On Non-local Convergence Analysis of Deep Linear Networks

Kun Chen [*1] **Dachao Lin** [*2] **Zhihua Zhang** [1]

Abstract

In this paper, we study the non-local convergence properties of deep linear networks. Specifically, under the quadratic loss, we consider optimizing deep linear networks in which there is at least a layer with only one neuron. We describe the convergent point of trajectories with an arbitrary balanced starting point under gradient flow, including the paths which converge to one of the saddle points. We also show specific convergence rates of trajectories that converge to the global minimizers by stages. We conclude that the rates vary from polynomial to linear. As far as we know, our results are the first to give a non-local analysis of deep linear neural networks with arbitrary balanced initialization, rather than the lazy training regime which has dominated the literature of neural networks or the restricted benign initialization.

1. Introduction

Deep neural networks have been successfully trained with simple gradient-based methods, which require optimizing highly non-convex functions. Many properties of the learning dynamic for deep neural networks are also present in the idealized and simplified case of deep linear networks. It is widely believed that deep linear networks could capture some important aspects of optimization in deep learning (Saxe et al., 2014). Therefore, many works have tried to study this issue in recent years (Hardt & Ma, 2017; Arora et al., 2018a;b; Bartlett et al., 2018; Shamir, 2019; Du & Hu, 2019; Hu et al., 2019; Zou et al., 2019; Eftekhari, 2020; Bah et al., 2021). However, previous understanding mainly focuses on local analysis or lazy training (Chizat et al., 2019), and there are few findings of the non-local analysis, even for linear networks.

[*]Equal contribution [1]School of Mathematical Sciences, Peking University. [2]Academy for Advanced Interdisciplinary Studies, Peking University.. Correspondence to: Zhihua Zhang <zhzhang@math.pku.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Local analysis of deep linear networks with quadratic loss. Several works analyzed linear networks with the quadratic loss. Bartlett et al. (2018) provided a linear convergence rate of gradient descent with identity initialization by assuming that the initial loss is small enough or the target is positive semi-definite. Bartlett et al. (2018) also showed the necessity of the positive definite target under identity initialization. Arora et al. (2018a) proved linear convergence rates of deep linear networks, by assuming that the initialization has a positive deficiency margin and is nearly balanced. Later on, a few works follow similar ideas with the neural tangent kernel (NTK) (Jacot et al., 2018) or lazy training (Chizat et al., 2019) to establish convergence analysis. Du & Hu (2019) demonstrated that if the width of hidden layers is all larger than the depth, gradient descent with Gaussian random initialization could then, with high probability, converge to a global minimum at a linear rate. Hu et al. (2019) improved the lower bound of width to be independent of depth, by utilizing orthogonal weight initialization but requiring each layer to have the same width. Moreover, Wu et al. (2019); Zou et al. (2019) obtained linear convergence for linear ResNet (He et al., 2016) with zero(-asymmetric) initialization, i.e., deep linear network with identity initialization. Specifically, Wu et al. (2019) adopted zero-asymmetric initialization requiring a zero-initialized output layer and identity initialization for the other layers. Such asymmetry also leads to a small variation of weight matrices[1], which is similar to local analysis. Zou et al. (2019) applied identity initialization (for deep linear networks), but still requiring a small initial loss or a lower bound for the width[2]. These works of local analysis have the feature of consistently bounded variation of weights[3] to make the entire trajectory stay in the benign local landscape. Additionally, such a small variation of weights can be satisfied by a slightly large width, a small initial loss, or a suitable initialization as previous works have shown.

Non-local analysis of deep linear networks with the quadratic loss. The non-local analysis requires a more comprehensive understanding. As far as we know, current works mainly focused on gradient flow, i.e., gradient descent

[1]See Wu et al. (2019, Eq. (4.7)) for detail.

[2]See Zou et al. (2019, Theorem 3.1 and Remark 3.2) for detail.

[3]For more examples, see the definitions of $\mathcal{C}(t)$ in Du & Hu (2019, Section 7) and Hu et al. (2019, Section 4.1).

with an infinitesimal small learning rate. From the manifold viewpoint, Bah et al. (2021) showed that the gradient flow always converges to a critical point of the underlying function. Moreover, they established that, for almost all initialization, the flow converges to a global minimum on the manifold of rank-k matrices, where k can be smaller than the largest possible rank of the induced weight. Hence, their work only ensured the convergence towards minimizers in a constrained subset, which is not necessarily the global minimizer. Additionally, they also provided a concrete example to display the existence of such rank unstable trajectories (see Bah et al. (2021, Remark 42)). Following Bah et al. (2021), Eftekhari (2020) provided a non-local convergence analysis for deep linear nets with quadratic loss. By assuming that there is a layer with only one neuron (including scalar output case) and the initialization is balanced, Eftekhari (2020) elaborated that gradient flow converges to global minimizers starting from a restricted set. Moreover, Eftekhari (2020) also confirmed that gradient flow could efficiently solve the problem by showing concrete linear convergence rates in the restricted set he defined.

In this work, we are interested in the *non-local analysis* of deep linear networks with the quadratic loss for *arbitrary balanced initialization*. To our knowledge, there was no non-local convergence analysis of gradient flow for deep linear nets in such a scheme.

1.1. Our Contributions

In this paper, we analyze gradient flow for deep linear networks with quadratic loss following the setting of Eftekhari (2020). The main contributions of this paper are summarized as follows:

- **Convergent result.** We first analyze the convergent behavior of trajectories. Compared to Eftekhari (2020), we define a more general rank-stable set of initialization to give *almost surely* convergence guarantee to the global minimizer (Theorem 3.4). Moreover, we also describe a more general global minimizer convergent set to guarantee convergence towards the global minimizer (Theorem 3.7).

  Furthermore, inherited from the above results, we introduce the indicator of *arbitrary* beginning point to decide the convergent point of the trajectory (Theorem 3.1). Our analysis is beyond the lazy training scheme, and does not require the constrained initialization region mentioned in Eftekhari (2020).

- **Convergence rate.** We also establish explicit convergence rates of the trajectories converging to the global minimizer. Our convergence rates are built on the fact that the singular value of the induced weight matrix goes through descending and ascending periods. In the case where the trajectory converges to the global minimizer, we show that in the worse case, the trajectory can be devided into three stages. The convergence rates vary from *polynomial to linear*. Our analysis is more comprehensive because Eftekhari (2020) only gave linear convergence rates for the last stage.

- We conduct numerical experiments to verify our findings. Though gradient descent seldom converges to strict saddle points (Lee et al., 2016), we find that our analysis of gradient flow reveals the long stuck period of trajectory under gradient descent in practice and the transition of the convergence rates for trajectories.

1.2. Additional Related work

Exponentially-tailed loss. There is much literature (Gunasekar et al., 2018; Nacson et al., 2019; Lyu & Li, 2019; Ji & Telgarsky, 2019; 2020) focusing on classification tasks under exponentially-tailed loss, such as logit loss or cross-entropy loss. Specifically, Gunasekar et al. (2018); Nacson et al. (2019) proved the convergence to a max-margin solution by assuming the loss converged to global optima. Lyu & Li (2019); Ji & Telgarsky (2019; 2020) also demonstrated the convergence to the max-margin solution under weaker assumptions that the initialization has zero classification error. These analyses focus on the final phase of training, which is still not a global analysis. Lin et al. (2021) showed a global analysis for directional convergence of deep linear networks. Their results also covered arbitrary initialization, but they required the spherically symmetric data assumption.

Global landscape analysis. Except for the non-local trajectory analysis, there is another line of works on non-local landscape analysis (see, e.g., Baldi & Hornik (1989); Kawaguchi (2016); Lu & Kawaguchi (2017); Safran & Shamir (2018); Laurent & Brecht (2018); Nguyen et al. (2018); Liang et al. (2018a;b; 2019); Venturi et al. (2019); Ding et al. (2019); Zhang (2019); Nouiehed & Razaviyayn (2021); Li et al. (2021); Achour et al. (2021) and the surveys (Sun et al., 2020; Sun, 2020)) which analyze the properties of stationary points, local minima, strict saddle points, etc. These works draw a whole picture of the benign landscape of deep networks, which provides a potential guarantee of the trajectory analysis, and motivates our work.

2. Preliminaries

In this paper, we consider the optimization of deep linear network under squared loss:

$$\min_{\boldsymbol{W}_1,\ldots,\boldsymbol{W}_N} L^N(\boldsymbol{W}_1,\ldots,\boldsymbol{W}_N):=\frac{1}{2}\left\|\boldsymbol{W}_N\cdots\boldsymbol{W}_1\boldsymbol{X}-\boldsymbol{Y}\right\|_F^2,$$

where data matrices are $\boldsymbol{X}\in\mathbb{R}^{d_x\times m}, \boldsymbol{Y}\in\mathbb{R}^{d_y\times m}$, and weight matrices are $\boldsymbol{W}_i\in\mathbb{R}^{d_i\times d_{i-1}}, i\in[N]$ with $d_0=$

$d_x, d_N = d_y$. The depth $N \geq 2$. We denote the induced weight matrix as $\boldsymbol{W} = \boldsymbol{W}_N \cdots \boldsymbol{W}_1 \in \mathbb{R}^{d_y \times d_x}$, and $\boldsymbol{W_N} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_N) \in \mathbb{R}^{d_1 \times d_0} \times \cdots \times \mathbb{R}^{d_N \times d_{N-1}}$ for brevity.

Notation. We denote vectors by lowercase bold letters (e.g., $\boldsymbol{u}, \boldsymbol{x}$), and matrices by capital bold letters (e.g., $\boldsymbol{W} = [w_{ij}]$). We use $(\boldsymbol{a})_i$ as the i-th entry of vector $\boldsymbol{a}$, and set $[a:b] = \{a, \ldots, b\}, [a] = [1:a], \forall a, b \in \mathbb{N}$. We denote by $s_i(\boldsymbol{Z})$ the i-th largest singular of $\boldsymbol{Z}$. We use $\|\cdot\|$ as the standard Euclidean norm for vectors, and $\|\cdot\|_F$ as the Frobenius norm for matrices. The convergence of vectors and matrices in this paper is defined under the standard Euclidean norm and Frobenius norm. We use the standard $\mathcal{O}(\cdot), \Omega(\cdot)$ and $\Theta(\cdot)$ notation to hide universal constant factors.

We integrate our assumptions in this paper below:

Assumption 2.1. Assume that the data, network, initialization and target satisfy:

- Data: $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{I}_{d_x}$.
- Network: $r := \min_{j \leq N} d_j = 1$.
- Initialization: $\boldsymbol{W}_i \boldsymbol{W}_i^\top = \boldsymbol{W}_{i+1}^\top \boldsymbol{W}_{i+1}, \forall i \in [N-1]$.
- Target: $\boldsymbol{Z} := \boldsymbol{Y}\boldsymbol{X}^\top \in \mathbb{R}^{d_y \times d_x}$ has different nonzero singular values, i.e., $s_1(\boldsymbol{Z}) > \cdots > s_d(\boldsymbol{Z}) > 0$, where $d = \operatorname{rank}(\boldsymbol{Y}\boldsymbol{X}^\top)$.

The first three assumptions are the same as Eftekhari (2020). The data assumption shows that the data is statistically whitened, which is common in the analysis of linear networks (Arora et al., 2018a; Bartlett et al., 2018). The network assumption includes the scalar output case. As mentioned in Eftekhari (2020), this case is significant as it corresponds to the popular spiked covariance model in statistics and signal processing (Eftekhari et al., 2019; Johnstone, 2001; Vershynin, 2012; Berthet & Rigollet, 2013; Deshpande & Montanari, 2014), to name a few. Moreover, the case $r = 1$ appears to be the natural beginning building block for understanding the behavior of trajectory. Finally, the third assumption is the common initialization technique for linear networks, which was used in Bartlett et al. (2018); Arora et al. (2018a;b; 2019); Eftekhari (2020); Zou et al. (2019). For the target, it is reasonable to assume different nonzero singular values in practice, because matrices with the same singular values have zero Lebesgue measure.

From the data assumption $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{I}_{d_x}$, we can simplify the problem as

$$\begin{aligned} &\min_{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_N} L^N(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_N) \\ &= \|\boldsymbol{W}_N \cdots \boldsymbol{W}_1 - \boldsymbol{Z}\|_F^2 + \|\boldsymbol{Y}\|_F^2 - \|\boldsymbol{Z}\|_F^2 . \end{aligned} \tag{1}$$

Hence, we call $\boldsymbol{Z}$ the target matrix. Moreover, we focus on the standard gradient flow method for all $i \in [N]$:

$$\dot{\boldsymbol{W}}_i(t) := \frac{d\boldsymbol{W}_i(t)}{dt} = -\nabla_{\boldsymbol{W}_i} L^N(\boldsymbol{W_N}(t)), t \geq 0. \tag{2}$$

Under the balanced initialization in Assumption 2.1, i.e., $\boldsymbol{W}_i \boldsymbol{W}_i^\top = \boldsymbol{W}_{i+1}^\top \boldsymbol{W}_{i+1}, \forall i \in [N-1]$, we have the induced weight flow of $\boldsymbol{W}(t) := \boldsymbol{W}_N(t) \cdots \boldsymbol{W}_1(t)$ following Arora et al. (2018b, Theorem 1):

$$\begin{aligned} \dot{\boldsymbol{W}}(t) &= -\mathcal{A}_{\boldsymbol{W}(t)}(\nabla L^1(\boldsymbol{W}(t))) \\ &= -\mathcal{A}_{\boldsymbol{W}(t)}(\boldsymbol{W}(t) - \boldsymbol{Z}), \end{aligned} \tag{3}$$

where we denote

$$\mathcal{A}_{\boldsymbol{W}}(\Delta) := \sum_{j=1}^{N} \left(\boldsymbol{W}\boldsymbol{W}^\top\right)^{\frac{N-j}{N}} \Delta \left(\boldsymbol{W}^\top \boldsymbol{W}\right)^{\frac{j-1}{N}} .$$

It is known that the induced flow in Eq. (3) admits an analytic singular value decomposition (SVD) (see Lemma 1 and Theorem 3 in Arora et al. (2019) for example). By $\operatorname{rank}(\boldsymbol{W}) \leq 1$ from Assumptions 2.1, we can denote the SVD of $\boldsymbol{W}(t)$ as $\boldsymbol{W}(t) \overset{\text{SVD}}{=\!=} s(t)\boldsymbol{u}(t)\boldsymbol{v}(t)^\top$ if $\boldsymbol{W}(t) \neq \boldsymbol{0}$. Here, $s(t), \boldsymbol{u}(t), \boldsymbol{v}(t)$ are all analytic functions of t. Moreover, $s(t) \in \mathbb{R}, \boldsymbol{u}(t) \in \mathbb{R}^{d_y}, \boldsymbol{v}(t) \in \mathbb{R}^{d_x}$, and $\|\boldsymbol{u}(t)\| = \|\boldsymbol{v}(t)\| = 1$. Previous work has already shown the dynamics of these terms:

$$\dot{\boldsymbol{u}}(t) = s(t)^{1-\frac{2}{N}} \left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right) \boldsymbol{Z}\boldsymbol{v}(t), \tag{4a}$$

$$\dot{\boldsymbol{v}}(t) = s(t)^{1-\frac{2}{N}} \left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right) \boldsymbol{Z}^\top \boldsymbol{u}(t), \tag{4b}$$

$$\dot{s}(t) = N s(t)^{2-\frac{2}{N}} \left(\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right) . \tag{4c}$$

Readers can find the derivation of $\dot{s}(t)$ in Arora et al. (2019, Theorem 3), and $\dot{\boldsymbol{u}}(t), \dot{\boldsymbol{v}}(t)$ in Eftekhari (2020, Eq. (139)) or Arora et al. (2019, Lemma 2) with some simplification. We also give the derivation of $\dot{\boldsymbol{u}}(t), \dot{\boldsymbol{v}}(t)$ in Lemma B.7.

To describe the solution obtained by flow, we also need the full SVD of target as $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \sum_{i=1}^{d} s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ with $s_1 > \cdots > s_d > 0$ by Assumption 2.1, orthogonal matrices $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{d_y}] \in \mathbb{R}^{d_y \times d_y}$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{d_x}] \in \mathbb{R}^{d_x \times d_x}$, and the rectangular-diagonal matrix $\boldsymbol{D} = \begin{pmatrix} \operatorname{diag}\{\boldsymbol{s}\} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \in \mathbb{R}^{d_y \times d_x}$, where $\boldsymbol{s} = (s_1, \ldots, s_d)^\top \in \mathbb{R}^d$. Thus, the best rank-one approximation matrix is $\boldsymbol{Z}_1 = s_1 \boldsymbol{u}_1 \boldsymbol{v}_1^\top$. Note that $\boldsymbol{Z}_1$ is the unique solution of problem (1), because $\boldsymbol{Z}$ has a nontrivial spectral gap by Assumption 2.1 (see Golub et al. (1987, Section 1)). For brevity, we define $s_k = 0, \forall k > d$. We adopt the projection length of $\boldsymbol{u}(t), \boldsymbol{v}(t)$ to each $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ as $a_i(t) = \boldsymbol{u}_i^\top \boldsymbol{u}(t), \forall i \in [d_y]$ and $b_j(t) = \boldsymbol{v}_j^\top \boldsymbol{v}(t), \forall j \in [d_x]$. Leaving the derivation in Appendix A, we have the gradient flow of each item:

$$\dot{a}_i(t) = s(t)^{1-\frac{2}{N}} \left(s_i b_i(t) - a_i(t) \sum_{k=1}^{d} [s_k a_k(t) b_k(t)] \right),$$

$$\dot{b}_j(t) = s(t)^{1-\frac{2}{N}}\left(s_j a_j(t) - b_j(t)\sum_{k=1}^{d}[s_k a_k(t) b_k(t)]\right),$$

where $i \in [d_y], j \in [d_x]$. Before we provide our results, we first state several useful invariance during the whole training dynamic as follows, which is crucial to our proofs.

Proposition 2.2. *If not mentioned specifically, we assume $s(0) > 0$. We have the following useful properties:*

1). If $s(0) > 0$, then $\forall t \geq 0, s(t) > 0$. Otherwise, $s(0) = 0$, then $\forall t \geq 0, s(t) = 0$ (i.e., $\boldsymbol{W}(t) = \mathbf{0}$).

2). $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t)$ is non-decreasing and converges.

3). $\boldsymbol{u}(t)^\top \boldsymbol{Z}_1\boldsymbol{v}(t)$ is non-decreasing and converges.

4). For all $t \geq 0$, $a_i(t) + b_i(t)$ has the same sign with $a_i(0) + b_i(0)$, i.e., $a_i(t) + b_i(t)$ is identically zero if $a_i(0) + b_i(0) = 0$, is positive if $a_i(0) + b_i(0) > 0$, and is negative if $a_i(0) + b_i(0) < 0, \forall i \in [d]$.

5). If for some $k \in [0 : d-1]$, $a_i(0) + b_i(0) = 0, \forall i \in [k]$ (if $k = 0$, then no such assumptions) and $a_{k+1}(0) + b_{k+1}(0) \neq 0$, then $|a_{k+1}(t) + b_{k+1}(t)|$ is non-decreasing, and $\lim_{t\to+\infty} a_{k+1}(t) + b_{k+1}(t)$ exists.

3. Convergent Behavior of Trajectories

We first give a glance of our main result for all initialization under Assumption 2.1. Our discovery is an extension of Eftekhari et al. (2019), since the description covers all trajectories including the ones converge to saddle points as well the global minimizers. The main conclusion is that the convergent point is decided by the indicator of initialization: $a_i(0) + b_i(0), i \in [d]$.

Theorem 3.1. *Under Assumption 2.1, and assume $s(0) > 0$. (I) If $k \in [0 : d-1], a_i(0) + b_i(0) = 0, \forall i \in [k]$ (if $k = 0$, then no such assumptions), and $a_{k+1}(0) + b_{k+1}(0) \neq 0$, then we have $\boldsymbol{W}(t) \to s_{k+1}\boldsymbol{u}_{k+1}\boldsymbol{v}_{k+1}^\top$. (II) Otherwise, i.e., $a_i(0) + b_i(0) = 0, \forall i \in [d]$, then we have $\boldsymbol{W}(t) \to \mathbf{0}$.*

We give a roadmap of the proof of Theorem 3.1 in Figure 3. We leave the complete proof in Section 3.3. Before we provide the details, we list some preparation built on the work of Eftekhari et al. (2019).

3.1. Rank-stable Set

By Bah et al. (2021, Theorem 5) (i.e., Theorem B.3), we know that $(\boldsymbol{W}_1(t), \ldots, \boldsymbol{W}_N(t))$ always converges to a critical point of L^N as $t \to +\infty$. Hence, we can define $\overline{\boldsymbol{W}} := \lim_{t\to+\infty} \boldsymbol{W}(t)$, and $\bar{s} := \lim_{t\to+\infty} s(t) = \lim_{t\to+\infty} \|\boldsymbol{W}(t)\|_F$. To specify the convergent point, we define rank-r set following Eftekhari (2020) as

$$\mathcal{M}_r = \{\boldsymbol{W} : \operatorname{rank}(\boldsymbol{W}) = r\}.$$

Furthermore, as mentioned in Eftekhari (2020, Lemma 3.3) (i.e., Lemma B.1), we have $\operatorname{rank}(\boldsymbol{W}(t)) = \operatorname{rank}(\boldsymbol{W}(0)) = 1, \forall t \geq 0$ if $\boldsymbol{W}(0) \neq \mathbf{0}$. However, the limit point $\overline{\boldsymbol{W}}(t)$ might not belong to $\mathcal{M}_1$ because $\mathcal{M}_1$ is not closed (see Eftekhari (2020, Lemma 3.4)). To exclude the zero matrix ($\bar{s} = 0$) as the limit point of gradient flow, Eftekhari (2020) introduced a restricted initialization set:

$$\mathcal{N}_\alpha(\boldsymbol{Z}) = \{\boldsymbol{W} : \boldsymbol{W} \overset{\text{SVD}}{=\!=} \boldsymbol{u}\cdot s\cdot \boldsymbol{v}^\top,\ s > \alpha s_1 - s_2 \geq 0,\ \boldsymbol{u}^\top \boldsymbol{Z}_1 \boldsymbol{v} > \alpha s_1\}, \alpha \in [s_2/s_1, 1).$$

While we find another rank-stable set $\mathcal{R}_b(\boldsymbol{Z})$ below with the similar rank-stable property shown in Lemma 3.2.

$$\mathcal{R}_b(\boldsymbol{Z}) = \{\boldsymbol{W} : \boldsymbol{W} \overset{\text{SVD}}{=\!=} \boldsymbol{u}\cdot s\cdot \boldsymbol{v}^\top,\ s > b, \boldsymbol{u}^\top \boldsymbol{Z}\boldsymbol{v} > b\}, b > 0.$$

Lemma 3.2 (Extension of Lemma 3.7 in Eftekhari (2020))**.** *Under Assumption 2.1, for gradient flow initialized at $\boldsymbol{W}(0) \in \mathcal{R}_b(\boldsymbol{Z})$, the limit point exists and satisfies $\overline{\boldsymbol{W}} = \lim_{t\to+\infty} \boldsymbol{W}(t) \in \mathcal{M}_1$, i.e., $\bar{s} > 0$.*

Proof. We only need to prove that $\boldsymbol{W}(t) \in \mathcal{R}_b(\boldsymbol{Z}), \forall t \geq 0$. Suppose $s(t_0) < b$ for some $t_0 > 0$. Since $\boldsymbol{W}(0) \in \mathcal{R}_b(\boldsymbol{Z})$, we have $s(0) > b$. Hence, by intermediate value theorem, $T := \sup\{t : s(t) = b, 0 \leq t \leq t_0\} > 0$. Since $s(t)$ is analytic, we obtain $s(T) = b$. Now we can conclude

$$s(t) < b, \forall T < t < t_0, s(T) = b. \tag{5}$$

Otherwise, if $s(t_1) \geq b$ for some $T < t_1 < t_0$, then by intermediate value theorem again, we can find a T' such that $T < t_1 \leq T' < t_2$, s.t., $s(T') = b$, which is a contradiction of the definition of T.

Moreover, since $\boldsymbol{W}(0) \in \mathcal{R}_b(\boldsymbol{Z})$, then $\boldsymbol{u}(0)^\top \boldsymbol{Z}\boldsymbol{v}(0) > b$. Thus, we have $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) > b, \forall t \geq 0$ by 2) in Proposition 2.2. Since $\boldsymbol{W}(0) \in \mathcal{R}_b(\boldsymbol{Z})$, then $s(0) > b > 0$. Thus, we have $s(T) > 0$ by 1) in Proposition 2.2. Therefore,

$$\begin{aligned}\dot{s}(T) &\overset{(4c)}{=} N s(T)^{2-\frac{2}{N}}\left(\boldsymbol{u}(T)^\top \boldsymbol{Z}\boldsymbol{v}(T) - s(T)\right)\\ &= N s(T)^{2-\frac{2}{N}}\left(\boldsymbol{u}(T)^\top \boldsymbol{Z}\boldsymbol{v}(T) - b\right) > 0,\end{aligned}$$

which is a contradiction with Eq. (5). □

Remark 3.3. Indeed, our rank-stable set $\mathcal{R}_b(\boldsymbol{Z})$ is more general than $\mathcal{N}_\alpha(\boldsymbol{Z})$. For any $\alpha \in [s_2/s_1, 1), \boldsymbol{W} \in \mathcal{N}_\alpha(\boldsymbol{Z})$, since $|\boldsymbol{u}^\top(\boldsymbol{Z} - \boldsymbol{Z}_1)\boldsymbol{v}| \leq s_2$, we get $\boldsymbol{u}^\top \boldsymbol{Z}\boldsymbol{v} \geq \boldsymbol{u}^\top \boldsymbol{Z}_1\boldsymbol{v} - s_2 > \alpha s_1 - s_2 \geq 0$. Hence, we can find some $b > 0$, such that $\boldsymbol{W} \in \mathcal{R}_b(\boldsymbol{Z})$. Moreover, when $\boldsymbol{W} = s_2\boldsymbol{u}_2\boldsymbol{v}_2^\top$, we can see $\boldsymbol{W} \in \mathcal{R}_{s_2/2}(\boldsymbol{Z})$, but $\boldsymbol{W} \notin \mathcal{N}_\alpha(\boldsymbol{Z}), \forall \alpha \in [s_2/s_1, 1)$, because $\boldsymbol{u}_2^\top \boldsymbol{Z}_1\boldsymbol{v}_2 = 0$. Additionally, we will see the necessity of our rank-stable set by showing counterexamples in Section 3.4.

Applying the same analysis as Eftekhari (2020, Theorem 3.8), we obtain almost surely convergence to the global minimizer from the initialization in our rank-stable set $\mathcal{R}_b(\boldsymbol{Z})$.

Theorem 3.4 (Extension of Theorem 3.8 in Eftekhari (2020))**.** *Under Assumption 2.1, gradient flow converges to a global minimizer of the original problem (1) from the initialization in $\boldsymbol{W}(0) \in \mathcal{R}_b(\boldsymbol{Z})$, outside of a subset with Lebesgue measure zero.*

3.2. Global Minimizer Convergent Set

Section 3.1 mainly analyzes the behavior of $s(t)$ to guarantee the rank of limit point is not degenerated. Though Theorem 3.4 ensures the almost surely convergence to the global minimizer, there are still some bad trajectories which converge to saddle points, such as $s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top, i \neq 1$. In this subsection, we move on to give another restricted initialization set to guarantee the global minimizer convergence without excluding a zero measure set. Our strategy mainly adopts singular vector analysis. In the following, we always assume $\boldsymbol{W}(0) \neq \boldsymbol{0}$ to ensure well-defined $\boldsymbol{u}(t)$ and $\boldsymbol{v}(t)$.

Lemma 3.5. *There exists a sequence $\{t_n\}$ with $t_n \to +\infty$, such that $\lim_{n\to+\infty} \left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t_n)\boldsymbol{u}(t_n)^\top\right) \boldsymbol{Z}\boldsymbol{v}(t_n) = \boldsymbol{0}$, and $\lim_{n\to+\infty} \left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t_n)\boldsymbol{v}(t_n)^\top\right) \boldsymbol{Z}^\top \boldsymbol{u}(t_n) = \boldsymbol{0}$. More specifically, we have*

$$\lim_{n\to+\infty} \left(\sum_{k=1}^{d} s_k a_k(t_n) b_k(t_n)\right) a_i(t_n) - s_i b_i(t_n) = 0, \quad (6)$$

$$\lim_{n\to+\infty} \left(\sum_{k=1}^{d} s_k a_k(t_n) b_k(t_n)\right) b_j(t_n) - s_j a_j(t_n) = 0, \quad (7)$$

for all $i \in [d_y]$ and $j \in [d_x]$. Furthermore, if there exists an $i_0 \in [d]$, such that $\lim_{n\to+\infty} a_{i_0}(t_n) + b_{i_0}(t_n)$ exists and the limit is not zero, then we could obtain

$$\lim_{n\to+\infty} \boldsymbol{u}(t_n)^\top \boldsymbol{Z}\boldsymbol{v}(t_n) = \lim_{n\to+\infty} \sum_{k=1}^{d} s_k a_k(t_n) b_k(t_n) = s_{i_0}. \quad (8)$$

Lemma 3.6. *Suppose there exists such a sequence $\{t_n\}_{n=0}^{\infty}$ that $t_n \to +\infty$, and for some $k < d, a_i(t_n) + b_i(t_n) = 0, \forall i \in [k], n \geq 0$ (if $k = 0$, then no such assumptions). Then if $\lim_{n\to+\infty} \sum_{j=1}^{d} s_j a_j(t_n) b_j(t_n) = s_{k+1}$, we have*

$$\lim_{n\to+\infty} \boldsymbol{u}(t_n)\boldsymbol{v}(t_n)^\top = \boldsymbol{u}_{k+1}\boldsymbol{v}_{k+1}^\top.$$

We underline that the result of Lemma 3.5 is the condition of Lemma 3.6, and the sub-sequence $\{t_n\}$ is enough to describe the convergent point because $\overline{\boldsymbol{W}}$ always exists.

Now we define the global minimizer convergent set as:

$$\mathcal{G}_b(\boldsymbol{Z}) = \{\boldsymbol{W} : \boldsymbol{W} \overset{\text{SVD}}{=\!=} \boldsymbol{u} \cdot s \cdot \boldsymbol{v}^\top,\ s > b, \boldsymbol{u}^\top \boldsymbol{Z}_1 \boldsymbol{v} > b\}, b > 0.$$

The only difference between $\mathcal{G}_b(\boldsymbol{Z})$ and $\mathcal{R}_b(\boldsymbol{Z})$ is that we replace target $\boldsymbol{Z}$ with the global minimizer $\boldsymbol{Z}_1$. We invoke Theorem 3.7 to conclude that the flow initialized from $\mathcal{G}_b(\boldsymbol{Z})$ could converge to a global minimizer.

Theorem 3.7 (Extension of Theorem 3.8 in Eftekhari (2020))**.** *Under Assumption 2.1, gradient flow converges to a global minimizer of the original problem (1) from the initialization $\boldsymbol{W}(0) \in \mathcal{G}_b(\boldsymbol{Z})$.*

Proof. Our proof is separated into the following steps.

Step 1. Since $\boldsymbol{W}(0) \in \mathcal{G}_b(\boldsymbol{Z})$, we get $s_1 a_1(0) b_1(0) = \boldsymbol{u}(0)^\top \boldsymbol{Z}_1 \boldsymbol{v}(0) > b > 0$. Hence, $|a_1(0) + b_1(0)| > 0$. Additionally, by 5) in Proposition 2.2, we get $|\lim_{t\to+\infty} a_1(t) + b_1(t)| \geq |a_1(0) + b_1(0)| > 0$. Thus, applying Lemma 3.5 with $i_0 = 1$, Eq. (8) holds, i.e., there exists a sequence $\{t_n\}$ with $t_n \to +\infty$, s.t., $\lim_{n\to+\infty} \boldsymbol{u}(t_n)^\top \boldsymbol{Z}\boldsymbol{v}(t_n) = \lim_{n\to+\infty} \sum_{k=1}^{d} s_k a_k(t_n) b_k(t_n) = s_1$.

Step 2. From **Step 1**, $\lim_{n\to+\infty} \sum_{k=1}^{d} s_k a_k(t_n) b_k(t_n) = s_1$. Then we can employ Lemma 3.6 with $k = 0$, showing that $\lim_{n\to+\infty} \boldsymbol{u}(t_n)\boldsymbol{v}(t_n)^\top = \boldsymbol{u}_1\boldsymbol{v}_1^\top$.

Step 3. From **Step 1**, we get $\boldsymbol{u}(t_n)^\top \boldsymbol{Z}\boldsymbol{v}(t_n) \to s_1 > 0$, leading to $\exists N > 0$, $\boldsymbol{u}(t_N)^\top \boldsymbol{Z}\boldsymbol{v}(t_N) > 0$. Moreover, since $\boldsymbol{W}(0) \in \mathcal{G}_b(\boldsymbol{Z})$, we have $s(0) > 0$. Thus, we have $s(t_N) > 0$ by 1) in Proposition 2.2. Hence, $\boldsymbol{W}(t_N) \in \mathcal{R}_b(\boldsymbol{Z})$ for some $b > 0$. Thus, by Lemma 3.2, we have $\bar{s} > 0$.

Step 4. Taking $t_n \to +\infty$ in Eq. (4c), we obtain $0 = N\bar{s}^{2-\frac{2}{N}}(s_1 - \bar{s})$. While by **Step 3**, $\bar{s} > 0$. Hence, $\bar{s} = s_1$.

Combining **Step 2** and **Step 4**, we obtain that $\boldsymbol{W}(t_n)$ converges to the global minimizer $s_1\boldsymbol{u}_1\boldsymbol{v}_1^\top$. Finally, we note that by Theorem 5 in Bah et al. (2021), $\boldsymbol{W}(t)$ converges. Hence, $\boldsymbol{W}(t) \to s_1\boldsymbol{u}_1\boldsymbol{v}_1^\top$, which is the global minimizer. □

Remark 3.8. We underline that, comparing to Theorem 3.4, Theorem 3.7 does not require to leave out a zero measure initialization set. Meanwhile, $\mathcal{G}_b(\boldsymbol{Z})$ is more general than $\mathcal{N}_\alpha(\boldsymbol{Z})$ as well, since we have less constraint for $\boldsymbol{u}^\top \boldsymbol{Z}_1 \boldsymbol{v}$. Moreover, we will see the necessity of our global minimizer convergent set by showing counter examples in Section 3.4.

3.3. Convergence Analysis for All Initialization

Now we turn back to the proof of Theorem 3.1.

Proof of Theorem 3.1. (I) For the first conclusion, the proof is separated into the following steps, which is similar as the proof of Theorem 3.7.

Step 1. In view of 4) in Proposition 2.2, we get $\forall t \geq 0, i \in [k], a_i(t) + b_i(t) = 0$ since $a_i(0) + b_i(0) = 0, \forall i \in [k]$. In view of 5) in Proposition 2.2, we could see $|\lim_{t\to+\infty} a_{k+1}(t) + b_{k+1}(t)| \geq |a_{k+1}(0) + b_{k+1}(0)| > 0, \forall t \geq 0$. Thus, applying Lemma 3.5, we obtain $\exists\{t_n\}$ with $t_n \to +\infty$, s.t., $\lim_{n\to+\infty} \boldsymbol{u}(t_n)^\top \boldsymbol{Z}\boldsymbol{v}(t_n) = \lim_{n\to+\infty} \sum_{i=1}^{d} s_i a_i(t_n) b_i(t_n) = s_{k+1}$.

Step 2. By **Step 1** and Lemma 3.6, we could obtain $\lim_{n\to+\infty} \boldsymbol{u}(t_n)\boldsymbol{v}(t_n)^\top = \boldsymbol{u}_{k+1}\boldsymbol{v}_{k+1}^\top$.

Step 3. By **Step 1**, $\boldsymbol{u}(t_n)^\top \boldsymbol{Z}\boldsymbol{v}(t_n) \to s_{k+1} > 0$, leading to $\exists N > 0$, $\boldsymbol{u}(t_N)^\top \boldsymbol{Z}\boldsymbol{v}(t_N) > 0$. Moreover, since $s(0) > 0$, we have $s(t_N) > 0$ by 1) in Proposition 2.2. Hence, $\boldsymbol{W}(t_N) \in \mathcal{R}_b(\boldsymbol{Z})$ for some $b > 0$. Thus, by Lemma 3.2, we have $\bar{s} > 0$.

Step 4. Finally, taking $t_n \to +\infty$ in Eq. (4c), by **Step 1**, we obtain $0 = N\bar{s}^{2-\frac{2}{N}}(s_{k+1} - \bar{s})$. While by **Step 3**, we have $\bar{s} > 0$. Hence, $\bar{s} = s_{k+1}$.

Combining **Step 2** and **Step 4**, we obtain that $\boldsymbol{W}(t_n)$ converges to $s_{k+1}\boldsymbol{u}_{k+1}\boldsymbol{v}_{k+1}^\top$. Finally, we note that by Theorem 5 in Bah et al. (2021), $\boldsymbol{W}(t)$ converges. Hence, $\boldsymbol{W}(t) \to s_{k+1}\boldsymbol{u}_{k+1}\boldsymbol{v}_{k+1}^\top$.

(II) For the second conclusion, by 4) in Proposition 2.2, we obtain $a_i(t) + b_i(t) = 0, \forall i \in [d], t \geq 0$. Hence, we get

$$\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \overset{(13)}{=} \sum_{j=1}^{d} s_j a_j(t) b_j(t) = -\sum_{j=1}^{d} s_j a_j^2(t) \leq 0,$$

which yields that

$$\dot{s}(t) \overset{(4c)}{=} Ns(t)^{2-\frac{2}{N}}\left(\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right) \leq -Ns(t)^{3-\frac{2}{N}}.$$

By solving the above flow, we derive that

$$s(t)^{2-\frac{2}{N}} \leq \left[s(0)^{\frac{2}{N}-2} + (2N-2)t\right]^{-1}.$$

Thus, $\bar{s} = \lim_{t\to+\infty} s(t) = 0$, i.e., $\boldsymbol{W}(t) \to \boldsymbol{0}$. □

Remark 3.9. We note that $s(0) = 0$ is a trivial case from 1) in Proposition 2.2. Moreover, we also give a convergence rate for the second conclusion, i.e., the rate for the convergence to the original point.

3.4. Some Intuitive Examples

Previous sections have shown the convergent behavior of arbitrary initialization. To give a better understanding of our results, we list some examples below.

Example 1. If $\boldsymbol{W}(0) = -s(0)\boldsymbol{u}_i\boldsymbol{v}_i^\top$ for some $i \in [\min\{d_y, d_x\}]$ and $s(0) > 0$, we have $\forall t \geq 0, \dot{\boldsymbol{u}}(t) = \boldsymbol{0}, \dot{\boldsymbol{v}}(t) = \boldsymbol{0}$ and $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) = -s_i$ from Eqs. (4a) and (4b). Thus, we obtain the ODE of $s(t)$ as follows

$$\dot{s}(t) \overset{(4c)}{=} -Ns(t)^{2-\frac{2}{N}}(s_i + s(t)) < 0.$$

So we could obtain $s(t) \to 0$. We see that $\boldsymbol{W}(t) \to \boldsymbol{0}$, which is a rank-unstable trajectory. Thus, the gradient flow of Eq. (2) does not converge to a global minimizer.

Example 2. Bah et al. (2021, Remark 42): If $\boldsymbol{Z} \succeq \boldsymbol{0}$, and $s(0) > 0, \boldsymbol{u}(0) = -\boldsymbol{v}(0)$. Then from Eqs. (4a) and (4b), we obtain $\dot{\boldsymbol{u}}(t) = -\dot{\boldsymbol{v}}(t)$ if $\boldsymbol{u}(t) = -\boldsymbol{v}(t)$. Hence, we get $\boldsymbol{u}(t) = -\boldsymbol{v}(t), \forall t \geq 0$. Thus, we obtain $\dot{s}(t)$ as follows:

$$\begin{aligned}\dot{s}(t) &\overset{(4c)}{=} Ns(t)^{2-\frac{2}{N}}\left(\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right)\\ &= Ns(t)^{2-\frac{2}{N}}\left(-\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{u}(t) - s(t)\right) < 0,\end{aligned}$$

leading to $s(t) \to 0$. We could see that $\boldsymbol{W}(t) \to \boldsymbol{0}$, which is a rank-unstable trajectory. Thus, the gradient flow of Eq. (2) does not converge to a global minimizer.

Example 3. If $\boldsymbol{W}(0) = s(0)\boldsymbol{u}_i\boldsymbol{v}_i^\top$ for some $i \in [d]$ and $s(0) > 0$, then from Eqs. (4a) and (4b), we obtain $\dot{\boldsymbol{u}}(t) = \boldsymbol{0}, \dot{\boldsymbol{v}}(t) = \boldsymbol{0}$ and $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) = s_i, \forall t \geq 0$. Thus, we obtain the ODE of $s(t)$ as follows

$$\dot{s}(t) \overset{(4c)}{=} Ns(t)^{2-\frac{2}{N}}(s_i - s(t)).$$

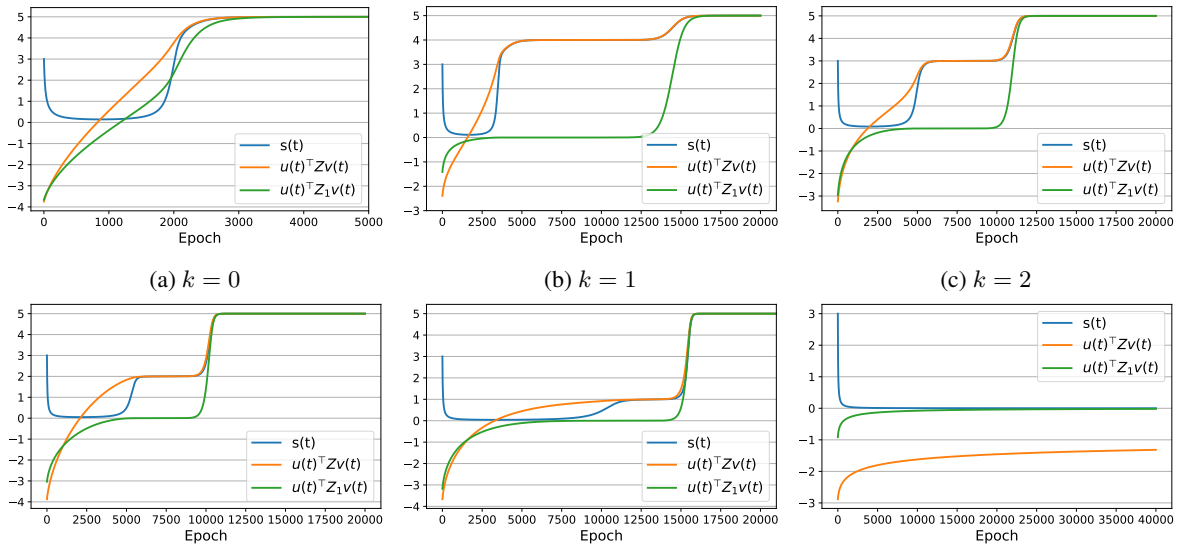
Hence, we obtain $s(t) \to s_i$. Once $i \neq 1$, we could see that $\boldsymbol{W}(t) \to s_i\boldsymbol{u}_i\boldsymbol{v}_i^\top$, i.e., the gradient flow of Eq. (1) does not converge to a global minimizer.

Remark 3.10. We note that our Theorem 3.1 covers Examples 1 and 2 by choosing $k = d$, and Example 3 by choosing $k = i - 1$. Moreover, we could not further improve our definition of $\mathcal{R}_b(\boldsymbol{Z})$ and $\mathcal{G}_b(\boldsymbol{Z})$ based on theses examples. We briefly show the reason here. Denote $\mathcal{R}(\boldsymbol{Z}) := \cup_{b>0}\mathcal{R}_b(\boldsymbol{Z}) = \{\boldsymbol{W}\colon \boldsymbol{W} \overset{\text{SVD}}{=\!=} \boldsymbol{u}\cdot s\cdot \boldsymbol{v}^\top, s > 0, \boldsymbol{u}^\top \boldsymbol{Z}\boldsymbol{v} > 0\}$. We obtain $\mathcal{R}(\boldsymbol{Z})^c := \mathcal{M}_1 \backslash \mathcal{R}(\boldsymbol{Z}) = \{\boldsymbol{W}\colon \boldsymbol{W} \overset{\text{SVD}}{=\!=} \boldsymbol{u}\cdot s\cdot \boldsymbol{v}^\top, s = 0 \text{ or } \boldsymbol{u}^\top \boldsymbol{Z}\boldsymbol{v} \leq 0\}$. Note that $s = 0$ in $\mathcal{R}(\boldsymbol{Z})^c$ is a trivial case corresponding to $\boldsymbol{W} = \boldsymbol{0}$. While in Example 1, $-s(0)\boldsymbol{u}_i\boldsymbol{v}_i^\top \notin \mathcal{R}(\boldsymbol{Z})$ because $-\boldsymbol{u}_i^\top \boldsymbol{Z}\boldsymbol{v}_i = -s_i \leq 0$, and in Example 2, $-s(0)\boldsymbol{u}(0)\boldsymbol{u}(0)^\top \notin \mathcal{R}(\boldsymbol{Z})$ because $-\boldsymbol{u}(0)^\top \boldsymbol{Z}\boldsymbol{u}(0) \leq 0$. Hence, through Examples 1 and 2, we find that certain initialization $s\boldsymbol{u}\boldsymbol{v}^\top$ with $\boldsymbol{u}^\top \boldsymbol{Z}\boldsymbol{v} \leq 0$ could indeed cause rank-unstable trajectories. Moreover, the equality ($\boldsymbol{u}^\top \boldsymbol{Z}\boldsymbol{v} = 0$) holds if in Example 1, $\text{rank}(\boldsymbol{Z}) = d < \min\{d_y, d_x\}$ and $i > d$. Therefore, the initialization set $\mathcal{R}_b(\boldsymbol{Z})$ could not be improved in the scope of s and $\boldsymbol{u}^\top \boldsymbol{Z}\boldsymbol{v}$. Additionally, we can get a similar argument of $\mathcal{G}_b(\boldsymbol{Z})$ by Example 3 since $s(0)\boldsymbol{u}_i\boldsymbol{v}_i^\top \notin \mathcal{G}_b(\boldsymbol{Z}), \forall i > 1$. Thus, we could not further improve our definition of $\mathcal{G}_b(\boldsymbol{Z})$ in the scope of s and $\boldsymbol{u}^\top \boldsymbol{Z}_1\boldsymbol{v}$ as well.

4. Convergence Rates to Global Minimizers

We briefly show the specific convergence rates in this section. We consider the rates to the global minimizers under Assumption 2.1, which is common in previous works. That is, from Theorem 3.1, we consider the initialization which satisfies $a_1(0) + b_1(0) \neq 0$ and $s(0) > 0$. Typically, for $N \geq 3$, the trajectories can be divided into three stages:

Stage 1. For $t \in [0, t_1]$, where $t_1 := \inf\{t : a_1(t)b_1(t) \geq$



(a) $k = 0$ (b) $k = 1$ (c) $k = 2$

(d) $k = 3$ (e) $k = 4$ (f) $k = 5$

Figure 1. We choose $N = 6, d = 5$ with hidden-layer width $(d_N, \dots, d_0) = (5, 4, 1, 10, 5, 3, 8)$, and set different $k \in [0 : d]$ in Theorem 3.1. The transparent horizontal lines are the singular value of $\boldsymbol{Z}$ in order (that we artificially set them to be $5, 4, 3, 2, 1$). The learning rate here is 5×10^{-4}. Legends present $s(t), \boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t), \boldsymbol{u}(t)^\top \boldsymbol{Z}_1\boldsymbol{v}(t)$. Here t is the running step in gradient descent.

$0\} < +\infty$, we have $a_1(t)b_1(t) \le 0$, and the rates are

$$1 - a_1(t)b_1(t) = \mathcal{O}([(N-2)t]^{-\frac{c_1}{N-2}}),$$
$$s(t) = \Omega([(N-2)t]^{-\frac{N}{N-2}}). \text{ [Theorem 4.5]}$$

Stage 2. For $t \in (t_1, t_2]$, where $t_2 := \inf\{t : \boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \ge s(t)\}$, we have $a_1(t)b_1(t) > 0, \dot{s}(t) \le 0$, and the rates are

$$1 - a_1(t)b_1(t) = \mathcal{O}([(N-2)t]^{-\frac{c_2}{N-2}}),$$
$$s(t) = \Omega([(N-2)t]^{-\frac{N}{N-2}}). \text{ [Theorem 4.7]}$$

Stage 3. For $t \in (\max\{t_1, t_2\}, +\infty)$, we have $a_1(t)b_1(t) > 0, \dot{s}(t) \ge 0$, and the rates are

$$1 - a_1(t)b_1(t) = \mathcal{O}\left(e^{-c_5 t}\right),$$
$$|s(t) - s_1| = \mathcal{O}(e^{-\min\{c_5, c_6\}t}). \text{ [Theorem 4.9]}$$

Here the c_is are constants and will be specified in theorems.

Remark 4.1. Although the rates in Stages 1 and 2 are similar, we adopt different proof techniques. Moreover, Stage 1 can be viewed as a warm-up period to enter $\mathcal{G}_b(\boldsymbol{Z})$.

Next we explain other minor cases: 1) If $t_1 = 0$, then Stage 1 vanishes; 2) If $t_1 \ge t_2$, then Stage 2 vanishes; 3) If $t_2 = +\infty$, then we have similar rates as Stage 3: $1 - a_1(t)b_1(t) = \mathcal{O}(e^{-c_3 t}), |s(t) - s_1| = \mathcal{O}(e^{-c_4 t})$ [Theorem 4.8]. However, this case is not suitable in our framework of the three stages.

Previous works (Arora et al., 2018a; Hu et al., 2019; Bartlett et al., 2018; Zou et al., 2019; Wu et al., 2019; Eftekhari, 2020) state linear rates under a local analysis or a restricted initialization set. However, we show the polynomial rates in the worse case for optimizing deep linear networks under arbitrary balanced initialization. Hence, we think that our results give a more general understanding of the trajectories optimizing deep linear networks.

4.1. Properties of t_1 and t_2

Before we give the detail of analysis, we need some preparation to better understand the choice of t_1 and t_2 in advance.

Lemma 4.2. *Let* $t_1 := \inf\{t : a_1(t)b_1(t) \ge 0\}$, $t_2 := \inf\{t : \boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \ge s(t)\}$ *and* $c_1(t) := a_1(t)b_1(t)$. *Then we have (I)* $t_1 < +\infty$ *if* $a_1(0) + b_1(0) \ne 0$, *and* $\dot{c}_1(t) \ge 0$ *for all* $t \ge 0$; *(II)* $\dot{s}(t) \le 0$ *for* $t \in [0, t_2)$, *and* $\dot{s}(t) \ge 0$ *for* $t \in [t_2, +\infty)$.

Remark 4.3. (I) in Lemma 4.2 tells us that the first stage, if exists, only appears a finite time in the beginning. (II) in Lemma 4.2 shows the induced weight norm ($\|\boldsymbol{W}(t)\|_F = s(t)$) goes through descending and ascending periods. If the initial induced weight norm starts with descending behavior, then it could descend forever, or it will change to ascending and continue increasing to s_1. If the initial induced weight norm begins with ascending behavior, then it would increase to s_1 directly. Such induced weight norm behavior also appears in deep linear networks with the logit loss (Lin et al., 2021).

4.2. Convergence Rates of $s(t)$: Stage 1 and Stage 2

In Stages 1 and 2, we have $\dot{s}(t) \leq 0$ from Lemma 4.2. Now we give global lower bounds for the singular value $s(t)$ within Stages 1 and 2.

Theorem 4.4. *Assuming $s(0) > 0$, we have $0 < s(t) \leq s_0 := \max\{s_1, s(0)\}$ for all $t \geq 0$. Further we have*

$$s(t) \geq s(0)e^{-2(s_1+s_0)t} \text{ for } N = 2, \tag{9a}$$

$$s(t) \geq [(s_1+s_0)(N-2)t + s(0)^{\frac{2}{N}-1}]^{-\frac{N}{N-2}} \text{ for } N \geq 3. \tag{9b}$$

We note that different lower bounds of $s(t)$ lead to different rates for the cases $N = 2$ and $N \geq 3$. For brevity, we only give the results for $N \geq 3$, and leave the simple case $N = 2$ in Appendix D.

4.3. Convergence Rates of $a_1(t)b_1(t)$: Stage 1

In the case where $a_1(0)b_1(0) < 0$, we prove that the case will reduce to the case $a_1(0)b_1(0) \geq 0$ in a finite time when $a_1(0)+b_1(0) \neq 0$. We further give an upper bound for time staying in Stage 1 and a positive lower bound of $a_1(t)b_1(t)$.

Theorem 4.5. *Suppose $N \geq 3$, $s(0) > 0, a_1(0)b_1(0) < 0$ and $a_1(0) + b_1(0) \neq 0$. Then we have*

$$1 - a_1(t)b_1(t) = \mathcal{O}([(N-2)t]^{-\frac{c_1}{N-2}}), \text{ for } 0 \leq t \leq t_1,$$

where $c_1 = 2/(s_1 + s_0)$ with $s_0 = \max\{s_1, s(0)\}$.

Furthermore, we have $t_1 \leq T_1$ with

$$T_1 = \frac{s(0)^{\frac{2}{N}-1}}{(s_1+s_0)(N-2)} \cdot \left(C_1^{-(s_1+s_0)(N-2)/2} - 1\right), \tag{10}$$

where $C_1 = |\frac{a_1(0)+b_1(0)}{a_1(0)-b_1(0)}|$. Additionally, we could obtain

$$a_1(t)b_1(t) = \Omega([a_1(0) + b_1(0)]^2/N), \tag{11}$$

if $t \geq \frac{s(0)^{2/N-1}}{(s_1+s_0)(N-2)} \cdot (e \cdot C_1^{-(s_1+s_0)(N-2)/2} - 1)$.

Remark 4.6. The upper bound of t_1 in Theorem 4.5 shows that if $a_1(0) + b_1(0) \approx 0$, then the Stage 1 would last for a long time according to Eq. (10). Moreover, Theorem 3.1 has already shown that once $a_1(0) + b_1(0) = 0$, the trajectory would not converge to the global minimizer. Hence, our finding in Theorem 4.5 is consistent with Theorem 3.1. Additionally, we also give guarantee of the trajectory to arrive at $\mathcal{G}_b(\boldsymbol{Z})$ for some $b > 0$ from Eq. (11). That is, the trajectory enters in our global minimizer convergent set.

4.4. Convergence Rates of $a_1(t)b_1(t)$: Stage 2

Based on Theorem 4.5, we can see once $a_1(0) + b_1(0) \neq 0$, then $a_1(t)b_1(t) > 0$ after finite time, that is, the trajectory enters in the global minimizer convergent set $\mathcal{G}_b(\boldsymbol{Z})$. In the following, we begin with $a_1(0)b_1(0) > 0$ for short. We find a similar polynomial convergence rate in Stage 2.

Theorem 4.7. *If $N \geq 3$, $s(0) > 0$ and $a_1(0)b_1(0) > 0$, then we get*

$$1 - a_1(t)b_1(t) = \mathcal{O}([(N-2)t]^{-\frac{c_2}{N-2}}),$$

where $c_2 = 2(s_1 - s_2)/(s_1 + s_0)$ with $s_0 = \max\{s_1, s(0)\}$.

4.5. Convergence Rates of $a_1(t)b_1(t)$ and $s(t)$: Stage 3

Before we start our analysis in Stage 3, we need to handle the minor case $t_2 := \inf\{t : \boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \geq s(t)\} = +\infty$.

Theorem 4.8. *If $N \geq 3$, $s(0) > 0, a_1(0)b_1(0) > 0$ and $t_2 = +\infty$, then we have*

$$1 - a_1(t)b_1(t) = \mathcal{O}\left(e^{-c_3 t}\right), |s(t) - s_1| = \mathcal{O}\left(e^{-c_4 t}\right),$$

where $c_3 = 2s_1^{1-\frac{2}{N}}(s_1 - s_2)$, $c_4 = Ns_1^{2-\frac{2}{N}}$.

Now we turn to the remaining case $t_2 < +\infty$. Additionally, we can assume $\dot{s}(0) \geq 0$ for short in Stage 3 by Lemma 4.2.

Theorem 4.9. *If $N \geq 3$, $s(0) > 0, a_1(0)b_1(0) > 0$ and $\dot{s}(0) \geq 0$, then we have*

$$1 - a_1(t)b_1(t) = \mathcal{O}(e^{-c_5 t}), |s(t) - s_1| = \mathcal{O}(e^{-\min\{c_5, c_6\}t}),$$

where $c_5 = 2s(0)^{1-\frac{2}{N}}(s_1 - s_2)$, $c_6 = Ns(0)^{2-\frac{2}{N}}$.

The difference between the minor case $t_2 = +\infty$ and Stage 3 is the constant above the exponent, and the proofs are similar between these two schemes. Thus, we combine them in a subsection.

Remark 4.10. Though we don't provide an upper bound of t_2 here, we still have a slower global convergence guarantee of $\boldsymbol{u}(t), \boldsymbol{v}(t)$ following Stage 2. Moreover, we discover the linear rate in Stage 3 only appears in the late training phase from experiments (see Section 5), and gives high precision guarantee of solution at last. Furthermore, Eftekhari (2020) also gave a linear rate in their restricted initialization set $\mathcal{N}_\alpha(\boldsymbol{Z})$. Thus, we mainly focus on the previous stages to highlight that our results cover a larger initialization set.

5. Experiments

In this section, we conduct simple numerical experiments to verify our findings. We first show the trajectories trapped into a saddle point $s_k \boldsymbol{u}_k \boldsymbol{v}_k^\top$ (or the global minimizer) guided by Theorem 3.1. Then we give some intuitive observation of the trajectories converged to the global minimizer.

Trajectories trapped into saddle points. We construct $\boldsymbol{u}(0) = \boldsymbol{U}\boldsymbol{\alpha}_1$ and $\boldsymbol{v}(0) = \boldsymbol{V}\boldsymbol{\alpha}_2$, where $\boldsymbol{\alpha}_1 \in \mathbb{R}^{d_y}$ and $\boldsymbol{\alpha}_2 \in \mathbb{R}^{d_x}$ have the inverse items until the k-th entry, i.e., $(\boldsymbol{\alpha}_1)_i + (\boldsymbol{\alpha}_2)_i = 0, \forall i \in [k]$ with $k \leq d$ and $(\boldsymbol{\alpha}_1)_{k+1} + (\boldsymbol{\alpha}_2)_{k+1} \neq 0$. Then we get $\forall i \in [k]$, $a_i(0) + b_i(0) = \boldsymbol{u}_i^\top \boldsymbol{u}(0) + \boldsymbol{v}_i^\top \boldsymbol{v}(0) = 0$ and $a_{k+1}(0) + b_{k+1}(0) \neq$

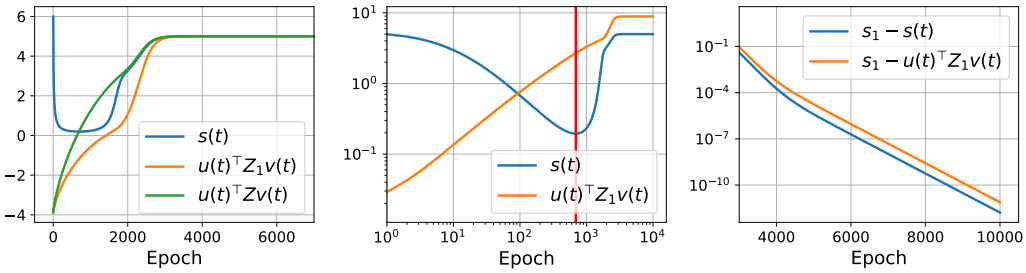


Figure 2. We choose $N = 6, d = 5$ with hidden-layer width $(d_N, \dots, d_0) = (5, 4, 1, 10, 5, 3, 8)$, and set $k = 0$ in Theorem 3.1, i.e., the trajectory converges to the global minimizer. Learning rate is 5×10^{-4}. Left: Dynamics of $s(t), \boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t), \boldsymbol{u}(t)^\top \boldsymbol{Z}_1\boldsymbol{v}(t)$ during the whole optimization. Middle: Polynomial convergence of $s(t)$ and $\boldsymbol{u}(t)^\top \boldsymbol{Z}_1\boldsymbol{v}(t)$ in Stages 1 and 2. Here we plot $\boldsymbol{u}(t)^\top \boldsymbol{Z}_1\boldsymbol{v}(t) - \boldsymbol{u}(0)^\top \boldsymbol{Z}_1\boldsymbol{v}(0)$ instead of $\boldsymbol{u}(t)^\top \boldsymbol{Z}_1\boldsymbol{v}(t)$ to avoid negative values. The red vertical line is the time that the monotonicity of $s(t)$ changes ($\dot{s}(t) = 0$). Right: Linear convergence of $s(t)$ and $\boldsymbol{u}(t)^\top \boldsymbol{Z}_1\boldsymbol{v}(t)$ in the final stage. Here t is the running step in gradient descent.

0. After $\boldsymbol{u}(0), \boldsymbol{v}(0)$ decided, we construct $\boldsymbol{W}_i(0) = s(0)^{1/N}\boldsymbol{h}_{i+1}\boldsymbol{h}_i^\top$ with $\|\boldsymbol{h}_i\| = 1, \forall i \in [N+1]$ and $\boldsymbol{h}_1 = \boldsymbol{v}(0), \boldsymbol{h}_{N+1} = \boldsymbol{u}(0)$ to obtain a balanced initialization $(\boldsymbol{W}_1(0), \dots, \boldsymbol{W}_N(0))$ and $\boldsymbol{W}(0) = s(0)\boldsymbol{u}(0)\boldsymbol{v}(0)^\top$. Finally, we run gradient descent (GD) for the problem (1) with a small learning rate 5×10^{-4}, and we artificially set $s_i = d + 1 - i, \forall i \in [d]$.

The simulations are shown in Figure 1. As Figure 1 depicts, $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t)$ and $\boldsymbol{u}(t)^\top \boldsymbol{Z}_1\boldsymbol{v}(t)$ are non-decreasing as our Proposition 2.2 shows, and $s(t)$ goes through descending and ascending periods as Lemma 4.2 mentioned. Additionally, we could see our construction gives a stuck region around s_{k+1} according to the choice of $k \le d$. Though our Theorem 3.1 shows that the gradient flow of $\boldsymbol{W}(t)$ would finally converge to $s_{k+1}\boldsymbol{u}_{k+1}\boldsymbol{v}_{k+1}^\top$, we find after a period of (long) time, gradient descent can escape from the saddle point around $s_{k+1}\boldsymbol{u}_{k+1}\boldsymbol{v}_{k+1}^\top$, and finally converges to a global minimizer. We consider the numerical error during optimization and unbalanced weight matrix caused by GD may lead to the inconsistent of gradient flow and its discrete version GD. Overall, we describe the possible convergence behavior of all initialization in the ideal setting.

Trajectories converged to the global minimizer. We also plot the trajectory converging to the global minimizer in detail shown in Figure 2. To give a more clear variation of stages, we adopt $s(0) = 6$ and a negative $\boldsymbol{u}(0)^\top \boldsymbol{Z}_1\boldsymbol{v}(0)$. As the left figure of Figure 2 shows, $s(t)$ first decreases, then increases. Additionally, the middle figure shows that $s(t)$ decreases and $\boldsymbol{u}(t)^\top \boldsymbol{Z}_1\boldsymbol{v}(t)$ increases with an approximate polynomial rate (noting the log scale in both x-axis and y-axis). Moreover, the left and middle figures also show that once $s(t)$ increases, $\boldsymbol{u}(t)^\top \boldsymbol{Z}_1\boldsymbol{v}(t)$ will increase much faster, and switch to another stage as we prove. Finally, we observe the final stage, that is, the linear convergence of both $s(t)$ and $\boldsymbol{u}(t)^\top \boldsymbol{Z}_1\boldsymbol{v}(t)$ to s_1 in the right graph of Figure 2. Overall, we conclude that our convergent rates match with the numerical experiments well[4].

6. Conclusion

In this work we have studied the training dynamic of deep linear networks which have a one-neuron layer. Specifically, we focus on the gradient flow methods under the quadratic loss and the balanced initialization. We have shown the convergent point of an arbitrary balanced starting point. Moreover, we have described the convergence rates of the trajectories towards the global minimizers, finding that the convergence goes through polynomial to linear rates. The behavior predicted by our theorems is also observed in numerical experiments. Though our analysis mainly focuses on gradient flow, a recent work (Elkabetz & Cohen, 2021) gives the conjecture that the theory of gradient flows will be central to unraveling mysteries behind deep learning, which makes our effort of flow analysis become practical. Although Elkabetz & Cohen (2021)'s results are not directly applicable due to the near-zero initialization assumption, the curvature analysis of flow trajectory may be a promising future work. Moreover, the analysis of linear networks without a one-neuron layer and other scalar-output linear networks (Woodworth et al., 2020; Yun et al., 2020) remain open problems. Overall, we hope that our findings of training trajectories would bring a better understanding of (linear) neural networks.

Acknowledgements

This work has been supported by the National Key Research and Development Project of China (No. 2018AAA0101004).

[4]Moreover, we also conduct experiments of the minor case ($t_2 = +\infty$) in Appendix E, which match with our findings as well.

References

Achour, E. M., Malgouyres, F., and Gerchinovitz, S. Global minimizers, strict and non-strict saddle points, and implicit regularization for deep linear neural networks. *arXiv preprint arXiv:2107.13289*, 2021.

Arora, S., Cohen, N., Golowich, N., and Hu, W. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2018a.

Arora, S., Cohen, N., and Hazan, E. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pp. 244–253. PMLR, 2018b.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32:7413–7424, 2019.

Bah, B., Rauhut, H., Terstiege, U., and Westdickenberg, M. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 02 2021. ISSN 2049-8772. doi: 10.1093/imaiai/iaaa039. URL https://doi.org/10.1093/imaiai/iaaa039. iaaa039.

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

Bartlett, P., Helmbold, D., and Long, P. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In *International conference on machine learning*, pp. 521–530. PMLR, 2018.

Berthet, Q. and Rigollet, P. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32:2937–2947, 2019.

Deshpande, Y. and Montanari, A. Information-theoretically optimal sparse pca. In *2014 IEEE International Symposium on Information Theory*, pp. 2197–2201. IEEE, 2014.

Ding, T., Li, D., and Sun, R. Sub-optimal local minima exist for neural networks with almost all non-linear activations. *arXiv preprint arXiv:1911.01413*, 2019.

Du, S. and Hu, W. Width provably matters in optimization for deep linear neural networks. In *International Conference on Machine Learning*, pp. 1655–1664. PMLR, 2019.

Eftekhari, A. Training linear neural networks: Non-local convergence and complexity results. In *International Conference on Machine Learning*, pp. 2836–2847. PMLR, 2020.

Eftekhari, A., Hauser, R. A., and Grammenos, A. Moses: A streaming algorithm for linear dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2901–2911, 2019.

Elkabetz, O. and Cohen, N. Continuous vs. discrete optimization of deep neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

Golub, G. H., Hoffman, A., and Stewart, G. W. A generalization of the eckart-young-mirsky matrix approximation theorem. *Linear Algebra and its applications*, 88:317–327, 1987.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems*, pp. 9461–9471, 2018.

Hardt, M. and Ma, T. Identity matters in deep learning. In *International Conference on Learning Representations*, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hu, W., Xiao, L., and Pennington, J. Provable benefit of orthogonal initialization in optimizing deep linear networks. In *International Conference on Learning Representations*, 2019.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Ji, Z. and Telgarsky, M. Gradient descent aligns the layers of deep linear networks. In *7th International Conference on Learning Representations, ICLR*, 2019.

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

Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pp. 295–327, 2001.

Kawaguchi, K. Deep learning without poor local minima. In *Advances in neural information processing systems*, pp. 586–594, 2016.

Laurent, T. and Brecht, J. Deep linear networks with arbitrary loss: All local minima are global. In *International Conference on Machine Learning*, pp. 2908–2913, 2018.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. In *Conference on learning theory*, pp. 1246–1257. PMLR, 2016.

Li, D., Ding, T., and Sun, R. On the benefit of width for neural networks: Disappearance of bad basins, 2021.

Liang, S., Sun, R., Lee, J. D., and Srikant, R. Adding one neuron can eliminate all bad local minima. In *Advances in Neural Information Processing Systems*, pp. 4355–4365, 2018a.

Liang, S., Sun, R., Li, Y., and Srikant, R. Understanding the loss surface of neural networks for binary classification. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2835–2843. PMLR, 10–15 Jul 2018b. URL https://proceedings.mlr.press/v80/liang18a.html.

Liang, S., Sun, R., and Srikant, R. Revisiting landscape analysis in deep neural networks: Eliminating decreasing paths to infinity, 2019.

Lin, D., Sun, R., and Zhang, Z. Faster directional convergence of linear neural networks under spherically symmetric data. *Advances in Neural Information Processing Systems*, 34, 2021.

Lu, H. and Kawaguchi, K. Depth creates no bad local minima. *arXiv preprint arXiv:1702.08580*, 2017.

Lyu, K. and Li, J. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2019.

Nacson, M. S., Lee, J., Gunasekar, S., Savarese, P. H. P., Srebro, N., and Soudry, D. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428. PMLR, 2019.

Nguyen, Q., Mukkamala, M. C., and Hein, M. On the loss landscape of a class of deep neural networks with no bad local valleys. In *International Conference on Learning Representations*, 2018.

Nouiehed, M. and Razaviyayn, M. Learning deep models: Critical points and local openness. *INFORMS Journal on Optimization*, 2021.

Safran, I. and Shamir, O. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pp. 4433–4441. PMLR, 2018.

Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Shamir, O. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. In *Conference on Learning Theory*, pp. 2691–2713. PMLR, 2019.

Sun, R., Li, D., Liang, S., Ding, T., and Srikant, R. The global landscape of neural networks: An overview. *IEEE Signal Processing Magazine*, 37(5):95–108, 2020.

Sun, R.-Y. Optimization for deep learning: An overview. *Journal of the Operations Research Society of China*, pp. 1–46, 2020.

Venturi, L., Bandeira, A. S., and Bruna, J. Spurious valleys in one-hidden-layer neural network optimization landscapes. *Journal of Machine Learning Research*, 20(133): 1–34, 2019. URL http://jmlr.org/papers/v20/18-674.html.

Vershynin, R. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Wu, L., Wang, Q., and Ma, C. Global convergence of gradient descent for deep linear residual networks. *Advances in Neural Information Processing Systems*, 32: 13389–13398, 2019.

Yun, C., Krishnan, S., and Mobahi, H. A unifying view on implicit bias in training linear neural networks. *arXiv preprint arXiv:2010.02501*, 2020.

Zhang, L. Depth creates no more spurious local minima. *arXiv preprint arXiv:1901.09827*, 2019.

Zou, D., Long, P. M., and Gu, Q. On the global convergence of training deep linear resnets. In *International Conference on Learning Representations*, 2019.

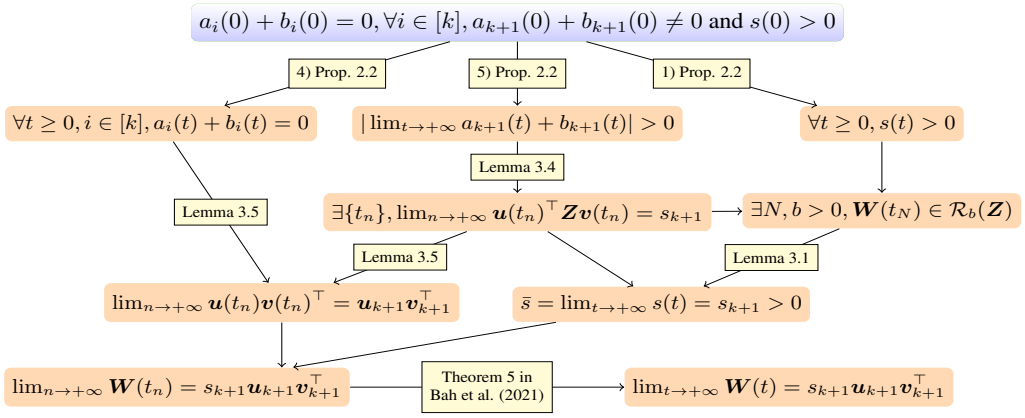


Figure 3. The roadmap of the proof of Theorem 3.1.

A. Missing Derivation in Section 2

We adopt the projection length of $\boldsymbol{u}(t), \boldsymbol{v}(t)$ to each $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ as $a_i(t)=\boldsymbol{u}_i^\top \boldsymbol{u}(t), \forall i\in[d_y]$ and $b_j(t)=\boldsymbol{v}_j^\top \boldsymbol{v}(t), \forall j\in[d_x]$. So we have

$$\boldsymbol{u}(t)=\sum_{i=1}^{d_y} a_i(t)\boldsymbol{u}_i,\ \boldsymbol{v}(t)=\sum_{j=1}^{d_x} b_j(t)\boldsymbol{v}_j,\ \sum_{i=1}^{d_y} a_i^2(t)=\sum_{j=1}^{d_x} b_j^2(t)=1. \tag{12}$$

Then we get

$$\begin{aligned}
&\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t)=\sum_{i=1}^{d_y}\sum_{j=1}^{d_x} a_i(t)b_j(t)\boldsymbol{u}_i^\top \boldsymbol{Z}\boldsymbol{v}_j=\sum_{i=j=1}^{d} a_i(t)b_j(t)\boldsymbol{u}_i^\top \boldsymbol{Z}\boldsymbol{v}_j=\sum_{k=1}^{d} s_k a_k(t)b_k(t),\\
&\boldsymbol{Z}\boldsymbol{v}(t)=\sum_{j=1}^{d_x} b_j(t)\boldsymbol{Z}\boldsymbol{v}_j=\sum_{j=1}^{d} s_j b_j(t)\boldsymbol{u}_j,\ \text{and}\ \boldsymbol{Z}^\top \boldsymbol{u}(t)=\sum_{i=1}^{d_y} a_i(t)\boldsymbol{Z}^\top \boldsymbol{u}_i=\sum_{i=1}^{d} s_i a_i(t)\boldsymbol{v}_i.
\end{aligned} \tag{13}$$

where we uses the fact that $\boldsymbol{Z}^\top \boldsymbol{u}_i=\boldsymbol{0}, \forall i>d$, $\boldsymbol{Z}\boldsymbol{v}_j=\boldsymbol{0}, \forall j>d$, and $\boldsymbol{u}_i^\top \boldsymbol{Z}\boldsymbol{v}_j=0, \boldsymbol{u}_i^\top \boldsymbol{Z}\boldsymbol{v}_i=s_i, \forall j\neq i\leq d$. Hence, we have the gradient flow of each item:

$$\begin{aligned}
&\dot{a}_i(t)\overset{(4a)}{=} s(t)^{1-\frac{2}{N}}\boldsymbol{u}_i^\top\left(\boldsymbol{I}_{d_y}-\boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right)\boldsymbol{Z}\boldsymbol{v}(t)\overset{(13)}{=} s(t)^{1-\frac{2}{N}}\left(s_i b_i(t)-a_i(t)\sum_{k=1}^{d}[s_k a_k(t)b_k(t)]\right), \forall i\in[d_y].\\
&\dot{b}_j(t)\overset{(4b)}{=} s(t)^{1-\frac{2}{N}}\boldsymbol{v}_j^\top\left(\boldsymbol{I}_{d_x}-\boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right)\boldsymbol{Z}^\top\boldsymbol{u}(t)\overset{(13)}{=} s(t)^{1-\frac{2}{N}}\left(s_j a_j(t)-b_j(t)\sum_{k=1}^{d}[s_k a_k(t)b_k(t)]\right), \forall j\in[d_x].
\end{aligned} \tag{14}$$

B. Auxiliary Results

B.1. Previous Results

Lemma B.1 (Lemma 3.3 in Eftekhari (2020)). *For the induced flow in Eq. (3), we have that* $\operatorname{rank}(\boldsymbol{W}(t))=\operatorname{rank}(\boldsymbol{W}(0)), \forall t\geq 0$, *provided that* $\boldsymbol{X}\boldsymbol{X}^\top$ *is invertible and the network depth* $N\geq 2$.

Lemma B.2 (Lemma 4 in Arora et al. (2019)). *Let* $\alpha\geq 1/2$ *and* $g:[0,\infty)\to\mathbb{R}$ *be a continuous function. Consider the initial value problem:*

$$s(0)=s_0, \dot{s}(t)=(s^2(t))^\alpha\cdot g(t), \forall t\geq 0,$$

where $s_0\in\mathbb{R}$. *Then, as long as it does not diverge to* $\pm\infty$, *the solution to this problem* $(s(t))$ *has the same sign as its initial value* (s_0). *That is,* $s(t)$ *is identically zero if* $s_0=0$, *is positive if* $s_0>0$, *and is negative if* $s_0<0$.

Theorem B.3 (Theorem 5 in Bah et al. (2021)). *Assume* $\boldsymbol{X}\boldsymbol{X}^\top$ *has full rank. Then the flows* $\boldsymbol{W}_i(t)$ *defined by Eq. (2) and* $\boldsymbol{W}(t)$ *given by Eq. (3) are defined and bounded for all* $t\geq 0$ *and* $(\boldsymbol{W}_1(t),\ldots,\boldsymbol{W}_N(t))$ *converges to a critical point of* L^N *as* $t\to+\infty$.

Definition B.4 (Definition 27 in Bah et al. (2021)). Let $(\mathcal{M}, g)$ be a Riemannian manifold with Levi-Civita connection ∇ and let $f : \mathcal{M} \to \mathbb{R}$ be a twice continuously differentiable function. A critical point $x_0 \in \mathcal{M}$, i.e., $\nabla^g f(x_0) = 0$ is called a strict saddle point if Hess $f(x)$ has a negative eigenvalue. We denote the set of all strict saddles of f by $\mathcal{X} = \mathcal{X}(f)$. We say that f has the strict saddle point property, if all critical points of f that are not local minima are strict saddle points.

The following theorem shows that flows avoid strict saddle points almost surely.

Theorem B.5 (Theorem 28 in Bah et al. (2021)). *Let $L : \mathcal{M} \to \mathbb{R}$ be a C^2-function on a second countable finite dimensional Riemannian manifold $(\mathcal{M}, g)$, where we assume that $\mathcal{M}$ is of class C^2 as a manifold and the metric g is of class C^1. Assume that $\phi_t(x_0)$ exists for all $x_0 \in \mathcal{M}$ and all $t \in [0, +\infty)$. Then the set*

$$\mathcal{S}_L := \{x_0 \in \mathcal{M} : \lim_{t \to +\infty} \phi_t(x_0) \in \mathcal{X} = \mathcal{X}(L)\}$$

of initial points such that the corresponding flow converges to a strict saddle point of L has measure zero.

Proposition B.6 (Proposition 33 in Bah et al. (2021)). *The function L^1 on $\mathcal{M}_k$ for $k \leq r$ satisfies the strict saddle point property. More precisely, all critical points of L^1 on $\mathcal{M}_k$ except for the global minimizers are strict saddle points.*

B.2. Auxiliary Lemmas

Lemma B.7 (Dynamic of $s(t), \boldsymbol{u}(t), \boldsymbol{v}(t)$). *We give the derivation of $\dot{\boldsymbol{u}}(t), \dot{\boldsymbol{v}}(t), \dot{s}(t)$ shown in the main context in this lemma:*

$$\begin{aligned}
\dot{\boldsymbol{u}}(t) &= s(t)^{1-\frac{2}{N}} \left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right) \boldsymbol{Z}\boldsymbol{v}(t), \\
\dot{\boldsymbol{v}}(t) &= s(t)^{1-\frac{2}{N}} \left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right) \boldsymbol{Z}^\top\boldsymbol{u}(t), \\
\dot{s}(t) &= Ns(t)^{2-\frac{2}{N}} \left(\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right).
\end{aligned}$$

Proof. $\dot{s}(t)$ directly follows Arora et al. (2019, Theorem 3). As for $\dot{\boldsymbol{u}}(t)$ and $\dot{\boldsymbol{v}}(t)$, we begin with $\boldsymbol{u}(t)^\top\boldsymbol{u}(t) = \boldsymbol{v}(t)^\top\boldsymbol{v}(t) = 1$. Then by taking the derivative of the identities, we get

$$\boldsymbol{u}(t)^\top\dot{\boldsymbol{u}}(t) = \boldsymbol{v}(t)^\top\dot{\boldsymbol{v}}(t) = 0, \forall t \geq 0. \tag{15}$$

By taking derivative of both sides of the SVD: $\boldsymbol{W}(t) = s(t)\boldsymbol{u}(t)\boldsymbol{v}(t)^\top$, we find that

$$\dot{\boldsymbol{W}}(t) = s(t)\dot{\boldsymbol{u}}(t)\boldsymbol{v}(t)^\top + s(t)\boldsymbol{u}(t)\dot{\boldsymbol{v}}(t)^\top + \dot{s}(t)\boldsymbol{u}(t)\boldsymbol{v}(t)^\top, \forall t \geq 0.$$

Hence, multiplying $\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top$ and $\boldsymbol{v}(t)$, we get

$$s(t)^{-1}\left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right)\dot{\boldsymbol{W}}(t)\boldsymbol{v}(t) = \left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right)\dot{\boldsymbol{u}}(t).$$

From Eq. (15), we know $\dot{\boldsymbol{u}}(t) \perp \boldsymbol{u}(t)$. Therefore, we obtain

$$\dot{\boldsymbol{u}}(t) = s(t)^{-1}\left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right)\dot{\boldsymbol{W}}(t)\boldsymbol{v}(t). \tag{16}$$

Similarly, we can find that

$$\dot{\boldsymbol{v}}(t) = s(t)^{-1}\left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right)\dot{\boldsymbol{W}}(t)^\top\boldsymbol{u}(t). \tag{17}$$

Now we replace $\dot{\boldsymbol{W}}(t)$ by Eq. (3) and $\boldsymbol{W}(t) = s(t)\boldsymbol{u}(t)\boldsymbol{v}(t)^\top$:

$$\begin{aligned}
\dot{\boldsymbol{W}}(t) \overset{(3)}{=} &- Ns(t)^{2-\frac{2}{N}}\boldsymbol{u}(t)\boldsymbol{u}(t)^\top\left[\boldsymbol{W}(t) - \boldsymbol{Z}\right]\boldsymbol{v}(t)\boldsymbol{v}(t)^\top \\
&- s(t)^{2-\frac{2}{N}}\left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right)\left[\boldsymbol{W}(t) - \boldsymbol{Z}\right]\boldsymbol{v}(t)\boldsymbol{v}(t)^\top \\
&- s(t)^{2-\frac{2}{N}}\boldsymbol{u}(t)\boldsymbol{u}(t)^\top\left[\boldsymbol{W}(t) - \boldsymbol{Z}\right]\left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right) \\
= &- Ns(t)^{1-\frac{2}{N}}\left(s(t) - \boldsymbol{u}(t)^\top\boldsymbol{Z}\boldsymbol{v}(t)\right)\boldsymbol{W}(t) \\
&+ s(t)^{2-\frac{2}{N}}\left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right)\boldsymbol{Z}\boldsymbol{v}(t)\boldsymbol{v}(t)^\top \\
&+ s(t)^{2-\frac{2}{N}}\boldsymbol{u}(t)\boldsymbol{u}(t)^\top\boldsymbol{Z}\left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right).
\end{aligned}$$

Substituting $\dot{\boldsymbol{W}}(t)$ back into Eqs. (16) and (17), we reach

$$\dot{\boldsymbol{u}}(t) = s(t)^{1-\frac{2}{N}} \left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right) \boldsymbol{Z}\boldsymbol{v}(t),\ \dot{\boldsymbol{v}}(t) = s(t)^{1-\frac{2}{N}} \left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right) \boldsymbol{Z}^\top\boldsymbol{u}(t).$$

□

Proposition B.8 (Stationary Singular Vector)**.** *If for time $T \geq 0$, $s(T) > 0, \dot{\boldsymbol{u}}(T) = \boldsymbol{0}, \dot{\boldsymbol{v}}(T) = \boldsymbol{0}$, then*

$$\boldsymbol{u}(T) = \pm\boldsymbol{u}_i, \boldsymbol{v}(T) = \pm\boldsymbol{v}_i, \text{for some } i \leq d, \text{or } \boldsymbol{u}(T) \perp \boldsymbol{u}_i, \boldsymbol{v}(T) \perp \boldsymbol{v}_i, \forall i \leq d.$$

Moreover, $\dot{\boldsymbol{u}}(t) = \boldsymbol{0}, \dot{\boldsymbol{v}}(t) = \boldsymbol{0}, \forall t \geq T$, that is, $\boldsymbol{u}(t) = \boldsymbol{u}(T), \boldsymbol{v}(t) = \boldsymbol{v}(T), \forall t \geq T$.

Proof. From $s(T) > 0, \dot{\boldsymbol{u}}(T) = \boldsymbol{0}, \dot{\boldsymbol{v}}(T) = \boldsymbol{0}$ and Eqs. (4a) and (4b), we obtain

$$\left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(T)\boldsymbol{u}(T)^\top\right) \boldsymbol{Z}\boldsymbol{v}(T) = \boldsymbol{0},\ \left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(T)\boldsymbol{v}(T)^\top\right) \boldsymbol{Z}^\top\boldsymbol{u}(T) = \boldsymbol{0}. \tag{18}$$

Hence, we could see

$$\boldsymbol{Z}\boldsymbol{v}(T) = \boldsymbol{u}(T)^\top\boldsymbol{Z}\boldsymbol{v}(T) \cdot \boldsymbol{u}(T),\ \boldsymbol{Z}^\top\boldsymbol{u}(T) = \boldsymbol{v}(T)^\top\boldsymbol{Z}^\top\boldsymbol{u}(T) \cdot \boldsymbol{v}(T), \tag{19}$$

showing that $\boldsymbol{Z}^\top\boldsymbol{Z}\boldsymbol{v}(T) = \left[\boldsymbol{u}(T)^\top\boldsymbol{Z}\boldsymbol{v}(T)\right]^2 \cdot \boldsymbol{v}(T)$, $\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{u}(T) = \left[\boldsymbol{u}(T)^\top\boldsymbol{Z}\boldsymbol{v}(T)\right]^2 \cdot \boldsymbol{u}(T)$. Thus, we can see $\boldsymbol{u}(T), \boldsymbol{v}(T)$ are the eigenvectors of $\boldsymbol{Z}^\top\boldsymbol{Z}, \boldsymbol{Z}\boldsymbol{Z}^\top$ with the same eigenvalue $\left[\boldsymbol{u}(T)^\top\boldsymbol{Z}\boldsymbol{v}(T)\right]^2$. Therefore, if $\boldsymbol{u}(T)^\top\boldsymbol{Z}\boldsymbol{v}(T) \neq 0$, we obtain $\boldsymbol{u}(T) = \pm\boldsymbol{u}_i, \boldsymbol{v}(T) = \pm\boldsymbol{v}_i$, for some $i \in [d]$. Otherwise, $\boldsymbol{u}(T)^\top\boldsymbol{Z}\boldsymbol{v}(T) = 0$. From Eq. (19), we obtain $\boldsymbol{Z}\boldsymbol{v}(T) = \boldsymbol{0}, \boldsymbol{Z}^\top\boldsymbol{u}(T) = \boldsymbol{0}$, showing that $\boldsymbol{u}(T) \perp \boldsymbol{u}_i, \boldsymbol{v}(T) \perp \boldsymbol{v}_i, \forall i \in [d]$.

Now we consider a substituted ODE started from T below

$$\dot{\tilde{s}}(t) = N\tilde{s}(t)^{2-\frac{2}{N}} \left(\boldsymbol{u}(T)^\top\boldsymbol{Z}\boldsymbol{v}(T) - \tilde{s}(t)\right), \tilde{s}(T) = s(T). \tag{20}$$

By Picard's existence and uniqueness theorem, such a solution $\tilde{s}^*(t)$ exists and is unique.

Recall the original ODE

$$\dot{\boldsymbol{u}}(t) = s(t)^{1-\frac{2}{N}} \left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right) \boldsymbol{Z}\boldsymbol{v}(t), \tag{21a}$$

$$\dot{\boldsymbol{v}}(t) = s(t)^{1-\frac{2}{N}} \left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right) \boldsymbol{Z}^\top\boldsymbol{u}(t), \tag{21b}$$

$$\dot{s}(t) = Ns(t)^{2-\frac{2}{N}} \left(\boldsymbol{u}(t)^\top\boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right). \tag{21c}$$

From Eq. (18), we could see $\boldsymbol{u}(t) = \boldsymbol{u}(T), \boldsymbol{v}(t) = \boldsymbol{v}(T), \forall t \geq T$ is a solution of Eqs. (21a) and (21b). Additionally, $\boldsymbol{u}(t)\boldsymbol{Z}\boldsymbol{v}(t) = \boldsymbol{u}(T)\boldsymbol{Z}\boldsymbol{v}(T), \forall t \geq T$. Hence, we can see $s(t) = \tilde{s}^*(t), \boldsymbol{u}(t) = \boldsymbol{u}(T), \boldsymbol{v}(t) = \boldsymbol{v}(T), \forall t \geq T$ is a solution of the original ODE. By the Picard's existence and uniqueness theorem, the solution is unique. i.e., we obtain that

$$\boldsymbol{u}(t) = \boldsymbol{u}(T), \boldsymbol{v}(t) = \boldsymbol{v}(T)$$

holds for all $t \geq T$. The proof is finished.

□

Lemma B.9. *If for certain $t \geq 0$, $\boldsymbol{u}(t)^\top\boldsymbol{Z}\boldsymbol{v}(t) = s(t) > 0$, and $\dot{\boldsymbol{u}}(t) \neq \boldsymbol{0}$ or $\dot{\boldsymbol{v}}(t) \neq \boldsymbol{0}$, then we have*

$$\frac{d\left(\boldsymbol{u}(t)^\top\boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right)}{dt} > 0.$$

Proof. Since $\dot{\boldsymbol{u}}(t) \neq \boldsymbol{0}$ or $\dot{\boldsymbol{v}}(t) \neq \boldsymbol{0}$, and $s(t) > 0$, by Eqs. (4a) and (4b), we obtain

$$\left\|\left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right) \boldsymbol{Z}\boldsymbol{v}(t)\right\|_2^2 + \left\|\left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right) \boldsymbol{Z}^\top\boldsymbol{u}(t)\right\|_2^2 > 0. \tag{22}$$

From the derivation of $d(\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - s(t))/dt$ and $ds(t)/dt$, we get

$$\begin{aligned}
&\frac{d\left(\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right)}{dt} \\
&\overset{\substack{(4c)\\(23)}}{=} s(t)^{1-\frac{2}{N}}\left[\left\|\left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right)\boldsymbol{Z}\boldsymbol{v}(t)\right\|_2^2 + \left\|\left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right)\boldsymbol{Z}^\top\boldsymbol{u}(t)\right\|_2^2 - Ns(t)\left(\boldsymbol{u}(t)^\top\boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right)\right] \\
&= s(t)^{1-\frac{2}{N}}\left[\left\|\left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right)\boldsymbol{Z}\boldsymbol{v}(t)\right\|_2^2 + \left\|\left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right)\boldsymbol{Z}^\top\boldsymbol{u}(t)\right\|_2^2\right] \overset{(22)}{>} 0,
\end{aligned}$$

where the second equality uses the assumption $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) = s(t)$. □

C. Missing Proofs

C.1. Proof of Proposition 2.2

Proof. 1). From Bah et al. (2021, Theorem 5) (i.e., Theorem B.3), we have $\boldsymbol{W}(t)$ converges. Thus $s(t) = \|\boldsymbol{W}(t)\|_F$ also converges, and not diverges to infinity. Applying Arora et al. (2019, Lemma 4) (i.e., Lemma B.2), we can see $s(t)$ obviously preserves the sign of its initial value.

2). $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t)$ is non-decreasing follows

$$\begin{aligned}
\frac{d\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t)}{dt} &= \frac{d\boldsymbol{u}(t)^\top}{dt}\boldsymbol{Z}\boldsymbol{v}(t) + \boldsymbol{u}(t)^\top \boldsymbol{Z}\frac{d\boldsymbol{v}(t)}{dt} \\
&\overset{(4a),(4b)}{=} s(t)^{1-\frac{2}{N}}\left[\left\|\left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right)\boldsymbol{Z}\boldsymbol{v}(t)\right\|_2^2 + \left\|\left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right)\boldsymbol{Z}^\top\boldsymbol{u}(t)\right\|_2^2\right] \geq 0.
\end{aligned} \tag{23}$$

Additionally, since $\|\boldsymbol{u}(t)\| = \|\boldsymbol{v}(t)\| = 1$, we have $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \leq s_1$. Hence, $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t)$ converges.

3). Using Eq. (14), we obtain

$$\begin{aligned}
\frac{da_1(t)b_1(t)}{dt} &= \frac{da_1(t)}{dt}\cdot b_1(t) + a_1(t)\cdot\frac{db_1(t)}{dt} \overset{(14)}{=} s(t)^{1-\frac{2}{N}}\left(s_1 b_1^2(t) + s_1 a_1^2(t) - 2a_1(t)b_1(t)\sum_{j=1}^d [s_j a_j(t)b_j(t)]\right) \\
&\geq s(t)^{1-\frac{2}{N}}\left(s_1 b_1^2(t) + s_1 a_1^2(t) - 2s_1|a_1(t)b_1(t)|\sum_{j=1}^d |a_j(t)b_j(t)|\right) \\
&\geq s(t)^{1-\frac{2}{N}}\left(s_1 b_1^2(t) + s_1 a_1^2(t) - 2s_1|a_1(t)b_1(t)|\right) = s_1 s(t)^{1-\frac{2}{N}}\left(|b_1(t)| - |a_1(t)|\right)^2 \geq 0,
\end{aligned}$$

where the second inequality uses Cauchy inequality:

$$\left(\sum_{j=1}^d |a_j(t)b_j(t)|\right)^2 \leq \left(\sum_{j=1}^d a_j^2(t)\right)\cdot\left(\sum_{j=1}^d b_j^2(t)\right) \leq \left(\sum_{j=1}^{d_y} a_j^2(t)\right)\cdot\left(\sum_{j=1}^{d_x} b_j^2(t)\right) = 1.$$

We note that $s_1 a_1(t)b_1(t) = s_1\cdot\boldsymbol{u}(t)^\top\boldsymbol{u}_1\cdot\boldsymbol{v}_1^\top\boldsymbol{v}(t) = \boldsymbol{u}(t)^\top\left(s_1\boldsymbol{u}_1\boldsymbol{v}_1^\top\right)\boldsymbol{v}(t) = \boldsymbol{u}(t)^\top\boldsymbol{Z}_1\boldsymbol{v}(t)$. Hence, we obtain $\boldsymbol{u}(t)^\top\boldsymbol{Z}_1\boldsymbol{v}(t)$ is non-decreasing. Moreover, since $\|\boldsymbol{u}(t)\| = \|\boldsymbol{v}(t)\| = 1$, we have $\boldsymbol{u}(t)^\top\boldsymbol{Z}_1\boldsymbol{v}(t) \leq s_1$. Hence, $\boldsymbol{u}(t)^\top\boldsymbol{Z}_1\boldsymbol{v}(t)$ also converges.

4). Using the derivation in the above, we obtain

$$\begin{aligned}
\frac{d\left(a_i(t) + b_i(t)\right)}{dt} &\overset{(14)}{=} s(t)^{1-\frac{2}{N}}\left[s_i\left(b_i(t) + a_i(t)\right) - \left(a_i(t) + b_i(t)\right)\sum_{j=1}^d [s_j a_j(t)b_j(t)]\right] \\
&= s(t)^{1-\frac{2}{N}}\left(a_i(t) + b_i(t)\right)\left(s_i - \sum_{j=1}^d [s_j a_j(t)b_j(t)]\right).
\end{aligned} \tag{24}$$

Moreover, $|a_i(t)+b_i(t)| = |\boldsymbol{v}_i^\top \boldsymbol{u}(t) + \boldsymbol{v}_i^\top \boldsymbol{v}(t)| \le 2$, showing that $a_i(t)+b_i(t)$ does not diverge to infinity. Hence, by Arora et al. (2019, Lemma 4), $a_i(t)+b_i(t)$ obviously preserves the sign of its initial value.

5). Since $a_i(0)+b_i(0)=0$, we get $a_i(t)+b_i(t)=0$ by 4), i.e.,

$$a_i(t) = -b_i(t), \forall i \in [k], t \ge 0. \tag{25}$$

Now we can bound

$$\begin{aligned}\sum_{j=k+1}^{d} s_j a_j(t) b_j(t) &\le s_{k+1} \sum_{j=k+1}^{d} |a_j(t) b_j(t)| \le s_{k+1} \sqrt{\sum_{j=k+1}^{d} a_j^2(t) \sum_{j=k+1}^{d} b_j^2(t)} \\ &\le s_{k+1} \sqrt{\left(1-\sum_{j=1}^{k} a_j^2(t)\right) \cdot \left(1-\sum_{j=1}^{k} b_j^2(t)\right)} \overset{(25)}{=} s_{k+1}\left(1-\sum_{j=1}^{k} a_j^2(t)\right).\end{aligned} \tag{26}$$

Hence, we obtain

$$\begin{aligned}s_{k+1} - \sum_{j=1}^{d} [s_j a_j(t) b_j(t)] &\overset{(25)}{=} s_{k+1} + \sum_{j=1}^{k} s_j a_j^2(t) - \sum_{j=k+1}^{d} [s_j a_j(t) b_j(t)] \\ &\overset{(26)}{\ge} s_{k+1}\left(1+\sum_{j=1}^{k} a_j^2(t)\right) - s_{k+1}\left(1-\sum_{j=1}^{k} a_j^2(t)\right) = 2 s_{k+1} \sum_{j=1}^{k} a_j^2(t) \ge 0.\end{aligned} \tag{27}$$

Now we consider the gradient of $a_{k+1}(t)+b_{k+1}(t)$:

$$\frac{d\left(a_{k+1}(t)+b_{k+1}(t)\right)}{dt} \overset{(24)}{=} s(t)^{1-\frac{2}{N}} \left(a_{k+1}(t)+b_{k+1}(t)\right)\left(s_{k+1} - \sum_{j=1}^{d} [s_j a_j(t) b_j(t)]\right). \tag{28}$$

If $a_{k+1}(t)+b_{k+1}(t) > 0$, from Eqs. (27) and (28) we can see $d\left(a_{k+1}(t)+b_{k+1}(t)\right)/dt \ge 0$. Thus, $a_{k+1}(t)+b_{k+1}(t)$ is non-decreasing. The case of $a_{k+1}(t)+b_{k+1}(t) < 0$ is similar. Therefore, we get $|a_{k+1}(t)+b_{k+1}(t)|$ is non-decreasing. Since $\|\boldsymbol{u}(t)\| = \|\boldsymbol{v}(t)\| = 1$, we have $|a_{k+1}(t)+b_{k+1}(t)| = |\boldsymbol{v}_{k+1}^\top \boldsymbol{u}(t) + \boldsymbol{v}_{k+1}^\top \boldsymbol{v}(t)| \le 2$. Hence, $|a_{k+1}(t)+b_{k+1}(t)|$ converges. Moreover, we note that from 4), $a_{k+1}(t)+b_{k+1}(t)$ preserves the sign of its initial value, showing that $\lim_{t\to+\infty} a_{k+1}(t)+b_{k+1}(t)$ exists. □

C.2. Proof of Theorem 3.4

Proof. From Lemma 3.2, we already know the convergent point $\overline{\boldsymbol{W}}(t)$ is still rank-one. Hence, using the facts shown in Bah et al. (2021, Theorem 28) that gradient flows avoid strict saddle points almost surely, and Bah et al. (2021, Proposition 33) that $L^1(\boldsymbol{W})$ on $\mathcal{M}_1$ satisfies the strict saddle point property, we could see gradient flow of $\boldsymbol{W}(t)$ converges to a global minimizer almost surely. □

C.3. Proof of Lemma 3.5

Proof. From 1) in Proposition 2.2 and $s(0) \neq 0$, we obtain $s(t) > 0, \forall t > 0$.

Case 1. If $\boldsymbol{u}(t_0)^\top \boldsymbol{Z} \boldsymbol{v}(t_0) > 0$ for some $t_0 \ge 0$, we get $\boldsymbol{W}(t_0) \in \mathcal{H}_b(\boldsymbol{Z})$ for some $b > 0$. Hence, by Lemma 3.2, we obtain $\bar{s} > 0$. From Eq. (4c), we obtain

$$0 = \lim_{t\to+\infty} N s(t)^{2-\frac{2}{N}} \left(\boldsymbol{u}(t)^\top \boldsymbol{Z} \boldsymbol{v}(t) - s(t)\right).$$

Using $s(t) \to \bar{s} > 0$ again, we obtain $\lim_{t\to+\infty} \boldsymbol{u}(t)^\top \boldsymbol{Z} \boldsymbol{v}(t)$ exists. Therefore,

$$0 = \lim_{t\to+\infty} \frac{d\boldsymbol{u}(t)^\top \boldsymbol{Z} \boldsymbol{v}(t)}{dt} \overset{(23)}{=} \lim_{t\to+\infty} s(t)^{1-\frac{2}{N}} \left[\left\|\left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right) \boldsymbol{Z} \boldsymbol{v}(t)\right\|_2^2 + \left\|\left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right) \boldsymbol{Z}^\top \boldsymbol{u}(t)\right\|_2^2\right].$$

By $s(t) \to \bar{s} > 0$, we obtain $\left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right) \boldsymbol{Z} \boldsymbol{v}(t) \to \boldsymbol{0}$, and $\left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right) \boldsymbol{Z}^\top \boldsymbol{u}(t) \to \boldsymbol{0}$. Thus, we can choose $t_n = n$ for example.

Case 2. If $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \le 0, \forall t \ge 0$. Then from Eq. (4c), we get $\dot{s}(t) \le 0, \forall t \ge 0$. Hence, $s(t) \le s(0)$. Moreover, by 2) in Proposition 2.2, we have $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \ge \boldsymbol{u}(0)^\top \boldsymbol{Z}\boldsymbol{v}(0)$. Therefore,

$$\dot{s}(t) \overset{(4c)}{=} N s(t)^{2-\frac{2}{N}} \left(\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right) \ge N s(t)^{2-\frac{2}{N}} \left(\boldsymbol{u}(0)^\top \boldsymbol{Z}\boldsymbol{v}(0) - s(0)\right). \tag{29}$$

Now we denote

$$C(a) := \inf_{t \ge a} \left\| \left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t)\boldsymbol{u}(t)^\top\right) \boldsymbol{Z}\boldsymbol{v}(t) \right\|_2^2 + \left\| \left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t)\boldsymbol{v}(t)^\top\right) \boldsymbol{Z}^\top \boldsymbol{u}(t) \right\|_2^2 \ge 0.$$

In the following, we show that $C(a) = 0, \forall a \ge 0$.

1) If $N = 2$, then we can see $\forall t \ge a$, by $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \le 0, \forall t \ge 0$,

$$\begin{aligned} -\boldsymbol{u}(a)^\top \boldsymbol{Z}\boldsymbol{v}(a) &\ge \boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - \boldsymbol{u}(a)^\top \boldsymbol{Z}\boldsymbol{v}(a) = \int_a^t \frac{d\boldsymbol{u}(x)^\top \boldsymbol{Z}\boldsymbol{v}(x)}{dx} dx \\ &\overset{(23)}{=} \int_a^t \left[\left\| \left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(x)\boldsymbol{u}(x)^\top\right) \boldsymbol{Z}\boldsymbol{v}(x) \right\|_2^2 + \left\| \left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(x)\boldsymbol{v}(x)^\top\right) \boldsymbol{Z}^\top \boldsymbol{u}(x) \right\|_2^2 \right] dx \ge C(a)(t-a). \end{aligned}$$

Taking $t \to +\infty$, we obtain $C(a) = 0, \forall a \ge 0$.

2) If $N > 2$, by solving Eq. (29), we get

$$\frac{N}{2-N} \cdot s(t)^{\frac{2}{N}-1} - \frac{N}{2-N} \cdot s(0)^{\frac{2}{N}-1} \ge N \left(\boldsymbol{u}(0)^\top \boldsymbol{Z}\boldsymbol{v}(0) - s(0)\right) t.$$

Therefore, we obtain

$$s(t)^{\frac{2}{N}-1} \le (N-2)\left(s(0) - \boldsymbol{u}(0)^\top \boldsymbol{Z}\boldsymbol{v}(0)\right) t + s(0)^{\frac{2}{N}-1} := A + Bt, A, B > 0. \tag{30}$$

Then we can see $\forall t \ge a$,

$$\begin{aligned} -\boldsymbol{u}(a)^\top \boldsymbol{Z}\boldsymbol{v}(a) \ &\ge \boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - \boldsymbol{u}(a)^\top \boldsymbol{Z}\boldsymbol{v}(a) = \int_a^t \frac{d\boldsymbol{u}(x)^\top \boldsymbol{Z}\boldsymbol{v}(x)}{dx} dx \\ &\overset{(23)}{=} \int_a^t s(x)^{1-\frac{2}{N}} \left[\left\| \left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(x)\boldsymbol{u}(x)^\top\right) \boldsymbol{Z}\boldsymbol{v}(x) \right\|_2^2 + \left\| \left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(x)\boldsymbol{v}(x)^\top\right) \boldsymbol{Z}^\top \boldsymbol{u}(x) \right\|_2^2 \right] dx \\ &\ge C(a) \int_a^t s(x)^{1-\frac{2}{N}} dx \overset{(30)}{\ge} C(a) \int_a^t \frac{1}{A+Bx} dx = \frac{C(a)}{B} \ln \frac{A+Bt}{A+Ba}. \end{aligned}$$

Taking $t \to +\infty$, we obtain $C(a) = 0, \forall a \ge 0$.

Therefore, combining 1) and 2), we conclude $C(a) = 0, \forall a \ge 0$. Hence, we can find a sequence $\{t_n\}$ with $t_n \to +\infty$, s.t.,

$$0 = \lim_{n \to +\infty} \left\| \left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t_n)\boldsymbol{u}(t_n)^\top\right) \boldsymbol{Z}\boldsymbol{v}(t_n) \right\|_2^2 + \left\| \left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t_n)\boldsymbol{v}(t_n)^\top\right) \boldsymbol{Z}^\top \boldsymbol{u}(t_n) \right\|_2^2.$$

Thus, we have

$$\lim_{n \to +\infty} \left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t_n)\boldsymbol{u}(t_n)^\top\right) \boldsymbol{Z}\boldsymbol{v}(t_n) = \boldsymbol{0}, \quad \lim_{n \to +\infty} \left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t_n)\boldsymbol{v}(t_n)^\top\right) \boldsymbol{Z}^\top \boldsymbol{u}(t_n) = \boldsymbol{0}. \tag{31}$$

Now we adopt the expansion following Eq. (12): $\boldsymbol{u}(t) = \sum_{i=1}^{d_y} a_i(t)\boldsymbol{u}_i,\ \boldsymbol{v}(t) = \sum_{i=1}^{d_x} b_i(t)\boldsymbol{v}_i$. Thus, we have

$$\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \overset{(13)}{=} \sum_{k=1}^{d} s_k a_k(t) b_k(t), \boldsymbol{Z}\boldsymbol{v}(t) \overset{(13)}{=} \sum_{j=1}^{d} s_j b_j(t) \boldsymbol{u}_j, \boldsymbol{Z}^\top \boldsymbol{u}(t) \overset{(13)}{=} \sum_{i=1}^{d} s_i a_i(t) \boldsymbol{v}_i.$$

Therefore, we obtain

$$\left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t_n)\boldsymbol{u}(t_n)^\top\right) \boldsymbol{Z}\boldsymbol{v}(t_n) = \sum_{i=1}^{d_y} \left[s_i b_i(t_n) - \left(\sum_{k=1}^{d} s_k a_k(t_n) b_k(t_n) \right) a_i(t_n) \right] \boldsymbol{u}_i,$$

where we utilize $s_i = 0, \forall i > d$. Since $\lim_{n\to+\infty} \left(\boldsymbol{I}_{d_y} - \boldsymbol{u}(t_n)\boldsymbol{u}(t_n)^\top\right) \boldsymbol{Z}\boldsymbol{v}(t_n) \overset{(31)}{=} \boldsymbol{0}$, and $\boldsymbol{u}_i$s are orthonormal basis, we obtain Eq. (6). Similarly, we could obtain Eq. (7) by $\lim_{n\to+\infty} \left(\boldsymbol{I}_{d_x} - \boldsymbol{v}(t_n)\boldsymbol{v}(t_n)^\top\right) \boldsymbol{Z}^\top \boldsymbol{u}(t_n) = \boldsymbol{0}$.

Finally, adding the equation in Eq. (6) and Eq. (7) with $i = j \in [d]$, we obtain

$$\lim_{n\to+\infty} \left(\sum_{k=1}^{d} s_k a_k(t_n) b_k(t_n) - s_i \right) (a_i(t_n) + b_i(t_n)) = 0, \forall i \in [d].$$

Since we have for some $i_0 \in [d]$ that $\lim_{n\to+\infty} a_{i_0}(t_n) + b_{i_0}(t_n)$ exists and not zero. Thus we obtain

$$\lim_{n\to+\infty} \boldsymbol{u}(t_n)^\top \boldsymbol{Z}\boldsymbol{v}(t_n) \overset{(13)}{=} \lim_{n\to+\infty} \sum_{k=1}^{d} s_k a_k(t_n) b_k(t_n) = s_{i_0}.$$

The proof is finished. □

C.4. Proof of Lemma 3.6

Proof. Since $a_i(t_n) + b_i(t_n) = 0, \forall i \in [k], n \geq 0$, we obtain

$$b_i(t_n) = -a_i(t_n), \forall i \in [k], n \geq 0, \tag{32}$$

and

$$\sum_{j=1}^{d} s_j a_j(t_n) b_j(t_n) = -\sum_{j=1}^{k} s_j a_j^2(t_n) + \sum_{j=k+1}^{d} s_j a_j(t_n) b_j(t_n) \overset{(26)}{\leq} s_{k+1} \left(1 - \sum_{j=1}^{k} a_j^2(t_n)\right).$$

Taking limit inferior in both sides and noting that $\lim_{n\to+\infty} \sum_{j=1}^{d} s_j a_j(t_n) b_j(t_n) = s_{k+1}$, we get

$$s_{k+1} \leq \liminf_{n\to+\infty} s_{k+1} \left(1 - \sum_{j=1}^{k} a_j^2(t_n)\right). \tag{33}$$

Moreover, naturally we have

$$\limsup_{n\to+\infty} s_{k+1} \left(1 - \sum_{j=1}^{k} a_j^2(t_n)\right) \leq s_{k+1}. \tag{34}$$

By Eq. (33) and Eq. (34), we obtain $\lim_{n\to+\infty} \sum_{j=1}^{k} a_j^2(t_n) = 0$, showing that

$$\lim_{n\to+\infty} -b_j(t_n) \overset{(32)}{=} \lim_{n\to+\infty} a_j(t_n) = 0, \forall j \in [k]. \tag{35}$$

Hence, we derive that

$$\lim_{n\to+\infty} \sum_{j=k+1}^{d} s_j a_j(t_n) b_j(t_n) = \lim_{n\to+\infty} \sum_{j=1}^{d} s_j a_j(t_n) b_j(t_n) - \lim_{n\to+\infty} \sum_{j=1}^{k} s_j a_j(t_n) b_j(t_n) \overset{(35)}{=} s_{k+1}. \tag{36}$$

Using Cauchy inequality, we have

$$\left[\sum_{j=k+1}^{d} s_j a_j(t_n) b_j(t_n)\right]^2 \leq \sum_{j=k+1}^{d} s_j^2 a_j^2(t_n) \cdot \sum_{j=k+1}^{d} b_j^2(t_n) \leq \left(\sum_{j=k+1}^{d} s_j^2 a_j^2(t_n)\right) \cdot \left(1 - \sum_{j=1}^{k} b_j^2(t_n)\right). \tag{37}$$

Since $\lim_{n\to+\infty} \sum_{j=1}^{k} b_j^2(t_n) \overset{(35)}{=} 0$, and $\sum_{j=1}^{d_y} a_j^2(t_n) \overset{(12)}{=} 1$, we obtain

$$\lim_{n\to+\infty} \sum_{j=k+1}^{d_y} s_{k+1}^2 a_j^2(t_n) \overset{(35)}{=} s_{k+1}^2 \overset{(36)}{=} \lim_{n\to+\infty} \left[\sum_{j=k+1}^{d} s_j a_j(t_n) b_j(t_n)\right]^2 \overset{(37)}{\leq} \liminf_{n\to+\infty} \sum_{j=k+1}^{d} s_j^2 a_j^2(t_n).$$

Noting that $s_j = 0, \forall j > d$, we get $0 \leq \liminf_{n\to+\infty} \sum_{j=k+2}^{d_y}(s_j^2 - s_{k+1}^2)a_j^2(t_n)$.

However, $s_j^2 - s_{k+1}^2 < 0, \forall j \geq k+2$ and $a_j^2(t_n) \geq 0$, showing that $\limsup_{n\to+\infty} \sum_{j=k+2}^{d_y}(s_j^2 - s_{k+1}^2)a_j^2(t_n) \leq 0$. Hence, we obtain $\lim_{n\to+\infty} a_j(t_n) = 0, \forall j \geq k+2$. The similar analysis holds for $b_j(t_n)$. Therefore, we obtain

$$\lim_{n\to+\infty} s_{k+1}a_{k+1}(t_n)b_{k+1}(t_n) = \lim_{n\to+\infty} \sum_{j=1}^{d} s_j a_j(t_n)b_j(t_n) = s_{k+1}.$$

Combining all the results, we derive that

$$\lim_{n\to+\infty} a_i(t_n) = \lim_{n\to+\infty} b_j(t_n) = 0, \forall i \in [d_y], j \in [d_x], i, j \neq k+1, \tag{38a}$$

$$\lim_{n\to+\infty} a_{k+1}(t_n)b_{k+1}(t_n) = 1. \tag{38b}$$

Finally, we have

$$\boldsymbol{u}_{k+1}\boldsymbol{v}_{k+1}^\top \overset{(38b)}{=} \lim_{n\to+\infty} a_{k+1}(t_n)b_{k+1}(t_n)\boldsymbol{u}_{k+1}\boldsymbol{v}_{k+1}^\top \overset{(38a)}{=} \lim_{n\to+\infty} \sum_{i,j} a_i(t_n)b_j(t_n)\boldsymbol{u}_i\boldsymbol{v}_j^\top \overset{(12)}{=} \lim_{n\to+\infty} \boldsymbol{u}(t_n)\boldsymbol{v}(t_n)^\top.$$

The proof is finished. □

C.5. Proof of Lemma 4.2

Proof. (I) The truth that $\dot{c}_1(t) \geq 0$ is direct from 3) in Proposition 2.2. Moreover, from Theorem 3.1 with $k = 0$, we have $c_1(t) = a_1(t)b_1(t) \to 1$. Thus, we obtain $t_1 < +\infty$.

(II) As for $s(t)$, if $t_2 = +\infty$, then $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \leq s(t), \forall t \in [0, +\infty)$. Thus by Eq. (4c), $\dot{s}(t) \leq 0$ for all $t \geq 0$. Now we consider the remaining case where $t_2 < +\infty$. Since $t_2 = \inf\{t : \boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \geq s(t)\}$, we have $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \leq s(t)$ when $t \in [0, t_2)$. Thus

$$\dot{s}(t) \overset{(4c)}{=} Ns(t)^{2-\frac{2}{N}} \left(\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right) \leq 0, \forall t \in [0, t_2).$$

Now from $t_2 < +\infty$, we get $\boldsymbol{u}(t_2)^\top \boldsymbol{Z}\boldsymbol{v}(t_2) = s(t_2)$. We denote $T := \inf\{t : \dot{\boldsymbol{u}}(t) = \boldsymbol{0}, \dot{\boldsymbol{v}}(t) = \boldsymbol{0}\}$.

(i) When $T > t_2$. Then for $t \in [t_2, T)$, we have $\dot{\boldsymbol{u}}(t) \neq \boldsymbol{0}$ or $\dot{\boldsymbol{v}}(t) \neq \boldsymbol{0}$. We also have $s(t) > 0, \forall t \geq 0$ from 1) in Proposition 2.2. Thus, applying Lemma B.9, we have

$$\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) = s(t) \Rightarrow d(\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - s(t))/dt > 0, \forall t \in [t_2, T).$$

By Lin et al. (2021, Lemma 10), we obtain $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \geq s(t), \forall t \in [t_2, T)$. Hence,

$$\dot{s}(t) = Ns(t)^{2-\frac{2}{N}} \left(\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right) \geq 0, \forall t \in [t_2, T).$$

And for $t \geq T$, we get $T < +\infty$. We obtain stationary singular vectors from time T by Proposition B.8. Thus, $\boldsymbol{u}(T)^\top \boldsymbol{Z}\boldsymbol{v}(T) = c, \forall t \geq T$ for a constant c, which reduce the variation of $s(t)$ as

$$\dot{s}(t) = Ns(t)^{2-\frac{2}{N}} \left(c - s(t)\right), \forall t \geq T.$$

Moreover, since $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \geq s(t), \forall t \in [t_2, T)$, we obtain $c = \boldsymbol{u}(T)^\top \boldsymbol{Z}\boldsymbol{v}(T) \geq s(T)$. Hence, we can see $\dot{s}(t) \geq 0, \forall t \geq T$.

(ii) When $T \leq t_2$, we have $T < +\infty$. We obtain stationary singular vectors from time T by Proposition B.8. Thus, $\boldsymbol{u}(T)^\top \boldsymbol{Z}\boldsymbol{v}(T) = c, \forall t \geq T$ for a constant c, which reduce the variation of $s(t)$ as

$$\dot{s}(t) = Ns(t)^{2-\frac{2}{N}} \left(c - s(t)\right), \forall t \geq T.$$

We note that $c = \boldsymbol{u}(t_2)^\top \boldsymbol{Z}\boldsymbol{v}(t_2) = s(t_2)$. Thus $\dot{s}(t) = 0, \forall t \geq t_2$.

Combining (i) and (ii), the proof is finished. □

C.6. Proof of Theorem 4.4

Proof. We first consider the upper bound. From $s(0) > 0$ and 1) in Proposition 2.2, we have $s(t) > 0$ for all $t \geq 0$. Moreover, we have

$$\dot{s}(t) \overset{(4c)}{=} Ns(t)^{2-\frac{2}{N}} \left(\sum_{j=1}^{d} s_j a_j(t) b_j(t) - s(t) \right) \overset{(26)}{\leq} Ns(t)^{2-\frac{2}{N}} \left(s_1 - s(t) \right). \tag{39}$$

Let $\tilde{s}(t)$ be the solution of the ODE

$$\dot{\tilde{s}}(t) = N\tilde{s}(t)^{2-\frac{2}{N}} \left(s_1 - \tilde{s}(t) \right), \quad \tilde{s}(0) = s(0).$$

Then we can see $s(t) \leq \tilde{s}(t)$ from Eq. (39).

If $s(0) > s_1$, then $\tilde{s}(t) \leq \tilde{s}(0) = s(0)$, showing that $s(t) \leq \tilde{s}(t) \leq s(0)$. Otherwise, $s(0) \leq s_1$, we get $\tilde{s}(t) \leq s_1$, showing that $s(t) \leq \tilde{s}(t) \leq s_1$. Therefore, we know $s(t) \leq s_0 := \max\{s_1, s(0)\}$ for all $t \geq 0$.

Now we consider the lower bound. Note that

$$\begin{aligned} \sum_{j=2}^{d} s_j a_j(t) b_j(t) &\overset{(26)}{\leq} s_2 \sqrt{(1 - a_1^2(t))(1 - b_1^2(t))} = s_2 \sqrt{a_1^2(t) b_1^2(t) - a_1^2(t) - b_1^2(t) + 1} \\ &\leq s_2 \sqrt{a_1^2(t) b_1^2(t) - 2|a_1(t) b_1(t)| + 1} = s_2 \left(1 - |a_1(t) b_1(t)| \right), \end{aligned} \tag{40}$$

where we use $|a_1(t) b_1(t)| = |\boldsymbol{u}_1^\top \boldsymbol{u}(t) \cdot \boldsymbol{v}_1^\top \boldsymbol{v}(t)| \leq 1$ in the last equality. Therefore, we derive that

$$\begin{aligned} \dot{s}(t) &\overset{(4c)}{=} Ns(t)^{2-\frac{2}{N}} \left(\sum_{j=1}^{d} s_j a_j(t) b_j(t) - s(t) \right) \overset{(40)}{\geq} Ns(t)^{2-\frac{2}{N}} \left[s_1 a_1(t) b_1(t) - s_2 \left(1 - |a_1(t) b_1(t)| \right) - s(t) \right] \\ &\geq Ns(t)^{2-\frac{2}{N}} \left[(s_2 - s_1) |a_1(t) b_1(t)| - s_2 - s(t) \right] \geq -N(s_1 + s(t)) s(t)^{2-\frac{2}{N}} \geq -N(s_1 + s_0) s(t)^{2-\frac{2}{N}}, \end{aligned} \tag{41}$$

where the last inequality uses $s(t) \leq s_0 := \max\{s_1, s(0)\}$, which is proved previously.

When $N = 2$, we solve Eq. (41) and get

$$\dot{s}(t) \overset{(41)}{\geq} -2(s_1 + s_0) s(t) \Rightarrow \frac{d\left(\ln s(t)\right)}{dt} \geq -2(s_1 + s_0) \Rightarrow \ln \frac{s(t)}{s(0)} \geq -2(s_1 + s_0) t \Rightarrow s(t) \geq s(0) e^{-2(s_1 + s_0)t}.$$

When $N \geq 3$, we solve Eq. (41) and get

$$\frac{N}{2 - N} \cdot \frac{d\left(s(t)^{\frac{2}{N} - 1} \right)}{dt} \overset{(41)}{\geq} -N(s_1 + s_0) \Rightarrow s(t)^{\frac{2}{N} - 1} - s(0)^{\frac{2}{N} - 1} \leq (s_1 + s_0)(N - 2) t.$$

Thus, we finally obtain

$$s(t) \geq \left[(s_1 + s_0)(N - 2) t + s(0)^{\frac{2}{N} - 1} \right]^{-\frac{N}{N-2}}.$$

The proof of Eqs. (9a) and (9b) is finished. □

C.7. Proof of Theorem 4.5

Proof. Since $a_1(0) b_1(0) < 0, a_1(0) + b_1(0) \neq 0$, without loss of generality, we suppose $a_1(0) > 0, b_1(0) < 0$ and $a_1(0) + b_1(0) > 0$. Note that

$$\dot{a}_1(t) - \dot{b}_1(t) \overset{(14)}{=} s(t)^{1-\frac{2}{N}} \left(b_1(t) - a_1(t) \right) \left(s_1 + \sum_{j=1}^{d} \left[s_j a_j(t) b_j(t) \right] \right).$$

By Arora et al. (2019, Lemma 4) and $|a_1(t) - b_1(t)| \leq 2$, we get that $a_1(t) - b_1(t)$ preserves the sign of its initial value:

$$a_1(t) - b_1(t) > 0, \forall t \geq 0. \tag{42}$$

Moreover, from 5) in Proposition 2.2 and $a_1(0) + b_1(0) > 0$, we obtain

$$a_1(t) + b_1(t) \geq a_1(0) + b_1(0) > 0, \forall t \geq 0. \tag{43}$$

Then we have

$$a_1(t) \overset{(42)}{\geq} \frac{a_1(t) + b_1(t)}{2} \overset{(43)}{\geq} \frac{a_1(0) + b_1(0)}{2} > 0, \forall t \geq 0, \tag{44}$$

and

$$-a_1(t) \overset{(43)}{\leq} b_1(t) \overset{(42)}{\leq} a_1(t), \forall t \geq 0 \overset{(44)}{\Rightarrow} -1 < \frac{b_1(t)}{a_1(t)} < 1, \forall t \geq 0. \tag{45}$$

Furthermore, we can derive that

$$\frac{d}{dt}\left(\frac{b_1(t)}{a_1(t)}\right) = \frac{\dot{b}_1(t)a_1(t) - \dot{a}_1(t)b_1(t)}{a_1^2(t)} \overset{(14)}{=} s(t)^{1-\frac{2}{N}}\left(1 - \left(\frac{b_1(t)}{a_1(t)}\right)^2\right) \overset{(45)}{>} 0.$$

Solving the above ODE, we obtain

$$d\left(\ln\sqrt{\frac{a_1(t) + b_1(t)}{a_1(t) - b_1(t)}}\right)/dt \geq s(t)^{1-\frac{2}{N}} \overset{(9b)}{\geq} \frac{1}{(s_1 + s_0)(N-2)t + s(0)^{\frac{2}{N}-1}}.$$

Therefore, we obtain

$$\begin{aligned}\ln\sqrt{\frac{a_1(t) + b_1(t)}{a_1(t) - b_1(t)}} - \ln\sqrt{\frac{a_1(0) + b_1(0)}{a_1(0) - b_1(0)}} &\geq \int_0^t \frac{dx}{(s_1 + s_0)(N-2)x + s(0)^{\frac{2}{N}-1}} \\ &= \frac{1}{(s_1 + s_0)(N-2)} \ln\left[1 + \frac{(s_1 + s_0)(N-2)t}{s(0)^{\frac{2}{N}-1}}\right].\end{aligned}$$

We hide constants related to initialization, and rewrite the inequality as

$$\frac{a_1(t) + b_1(t)}{a_1(t) - b_1(t)} \geq C_1(1 + A_1 t)^{B_1}, \tag{46}$$

where $A_1 := \frac{(s_1+s_0)(N-2)}{s(0)^{\frac{2}{N}-1}}$, $B_1 := \frac{2}{(s_1+s_0)(N-2)}$, $1 > C_1 := \frac{a_1(0)+b_1(0)}{a_1(0)-b_1(0)} \overset{(45)}{>} 0$. Hence, we obtain

$$a_1(t)b_1(t) \overset{(46),(44)}{\geq} \frac{C_1(1 + A_1 t)^{B_1} - 1}{C_1(1 + A_1 t)^{B_1} + 1} \cdot a_1^2(t). \tag{47}$$

Then we can see $a_1(t)b_1(t) \geq 0$ provided $C_1(1 + A_1 t)^{B_1} > 1$, i.e.,

$$t \geq T_1 = \frac{C_1^{-1/B_1} - 1}{A_1} = \frac{s(0)^{\frac{2}{N}-1}}{(s_1 + s_0)(N-2)} \cdot \left[\left(\frac{a_1(0) - b_1(0)}{a_1(0) + b_1(0)}\right)^{(s_1+s_0)(N-2)/2} - 1\right].$$

Therefore, we obtain $t_1 \leq T_1$. Moreover, when $t \leq T_1$, by $a_1^2(t) \leq 1$, we have

$$a_1(t)b_1(t) \overset{(47)}{\geq} \frac{C_1(1 + A_1 t)^{B_1} - 1}{C_1(1 + A_1 t)^{B_1} + 1}.$$

That is, $1 - a_1(t)b_1(t) = \mathcal{O}((1 + A_1 t)^{-B_1}) = \mathcal{O}([(N-2)t]^{-\frac{c_1}{N-2}})$.

Additionally, when $A_1 t \geq e \cdot C_1^{-1/B_1} - 1$, we have

$$A_1 t \geq e \cdot C_1^{-1/B_1} - 1 \geq \left(\frac{1 + B_1}{C_1}\right)^{1/B_1} - 1 \Rightarrow C_1(1 + A_1 t)^{B_1} \geq 1 + B_1. \tag{48}$$

Thus, we get

$$a_1(t)b_1(t) \overset{(47)}{\geq} \frac{C_1(1 + A_1 t)^{B_1} - 1}{C_1(1 + A_1 t)^{B_1} + 1} \cdot a_1^2(t) \overset{(44),(48)}{\geq} \frac{B_1}{2 + B_1} \cdot \frac{(a_1(0) + b_1(0))^2}{4} = \Theta\left(\frac{(a_1(0) + b_1(0))^2}{N}\right) > 0.$$

□

C.8. Proof of Theorem 4.7

Proof. Since $a_1(0)b_1(0) \geq 0$, then by 3) in Proposition 2.2, we know $a_1(t)b_1(t) \geq 0$ for all $t \geq 0$. Now we consider the flow of $a_1(t)b_1(t)$.

$$\begin{aligned}\frac{da_1(t)b_1(t)}{dt} &\overset{(14)}{=} s(t)^{1-\frac{2}{N}}\left(s_1 b_1(t)^2 + s_1 a_1(t)^2 - 2a_1(t)b_1(t)\sum_{j=1}^{d}\left[s_j a_j(t) b_j(t)\right]\right)\\ &\overset{(40)}{\geq} s(t)^{1-\frac{2}{N}}\left[2s_1 a_1(t)b_1(t) - 2a_1(t)b_1(t)\left(s_1 a_1(t)b_1(t) + s_2\left(1 - |a_1(t)b_1(t)|\right)\right)\right]\\ &= 2a_1(t)b_1(t)s(t)^{1-\frac{2}{N}}\left[s_1 - s_1 a_1(t)b_1(t) - s_2\left(1 - a_1(t)b_1(t)\right)\right]\\ &= 2a_1(t)b_1(t)s(t)^{1-\frac{2}{N}}\left(s_1 - s_2\right)\left(1 - a_1(t)b_1(t)\right).\end{aligned} \tag{49}$$

By the lower bound of $s(t)$ in Theorem 4.4, we obtain

$$\frac{da_1(t)b_1(t)}{dt} \overset{(49)}{\geq} 2a_1(t)b_1(t)s(t)^{1-\frac{2}{N}}\left(s_1 - s_2\right)\left(1 - a_1(t)b_1(t)\right) \overset{(9b)}{\geq} \frac{2a_1(t)b_1(t)\left(s_1 - s_2\right)\left(1 - a_1(t)b_1(t)\right)}{(s_0 + s_1)(N-2)t + s(0)^{\frac{2}{N}-1}}.$$

Denoting $c_1(t) := a_1(t)b_1(t) \leq 1$, we get

$$\ln\frac{c_1(t)}{1-c_1(t)} - \ln\frac{c_1(0)}{1-c_1(0)} \geq \frac{2(s_1 - s_2)}{(s_1 + s_0)(N-2)}\ln\left(1 + \frac{(s_1 + s_0)(N-2)t}{s(0)^{\frac{2}{N}-1}}\right).$$

Further we can rewrite the bound as

$$c_1(t) \geq 1 - \frac{1}{A(1 + B(N-2)t)^{\frac{c_2}{N-2}} + 1},$$

where $A = \frac{c_1(0)}{1-c_1(0)} > 0, B = (s_1 + s_0)s(0)^{1-\frac{2}{N}} > 0, c_2 = \frac{2(s_1 - s_2)}{s_1 + s_0} > 0$. Then we have $1 - a_1(t)b_1(t) = \mathcal{O}\left([(N-2)t]^{-c_2/(N-2)}\right)$. The proof is finished. □

C.9. Proof of Theorem 4.8

Proof. When $t_2 = +\infty$, we have $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) < s(t), \forall t \geq 0$. Thus, we obtain

$$\dot{s}(t) \overset{(4c)}{=} N s(t)^{2-\frac{2}{N}}\left(\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right) \leq 0. \tag{50}$$

We note that by Theorem 3.1, $s(t) \to s_1$. Thus, we conclude

$$s(t) \overset{(50)}{\geq} \lim_{t\to+\infty} s(t) = s_1. \tag{51}$$

Then we have

$$\begin{aligned}\frac{da_1(t)b_1(t)}{dt} &\overset{(49)}{\geq} 2a_1(t)b_1(t)s(t)^{1-\frac{2}{N}}\left(s_1 - s_2\right)\left(1 - a_1(t)b_1(t)\right)\\ &\overset{(51)}{\geq} 2a_1(t)b_1(t)s_1^{1-\frac{2}{N}}\left(s_1 - s_2\right)\left(1 - a_1(t)b_1(t)\right).\end{aligned}$$

Setting $c_1(t) := a_1(t)b_1(t)$ and solving the ODE in the above, we get

$$\ln\frac{c_1(t)}{1-c_1(t)} - \ln\frac{c_1(0)}{1-c_1(0)} \geq 2s_1^{1-\frac{2}{N}}\left(s_1 - s_2\right)t.$$

We rewrite the bound to

$$1 - c_1(t) \leq \left(1 + \frac{c_1(0)}{1-c_1(0)}\cdot e^{2s_1^{1-\frac{2}{N}}(s_1 - s_2)t}\right)^{-1}, i.e., 1 - c_1(t) = \mathcal{O}\left(e^{-c_3 t}\right).$$

To obtain the bound of $s(t)$, we notice that

$$\dot{s}(t) \overset{(4c)}{=} Ns(t)^{2-\frac{2}{N}} \left(\sum_{j=1}^{d} s_j a_j(t) b_j(t) - s(t) \right) \overset{(26)}{\leq} Ns(t)^{2-\frac{2}{N}} (s_1 - s(t)) \leq Ns_1^{2-\frac{2}{N}} (s_1 - s(t)) .$$

We can obtain the upper bound of the evolution $s(t)$ as

$$d\left(\ln\left(s(t) - s_1\right)\right)/dt \leq -Ns_1^{2-\frac{2}{N}} \Rightarrow s(t) \leq s_1 + (s(0) - s_1) e^{-Ns_1^{2-\frac{2}{N}} t}, i.e., s(t) - s_1 = \mathcal{O}\left(e^{-c_4 t}\right).$$

Finally, noting that $s(t) \overset{(51)}{\geq} s_1$, we obtain $|s(t) - s_1| = \mathcal{O}\left(e^{-c_4 t}\right)$. The proof is finished. □

C.10. Proof of Theorem 4.9

Proof. By Lemma 4.2, we have $\dot{s}(t) \geq 0, \forall t \geq 0$. Thus, we can lower bound $s(t) \geq s(0) > 0$. Then we obtain

$$\frac{da_1(t)b_1(t)}{dt} \overset{(49)}{\geq} 2a_1(t)b_1(t)s(t)^{1-\frac{2}{N}} (s_1 - s_2)(1 - a_1(t)b_1(t)) \geq 2a_1(t)b_1(t)s(0)^{1-\frac{2}{N}} (s_1 - s_2)(1 - a_1(t)b_1(t)) .$$

Setting $c_1(t) := a_1(t)b_1(t)$ and solving the ODE in the above, we get

$$1 - c_1(t) \leq \left(1 + \frac{c_1(0)}{1 - c_1(0)} \cdot e^{2s(0)^{1-\frac{2}{N}}(s_1 - s_2)t}\right)^{-1}, i.e., 1 - c_1(t) = \mathcal{O}\left(e^{-c_5 t}\right). \tag{52}$$

Next we derive the bound for $s(t)$. We continue from

$$\begin{aligned}
\dot{s}(t) &\overset{(4c)}{=} Ns(t)^{2-\frac{2}{N}} \left(\sum_{j=1}^{d} s_j a_j(t) b_j(t) - s(t) \right) \overset{(40)}{\geq} Ns(t)^{2-\frac{2}{N}} \left[s_1 a_1(t) b_1(t) - s_2 \left(1 - |a_1(t)b_1(t)|\right) - s(t) \right] \\
&\geq Ns(t)^{2-\frac{2}{N}} \left[(s_1 + s_2) a_1(t) b_1(t) - s_2 - s(t) \right] \overset{(52)}{\geq} Ns(0)^{2-\frac{2}{N}} \left[-(s_1 + s_2)\left(1 + Ae^{c_5 t}\right)^{-1} + s_1 - s(t) \right],
\end{aligned}$$

where $A = \frac{c_1(0)}{1 - c_1(0)} > 0$. Solving the above ODE, we arrive at

$$\frac{d(s(t)e^{c_6 t})}{dt} \geq c_6 e^{c_6 t} \left[s_1 - (s_1 + s_2)\left(1 + Ae^{c_5 t}\right)^{-1} \right].$$

Hence, we get

$$\begin{aligned}
&s(t)e^{c_6 t} - s(0) \geq s_1 \left(e^{c_6 t} - 1\right) - \int_0^t c_6 e^{c_6 x} (s_1 + s_2)\left(1 + Ae^{c_5 x}\right)^{-1} dx \\
&\geq s_1 \left(e^{c_6 t} - 1\right) - e^{c_6 t} \int_0^t c_6 (s_1 + s_2)\left(1 + Ae^{c_5 x}\right)^{-1} dx = s_1 \left(e^{c_6 t} - 1\right) - \frac{(s_1 + s_2) c_6 e^{c_6 t}}{c_5} \cdot \ln\left(1 + A^{-1} e^{-c_5 t}\right) \\
&\geq s_1 \left(e^{c_6 t} - 1\right) - \frac{(s_1 + s_2) c_6 e^{c_6 t}}{A c_5} \cdot e^{-c_5 t}.
\end{aligned}$$

Therefore, we obtain

$$s(t) \geq s_1 - (s(0) + s_1) e^{-c_6 t} - \frac{(s_1 + s_2) c_6}{A c_5} \cdot e^{-c_5 t}.$$

After hiding the constants and noting that $s(t)$ is non-decreasing, we obtain

$$s_1 - s(t) = \mathcal{O}\left(e^{-\min\{c_5, c_6\} t}\right).$$

We note that by Theorem 3.1, $s(t) \to s_1$. Since $s(t)$ is non-decreasing, we conclude $s(t) \leq \lim_{t \to +\infty} s(t) = s_1$. Then we obtain

$$|s_1 - s(t)| = \mathcal{O}\left(e^{-\min\{c_5, c_6\} t}\right).$$

The proof is finished. □

D. Convergence Rates: $N = 2$

We provide convergence rates of the case $N = 2$ in this section. Corresponding to the case $N \geq 3$, we list the rates of three stages as below:

Stage 1. For $t \in [0, t_1]$, where $t_1 := \inf\{t : a_1(t)b_1(t) \geq 0\} < +\infty$, we have $a_1(t)b_1(t) \leq 0$, and the rates are

$$1 - a_1(t)b_1(t) = \mathcal{O}(e^{-2t}), \quad s(t) = \Omega\left(e^{-2(s_1+s_0)t}\right).$$

Stage 2. For $t \in (t_1, t_2]$, where $t_2 := \inf\{t : \boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \geq s(t)\}$, we have $a_1(t)b_1(t) > 0, \dot{s}(t) \leq 0$, and

$$1 - a_1(t)b_1(t) = \mathcal{O}\left(e^{-2(s_1-s_2)t}\right), \quad s(t) = \Omega\left(e^{-2(s_1+s_0)t}\right).$$

Stage 3. For $t \in (\max\{t_1, t_2\}, +\infty)$, we have $a_1(t)b_1(t) > 0$ and $\dot{s}(t) \geq 0$, and the rates are

$$1 - a_1(t)b_1(t) = \mathcal{O}\left(e^{-2(s_1-s_2)t}\right), \quad |s_1 - s(t)| = \mathcal{O}\left(e^{-\min\{2(s_1-s_2), 2s(0)\}t}\right).$$

We have shown that the definitions of the stages are well defined by Lemma 4.2. The convergence rates of $s(t)$, i.e. $s(t) = \Omega\left(e^{-2(s_1+s_0)t}\right)$ in Stage 1 and Stage 2 are given by Theorem 4.4.

Convergence Rates of $a_1(t)b_1(t)$: Stage 1

Theorem D.1. *Suppose $N = 2$, $s(0) > 0, a_1(0)b_1(0) < 0$ and $a_1(0) + b_1(0) \neq 0$. Then we have*

$$1 - a_1(t)b_1(t) = \mathcal{O}(e^{-2t}), 0 \leq t \leq t_1.$$

Furthermore, we have the upper bound of t_1 below:

$$t_1 \leq \frac{1}{2} \ln \left| \frac{a_1(0) - b_1(0)}{a_1(0) + b_1(0)} \right|. \tag{53}$$

Additionally, we could obtain

$$a_1(t)b_1(t) = \Omega\left((a_1(0) + b_1(0))^2\right), \text{ if } t \geq \frac{1}{2} \ln \left| \frac{2(a_1(0) - b_1(0))}{a_1(0) + b_1(0)} \right|. \tag{54}$$

Proof. Since $a_1(0)b_1(0) < 0, a_1(0) + b_1(0) \neq 0$, without loss of generality, we suppose $a_1(0) > 0, b_1(0) < 0$ and $a_1(0) + b_1(0) > 0$. Note that

$$\dot{a}_1(t) - \dot{b}_1(t) \overset{(14)}{=} s(t)^{1-\frac{2}{N}} (b_1(t) - a_1(t)) \left(s_1 + \sum_{j=1}^{d} [s_j a_j(t) b_j(t)]\right) = (b_1(t) - a_1(t)) \left(s_1 + \sum_{j=1}^{d} [s_j a_j(t) b_j(t)]\right).$$

By Arora et al. (2019, Lemma 4) and $|a_1(t) - b_1(t)| \leq 2$, we get that $a_1(t) - b_1(t)$ preserves the sign of its initial value:

$$a_1(t) - b_1(t) > 0, \forall t \geq 0. \tag{55}$$

Moreover, from 5) in Proposition 2.2 and $a_1(0) + b_1(0) > 0$, we obtain

$$a_1(t) + b_1(t) \geq a_1(0) + b_1(0) > 0, \forall t \geq 0. \tag{56}$$

Then we have

$$a_1(t) \overset{(55)}{\geq} \frac{a_1(t) + b_1(t)}{2} \overset{(56)}{\geq} \frac{a_1(0) + b_1(0)}{2} > 0, \forall t \geq 0, \tag{57}$$

and

$$-a_1(t) \overset{(56)}{<} b_1(t) \overset{(55)}{<} a_1(t), \forall t \geq 0 \overset{(57)}{\Rightarrow} -1 < \frac{b_1(t)}{a_1(t)} < 1, \forall t \geq 0. \tag{58}$$

Furthermore, we can derive that

$$\frac{d}{dt}\left(\frac{b_1(t)}{a_1(t)}\right) = \frac{\dot{b}_1(t)a_1(t) - \dot{a}_1(t)b_1(t)}{a_1^2(t)} \overset{(14)}{=} \left(1 - \left(\frac{b_1(t)}{a_1(t)}\right)^2\right) \overset{(58)}{>} 0.$$

Then we have

$$d\left(\ln\sqrt{\frac{a_1(t)+b_1(t)}{a_1(t)-b_1(t)}}\right)/dt = 1 \Rightarrow \frac{b_1(t)}{a_1(t)} = \frac{e^{2t}\left[\frac{a_1(0)+b_1(0)}{a_1(0)-b_1(0)}\right]-1}{e^{2t}\left[\frac{a_1(0)+b_1(0)}{a_1(0)-b_1(0)}\right]+1}. \tag{59}$$

Thus, we get

$$a_1(t)b_1(t) = a_1^2(t)\cdot\frac{b_1(t)}{a_1(t)} \overset{(59)}{=} \frac{A_2e^{2t}-1}{A_2e^{2t}+1}\cdot a_1^2(t), A_2 := \frac{a_1(0)+b_1(0)}{a_1(0)-b_1(0)}. \tag{60}$$

Then we can see $a_1(t)b_1(t) \geq 0$ provided $A_2e^{2t} \geq 1$, i.e.,

$$t \geq T_2 := \frac{1}{2}\ln\frac{a_1(0)-b_1(0)}{a_1(0)+b_1(0)}.$$

Therefore, Eq. (53) is proved. Moreover, when $t \leq T_2$, by $a_1^2(t) \leq 1$, we have

$$a_1(t)b_1(t) \overset{(60)}{\geq} \frac{A_2e^{2t}-1}{A_2e^{2t}+1}.$$

That is, $1 - a_1(t)b_1(t) = \mathcal{O}(e^{-2t})$.

Additionally, when $t \geq \frac{1}{2}\ln\frac{2(a_1(0)-b_1(0))}{a_1(0)+b_1(0)}$, we have $A_2e^{2t} \geq 2$. Hence, we derive that

$$a_1(t)b_1(t) \overset{(60)}{=} \frac{A_2e^{2t}-1}{A_2e^{2t}+1}\cdot a_1^2(t) \overset{(57)}{\geq} \frac{(a_1(0)+b_1(0))^2}{12} = \Theta\left((a_1(0)+b_1(0))^2\right) > 0.$$

Thus, Eq. (54) is proved. □

Convergence Rates of $a_1(t)b_1(t)$: Stage 2 and Stage 3

Theorem D.2. *Assume $N = 2$, $s(0) > 0, a_1(0)b_1(0) > 0$. Then we have*

$$1 - a_1(t)b_1(t) = \mathcal{O}\left(e^{-2(s_1-s_2)t}\right).$$

Proof. Since $a_1(0)b_1(0) > 0$, then by 3) in Proposition 2.2, we know $a_1(t)b_1(t) > 0$ for all $t \geq 0$. Now we consider the flow of $a_1(t)b_1(t)$.

$$\begin{aligned}
\frac{da_1(t)b_1(t)}{dt} &\overset{(14)}{=} s(t)^{1-\frac{2}{N}}\left(s_1b_1(t)^2 + s_1a_1(t)^2 - 2a_1(t)b_1(t)\sum_{j=1}^{d}\left[s_ja_j(t)b_j(t)\right]\right) \\
&\overset{(40)}{\geq} s(t)^{1-\frac{2}{N}}\left[2s_1a_1(t)b_1(t) - 2a_1(t)b_1(t)\left(s_1a_1(t)b_1(t) + s_2\left(1 - |a_1(t)b_1(t)|\right)\right)\right] \\
&= 2a_1(t)b_1(t)\left[s_1 - s_1a_1(t)b_1(t) - s_2\left(1 - a_1(t)b_1(t)\right)\right] \\
&= 2a_1(t)b_1(t)\left(s_1 - s_2\right)\left(1 - a_1(t)b_1(t)\right).
\end{aligned} \tag{61}$$

Denoting $c_1(t) := a_1(t)b_1(t)$, by solving the ODE above, we obtain

$$\ln\frac{c_1(t)}{1-c_1(t)} - \ln\frac{c_1(0)}{1-c_1(0)} \overset{(61)}{\geq} 2(s_1-s_2)t.$$

Further we can rewrite the bound as

$$c_1(t) \geq 1 - \frac{1}{\frac{c_1(0)}{1-c_1(0)}e^{2(s_1-s_2)t}+1}. \tag{62}$$

Then we have $1 - a_1(t)b_1(t) = \mathcal{O}\left(e^{-2(s_1-s_2)t}\right)$. The proof is finished. □

Convergence Rates of $s(t)$: Stage 3

Similarly, before we start our analysis in Stage 3, we need to handle the minor case $t_2 := \inf\{t : \boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) \geq s(t)\} = +\infty$.

Theorem D.3. *Suppose $N = 2$, $s(0) > 0, a_1(0)b_1(0) > 0$ and $t_2 = +\infty$. Then we have*

$$|s(t) - s_1| = \mathcal{O}\left(e^{-2s_1 t}\right).$$

Proof. When $t_2 = +\infty$, by the defination of t_2, we have $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) < s(t), \forall t \geq 0$. Thus, we obtain

$$\dot{s}(t) \overset{(4c)}{=} N s(t)^{2-\frac{2}{N}} \left(\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right) = 2s(t)\left(\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t) - s(t)\right) \leq 0. \tag{63}$$

We note that by Theorem 3.1, $s(t) \to s_1$. Thus, we conclude

$$s(t) \overset{(63)}{\geq} \lim_{t\to+\infty} s(t) = s_1. \tag{64}$$

To obtain the bound of $s(t)$, we notice that

$$\begin{aligned}\dot{s}(t) &\overset{(4c)}{=} N s(t)^{2-\frac{2}{N}} \left(\sum_{j=1}^{d} s_j a_j(t) b_j(t) - s(t)\right) \overset{(26)}{\leq} N s(t)^{2-\frac{2}{N}} \left(s_1 - s(t)\right) \\ &\leq N s_1^{2-\frac{2}{N}} \left(s_1 - s(t)\right) = 2s_1 \left(s_1 - s(t)\right).\end{aligned}$$

By solving the ODE above, we can obtain the upper bound of the evolution $s(t)$ as

$$d\left(\ln\left(s(t) - s_1\right)\right)/dt \leq -2s_1 \Rightarrow s(t) \leq s_1 + (s(0) - s_1)e^{-2s_1 t}, i.e., s(t) - s_1 = \mathcal{O}\left(e^{-2s_1 t}\right).$$

Finally, noting that $s(t) \overset{(64)}{\geq} s_1$, we obtain $|s(t) - s_1| = \mathcal{O}\left(e^{-2s_1 t}\right)$. The proof is finished. □

Now we turn to the case $t_2 < +\infty$. We assume $a_1(0)b_1(0) > 0$ and $\dot{s}(0) \geq 0$ for short in Stage 3.

Theorem D.4. *Assume $N = 2$, $s(0) > 0, a_1(0)b_1(0) > 0$, and $\dot{s}(0) \geq 0$. Then we have*

$$|s_1 - s(t)| = \mathcal{O}\left(e^{-\min\{2(s_1 - s_2), 2s(0)\}t}\right).$$

Proof. By Lemma 4.2, we have $\dot{s}(t) \geq 0, \forall t \geq 0$. Thus, we can lower bound $s(t) \geq s(0) > 0$. Furthermore, we have

$$\begin{aligned}\dot{s}(t) &\overset{(4c)}{=} N s(t)^{2-\frac{2}{N}} \left(\sum_{j=1}^{d} s_j a_j(t) b_j(t) - s(t)\right) \overset{(40)}{\geq} N s(t)^{2-\frac{2}{N}} \left[s_1 a_1(t) b_1(t) - s_2\left(1 - |a_1(t)b_1(t)|\right) - s(t)\right] \\ &\geq N s(t)^{2-\frac{2}{N}} \left[(s_1 + s_2) a_1(t) b_1(t) - s_2 - s(t)\right] \overset{(62)}{\geq} 2s(0)\left[-(s_1 + s_2)\left(1 + Ae^{2(s_1 - s_2)t}\right)^{-1} + s_1 - s(t)\right],\end{aligned}$$

where $A = \frac{c_1(0)}{1 - c_1(0)}$. By solving the above ODE, we get

$$\frac{d(s(t)e^{2s(0)t})}{dt} \geq 2s(0)e^{2s(0)t}\left[s_1 - (s_1 + s_2)\left(1 + Ae^{2(s_1 - s_2)t}\right)^{-1}\right].$$

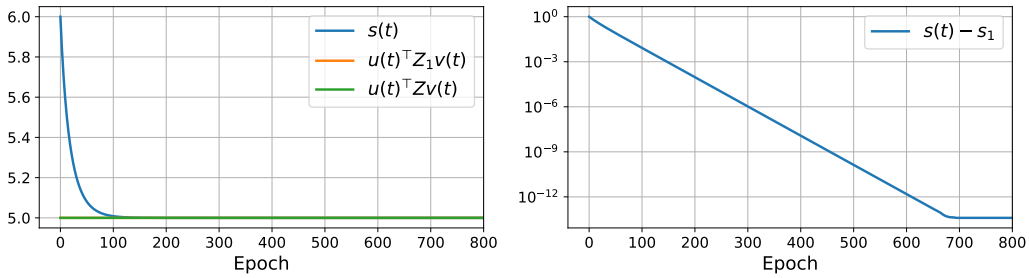


Figure 4. We choose $N = 6, d = 5$ with hidden-layer width $(d_N, \ldots, d_0) = (5, 4, 1, 10, 5, 3, 8)$, and set $k = 0$ in Theorem 3.1, i.e., the trajectory converges to the global minimizer. Learning rate is 5×10^{-4}. We choose $\boldsymbol{u}(0) = \boldsymbol{u}_1, \boldsymbol{v}(0) = \boldsymbol{v}_1, s(0) = 6$ to guarantee the minor case $t_2 = +\infty$. Left: Dynamics of $s(t), \boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t), \boldsymbol{u}(t)^\top \boldsymbol{Z}_1\boldsymbol{v}(t)$ during the whole optimization. (The lines of $\boldsymbol{u}(t)^\top \boldsymbol{Z}\boldsymbol{v}(t)$ and $\boldsymbol{u}(t)^\top \boldsymbol{Z}_1\boldsymbol{v}(t)$ are totally overlapped.) Right: Linear convergence of $s(t) - s_1$. Here t is the running step in gradient descent.

Hence, we get

$$\begin{aligned} s(t)e^{2s(0)t} - s(0) &\geq s_1\left(e^{2s(0)t} - 1\right) - \int_0^t 2s(0)e^{2s(0)x}(s_1 + s_2)\left(1 + Ae^{2(s_1 - s_2)x}\right)^{-1} dx \\ &\geq s_1\left(e^{2s(0)t} - 1\right) - e^{2s(0)t}\int_0^t 2s(0)(s_1 + s_2)\left(1 + Ae^{2(s_1 - s_2)x}\right)^{-1} dx \\ &= s_1\left(e^{2s(0)t} - 1\right) - \frac{2s(0)(s_1 + s_2)e^{2s(0)t}}{2(s_1 - s_2)} \cdot \ln\left(1 + A^{-1}e^{-2(s_1 - s_2)t}\right) \\ &\geq s_1\left(e^{2s(0)t} - 1\right) - \frac{s(0)(s_1 + s_2)e^{2s(0)t}}{A(s_1 - s_2)} \cdot e^{-2(s_1 - s_2)t}. \end{aligned}$$

Therefore, we obtain

$$s(t) \geq s_1 - (s(0) + s_1)e^{-2s(0)t} - \frac{s(0)(s_1 + s_2)}{A\,(s_1 - s_2)} \cdot e^{-2(s_1 - s_2)t}.$$

After hiding the constants, we obtain

$$s_1 - s(t) = \mathcal{O}\left(e^{-\min\{2(s_1 - s_2), 2s(0)\}t}\right).$$

We note that by Theorem 3.1, $s(t) \to s_1$. Since $s(t)$ is non-decreasing, we conclude

$$s(t) \leq \lim_{t \to +\infty} s(t) = s_1.$$

Then we obtain $|s_1 - s(t)| = \mathcal{O}\left(e^{-\min\{2(s_1 - s_2), 2s(0)\}t}\right)$. The proof is finished. □

E. Missing Experiments

To show the minor case of $t_2 = +\infty$ in Theorem 4.8, we adopt $\boldsymbol{u}(0) = \boldsymbol{u}_1, \boldsymbol{v}(0) = \boldsymbol{v}_1, s(0) > s_1$, which guarantees $\boldsymbol{u}(t) = \boldsymbol{u}_1, \boldsymbol{v}(t) = \boldsymbol{v}_1, \forall t \geq 0$. Thus, we only need to consider the variation of $s(t)$:

$$\dot{s}(t) \overset{(4c)}{=} Ns(t)^{2 - \frac{2}{N}}\left(s_1 - s(t)\right).$$

Then we obtain $s(t) > s_1, \forall t \geq 0$, leading to $\boldsymbol{u}(t)\boldsymbol{Z}\boldsymbol{v}(t) = s_1 < s(t), \forall t \geq 0$. Hence, $t_2 = +\infty$. The numerical results are shown in Figure 4, which match with the linear convergence of $s(t) - s_1$ well.