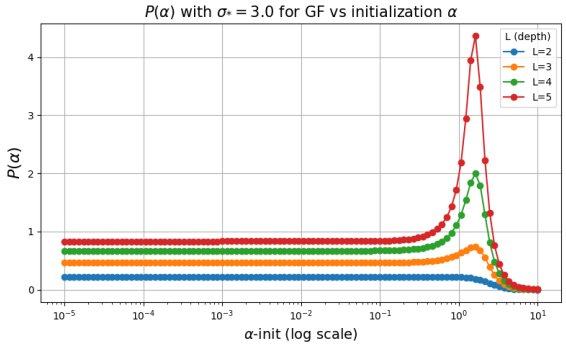


Figure 26: $P(\alpha)$ for GF has a unique maxima at $\alpha = \frac{\sigma_*^{\frac{1}{L}}}{\left(\frac{1}{\sqrt{L(L-2)}}\left(1+\frac{1}{L(L-2)}\right)^{\frac{L-1}{2}}\right)^{\frac{1}{L}}}$ for $L > 2$.

solving exactly, we can find a differential equation $\frac{d\Psi}{\Psi(\alpha)} = F(\alpha)\alpha^3 d\alpha$ with $F(\alpha) > P(\alpha)$ such that $F(\alpha)$ is more increasing, and use it to solve the PDE instead. Though, note that the initialization limit on α that would be found after solving the surrogate PDE $\frac{d\Psi}{\Psi(\alpha)} = F(\alpha)\alpha^3 d\alpha$ would be smaller than the α if it was found using the original PDE $\frac{d\Psi}{\Psi(\alpha)} = P(\alpha)\alpha^3 d\alpha$.

Step 3: Finding an upper bound function and solving for initialization.
Note that the original coefficient in $\frac{d\Psi}{\Psi(\alpha)} = P(\alpha)\alpha^3 d\alpha$ is of the form

$$P(\alpha) = \frac{2}{(\hat{x}^2 + \frac{\alpha^2}{L})^2}\left(\frac{1}{L} - \frac{1}{L^2}\right) \tag{53}$$

Let us consider the two corner cases for α. We showed before that $\lim_{\alpha\to 0} \hat{x}(\alpha) = \sigma_\star^{\frac{1}{L}}$, so

$$\lim_{\alpha\to 0} P(\alpha) = \frac{2\left(\frac{1}{L} - \frac{1}{L^2}\right)}{\sigma_\star^{\frac{4}{L}}}$$

As $\alpha \to \infty$, we have $\lim_{\alpha\to\infty} P(\alpha) \to 0$ since $\lim_{\alpha\to\infty} \hat{x} = 0$. Furthermore, we have

$$\begin{aligned}
P'(\alpha) &= \frac{-4}{(\hat{x}^2 + \frac{\alpha^2}{L})^3}\left(\frac{1}{L} - \frac{1}{L^2}\right)(2\hat{x}\frac{d\hat{x}}{d\alpha} + \frac{2\alpha}{L}) \\
&= \frac{-4\left(\frac{1}{L} - \frac{1}{L^2}\right)}{(\hat{x}^2 + \frac{\alpha^2}{L})^3}\left(2\hat{x}\left(\frac{-\alpha}{\left(\frac{L}{L-1}\hat{x} + \frac{\alpha^2}{L-1}\frac{1}{\hat{x}}\right)}\right) + \frac{2\alpha}{L}\right) \\
&= \frac{8\alpha\left(\frac{1}{L} - \frac{1}{L^2}\right)}{(\hat{x}^2 + \frac{\alpha^2}{L})^3}\left(\frac{L-1}{L + \frac{\alpha^2}{\hat{x}^2}} - \frac{1}{L}\right)
\end{aligned}$$

For $L = 2$, we always have $P'(\alpha) < 0$. Hence, choosing $F(\alpha) = \lim_{\alpha\to 0} P(\alpha) = \frac{2\left(\frac{1}{L} - \frac{1}{L^2}\right)}{\sigma_\star^{\frac{4}{L}}}$, will serve as the correct upper bound.

Then, let us consider $L > 2$. We can see that α at which $\hat{x}(\alpha) = \frac{\alpha}{\sqrt{L(L-2)}}$ is the critical point of $P(\alpha)$. Further, we note that when $\hat{x}(\alpha) < \frac{\alpha}{\sqrt{L(L-2)}}$, $P'(\alpha) < 0$ meaning $P(\alpha)$ is decreasing. Since $\hat{x}(\alpha)$ is itself decreasing in α, this states there for any $\alpha > \alpha_{crit}$, $P(\alpha)$ is decreasing. α_{crit} is the solution of $\hat{x}(\alpha) = \frac{\alpha}{\sqrt{L(L-2)}}$.

For any $\alpha < \alpha_{crit}$, $P'(\alpha) > 0$, so $P(\alpha)$ is increasing. So, $P(\alpha_{crit})$ corresponds to the maximum of P in α. Choosing $F(\alpha) = P(\alpha_{crit})$, a constant allows us to find an upper bound function for the

function in Equation (53). Furthermore, note that since $\frac{d\hat{x}}{d\alpha} < 0$ and $\hat{x} > 0$, $\alpha > 0$, there must be only one critical point of $P(\alpha)$ which is at α_{crit}.

Hence, we have $\hat{x}(\alpha_{crit}) = \frac{\alpha_{crit}}{\sqrt{L(L-2)}}$. From Equation (44), we also get $\hat{y} = \alpha_{crit}\sqrt{1 + \frac{1}{L(L-2)}}$ and

$$\left(\frac{\alpha_{\text{crit}}}{\sqrt{L(L-2)}}\right)\left(\alpha_{\text{crit}}\sqrt{1+\frac{1}{L(L-2)}}\right)^{L-1} = \sigma_*$$
$$\implies \alpha_{\text{crit}} = \frac{\sigma_*^{\frac{1}{L}}}{\left(\frac{1}{\sqrt{L(L-2)}}\left(1+\frac{1}{L(L-2)}\right)^{\frac{L-1}{2}}\right)^{\frac{1}{L}}}$$

Using α_{crit}, we obtain the maximum of $P(\alpha)$ to be

$$\begin{aligned} P(\alpha_{crit}) &= \frac{2}{(\hat{x}(\alpha_{crit})^2 + \frac{\alpha_{crit}^2}{L})^2}\left(\frac{1}{L} - \frac{1}{L^2}\right) \\ &= \frac{2}{\left(\alpha_{crit}^2(\frac{1}{L(L-2)} + \frac{1}{L})\right)^2}\left(\frac{1}{L} - \frac{1}{L^2}\right) \\ &= \frac{2}{\sigma_*^{\frac{4}{L}}} g(L) \end{aligned}$$

where $g(L) = \frac{(\frac{1}{L} - \frac{1}{L^2})[\frac{1}{\sqrt{L(L-2)}}\left(1+\frac{1}{L(L-2)}\right)^{\frac{L-1}{2}}]^{\frac{4}{L}}}{(\frac{1}{L(L-2)}+\frac{1}{L})^2}$.

Now, chosing $F(\alpha) = P(\alpha_{crit})$, we integrate the upper bound function as

$$\begin{aligned} \int \frac{d\Psi}{\Psi} &= \frac{2g(L)}{\sigma_\star^{\frac{4}{L}}} \int \alpha^3 \, d\alpha \\ \implies \ln(\frac{\Psi}{\Psi_0}) &= \frac{g(L)}{2\sigma_\star^{\frac{4}{L}}}(\alpha^4) \\ \implies \Psi &= \Psi_0 \exp\left(\frac{g(L)}{2\sigma_\star^{\frac{4}{L}}}\alpha^4\right) \end{aligned}$$

where $\Psi_0 = \lim_{\alpha\to 0}\Psi = L\sigma_\star^{2-\frac{2}{L}}$. We verify this upper bound empirically from Figure 25, where we see a near exponential growth in sharpness as function of α.

Now, note that the function $\Psi = \Psi_0 \exp\left(\frac{g(L)}{2\sigma_\star^{\frac{4}{L}}}\alpha^4\right)$ acts an upper bound to the original sharpness function of α and both are increasing in α (Step 1). So, solving for an initialization α-upper limit with $\Psi = \frac{2\sqrt{2}}{\eta}$ would mean that the original sharpness with this initialization would be less than $\frac{2\sqrt{1+c}}{\eta}$ for some $0 < c < 1$. Hence, α is restricted to

$$\alpha < \left(\ln\left(\frac{\frac{2\sqrt{2}}{\eta}}{L\sigma_\star^{2-\frac{2}{L}}}\right)\cdot\frac{2\sigma_\star^{\frac{4}{L}}}{g(L)}\right)^{\frac{1}{4}}.$$

We can simplify the bound further by finding an upper bound on $g(L)$ [3]:

$$g(L) \leq \frac{\left(\left(1+\frac{1}{L(L-2)}\right)^{\frac{L-1}{2}}\right)^{\frac{4}{L}}}{\left(\frac{1}{L(L-2)}+\frac{1}{L}\right)^2} \leq L^2 \cdot \left(\left(1+\frac{1}{L(L-2)}\right)^{\frac{L-1}{2}}\right)^{\frac{4}{L}} \leq L^2 \cdot 2^{\frac{2(L-1)}{L}}.$$

Then, we get obtain a lower bound on α:

$$\alpha < \left(\ln\left(\frac{2\sqrt{2}}{\eta L \sigma_\star^{2-\frac{2}{L}}}\right) \cdot \frac{2\sigma_\star^{\frac{4}{L}}}{L^2 \cdot 2^{\frac{2(L-1)}{L}}}\right)^{\frac{1}{4}} = \left(\ln\left(\frac{2\sqrt{2}}{\eta L \sigma_\star^{2-\frac{2}{L}}}\right) \cdot \frac{\sigma_\star^{\frac{4}{L}}}{L^2 \cdot 2^{\frac{2L-3}{L}}}\right)^{\frac{1}{4}}$$

Hence, as long as α satisfies this upper bound, we will have balancing. This completes the proof.

□

Lemma 6. *Consider the minimizing the loss*

$$\mathcal{L}\left(\{\sigma_\ell\}_{\ell=1}^L\right) = \frac{1}{2}\left(\prod_{\ell=1}^L \sigma_\ell - \sigma_\star\right)^2,$$

using gradient flow. Then, the balancedness between two singular values defined by $\sigma_\ell^2(t) - \sigma_m^2(t)$ for all $m, \ell \in [L]$ is constant for all t.

Proof. Notice that the result holds specifically for gradient flow and not descent. The dynamics of each scalar factor for gradient flow can be written as

$$\dot{\sigma}_\ell(t) = -\left(\prod_{\ell=1}^L \sigma_\ell(t) - \sigma_\star\right) \cdot \prod_{i\neq\ell}^L \sigma_i(t)$$

Then, the time derivative of balancing is given as

$$\begin{aligned}
&\frac{\partial}{\partial t}(\sigma_\ell^2(t) - \sigma_m^2(t)) = \sigma_\ell(t)\dot{\sigma}_\ell(t) - \sigma_m(t)\dot{\sigma}_m(t) \\
&= -\sigma_\ell(t)\left(\prod_{\ell=1}^L \sigma_\ell(t) - \sigma_\star\right) \cdot \prod_{i\neq\ell}^L \sigma_i(t) + \sigma_m(t)\left(\prod_{m=1}^L \sigma_\ell(t) - \sigma_\star\right) \cdot \prod_{j\neq m}^L \sigma_j(t). \\
&= 0.
\end{aligned}$$

Hence, the quantity $\sigma_\ell^2(t) - \sigma_m^2(t)$ remains constant for all time t, hence preserving unbalancedness.

□

Lemma 7. *Consider the scalar loss*

$$\mathcal{L}(\{\sigma_i\}_{i=1}^d) = \frac{1}{2}\left(\prod_{i=1}^L \sigma_i - \sigma_\star\right)^2,$$

The sharpness at the global minima is given as $\|\nabla^2\mathcal{L}\|_2 = \sum_{i=1}^L \frac{\sigma_\star^2}{\sigma_i^2}$.

Proof. The gradient is given by

$$\nabla_{\sigma_i}\mathcal{L} = \left(\prod_{\ell=1}^L \sigma_\ell(t) - \sigma_\star\right)\prod_{j\neq i}^L \sigma_j(t).$$

[3] This includes the case for $L = 2$ since $(\frac{1}{L} - \frac{1}{L^2}) < L^2 \cdot 2^{\frac{2(L-1)}{L}}$ for $L = 2$.

Then,

$$\nabla_{\sigma_j}\nabla_{\sigma_i}\mathcal{L} = \prod_{\ell\neq i}^{L}\sigma_\ell(t)\prod_{\ell\neq j}^{L}\sigma_\ell(t) + \left(\prod_{\ell=1}^{L}\sigma_\ell(t) - \sigma_\star\right)\prod_{\ell\neq j,\ell\neq i}^{L}\sigma_\ell(t)$$

Let $\pi(t) = \prod_{i=1}^{L}\sigma_i(t)$. Then, at the global minima, we have

$$\nabla_{\sigma_j}\nabla_{\sigma_i}\mathcal{L} = \frac{\pi^2}{\sigma_i\sigma_j} = \frac{\sigma_\star^2}{\sigma_i\sigma_j}$$

Thus, the sharpness of the largest eigenvalue is given as $\|\nabla^2\mathcal{L}\|_2 = \sum_{i=1}^{L}\frac{\sigma_\star^2}{\sigma_i^2}$. □

Lemma 8. *Consider the loss*

$$\mathcal{L}\left(\{\sigma_i\}_{i=1}^{L}\right) = \frac{1}{2}\left(\prod_{i=1}^{L}\sigma_i - \sigma_\star\right)^2.$$

The balanced minimum $\sigma_i = \sigma_\star^{1/L}$ has the smallest sharpness amongst all global minima with a value of $\|\nabla^2\mathcal{L}\|_2 = L\sigma_\star^{2-2/L}$.

Proof. By Lemma 7, recall that the sharpness at the global minima is given in the form

$$\|\nabla^2\mathcal{L}\|_2 = \sum_{i=1}^{L}\frac{\sigma_\star^2}{\sigma_i^2}.$$

To show that the balanced minimum is the flattest (i.e., it has the smallest sharpness amongst all global minima), we will show that KKT stationarity condition of the constrained objective

$$\min_{\{\sigma_i\}_{i=1}^{L}}\sum_{i=1}^{L}\frac{\sigma_\star^2}{\sigma_i^2} \quad \text{s.t. } \prod_{i=1}^{L}\sigma_i = \sigma_\star,$$

are only met at the balanced minimum, which gives us the sharpness value $\|\nabla^2\mathcal{L}\|_2 = L\sigma_\star^{2-2/L}$. The Lagrangian is given by

$$L(\sigma_1,\ldots,\sigma_L,\mu) = \sum_{i=1}^{L}\frac{\sigma_\star^2}{\sigma_i^2} + \mu\left(\prod_{i=1}^{L}\sigma_i - \sigma_\star\right).$$

Then, the stationary point conditions of the Langrangian is given by

$$\frac{\partial L}{\partial\sigma_i} = -\frac{2\sigma_\star^2}{\sigma_i^3} + \mu\prod_{j\neq i}\sigma_j = 0, \tag{54}$$

$$\frac{\partial L}{\partial\mu} = \prod_{i=1}^{L}\sigma_i - \sigma_\star = 0. \tag{55}$$

From Equation (54), the solution of the stationary point gives

$$\frac{2\sigma_\star^2}{\sigma_i^3} = \mu\prod_{j\neq i}\sigma_j \implies \mu = \frac{2\sigma_\star^2}{\sigma_i^3\prod_{j\neq i}\sigma_j} = \frac{2\sigma_\star^2}{\sigma_i^2\sigma_\star} = \frac{2\sigma_\star}{\sigma_i^2}.$$

This also indicates that at the stationary point, $\sigma_i = \sqrt{\frac{2\sigma_\star}{\mu}}$ for all $i \in [L]$, which means that the condition is *only* satisfied at the balanced minimum, i.e, $\sigma_i = \sigma_\star^{1/L}$. Furthermore, notice that

$$\nabla^2 f(\sigma_i) = 6\sigma_\star^2\cdot\text{Diag}\left(\frac{1}{\sigma_i^4}\right) \succ \mathbf{0},$$

where $f(\sigma_i) = \sum_{i=1}^{L}\frac{\sigma_\star^2}{\sigma_i^2}$, indicating that f only has a minimum. Notice that Equation (55) holds immediately. Thus, the balanced minimum has the smallest shaprness (flattest), which plugging into f gives a sharpness of $\|\nabla^2\mathcal{L}\|_2 = L\sigma_\star^{2-2/L}$.

□

Theorem 3 (Subspace Oscillation for Diagonal Linear Networks)**.** *Consider an L-layer diagonal linear network on the loss*

$$\mathcal{L}\left(\{\mathbf{s}_\ell\}_{\ell=1}^L\right) \coloneqq \frac{1}{2}\|\mathbf{s}_1 \odot \ldots \odot \mathbf{s}_L - \mathbf{s}_\star\|_2^2, \tag{56}$$

where $\mathbf{s}_\star \in \mathbb{R}^d$ be an r-sparse vector with ordered coordinates such that $s_{\star,1} > \ldots > s_{\star,d}$ and define $S_p \coloneqq Ls_{\star,p}^{2-\frac{2}{L}}$ and $\alpha' \coloneqq \left(\ln\left(\frac{2\sqrt{2}}{\eta L s_{\star,1}^{2-\frac{2}{L}}}\right)\cdot\frac{s_{\star,1}^{\frac{4}{L}}}{L^2\cdot 2^{\frac{2L-3}{L}}}\right)^{\frac{1}{4}}$. For any $p < r-1$ and $\alpha < \alpha'$, suppose we run GD on Equation (5) with learning rate $\eta = \frac{2}{K}$, where $S_p \geq K > S_{p+1}$ with initialization $\mathbf{s}_\ell = \alpha\mathbf{1}_d$ for all $\ell \in [L-1]$ and $\mathbf{s}_L = \mathbf{0}_d$. Then, under strict balancing, the top-p coordinates of $\mathbf{s}_\ell$ oscillate within a 2-period fixed orbit around the minima in the form

$$s_{\ell,i}(t) = \rho_{i,j}(t), \quad \forall i < p, \forall \ell \in [L],$$

where $\rho_{i,j}(t) \in \{\rho_{i,1}, \rho_{i,2}\}$, $\rho_{i,1} \in \left(0, s_{\star,i}^{1/L}\right)$ and $\rho_{i,2} \in \left(s_{\star,i}^{1/L}, (2s_{\star,i})^{1/L}\right)$ are two real roots of the polynomial $h(\rho) = 0$:

$$h(\rho) = \rho^L \cdot \frac{1 + \left(1 + \eta L(s_{\star,i} - \rho^L)\cdot\rho^{L-2}\right)^{2L-1}}{1 + \left(1 + \eta L(s_{\star,i} - \rho^L)\cdot\rho^{L-2}\right)^{L-1}} - s_{\star,i}.$$

Proof. This proof essentially mimics that of the DLN proof from Theorem 2, in that we will

(i) Compute the eigenvalues and eigenvectors of the flattened training loss Hessian of the diagonal linear network at convergence under the balancing assumption.

(ii) Show that in 1D cross-section of the eigenvector, the stable condition oscillation $3[f_{\Delta_i}^{(3)}]^2 - f_{\Delta_i}^{(2)} f_{\Delta_i}^{(4)} > 0$ is satisfied, where f_{Δ_i} denotes the 1D cross-section function at the i-th eigenvector direction.

With a slight abuse in notation, let $\mathbf{s} \coloneqq \{\mathbf{s}_\ell\}_{\ell=1}^L$. Let us first derive the Hessian at convergence by considering each block of the flattened Hessian matrix denoted by $\mathbf{H}_{m,\ell}$:

$$\mathbf{H}_{\ell,\ell} = \begin{bmatrix} \prod_{k\neq\ell} s_{k,1} & & \\ & \ddots & \\ & & \prod_{k\neq\ell} s_{k,d} \end{bmatrix} \qquad \text{if } m = \ell \tag{57}$$

$$\mathbf{H}_{m,\ell} = \begin{bmatrix} \gamma_1 & & \\ & \ddots & \\ & & \gamma_d \end{bmatrix} \qquad \text{if } m \neq \ell, \tag{58}$$

where

$$\gamma_i \coloneqq \left(\prod_{k\neq l} s_{k,i}\right)\left(\prod_{k\neq m} s_{k,i}\right) + \left(\prod_k s_{k,i} - s_{\star,i}\right)\left(\prod_{k\neq l, k\neq m} s_{k,i}\right).$$

Then, under Lemma 2, at convergence (i.e. the gradient is zero), we have

$$\left(\prod_k s_{k,i} - s_{\star,i}\right) = 0 \implies s_{k,i} = s_{\star,i}^{\frac{1}{L}},$$

which means that at convergence, the Hessian is given by

$$\mathbf{H} = \begin{bmatrix} \mathbf{A} & \ldots & \mathbf{A} & \mathbf{A} \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{A} & \ldots & \mathbf{A} & \mathbf{A} \\ \mathbf{A} & \ldots & \mathbf{A} & \mathbf{B} \end{bmatrix} \in \mathbb{R}^{dL\times dL},$$

where

$$\mathbf{A} \coloneqq \begin{bmatrix} s_{\star,1}^{2-\frac{2}{L}} & & & \\ & \ddots & & \\ & & s_{\star,r}^{2-\frac{2}{L}} & \\ & & & \mathbf{0}_{d-r} \end{bmatrix} \in \mathbb{R}^{d\times d}, \qquad \mathbf{B} \coloneqq \begin{bmatrix} s_{\star,1}^{2-\frac{2}{L}} & & & \\ & \ddots & & \\ & & s_{\star,r}^{2-\frac{2}{L}} & \\ & & & \alpha^{2(L-1)}\cdot \mathbf{I}_{d-r} \end{bmatrix} \in \mathbb{R}^{d\times d}.$$

To compute the eigenvalues of $\mathbf{H}$, we can block diagonalize $\mathbf{H}$ into the form $\mathbf{C} = \mathbf{P}\mathbf{H}\mathbf{P}^\top$, where $\mathbf{P}$ is a permutation matrix and

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 & & \\ & \ddots & \\ & & \mathbf{C}_d \end{bmatrix} \in \mathbb{R}^{dL\times dL},$$

where each (i,j)-th entry of $\mathbf{C}_k \in \mathbb{R}^{L\times L}$ is the k-th diagonal element of $\mathbf{H}_{i,j}$. Then, since $\mathbf{C}$ is a block diagonal matrix, its eigenvalues are the union of each of the eigenvalues of its blocks. Then, notice that

$$\mathbf{C}_j = s_{\star,k}^{2-\frac{2}{L}}\cdot \mathbf{1}_L\mathbf{1}_L^\top,\ \forall j\in[r] \qquad \mathbf{C}_k = \begin{bmatrix} 0 & \dots & 0 & \alpha^{2(L-1)} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & \alpha^{2(L-1)} \\ \alpha^{2(L-1)} & \dots & \alpha^{2(L-1)} & \alpha^{2(L-1)} \end{bmatrix},\ \forall k\in[r+1,d].$$

Hence, the eigenvalues of $\mathbf{C}$ (and the eigenvalues of $\mathbf{H}$) are given by

$$\lambda_{\mathbf{H}} = \left\{ Ls_{\star,i}^{2-\frac{2}{L}}, \underbrace{0}_{\text{multiplicity } L-1} \right\}_{i=1}^{r} \bigcup \left\{ \underbrace{\frac{-\alpha^{2(L-1)} \pm \sqrt{(4L-3)\cdot\alpha^{4(L-1)^2}}}{-2}}_{\text{multiplicity } d-r}, \underbrace{0}_{\text{multiplicity } (d-r)(L-2)} \right\},$$

which can be computed using co-factor expansion. For the eigenvectors, notice that we can write

$$\mathbf{C}\mathbf{v} = \mathbf{P}\mathbf{H}\mathbf{P}^\top\mathbf{v} = \lambda\mathbf{v} \implies \mathbf{H}\mathbf{P}^\top\mathbf{v} = \lambda\mathbf{P}^\top\mathbf{v}.$$

Hence, we can find the eigenvectors of the block diagonal matrix $\mathbf{C}$, and left multiply them by $\mathbf{P}^\top$ to obtain the eigenvectors of the Hessian $\mathbf{H}$. This yields the eigenvector and eigenvalue pairs

$$\Delta_1 = \mathbf{P}^\top \mathrm{vec}\left(\frac{1}{\sqrt{L}}\mathbf{1}_L, \mathbf{0}, \dots, \mathbf{0}\right), \qquad \lambda_1 = Ls_{\star,1}^{2-\frac{2}{L}} \tag{59}$$

$$\Delta_2 = \mathbf{P}^\top \mathrm{vec}\left(\mathbf{0}, \frac{1}{\sqrt{L}}\mathbf{1}_L, \dots, \mathbf{0}\right), \qquad \lambda_2 = Ls_{\star,2}^{2-\frac{2}{L}} \tag{60}$$

$$\vdots \qquad\qquad \vdots \tag{61}$$

$$\Delta_r = \mathbf{P}^\top \mathrm{vec}\left(\mathbf{0}, \dots, \frac{1}{\sqrt{L}}\mathbf{1}_L, \dots, \mathbf{0}\right), \quad \lambda_r = Ls_{\star,r}^{2-\frac{2}{L}} \tag{62}$$

$$\vdots \qquad\qquad \vdots \tag{63}$$

$$\Delta_{r+j} = \mathbf{P}^\top \mathrm{vec}\left(\mathbf{0}, \dots, \mathbf{e}_{r+j}, \dots, \mathbf{0}\right), \qquad \lambda_{r+j} = \frac{-\alpha^{2(L-1)} \pm \sqrt{(4L-3)\cdot\alpha^{4(L-1)^2}}}{-2}, \tag{64}$$

where $\mathbf{e}_i$ is an i-th elementary basis vector.

Then, in each 1-D eigenvector direction, we can analyze the loss and verify if it satisfies the stability condition. Notice that we can consider the scalar loss

$$\mathcal{L}_i(\mathbf{s}) = \frac{1}{2}(s_{1,i}\odot\dots\odot s_{L,i} - s_{\star,i})^2 = \frac{1}{2}(s_i^L - s_{\star,i})^2. \qquad \text{(By Lemma 2)}$$

Using Corollary 5 by [16] or restated Lemma 12 on the 1D scalar function, this 1D loss is amenable to stable oscillation when learning rate $\eta > \frac{2}{\lambda_i}$. Finally, to prove the uniqueness and existence of

two period orbit fixed point for $\eta > \frac{2}{\lambda_i}$, we show that the polynomial obtained by solving two step fixed point has a real root. This is the same loss we analyzed in Theorem 1, where we showed that the oscillations are real roots of the polynomial

$$\sigma_{\star,1} = \rho^L \frac{1 + z^{2L-1}}{1 + z^{L-1}}, \quad \text{where } z \coloneqq \left(1 + \eta L(\sigma_{\star,1} - \rho^L) \cdot \rho^{L-2}\right).$$

and $\rho_1 \in \left(0, \sigma_{\star,1}^{1/L}\right)$ and $\rho_2 \in \left(\sigma_{\star,1}^{1/L}, (2\sigma_{\star,1})^{1/L}\right)$ are the two real roots of the polynomial which exists and are unique. Hence, whenever the learning rate η lies between $[2/\lambda_p, 2/\lambda_{p+1}]$, we will have oscillations in all of the p eigenvector directions. This completes the proof.

□

C.2. Deferred Proofs for Singular Vector Invariance

Proposition 2. *Let $\mathbf{M}_\star = \mathbf{U}_\star \mathbf{\Sigma}_\star \mathbf{V}_\star^\top$ denote the SVD of the target matrix. The initialization in Equation (**??**) is a member of the singular vector stationary set in Proposition 1, where $\mathbf{Q}_L = \ldots = \mathbf{Q}_2 = \mathbf{V}_\star$.*

Proof. Recall that the initialization is given by

$$\mathbf{W}_L(0) = 0 \quad \text{and} \quad \mathbf{W}_\ell(0) = \alpha \mathbf{I}_d \quad \forall \ell \in [L-1].$$

We will show that under this initialization, each weight matrix admits the following decomposition for all $t \geq 1$:

$$\mathbf{W}_L(t) = \mathbf{U}_\star \begin{bmatrix} \widetilde{\mathbf{\Sigma}}_L(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}_\star^\top, \qquad \mathbf{W}_\ell(t) = \mathbf{V}_\star \begin{bmatrix} \widetilde{\mathbf{\Sigma}}(t) & \mathbf{0} \\ \mathbf{0} & \alpha \mathbf{I}_{d-r} \end{bmatrix} \mathbf{V}_\star^\top, \quad \forall \ell \in [L-1], \tag{65}$$

where

$$\widetilde{\mathbf{\Sigma}}_L(t) = \widetilde{\mathbf{\Sigma}}_L(t-1) - \eta \cdot \left(\widetilde{\mathbf{\Sigma}}_L(t-1) \cdot \widetilde{\mathbf{\Sigma}}^{L-1}(t-1) - \mathbf{\Sigma}_{\star,r} \right) \cdot \widetilde{\mathbf{\Sigma}}^{L-1}(t-1)$$
$$\widetilde{\mathbf{\Sigma}}(t) = \widetilde{\mathbf{\Sigma}}(t-1) \cdot \left(\mathbf{I}_r - \eta \cdot \widetilde{\mathbf{\Sigma}}_L(t-1) \cdot \left(\widetilde{\mathbf{\Sigma}}_L(t-1) \cdot \widetilde{\mathbf{\Sigma}}^{L-1}(t-1) - \mathbf{\Sigma}_{\star,r} \right) \cdot \widetilde{\mathbf{\Sigma}}^{L-3}(t-1) \right),$$

where $\widetilde{\mathbf{\Sigma}}_L(t), \widetilde{\mathbf{\Sigma}}(t) \in \mathbb{R}^{r \times r}$ is a diagonal matrix with $\widetilde{\mathbf{\Sigma}}_L(1) = \eta \alpha^{L-1} \cdot \mathbf{\Sigma}_{r,\star}$ and $\widetilde{\mathbf{\Sigma}}(1) = \alpha \mathbf{I}_r$.

This will prove that the singular vectors are stationary with $\mathbf{\Sigma}_L = \ldots = \mathbf{\Sigma}_2 = \mathbf{V}_\star$. We proceed with mathematical induction.

Base Case. For the base case, we will show that the decomposition holds for each weight matrix at $t = 1$. The gradient of $f(\mathbf{\Theta})$ with respect to $\mathbf{W}_\ell$ is

$$\nabla_{\mathbf{W}_\ell} f(\mathbf{\Theta}) = \mathbf{W}_{L:\ell+1}^\top \cdot (\mathbf{W}_{L:1} - \mathbf{M}_\star) \cdot \mathbf{W}_{\ell-1:1}^\top.$$

For $\mathbf{W}_L(1)$, we have

$$\begin{aligned} \mathbf{W}_L(1) &= \mathbf{W}_L(0) - \eta \cdot \nabla_{\mathbf{W}_L} f(\mathbf{\Theta}(0)) \\ &= \mathbf{W}_L(0) - \eta \cdot (\mathbf{W}_{L:1}(0) - \mathbf{M}_\star) \cdot \mathbf{W}_{L-1:1}^\top(0) \\ &= \eta \alpha^{L-1} \mathbf{\Sigma}_\star \\ &= \mathbf{U}_\star \cdot \left(\eta \alpha^{L-1} \cdot \mathbf{\Sigma}_\star \right) \cdot \mathbf{V}_\star^\top \\ &= \mathbf{U}_\star \begin{bmatrix} \widetilde{\mathbf{\Sigma}}_L(1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}_\star^\top. \end{aligned}$$

Then, for each $\mathbf{W}_\ell(1)$ in $\ell \in [L-1]$, we have

$$\begin{aligned} \mathbf{W}_\ell(1) &= \mathbf{W}_\ell(0) - \eta \cdot \nabla_{\mathbf{W}_\ell} f(\mathbf{\Theta}(0)) \\ &= \alpha \mathbf{I}_d, \end{aligned}$$

where the last equality follows from the fact that $\mathbf{W}_L(0) = \mathbf{0}$. Finally, we have

$$\mathbf{W}_\ell(1) = \alpha \mathbf{V}_\star \mathbf{V}_\star^\top = \mathbf{V}_\star \begin{bmatrix} \widetilde{\mathbf{\Sigma}}(1) & \mathbf{0} \\ \mathbf{0} & \alpha \mathbf{I}_{d-r} \end{bmatrix} \mathbf{V}_\star^\top, \quad \forall \ell \in [L-1].$$

Inductive Step. By the inductive hypothesis, suppose that the decomposition holds. Then, notice that we can simplify the end-to-end weight matrix to

$$\mathbf{W}_{L:1}(t) = \mathbf{U}_\star \begin{bmatrix} \widetilde{\mathbf{\Sigma}}_L(t) \cdot \widetilde{\mathbf{\Sigma}}^{L-1}(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}_\star^\top,$$

for which we can simplify the gradients to

$$\begin{aligned} \nabla_{\mathbf{W}_L} f(\mathbf{\Theta}(t)) &= \left(\mathbf{U}_\star \begin{bmatrix} \widetilde{\mathbf{\Sigma}}_L(t) \cdot \widetilde{\mathbf{\Sigma}}^{L-1}(t) - \mathbf{\Sigma}_{\star,r} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}_\star^\top \right) \cdot \mathbf{V}_\star \begin{bmatrix} \widetilde{\mathbf{\Sigma}}^{L-1}(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}_\star^\top \\ &= \mathbf{U}_\star \begin{bmatrix} \left(\widetilde{\mathbf{\Sigma}}_L(t) \cdot \widetilde{\mathbf{\Sigma}}^{L-1}(t) - \mathbf{\Sigma}_{\star,r} \right) \cdot \widetilde{\mathbf{\Sigma}}^{L-1}(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}_\star^\top, \end{aligned}$$

for the last layer matrix, and similarly,

$$\nabla_{\mathbf{W}_\ell} f(\mathbf{\Theta}(t)) = \mathbf{V}_\star \begin{bmatrix} \widetilde{\mathbf{\Sigma}}_L(t) \cdot \left(\widetilde{\mathbf{\Sigma}}_L(t) \cdot \widetilde{\mathbf{\Sigma}}^{L-1}(t) - \mathbf{\Sigma}_{\star,r} \right) \cdot \widetilde{\mathbf{\Sigma}}^{L-2}(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}_\star^\top, \quad \ell \in [L-1],$$

for all other layer matrices. Thus, for the next GD iteration, we have

$$\begin{aligned} \mathbf{W}_L(t+1) &= \mathbf{W}_L(t) - \eta \cdot \nabla_{\mathbf{W}_L}(\mathbf{\Theta}(t)) \\ &= \mathbf{U}_\star \begin{bmatrix} \widetilde{\mathbf{\Sigma}}_L(t) - \eta \cdot \left(\widetilde{\mathbf{\Sigma}}_L(t) \cdot \widetilde{\mathbf{\Sigma}}^{L-1}(t) - \mathbf{\Sigma}_{\star,r} \right) \cdot \widetilde{\mathbf{\Sigma}}^{L-1}(t) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}_\star^\top \\ &= \mathbf{U}_\star \begin{bmatrix} \widetilde{\mathbf{\Sigma}}_L(t+1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{V}_\star^\top. \end{aligned}$$

Similarly, we have

$$\begin{aligned} \mathbf{W}_\ell(t+1) &= \mathbf{W}_\ell(t) - \eta \cdot \nabla_{\mathbf{W}_\ell}(\mathbf{\Theta}(t)) \\ &= \mathbf{V}_\star \begin{bmatrix} \widetilde{\mathbf{\Sigma}}(t) - \eta \cdot \widetilde{\mathbf{\Sigma}}_L(t) \cdot \left(\widetilde{\mathbf{\Sigma}}_L(t) \cdot \widetilde{\mathbf{\Sigma}}^{L-1}(t) - \mathbf{\Sigma}_{\star,r} \right) \cdot \widetilde{\mathbf{\Sigma}}^{L-2}(t) & \mathbf{0} \\ \mathbf{0} & \alpha \mathbf{I}_{d-r} \end{bmatrix} \mathbf{V}_\star^\top \\ &= \mathbf{V}_\star \begin{bmatrix} \widetilde{\mathbf{\Sigma}}(t) \cdot \left(\mathbf{I}_r - \eta \cdot \widetilde{\mathbf{\Sigma}}_L(t) \cdot \left(\widetilde{\mathbf{\Sigma}}_L(t) \cdot \widetilde{\mathbf{\Sigma}}^{L-1}(t) - \mathbf{\Sigma}_{\star,r} \right) \cdot \widetilde{\mathbf{\Sigma}}^{L-3}(t) \right) & \mathbf{0} \\ \mathbf{0} & \alpha \mathbf{I}_{d-r} \end{bmatrix} \mathbf{V}_\star^\top \\ &= \mathbf{V}_\star \begin{bmatrix} \widetilde{\mathbf{\Sigma}}(t+1) & \mathbf{0} \\ \mathbf{0} & \alpha \mathbf{I}_{d-r} \end{bmatrix} \mathbf{V}_\star^\top, \end{aligned}$$

for all $\ell \in [L-1]$. This completes the proof. □

Proposition 3. *Let $\mathbf{M}_\star = \mathbf{V}_\star \mathbf{\Sigma}_\star \mathbf{V}_\star^\top \in \mathbb{R}^{d \times d}$ denote the SVD of the target matrix. The balanced initialization in Equation (**??**) is a member of the singular vector stationary set in Proposition 1, where $\mathbf{U}_L = \mathbf{Q}_L = \ldots = \mathbf{Q}_2 = \mathbf{V}_1 = \mathbf{V}_\star$.*

Proof. Using mathematical induction, we will show that with the balanced initialization in Equation (**??**), each weight matrix admits a decomposition of the form

$$\mathbf{W}_\ell(t) = \mathbf{V}_\star \mathbf{\Sigma}_\ell(t) \mathbf{V}_\star^\top, \tag{66}$$

which implies that the singular vectors are stationary for all t such that $\mathbf{U}_L = \mathbf{Q}_L = \ldots = \mathbf{Q}_2 = \mathbf{V}_1 = \mathbf{V}_\star$.

Base Case. Consider the weights at iteration $t = 0$. By the initialization scheme, we can write each weight matrix as

$$\mathbf{W}_\ell(0) = \alpha \mathbf{I}_d \implies \mathbf{W}_\ell(0) = \alpha \mathbf{V}_\star \mathbf{V}_\star^\top,$$

which implies that $\mathbf{W}_\ell(0) = \mathbf{V}_\star \mathbf{\Sigma}_\ell(0) \mathbf{V}_\star^\top$ with $\mathbf{\Sigma}_\ell(0) = \alpha \mathbf{I}_d$.

Inductive Step. By the inductive hypothesis, assume that the decomposition holds for all $t \geq 0$. We will show that it holds for all iterations $t+1$. Recall that the gradient of $f(\mathbf{\Theta})$ with respect to $\mathbf{W}_\ell$ is

$$\nabla_{\mathbf{W}_\ell} f(\mathbf{\Theta}) = \mathbf{W}_{L:\ell+1}^\top \cdot (\mathbf{W}_{L:1} - \mathbf{M}_\star) \cdot \mathbf{W}_{\ell-1:1}^\top.$$

Then, for $\mathbf{W}_\ell(t+1)$, we have

$$\begin{aligned} \mathbf{W}_\ell(t+1) &= \mathbf{W}_\ell(t) - \eta \cdot \nabla_{\mathbf{W}_L} f(\mathbf{\Theta}(t)) \\ &= \mathbf{V}_\star \mathbf{\Sigma}_\ell(t) \mathbf{V}_\star^\top - \eta \mathbf{W}_{L:\ell+1}^\top(t) \cdot (\mathbf{W}_{L:1}(t) - \mathbf{M}_\star) \cdot \mathbf{W}_{\ell-1:1}^\top(t) \\ &= \mathbf{V}_\star \mathbf{\Sigma}_\ell(t) \mathbf{V}_\star^\top - \eta \mathbf{V}_\star \cdot \left(\mathbf{\Sigma}_\ell^{L-\ell}(t) \cdot \left(\mathbf{\Sigma}_\ell^L(t) - \mathbf{\Sigma}_\star \right) \cdot \mathbf{\Sigma}_\ell^{\ell-1}(t) \right) \cdot \mathbf{V}_\star^\top \\ &= \mathbf{V}_\star \cdot \left(\mathbf{\Sigma}_\ell(t) - \eta \cdot \mathbf{\Sigma}_\ell^{L-\ell}(t) \cdot \left(\mathbf{\Sigma}_\ell^L(t) - \mathbf{\Sigma}_\star \right) \cdot \mathbf{\Sigma}_\ell^{\ell-1}(t) \right) \cdot \mathbf{V}_\star^\top \\ &= \mathbf{V}_\star \mathbf{\Sigma}(t) \mathbf{V}_\star^\top, \end{aligned}$$

where $\mathbf{\Sigma}(t) = \mathbf{\Sigma}_\ell(t) - \eta \cdot \mathbf{\Sigma}_\ell^{L-\ell}(t) \cdot \left(\mathbf{\Sigma}_\ell^L(t) - \mathbf{\Sigma}_\star \right) \cdot \mathbf{\Sigma}_\ell^{\ell-1}(t)$. This completes the proof. □

C.3. Auxiliary Results

Lemma 9. *Let $\{\mathbf{R}_\ell\}_{\ell=1}^L \in \mathbb{R}^{n\times n}$ be orthogonal matrices and $\mathbf{H}_{i,j} \in \mathbb{R}^{n^2\times n^2}$ be diagonal matrices. Consider the two following block matrices:*

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{1,1} & \mathbf{H}_{1,2} & \dots & \mathbf{H}_{L,1} \\ \mathbf{H}_{2,1} & \mathbf{H}_{2,2} & \dots & \mathbf{H}_{L,2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{1,L} & \mathbf{H}_{2,L} & \dots & \mathbf{H}_{L,L} \end{bmatrix}$$

$$\widetilde{\mathbf{H}} = \begin{bmatrix} \mathbf{R}_L\mathbf{H}_{1,1}\mathbf{R}_L^\top & \mathbf{R}_L\mathbf{H}_{1,2}\mathbf{R}_{L-1}^\top & \dots & \mathbf{R}_L\mathbf{H}_{1,L}\mathbf{R}_1^\top \\ \mathbf{R}_{L-1}\mathbf{H}_{2,1}\mathbf{R}_L^\top & \mathbf{R}_{L-1}\mathbf{H}_{2,2}\mathbf{R}_{L-1}^\top & \dots & \mathbf{R}_{L-1}\mathbf{H}_{2,L}\mathbf{R}_1^\top \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_1\mathbf{H}_{L,1}\mathbf{R}_L^\top & \mathbf{R}_1\mathbf{H}_{L,2}\mathbf{R}_{L-1}^\top & \dots & \mathbf{R}_1\mathbf{H}_{L,L}\mathbf{R}_1^\top \end{bmatrix}.$$

Then, the two matrices $\mathbf{H}$ and $\widetilde{\mathbf{H}}$ are similar, in the sense that they have the same eigenvalues.

Proof. It suffices to show that $\mathbf{H}$ and $\widetilde{\mathbf{H}}$ have the same characteristic polynomials. Let us define

$$\widetilde{\mathbf{H}} \coloneqq \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix},$$

where

$$\mathbf{A} \coloneqq \mathbf{R}_L\mathbf{H}_{1,1}\mathbf{R}_L^\top \qquad \mathbf{B} \coloneqq \begin{bmatrix} \mathbf{R}_L\mathbf{H}_{1,2}\mathbf{R}_{L-1}^\top & \dots & \mathbf{R}_L\mathbf{H}_{1,L}\mathbf{R}_1^\top \end{bmatrix} \tag{67}$$

$$\mathbf{C} \coloneqq \begin{bmatrix} \mathbf{R}_{L-1}\mathbf{H}_{2,1}\mathbf{R}_L^\top \\ \vdots \\ \mathbf{R}_1\mathbf{H}_{L,1}\mathbf{R}_L^\top \end{bmatrix} \qquad \mathbf{D} \coloneqq \begin{bmatrix} \mathbf{R}_{L-1}\mathbf{H}_{2,2}\mathbf{R}_{L-1}^\top & \dots & \mathbf{R}_{L-1}\mathbf{H}_{2,L}\mathbf{R}_1^\top \\ \vdots & \ddots & \vdots \\ \mathbf{R}_1\mathbf{H}_{L,2}\mathbf{R}_{L-1}^\top & \dots & \mathbf{R}_1\mathbf{H}_{L,L}\mathbf{R}_1^\top \end{bmatrix}. \tag{68}$$

Then, we have

$$\begin{aligned} \det(\widetilde{\mathbf{H}} - \lambda\mathbf{I}) &= \det\left(\begin{bmatrix} \mathbf{A}-\lambda\mathbf{I} & \mathbf{B} \\ \mathbf{C} & \mathbf{D}-\lambda\mathbf{I} \end{bmatrix}\right) \\ &= \det(\mathbf{A}-\lambda\mathbf{I}) \cdot \det((\mathbf{D}-\lambda\mathbf{I}) - \mathbf{C}(\mathbf{A}-\lambda\mathbf{I})^{-1}\mathbf{B}), \end{aligned}$$

where the second equality is by the Schur complement. Notice that

$$\begin{aligned} (\mathbf{A}-\lambda\mathbf{I})^{-1} = (\mathbf{R}_L\mathbf{H}_{1,1}\mathbf{R}_L^\top - \lambda\mathbf{I})^{-1} &= (\mathbf{R}_L\mathbf{H}_{1,1}\mathbf{R}_L^\top - \lambda\mathbf{R_L}\mathbf{R_L}^\top)^{-1} \\ &= \mathbf{R}_L \cdot (\mathbf{H}_{1,1}-\lambda\mathbf{I})^{-1} \cdot \mathbf{R}_L^\top. \end{aligned}$$

Then, we also see that,

$$\mathbf{C}(\mathbf{A}-\lambda\mathbf{I})^{-1}\mathbf{B} = \underbrace{\begin{bmatrix} \mathbf{R}_{L-1} & & \\ & \ddots & \\ & & \mathbf{R}_1 \end{bmatrix}}_{=:\widehat{\mathbf{V}}} \cdot \mathbf{E} \cdot \underbrace{\begin{bmatrix} \mathbf{R}_{L-1}^\top & & \\ & \ddots & \\ & & \mathbf{R}_1^\top \end{bmatrix}}_{=:\widehat{\mathbf{V}}^\top}.$$

where

$$\mathbf{E} \coloneqq \begin{bmatrix} \mathbf{H}_{2,1}\cdot(\mathbf{H}_{1,1}-\lambda\mathbf{I})^{-1}\cdot\mathbf{H}_{1,2} & \dots & \mathbf{H}_{2,1}\cdot(\mathbf{H}_{1,1}-\lambda\mathbf{I})^{-1}\cdot\mathbf{H}_{1,L} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{L,1}\cdot(\mathbf{H}_{1,1}-\lambda\mathbf{I})^{-1}\cdot\mathbf{H}_{1,2} & \dots & \mathbf{H}_{L,1}\cdot(\mathbf{H}_{1,1}-\lambda\mathbf{I})^{-1}\cdot\mathbf{H}_{1,L} \end{bmatrix}.$$

Similarly, we can write $\mathbf{D}$ as

$$\mathbf{D} = \widehat{\mathbf{V}} \underbrace{\begin{bmatrix} \mathbf{H}_{2,2} & \dots & \mathbf{H}_{2,L} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{L,2} & \dots & \mathbf{H}_{L,L} \end{bmatrix}}_{=:\mathbf{F}} \widehat{\mathbf{V}}^\top.$$

Then, we have

$$\begin{aligned}\det(\widetilde{\mathbf{H}} - \lambda\mathbf{I}) &= \det(\mathbf{R}_L \cdot (\mathbf{H}_{1,1} - \lambda\mathbf{I}) \cdot \mathbf{R}_L^\top) \cdot \det\left(\widehat{\mathbf{V}} \cdot (\mathbf{E} - \mathbf{F}) \cdot \widehat{\mathbf{V}}^\top\right) \\ &= \det(\mathbf{H}_{1,1} - \lambda\mathbf{I}) \cdot \det(\mathbf{E} - \mathbf{F}),\end{aligned}$$

which is not a function of $\mathbf{U}, \mathbf{V}, \{\mathbf{R}_\ell\}_{\ell=1}^L$. By doing the same for $\mathbf{H}$, we can show that both $\widetilde{\mathbf{H}}$ and $\mathbf{H}$ have the same characteristic polynomials, and hence the same eigenvalues. This completes the proof.

□

Lemma 10. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ be two orthogonal matrices. Then, the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ is also an orthogonal matrix:*

$$(\mathbf{A}\otimes\mathbf{B})^\top(\mathbf{A}\otimes\mathbf{B}) = (\mathbf{A}\otimes\mathbf{B})(\mathbf{A}\otimes\mathbf{B})^\top = \mathbf{I}_{d^2}.$$

Proof. We prove this directly by using properties of Kronecker products:

$$\begin{aligned}(\mathbf{A}\otimes\mathbf{B})^\top(\mathbf{A}\otimes\mathbf{B}) &= \mathbf{A}^\top\mathbf{A}\otimes\mathbf{B}^\top\mathbf{B} \\ &= \mathbf{I}_d \otimes \mathbf{I}_d = \mathbf{I}_{d^2}.\end{aligned}$$

Similarly, we have

$$\begin{aligned}(\mathbf{A}\otimes\mathbf{B})(\mathbf{A}\otimes\mathbf{B})^\top &= \mathbf{A}\mathbf{A}^\top\otimes\mathbf{B}\mathbf{B}^\top \\ &= \mathbf{I}_d \otimes \mathbf{I}_d = \mathbf{I}_{d^2}.\end{aligned}$$

This completes the proof. □

Lemma 11. *Let $\{a(t)\}_{t=1}^N$ be a sequence such that $a(t) \geq 0$ for all t. If there exists a constant $c \in (0,1)$ such that $a(t+1) < c \cdot a(t)$ for all t, then $\lim_{t\to\infty} a(t) = 0$.*

Proof. We prove this by direct reasoning. From the assumption $a(t+1) < c\cdot a(t)$ for some $c \in (0,1)$, we can iteratively expand this inequality:

$$a(t+1) < c\cdot a(t), \quad a(t+2) < c\cdot a(t+1) < c^2\cdot a(t),$$

and, more generally, by induction:

$$a(t+k) < c^k \cdot a(t), \quad \text{for all } k \geq 0.$$

Since $c \in (0,1)$, the sequence $\{c^k\}_{k=0}^\infty$ converges to 0 as $k \to \infty$. Hence:

$$0 \leq \lim_{k\to\infty} a(t+k) \leq \lim_{k\to\infty} c^k \cdot a(t) = 0.$$

Therefore, by the squeeze theorem, the sequence $\{a(t)\}$ converges to 0 as $t \to \infty$. □

Lemma 12 ([16])**.** *Consider any 1-D differentiable function $f(x)$ around a local minima $\bar{x}$, satisfying (i) $f^{(3)}(\bar{x}) \neq 0$, and (ii) $3[f^{(3)}]^2 - f''f^{(4)} > 0$ at $\bar{x}$. Then, there exists ϵ with sufficiently small $|\epsilon|$ and $\epsilon\cdot f^{(3)} > 0$ such that: for any point x_0 between $\bar{x}$ and $\bar{x}-\epsilon$, there exists a learning rate η such that $F_\eta^2(x_0) = x_0$, and*

$$\frac{2}{f''(\bar{x})} < \eta < \frac{2}{f''(\bar{x}) - \epsilon\cdot f^{(3)}(\bar{x})}.$$