ON THE EXPLICIT ROLE OF INITIALIZATION ON THE CONVERGENCE AND GENERALIZATION PROPERTIES OF OVERPARAMETRIZED LINEAR NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks trained via gradient descent with random initialization and without any regularization enjoy good generalization performance in practice despite being highly overparametrized. A promising direction to explain this phenomenon is the *Neural Tangent Kernel* (NTK), which characterizes the implicit regularization effect of gradient flow/descent on infinitely wide neural networks with random initialization. However, a non-asymptotic analysis that connects generalization performance, initialization, and optimization for finite width networks remains elusive. In this paper, we present a novel analysis of overparametrized single-hidden layer linear networks, which formally connects initialization, optimization, and overparametrization with generalization performance. We exploit the fact that gradient flow preserves a certain matrix that characterizes the *imbalance* of the network weights, to show that the squared loss converges exponentially at a rate that depends on the level of imbalance of the initialization. Such guarantees on the convergence rate allow us to show that large hidden layer width, together with (properly scaled) random initialization, implicitly constrains the dynamics of the network parameters to be close to a low-dimensional manifold. In turn, minimizing the loss over this manifold leads to solutions with good generalization, which correspond to the min-norm solution in the linear case. Finally, we derive a novel $\mathcal{O}(h^{-1/2})$ upper-bound on the operator norm distance between the trained network and the min-norm solution, where h is the hidden layer width.

1 INTRODUCTION

Neural networks have shown excellent empirical performance in many application domains such as vision (Krizhevsky et al., 2012; Rawat & Wang, 2017), speech (Hinton et al., 2012; Graves et al., 2013) and video games (Silver et al., 2016; Vinyals et al., 2017). Among the many unexplained mysteries behind this success is the fact that gradient descent with random initialization and without explicit regularization enjoys good generalization performance despite being highly overparametrized.

A promising attempt to explain this phenomena is the *Neural Tangent Kernel* (NTK) (Jacot et al., 2018), which characterizes the implicit regularization effect of gradient flow/descent on infinitely wide neural networks with random initialization. Precisely, under this infinite width assumption, a proper initialization, together with gradient flow training, can be understood as a kernel gradient flow (NTK flow) of a functional that is constrained on a manifold that guarantees good generalization performance. The analysis further admits extensions to the finite width (Arora et al., 2019b; Buchanan et al., 2020). The core of the argument, illustrated in Figure 1, amounts to showing that (properly scaled) random initialization of networks with sufficiently large width, leads to trajectories that are, in some sense initialized close to the aforementioned manifold. Thus, approximately good initialization, together with acceleration, ensures that the training stays close to the NTK Flow.

Such analysis, however, leads to bounds on the network width that significantly exceed practical values (Buchanan et al., 2020), and seems to suggest the need of acceleration to achieve generalization. This motivates the following questions:

- *Is the kernel regime, which requires impractical bounds on the network width, necessary to achieve good generalization?*

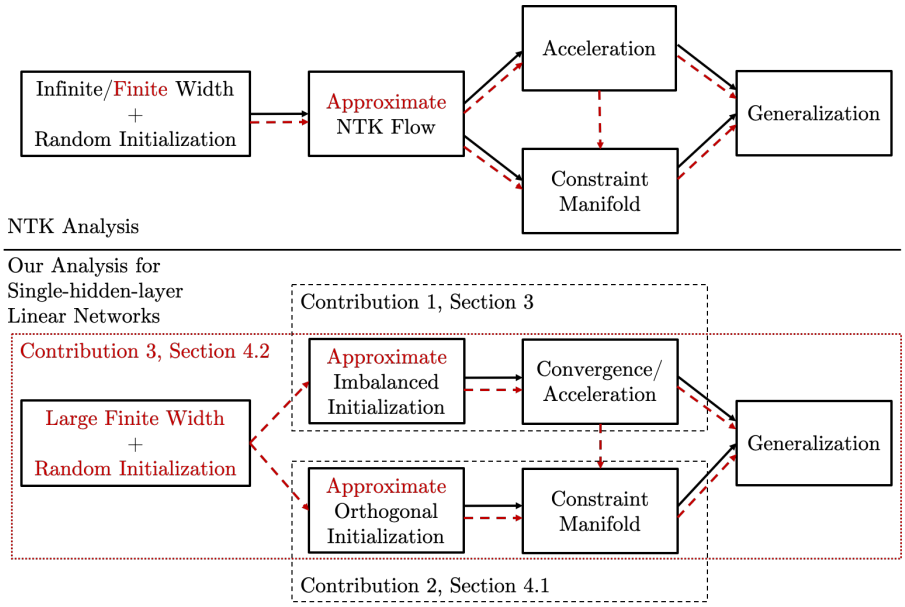

Figure 1: Comparing our analysis to asymptotic/non-asymptotic NTK analysis.

- *Does generalization depends explicitly on acceleration? Or is acceleration required only due to the choosing an initialization outside the good generalization manifold?*

For the simplified, yet certainly non-trivial, single-hidden layer linear network setting, this paper finds an answer to these questions.

Contributions. We present a novel analysis of the gradient flow dynamics of overparametrized single-hidden layer linear networks, which unveils the explicit and disentangled role of initialization on acceleration and generalization. Specifically, we show that imbalanced initialization ensures acceleration, while orthogonal initialization ensures that trajectories remain close to the generalization manifold. Interestingly, properly scaled random initialization of moderately wide networks is sufficient to ensure that initialization is approximately imbalanced and orthogonal, yet it is not necessary for either. More specifically, as illustrated in Figure 1, this paper makes the following contributions:

1. We show first that gradient flow on the squared-l_2 loss preserves a certain matrix-valued quantity, akin to constants of motion in mechanics or conservation laws of physics, that measures the *imbalance* of the network weights. Notably, the exponential rate of convergence of the loss directly depends on the level of imbalance, measured by certain singular value of the imbalance matrix, which is defined at initialization. Our analysis is non-probabilistic and valid under very mild assumptions, satisfied by moderately wide single-hidden-layer linear networks.

2. We characterize the existence of a low-dimensional manifold defined by a specific orthogonality condition on the parameters, which is invariant under the gradient flow. All trajectories within this manifold lead to a unique (w.r.t the end-to-end function) minimizer with good generalization performance, which corresponds to the min-norm solution. As a result, initializing the network within this manifold guarantees good generalization.

3. We further show that by randomly initializing the network weights using $\mathcal{N}(0, 1/h)$ (where h is the hidden layer width), i.e., the kernel regime, one can approximately satisfy both our sufficient imbalance and orthogonality conditions with high probability. Notably, the inaccurate initialization relative to the good generalization manifold requires acceleration to control the generalization error. In the context of NTK, our result further provide for linear networks a novel $\mathcal{O}\left(h^{-1/2}\right)$ upper-bound on the operator norm distance between the trained network and the min-norm solution. To the best of our knowledge, this is the first non-asymptotic bound regarding the generalization of linear networks in the global sense.

Notation. For a matrix A, we let A^T denote its transpose, $\mathrm{tr}(A)$ denote its trace, $\lambda_i(A)$ and $\sigma_i(A)$ denote its i-th eigenvalue and i-th singular value, respectively, in decreasing order (when adequate). We let $[A]_{ij}$, $[A]_{i,:}$, and $[A]_{:,j}$ denote the (i,j)-th element, the i-th row and the j-th column of A, respectively. We also let $\|A\|_2$ and $\|A\|_F$ denote the spectral norm and the Frobenius norm of A, respectively. For a scalar-valued or matrix-valued function of time, $F(t)$, we let $\dot{F} = \dot{F}(t) = \frac{d}{dt}F(t)$ denote its time derivative. Additionally, we let I_n denote the identity matrix of order n and $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution with mean μ and variance σ^2.

2 Related Work

Wide neural networks. For wide neural networks with random initialization, there is a line of work regarding the convergence (Du et al., 2019b;a; Du & Hu, 2019; Allen-Zhu et al., 2019b) and generalization (Allen-Zhu et al., 2019a; Arora et al., 2019a;b; Li & Liang, 2018; Cao & Gu, 2019; Buchanan et al., 2020). The behavior of such networks in their infinite width limit can be characterized by the *Neural Tangent Kernel* (NTK) (Jacot et al., 2018). With the concept of NTK, heuristically, training wide neural networks can be approximately viewed as kernel regression under gradient flow/descent (Arora et al., 2019b), hence the convergence and generalization can be understood by studying the non-asymptotic results regarding the equivalence of finite width networks to their infinite limit (Du et al., 2019b;a; Du & Hu, 2019; Allen-Zhu et al., 2019b; Arora et al., 2019a;b; Buchanan et al., 2020). More generally, such non-asymptotic results are related to the "lazy training" (Chizat et al., 2019; Du et al., 2019a; Allen-Zhu et al., 2019b), where the network weights do not deviate too much from its initialization during training. Our results for wide linear networks presented in Section 4.2 do not follow the NTK analysis, but provide an alternative, presumably more general view on the effect of random initialization when the hidden layer is sufficiently wide.

Convergence of linear networks. Convergence in overparametrized linear networks has been studied for both gradient flow (Saxe et al., 2014) and gradient descent (Bartlett et al., 2018; Arora et al., 2018a;b). In the kernel regime, Du & Hu (2019) applied the analysis of convergence of wide neural networks (Du et al., 2019b) to deep linear networks. Aside from the large hidden layer width and random initialization assumptions, Saxe et al. (2014) analyzed the trajectory of network parameters under spectral initialization, while Bartlett et al. (2018) studied the case of identity initialization. Although the fact that the imbalance is conservative under gradient flow has been exploited in Arora et al. (2018b;a), they consider the case of balanced initialization only to simplify the learning dynamics, and additional conditions are required for convergence. Most works mentioned above considered specific types of initialization that simplify the learning dynamics, while our convergence result with imbalance does not rely on simplified trajectory analysis thus works for much broader range of initialization schemes.

Min-norm solution in high-dimension linear regression. For high-dimensional under-determined linear regression, the asymptotic generalization error of min-norm solution has been studied in Hastie et al. (2019). In Bartlett et al. (2020); Mei & Montanari (2019), the min-norm solution was proved to have near-optimal generalization performance under mild assumptions on the data model. For our purpose, we study the generalization of a trained linear network as its distance to the min-norm solution. In this way, we refer to solutions with good generalization performance, as those with small distance to the min-norm solution.

3 Convergence Rate of Gradient Flow for Single-Hidden-Layer Linear Networks

In this section we study the convergence of gradient flow on squared l_2-loss for single-hidden-layer linear networks. Given training data of n samples $\{x^{(i)}, y^{(i)}\}_{i=1}^n$ with $x^{(i)} \in \mathbb{R}^D, y^{(i)} \in \mathbb{R}^m$, we aim to solve the linear regression problem

$$\min_{\Theta \in \mathbb{R}^{D\times m}} \mathcal{L} = \frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - \Theta^T x^{(i)})^2 \,, \tag{1}$$

by training a single-hidden-layer linear network $y = f(x; V, U) = VU^T x, V = \mathbb{R}^{m\times h}, U \in \mathbb{R}^{D\times h}$ with gradient flow, or equivalently, gradient descent with "infinitesimal step size", here h is the hidden layer width. We consider an *overparametrized* model such that $h \geq \min\{m, D\}$.

We rewrite the loss function with respect to our parameters V, U as

$$\mathcal{L}(V,U) = \frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - VU^T x^{(i)})^2 = \frac{1}{2}\|Y - XUV^T\|_F^2 \,, \tag{2}$$

where $Y = [y^{(1)}, \cdots, y^{(n)}]^T, X = [x^{(1)}, \cdots, x^{(n)}]^T$ are concatenations of the training data in rows. Assuming that the input data X has full rank, we consider the under-determined case $D > n$ for our regression problem, i.e., there are infinitely many solutions Θ^* to achieve zero loss of (1).

We will show that under certain conditions, the trajectory of the loss function $\mathcal{L}(t) = \mathcal{L}(V(t), U(t))$ under gradient flow of (2), i.e.,

$$\dot{V}(t) = -\frac{\partial}{\partial V}\mathcal{L}(V(t),U(t)) \,, \dot{U}(t) = -\frac{\partial}{\partial U}\mathcal{L}(V(t),U(t)) \,, \tag{3}$$

converges to 0 exponentially, and that proper initialization of $U(0), V(0)$ controls the convergence rate via a time-invariant matrix-valued term, the *imbalance* of the network.

3.1 Reparametrization of Gradient Flow

Given $D > n = \mathrm{rank}(X)$, the singular value decomposition of X is given by

$$X = W \begin{bmatrix} \Sigma_x^{1/2} & 0 \end{bmatrix} \begin{bmatrix} \Phi_1^T \\ \Phi_2^T \end{bmatrix} , W \in \mathbb{R}^{n\times n} \,, \Phi_1 \in \mathbb{R}^{D\times n} \,, \Phi_2 \in \mathbb{R}^{D\times(D-n)} \,, \tag{4}$$

with $WW^T = W^TW = \Phi_1^T\Phi_1 = I_n \,, \Phi_2^T\Phi_2 = I_{D-n} \,, \Phi_1^T\Phi_2 = 0$, and $\Phi_1\Phi_1^T + \Phi_2\Phi_2^T = I_D$.

Notice that

$$U = I_D U = (\Phi_1\Phi_1^T + \Phi_2\Phi_2^T)U = \Phi_1\Phi_1^T U + \Phi_2\Phi_2^T U \,,$$

hence we can reparametrize U as (U_1, U_2) using the bijection $g : \mathbb{R}^{n\times h} \times \mathbb{R}^{(D-n)\times h} \to \mathbb{R}^{D\times h} \,, U = g(U_1, U_2) = \Phi_1 U_1 + \Phi_2 U_2$, with inverse $(U_1, U_2) = g^{-1}(U) = (\Phi_1^T U, \Phi_2^T U)$.

We write the gradient flow in (3) explicitly as

$$\dot{V}(t) = \left(Y - XU(t)V^T(t)\right)^T XU(t) = E^T(t)\Sigma_x^{1/2}\Phi_1^T U(t) \,, \tag{5a}$$

$$\dot{U}(t) = X^T\left(Y - XU(t)V^T(t)\right)V(t) = \Phi_1\Sigma_x^{1/2}E(t)V(t) \,, \tag{5b}$$

based on the SVD of data X in (4), where

$$E(t) = E(V(t), U_1(t)) = W^TY - \Sigma_x^{1/2}U_1(t)V^T(t) \,, \tag{6}$$

is defined to be the *error*. Then from (5a)(5b) we obtain the dynamics in parameter space (V, U_1, U_2) as

$$\dot{V}(t) = E^T(t)\Sigma_x^{1/2}U_1(t) \,,\ \dot{U}_1(t) = \Sigma_x^{1/2}E(t)V(t) \,,\ \dot{U}_2(t) = 0 \,. \tag{7}$$

Notice that since W is orthogonal, we have

$$\mathcal{L}(t) = \frac{1}{2}\|Y - XU(t)V^T(t)\|_F^2 = \frac{1}{2}\|WE(t)\|_F^2 = \frac{1}{2}\|E(t)\|_F^2 \,. \tag{8}$$

Therefore it suffices to analyze the convergence rate of the error $E(t)$ under the dynamics of $V(t), U_1(t)$ in (7). As we mentioned in Section 1, the exponential convergence of $E(t)$, or equivalently the loss function $\mathcal{L}(t)$ is crucial for our analysis for generalization, in the sense that exponential convergence ensures the parameters do not deviate much away from its initialization, so that good properties from the initialization are approximately preserved during training.

3.2 Imbalance and Convergence Rate of the Error

We define the *imbalance* of the single-hidden-layer linear network under input data X as

$$\textit{Imbalance}:\ U_1^TU_1 - V^TV \,. \tag{9}$$

One can easily check that the imbalance term is time-invariant under the gradient flow (7), i.e. $\frac{d}{dt}[U_1^T(t)U_1(t) - V^T(t)V(t)] \equiv 0$, hence the imbalance is fully determined by its initialization.

The imbalance is a $h \times h$ matrix with rank at most $m + n$, the rank of imbalance characterizes how much the row spaces of U_1 and V are misaligned. We show that a rank-$(m + n - 1)$ imbalance is sufficient for exponential convergence of the error $E(t)$, or equivalently, the loss function.

Now we present the result regarding convergence of the error. (see Appendix C for the proof).

Theorem 1 (Convergence of linear networks with sufficient rank of imbalance)**.** *Suppose* $h \geq m + n - 1$. *Let* $V(t), U_1(t), t > 0$ *be the trajectory of continuous dynamics* (7) *starting from some* $V(0), U_1(0)$. *If*

$$\sigma_{n+m-1}\left(U_1^T(0)U_1(0) - V^T(0)V(0)\right) = c > 0\,, \tag{10}$$

then for $E(t)$ *defined in* (6)*, we have*

$$\|E(t)\|_F^2 \leq \exp\left(-2\sigma_n(\Sigma_x)ct\right)\|E(0)\|_F^2,\ \forall t > 0\,. \tag{11}$$

Additionally, $V(t), U_1(t), t > 0$ *converges to some equilibrium point* $(V(\infty), U_1(\infty))$ *such that* $E(V(\infty), U_1(\infty)) = 0$.

The result is simple but surprising: If we initialize our parameters deliberately to ensure sufficient rank of imbalance, then the exponential convergence of the loss function $\mathcal{L}(t)$ is guaranteed with convergence rate depending on singular values of the data and the imbalance.

The fact that the imbalance is preserved under gradient flow has been exploited in Arora et al. (2018a;b), where imbalance is assumed to be zero (or small), such that the learning dynamics can be expressed in closed form with respect to the end-to-end matrix. This analysis, requires, however, additional assumptions on the initialization of the end-to-end matrix for acceleration. Similarly, though in a more general setting, Du et al. (2018) showed that the imbalance is preserved, and proves convergence under a small imbalance assumption. Acceleration (exponential rate), however, is not guaranteed. Exploiting imbalance for guaranteeing acceleration was first presented in Saxe et al. (2014), under a spectral initialization assumption. In contrast, Theorem 1 shows acceleration without the spectral initialization condition. Rather, we show that under very mild conditions on the alignment between the initialization and the data, acceleration is achieved.

Such good choice of initialization, at first glance, seems to largely depend on the data given the definition of the imbalance. However, we show in the next section that for sufficiently wide networks with random initialization, the imbalance has rank at least $n + m$ with high probability, for any data matrix X, hence exponential convergence is almost guaranteed when training such networks. Later we will illustrate how such convergence also affects the generalization of the trained network.

The dependence on $\sigma_n(\Sigma_x)$ has been also shown in Du & Hu (2019). There it was shown the linear convergence of gradient descent in a multi-layer linear networks. Their proof followed the same high-level procedure as in showing convergence for networks with nonlinear activation (Du et al., 2019b;a), which relied on showing the Gram matrix is close to its initialization during training. For our result, although it is provided for single-hidden-layer linear networks, we essentially lower bound the smallest eigenvalue of the Gram matrix at any time t by a fixed constant that only depends on the initialization.

We end the section by noting that our result is not limited to extremely wide networks with random initialization, but rather provide a universal condition on the initialization that enforces exponential convergence, provided a mild requirement on the hidden layer width $h \geq n + m - 1$. While we focus on training for under-determined regression problems $D \geq n$ and X is assumed to be full rank, we show that similar results hold for other cases as well, with minor reformulations. Please, see Appendix A.

4 Generalization of Single-hidden-layer Linear Network

In this section, we study the generalization properties of trained single-hidden-layer linear networks under gradient flow. Assuming that $D > n = \text{rank}(X)$, the regression problem (1) has infinitely many solutions Θ^* that achieve zero loss. Among all these solutions, one that is of particular interest in high-dimensional linear regression is the *minimum norm solution* (min-norm solution)

$$\hat{\Theta} = \underset{\Theta \in \mathbb{R}^{D\times m}}{\arg\min}\{\|\Theta\|_2 : Y - X\Theta = 0\} = X^T(XX^T)^{-1}Y, \tag{12}$$

which has near-optimal generalization error for suitable data models, as shown in (Bartlett et al., 2020; Mei & Montanari, 2019). Here, we study conditions under which our trained network is equal or close to the min-norm solution by showing how the initialization explicitly controls the trajectory of the training parameters to be exactly (or approximately) confined within some low-dimensional manifold. In turn, minimizing the loss over this manifold leads to the min-norm solution.

Moreover, our analysis on constrained learning applies to wide single-hidden-layer linear networks with random initialization, whose infinite width limit is equivalent to the kernel predictor with linear kernel $K(x, x') = x^T x'$, as suggested by Jacot et al. (2018). One can easily check that such a kernel predictor is the min-norm solution $\hat{\Theta}$. In addition, we show that the operator norm distance between trained finite width single-hidden-layer linear network and the min-norm solution is upper bounded by a $\mathcal{O}(h^{-1/2})$ term with high probability over random initialization.

4.1 Decomposition of Trained Network

To begin with, notice that the linear operator $UV^T \in \mathbb{R}^{D\times m}$ associated with the single-hidden-layer linear network can be decomposed according to the data matrix X as

$$UV^T = (\Phi_1\Phi_1^T + \Phi_2\Phi_2^T)UV^T = \Phi_1 U_1 V^T + \Phi_2 U_2 V^T\,, \tag{13}$$

where Φ_1, Φ_2, U_1, U_2 are previously defined in Section 3. Here $[UV^T]_{:,j}$, i.e., the j-th column of UV^T, is the linear predictor for the j-th output y_j, and is decomposed into two components within complementary subspaces $\mathrm{span}(\Phi_1)$ and $\mathrm{span}(\Phi_2)$. Moreover $[U_1V^T]_{:,j}$ is the coordinate of $[UV^T]_{:,j}$ w.r.t. the orthonormal basis as the columns of Φ_1, and similarly $[U_2V^T]_{:,j}$ is the coordinate w.r.t. basis Φ_2. Clearly, under gradient flow (3), the trajectory $(U(t)V(t)^T, t >)0$ is fully determined by the trajectory $(U_1(t)V^T(t), U_2(t)V^T(t), t > 0)$, which is governed by the dynamics (7).

Convergence of Training Parameters. We have derived useful results regarding $U_1(t)V^T(t)$ for $t > 0$ in Section 3. By Theorem 1, if the rank of the imbalance matrix is large enough, $U_1(t)V^T(t)$ converges to some $U_1(\infty)V^T(\infty)$ and the stationary point satisfies $W^TY - \Sigma_x^{1/2}U_1(\infty)V^T(\infty) = 0$, which implies $U_1(\infty)V^T(\infty) = \Sigma_x^{-1/2}W^TY$. Then it is easy to check that

$$\Phi_1 U_1(\infty)V^T(\infty) = \Phi_1\Sigma_x^{-1/2}W^TY = X^T(XX^T)^{-1}Y = \hat{\Theta}\,. \tag{14}$$

In other words, the projected trajectory (in columns) of $U(t)V^T(t)$ onto $\mathrm{span}(\Phi_1)$ converges exactly to the min-norm solution.

For $U_2(t)V^T(t)$, notice that $\dot{U}_2(t) = 0$ in dynamics (7), hence $U_2(t) = U_2(0), \forall t > 0$. Then under sufficient rank of imbalance, $U(t)V^T(t)$ converges to some $U(\infty)V^T(\infty)$ and

$$U(\infty)V^T(\infty) = \Phi_1 U_1(\infty)V^T(\infty) + \Phi_2 U_2(0)V^T(\infty) = \hat{\Theta} + \Phi_2 U_2(0)V^T(\infty)\,.$$

Therefore $U_2(0)V^T(\infty)$ quantifies how much the trained network $U(\infty)V^T(\infty)$ is deviated from the min-norm solution $\hat{\Theta}$, and since $\Phi_2^T\Phi_2 = I_{D-n}$, we have

$$\|U(\infty)V^T(\infty) - \hat{\Theta}\|_2 = \|\Phi_2 U_2(0)V^T(\infty)\|_2 = \|U_2(0)V^T(\infty)\|_2\,. \tag{15}$$

Constrained Training via Initialization. Based on our analysis above, initializing $U_2(0)$ such that $U_2(0)V^T(\infty) = 0$ in the limit, guarantees convergence to the min-norm solution via (15). However, this is not easily achievable, as one needs to know a priori $V(\infty)$. Instead, we can show that by choosing a proper initialization, one can constrain the trajectory of the matrix $U(t)V^T(t)$ to lie identically in the set $\Phi_2^T U(t)V^T(t) \equiv 0$ for all $t \geq 0$, which is equivalent to saying that the columns of $U(t)V^T(t)$ lie in $\mathrm{span}(\Phi_1)$. Indeed, using the fact that for all $t \geq 0$, $U_2(t) = U_2(0)$ and $\Phi_2^T U(t)V^T(t) \equiv 0$ we obtain

$$0 = \Phi_2^T U(t)V^T(t) = \Phi_2^T U(\infty)V^T(\infty) = U_2(\infty)V^T(\infty) = U_2(0)V^T(\infty)\,. \tag{16}$$

Therefore, using (15), it follows that $U(\infty)V^T(\infty) = \hat{\Theta}$, as desired.

Here, the constraint that all columns of $U(t)V^T(t)$ lie in $\mathrm{span}(\Phi_1)$ is equivalent to the constraint that (V, U) is within some low-dimensional manifold in the parameter space. More importantly, such constraint on UV^T is relevant to generalization: When a column of UV^T is in $\mathrm{span}\,(\Phi_2)$, predictions are made based on features in $\mathrm{span}(\Phi_2)$. However, those features are not present in the data X that spans Φ_1, thus, intuitively, hurting the generalization performance.

To enforce the constraint $\Phi_2^T U(t)V^T(t) \equiv 0$, consider the dynamics of $U_2(0)V^T(t)$, or equivalently $V(t)U_2^T(0)$. From (7) we have

$$\frac{d}{dt}\begin{bmatrix} V(t)U_2^T(0) \\ U_1(t)U_2^T(0)\end{bmatrix} = \begin{bmatrix} 0 & E^T(t)\Sigma_x^{1/2} \\ \Sigma_x^{1/2}E(t) & 0\end{bmatrix}\begin{bmatrix} V(t)U_2^T(0) \\ U_1(t)U_2^T(0)\end{bmatrix}\,. \tag{17}$$

The most straightforward way to enforce $V(t)U_2^T(0) = 0, \forall t > 0$ is to properly initialize $V(0), U(0)$ such that $V(0)U_2^T(0) = 0$ and $U_1(0)U_2^T(0) = 0$. To have such a proper initialization, one can

- Initialize columns of $U(0)$ in $\mathrm{span}(\Phi_1)$, which leads to $U_2(0) = 0$;
- Initialize $U(0)$ and $V(0)$ to enforce the orthogonality condition on the rows, i.e. $V(0)U^T(0) = 0$ and $U(0)U^T(0) = I_D$.

Such initialization guarantees that gradient flow is constrained within some low-dimensional manifold in the parameter space, such that any global minimizer of the loss in that manifold corresponds to the min-norm solution. Therefore, whenever the network parameters in this manifold converge, and $E(\infty) = 0$, then the solution must be the minimum-norm one. While in practice we can make the initialization exactly as above, such choice is data-dependent and requires the SVD of the data matrix X.

In the next section, we show that under random initialization and sufficiently large hidden layer width h, these conditions on initialization are approximately satisfied, i.e., with high probability the rank of imbalance is $m + n$ and $\|V(0)U_2^T(0)\|_F, \|U_1(0)U_2^T(0)\|_F \sim \mathcal{O}(h^{-1/2})$. So that the trajectory $U(t)V^T(t), t > 0$ will be approximately constrained in the subspace as we mentioned above.

4.2 Wide single-hidden-layer Linear Network

We will now discuss the generalization of wide single-hidden-layer linear network with random initialization. In particular, we will show how the previously mentioned conditions for convergence and good generalization, i.e., high imbalance and orthogonality, are approximately satisfied with high probability under the following initialization

$$[U(0)]_{ij} \sim \mathcal{N}\left(0, \frac{1}{h}\right),\ 1 \le i \le D, 1 \le j \le h\,,$$
$$[V(0)]_{ij} \sim \mathcal{N}\left(0, \frac{1}{h}\right),\ 1 \le i \le m, 1 \le j \le h\,,$$

where all the entries are independent. Our analysis indeed highlights the need, within this regime, of exponential convergence to ensure good generalization.

Previous works Jacot et al. (2018) have suggested in the limit $h \to \infty$, the trained network is equivalent to the kernel predictor with NTK. For linear networks, the NTK is the linear kernel $K(x, x') = x^T x'$ whose corresponding kernel predictor is the min-norm solution. Therefore, as $h \to \infty$, we should expect the trained network to converge to the min-norm solution. Combining what we have discussed regarding the convergence and generalization of linear networks with basic random matrix theory, we are able to derive the non-asymptotic $\mathcal{O}(h^{-1/2})$ bound on the operator norm distance between a trained h-width network under random initialization and the min-norm solution.

Recall in the last section, one can obtain exactly min-norm solution via proper initialization of the single-hidden-layer network. In particular, it requires 1) convergence of the error $E(t)$ to zero; and 2) the orthogonality conditions $V(0)U_2^T(0) = 0$ and $U_1(0)U_2^T(0) = 0$. Under random initialization and sufficiently large hidden layer width h, these two conditions are approximately satisfied. Using basic random matrix theory, one can show that with high probability, the rank of imbalance is $m+n$, which leads to (exponential) convergence of $E(t)$, and $\|V(0)U_2^T(0)\|_F, \|U_1(0)U_2^T(0)\|_F \sim \mathcal{O}(h^{-1/2})$. For the formal statement, see Lemma D.2 in Appendix D.

Clearly, $\|V(0)U_2^T(0)\| \sim \mathcal{O}(h^{-1/2})$ alone is insufficient to have the final bound on the trained network. However, from (17) we show that as long as the error $E(t)$ converges to 0 exponentially, which is guaranteed by sufficient rank of imbalance with high probability, the final deviation from min-norm solution $\|V(\infty)U_2^T(0)\|_F$ can not exceed $C(\|V(0)U_2^T(0)\|_F + \|U_1(0)U_2^T(0)\|_F)$ for some constant C that depends on the data and the convergence rate of $E(t)$, leading to the desired bound. The formal statement is summarized in the following theorem.

Theorem 2 (Generalization of wide single-hidden-layer linear network)**.** *Let* $(V(t), U(t), t > 0)$ *be a trajectory of continuous dynamics* (7)*. Then,* $\exists C > 0$*, such that* $\forall \delta \in (0,1), \forall h > h_0 = poly\left(m, D, \frac{1}{\delta}, \frac{\sigma_1(\Sigma_x)}{\sigma_n^3(\Sigma_x)}\right)$*, with probability* $1-\delta$ *over random initializations, we have*

$$\|U(\infty)V^T(\infty) - \hat{\Theta}\|_2 \leq C\sqrt{m+n}\frac{\sqrt{m+D} + \frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}}\,, \tag{18}$$

where C *depends on the data* X, Y*.*

The proof is shown in Appendix D. This is, to our best knowledge, the first non-asymptotic bound in the global (operator) sense regarding the trained wide neural networks. Although we understand that this can not be directly compared to previous works (Arora et al., 2019b; Buchanan et al., 2020) that show non-asymptotic results connecting a trained network to the kernel predictor by the NTK, using more general network structure and activation functions than that of the linear network, we believe this theorem is a clear illustration of how overparametrization, in particular the hidden layer width, together with random initialization affects the convergence and generalization, beyond the kernel regime.

To be specific, regarding the non-asymptotic analysis for wide neural networks, the concept of constrained learning presented in this section and used to prove Theorem 2 is more general than previous works (Arora et al., 2019b), where one requires sufficiently large hidden layer width such that the trajectory of the network in the function space approximately matches the trajectory in the infinite width limit. Loosely speaking, such a large width h enforces the trajectory to be approximately confined within a one-dimension manifold, parametrized only by t, and independent of the initialization. We have shown here, however, that even for relative smaller width h, there is a larger dimensional manifold that provides good generalization performance. This shows that, while the kernell regime maybe sufficient, it is certainly, at least for the linear single-hiden layer network, not necessary to guarantee good generalization.

5 Conclusion

In this paper, we study the explicit role of initialization on controlling the convergence and generalization of single-hidden-layer linear networks trained under gradient flow. First of all, initializing the imbalance to have sufficient rank leads to the exponential convergence of the loss. Then proper initialization enforces the trajectory of network parameters to be exactly (or approximately) constrained in a low-dimensional manifold, over which minimizing the loss yields the min-norm solution. Combining those results, we obtain $\mathcal{O}(h^{-1/2})$ non-asymptotic bound regarding the equivalence of trained wide linear networks under random initialization to the min-norm solution. Our analysis, although on a simpler overparametrized model, formally connects overparametrization, initialization, and optimization with generalization performance. We think it is promising to translate some of the concepts such as the imbalance, and the constrained learning to multi-layer linear networks, and eventually to neural networks with nonlinear activations.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in neural information processing systems*, pp. 6158–6169, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019b.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2018a.

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

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pp. 322–332, 2019a.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pp. 8141–8150, 2019b.

Peter Bartlett, Dave Helmbold, and Philip Long. 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.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Sam Buchanan, Dar Gilboa, and John Wright. Deep networks and the multiple manifold problem. *arXiv preprint arXiv:2008.11245*, 2020.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 10836–10846, 2019.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pp. 2937–2947, 2019.

Kenneth R Davidson and Stanislaw J Szarek. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

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

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685, 2019a.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations(ICLR), 2019*, 2019b.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649. IEEE, 2013.

T. H. Grönwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919. ISSN 0003486X.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

Morris W Hirsch, Robert L Devaney, and Stephen Smale. *Differential equations, dynamical systems, and linear algebra*, volume 60. Academic press, 1974.

Roger A Horn and Charles R Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, New York, NY, USA, 2nd edition, 2012. ISBN 0521548233, 9780521548236.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pp. 8157–8166, 2018.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

Andrew M Saxe, James L Mcclelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural network. In *International Conference on Learning Representations*, 2014.

Sheng-De Wang, Te-Son Kuo, and Chen-Fa Hsu. Trace bounds on the solution of the algebraic matrix riccati and lyapunov equation. *IEEE Transactions on Automatic Control*, 31(7):654–656, 1986.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782*, 2017.

A Convergence Rate Analysis for Linear Regression: General Case

Suppose the input data matrix X has rank $d \leq \min\{D, n\}$, we write the compact SVD of X as

$$X = W\Sigma_x\Phi_1^T, W \in \mathbb{R}^{n\times d}, \Phi_1 \in \mathbb{R}^{D\times d},$$

and in Section 3 we assume that $d = n < D$. Notice that we always have $W^TW = I$.

Given the compact SVD, we still define $U_1 = \Phi_1^T U$, and write the loss function as

$$\begin{aligned}
\mathcal{L}(V,U) = \frac{1}{2}\|Y - XUV^T\|_F^2 = &\ \frac{1}{2}\|Y - W\Sigma_x U_1 V^T\|_F^2 \\
= &\ \frac{1}{2}\|(I_d - WW^T + WW^T)Y - W\Sigma_x U_1 V^T\|_F^2 \\
= &\ \frac{1}{2}\|(I_d - WW^T)Y + W(W^TY - \Sigma_x U_1 V^T)\|_F^2 \\
= &\ \frac{1}{2}\|(I_d - WW^T)Y\|_F^2 + \frac{1}{2}\|W(W^TY - \Sigma_x U_1 V^T)\|_F^2 \\
&\quad + \langle (I_d - WW^T)Y, W(W^TY - \Sigma_x U_1 V^T)\rangle_F \\
= &\ \frac{1}{2}\|(I_d - WW^T)Y\|_F^2 + \frac{1}{2}\|W(W^TY - \Sigma_x U_1 V^T)\|_F^2 \\
= &\ \frac{1}{2}\|(I_d - WW^T)Y\|_F^2 + \frac{1}{2}\|W^TY - \Sigma_x U_1 V^T\|_F^2,
\end{aligned}$$

where the last equality is because that $W^TW = I$, and the second last equality is because that the cross terms equal zero due to $W^T(I_d - WW^T) = 0$.

It is easy to see that $\min_{V,U} \mathcal{L}(V,U) = \frac{1}{2}\|(I_d - WW^T)Y\|_F^2 := \mathcal{L}^*$, which is usually referred as the residue. Similarly to Section.3, we still define the error as $E = W^TY - \Sigma_x U_1 V^T$ and one can check that the gradient flow on $\mathcal{L}(V,U)$ yields

$$\dot{V}(t) = E^T(t)\Sigma_x U_1(t)\,,\ \dot{U}_1(t) = \Sigma_x E(t) V(t)\,,$$

in parameter space (V, U_1). Apply Theorem 1 we have exponential convergence of $\mathcal{L}(t) - \mathcal{L}^* = \frac{1}{2}\|E(t)\|_F^2$ given sufficient rank on the imbalance $(\sigma_{m+d-1}(U_1(0)^TU_1(0) - V^T(0)V(0)) > 0)$. Here it requires that $h > d + m - 1$ for the hidden layer width.

To Summarize, for any linear regression problem, Theorem 1 shows that sufficient rank of imbalance guarantees exponential convergence of $\mathcal{L}(t) - \mathcal{L}^*$, where $\mathcal{L}^* = \frac{1}{2}\|(I_d - WW^T)Y\|_F^2$.

B Useful Lemmas

Before proving Theorem 1 and 2, we state several Lemmas that will be used in the proof.

The first Lemma is the Grönwall's inequality (Grönwall, 1919) in the differential form.

Lemma B.1 (Grönwall's inequality)**.** *Let $u(t), \beta(t) : [0, +\infty) \to \mathbb{R}$ be continuous, and $u(t)$ differentiable on $(0, +\infty)$. If*

$$\frac{d}{dt}u(t) \le \beta(t)u(t),\ \forall t > 0\,,$$

then

$$u(t) \le u(0)\exp\left(\int_0^t \beta(\tau)d\tau\right),\ \forall t > 0\,.$$

The next Lemma is known as Weyl's Inequality for singular values.

Lemma B.2 (Weyl's inequality for singular values)**.** *Let $A, B \in \mathbb{R}^{n\times m}$, let $q = \min\{n, m\}$, then*

$$\begin{aligned}\sigma_{i+j-1}(A+B) &\le \sigma_i(A) + \sigma_j(B)\\ \sigma_{i+j-1}(AB^T) &\le \sigma_i(A)\sigma_j(B^T)\,,\end{aligned}$$

for any i, j satisfying $1 \le i, j \le q$ and $i + j - 1 \le q$.

The proof can be found in Horn & Johnson (1994, Theorem 3.3.16). Using Weyl's inequality, we state and prove a lemma that is used for proving Theorem 2.

Lemma B.3. *Let $A \in \mathbb{R}^{k\times n}, B \in \mathbb{R}^{n\times m}$. Suppose $n \le m$, then*

$$\sigma_i(A)\sigma_n(B) \le \sigma_i(AB)\,,$$

for $1 \le i \le \min\{k, n\}$.

Proof. We start with the case where $k = n$. When $\sigma_n(B^T) = 0$, the result is trivial. When $\sigma_n(B^T) \ne 0$, we have $BB^\dagger = I$, where $B^\dagger$ is the Moore–Penrose inverse of B. By Lemma B.2, it follows that

$$\sigma_i(A) \le \sigma_i(AB)\sigma_1(B^\dagger),\ \forall 1 \le i \le n\,.$$

Since $\sigma_1(B^\dagger) = \sigma_n^{-1}(B)$, we get the desired inequality.

When $k > n$, we have

$$\sigma_i(A) = \sigma_i\left(\begin{bmatrix}A & 0_{k\times(k-n)}\end{bmatrix}\right) \le \sigma_i(AB)\,\sigma_1(\begin{bmatrix}B^\dagger & 0_{m\times(k-n)}\end{bmatrix}) = \sigma_i(AB)\sigma_1(B^\dagger),\ \forall 1 \le i \le n\,,$$

which still leads to the desired result.

When $k < n$, consider replacing A with $\begin{bmatrix}A\\ 0_{(n-k)\times n}\end{bmatrix}$, we have

$$\sigma_i(A)\sigma_n(B) = \sigma_i\left(\begin{bmatrix}A\\ 0_{(n-k)\times n}\end{bmatrix}\right)\sigma_n(B) \le \sigma_i\left(\begin{bmatrix}AB\\ 0_{(n-k)\times m}\end{bmatrix}\right) = \sigma_i(AB),\ \forall 1 \le i \le k\,.$$

□

We also state a trace inequality widely using for solving control problems

Lemma B.4. *Suppose for $A, B \in \mathbb{R}^{n\times n}$, A is symmetric and B is positive semidefinite, then*

$$\lambda_n(A)\,\mathrm{tr}(B) \le \mathrm{tr}(AB) \le \lambda_1(A)\,\mathrm{tr}(B)\,.$$

If both A, B are positive semidefinite, then

$$\sigma_n(A)\,\mathrm{tr}(B) \le \mathrm{tr}(AB) \le \sigma_1(A)\,\mathrm{tr}(B)\,.$$

The proof can be found in Sheng-De Wang et al. (1986, Lemma 1).

C Proof of Theorem 1

We begin with restating the Theorem.

Theorem 1 (Convergence of linear networks with sufficient rank of imbalance,restated)**.** *Suppose $h \ge m+n-1$. Let $V(t), U_1(t), t > 0$ be the trajectory of continuous dynamics* (7) *starting from some $V(0), U_1(0)$. If*

$$\sigma_{n+m-1}\left(U_1^T(0)U_1(0) - V^T(0)V(0)\right) = c > 0\,,$$

then for $E(t)$ defined in (6)*, we have*

$$\|E(t)\|_F^2 \le \exp\left(-2\sigma_n(\Sigma_x)ct\right)\|E(0)\|_F^2,\ \forall t > 0\,.$$

Additionally, $V(t), U_1(t), t > 0$ converges to some equilibrium point $(V(\infty), U_1(\infty))$ such that $E(V(\infty), U_1(\infty)) = 0$.

Proof. For readability we simply write $V(t), U_1(t), E(t)$ as V, U_1, E for most of the proof.

Under (7), the time derivative of error is given by

$$\dot{E} = -\Sigma_x^{1/2}U_1U_1^T\Sigma_x^{1/2}E - \Sigma_x EVV^T\,.$$

Consider the time derivative of $\|E\|_F^2$,

$$\begin{aligned}\frac{d}{dt}\|E\|_F^2 =&\ \frac{d}{dt}\mathrm{tr}(E^TE)\\ =&\ -2\,\mathrm{tr}\left(E^T\Sigma_x^{1/2}U_1U_1^T\Sigma_x^{1/2}E + E^T\Sigma_x EVV^T\right)\,.\end{aligned} \tag{19}$$

Use the trace inequality in Lemma B.4 to get the lower bound the trace of two matrices respectively as

$$\begin{aligned}\mathrm{tr}\left(E^T\Sigma_x^{1/2}U_1U_1^T\Sigma_x^{1/2}E\right) =&\ \mathrm{tr}\left(\Sigma_x^{1/2}EE^T\Sigma_x^{1/2}U_1U_1^T\right)\\ \ge&\ \sigma_n(U_1U_1^T)\,\mathrm{tr}\left(\Sigma_x^{1/2}EE^T\Sigma_x^{1/2}\right)\\ =&\ \sigma_n(U_1U_1^T)\,\mathrm{tr}\left(\Sigma_x EE^T\right)\\ \ge&\ \sigma_n(U_1U_1^T)\sigma_n(\Sigma_x)\,\mathrm{tr}(EE^T) = \sigma_n(U_1U_1^T)\sigma_n(\Sigma_x)\|E\|_F^2\,,\end{aligned} \tag{20}$$

and

$$\begin{aligned}\mathrm{tr}\left(E^T\Sigma_x EVV^T\right) \ge&\ \sigma_m(VV^T)\,\mathrm{tr}\left(E^T\Sigma_x E\right)\\ =&\ \sigma_m(VV^T)\,\mathrm{tr}\left(\Sigma_x EE^T\right)\\ \ge&\ \sigma_m(VV^T)\sigma_n(\Sigma_x)\,\mathrm{tr}(EE^T) = \sigma_m(VV^T)\sigma_n(\Sigma_x)\|E\|_F^2\,.\end{aligned} \tag{21}$$

Combine (19) with (20)(21), we have

$$\frac{d}{dt}\|E\|_F^2 \le -2\sigma_n(\Sigma_x)\left(\sigma_n(U_1U_1^T) + \sigma_m(VV^T)\right)\|E\|_F^2 \tag{22}$$

Moreover, we have

$$\sigma_n(U_1U_1^T) + \sigma_m(VV^T)$$

$$
\begin{aligned}
&= \sigma_n(U_1^T U_1) + \sigma_m(V^T V) \\
&= \sigma_n(U_1^T U_1) + \sigma_m(-V^T V) \\
\text{(Lemma } B.2) \ &\geq \sigma_{n+m-1}(U_1^T U_1 - V^T V) \\
\text{(Imbalance is time-invariant)} \ &= \sigma_{n+m-1}(U_1^T(0) U_1(0) - V^T(0) V(0)) = c\,,
\end{aligned}
$$

where the first equality uses the fact that $U_1 U_1^T (VV^T$ resp.) has the same non-zero singular values as $U_1^T U_1 (V^T V$ resp.). Finally we have

$$
\frac{d}{dt}\|E\|_F^2 \leq -2\sigma_n(\Sigma_x) c \|E\|_F^2\,.
$$

The result follows by applying Grönwall's inequality, Lemma B.1, which leads to

$$
\|E(t)\|_F^2 \leq \exp\left(-2\sigma_n(\Sigma_x) ct\right) \|E(0)\|_F^2,\ \forall t > 0\,, \tag{23}
$$

then the exponential convergence of $E(t)$ is proved.

Regarding the second statement, for the gradient system (7), the parameters $(U_1(t), V(t))$ converge either to an equilibrium point which minimizes the potential $\|E(t)\|_F^2$ or to infinity (Hirsch et al., 1974).

Consider the following dynamics

$$
\frac{d}{dt}\begin{bmatrix} V(t) \\ U_1(t) \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & E^T(t)\Sigma_x^{1/2} \\ \Sigma_x^{1/2} E(t) & 0 \end{bmatrix}}_{:=A_Z(t)} \underbrace{\begin{bmatrix} V(t) \\ U_1(t) \end{bmatrix}}_{:=Z(t)}, \tag{24}
$$

which is a time-variant linear system. Notice that by Horn & Johnson (2012, Theorem 7.3.3), we have $\|A_Z(t)\|_2 = \|\Sigma_x^{1/2} E(t)\|_2$.

From (24), we have

$$
\begin{aligned}
\frac{d}{dt}\|Z(t)\|_F^2 =&\ 2\,\mathrm{tr}\left(Z^T(t) A_Z(t) Z(t)\right) \\
=&\ 2\,\mathrm{tr}\left(Z(t) Z^T(t) A_Z(t)\right) \\
\leq&\ 2\|A_Z(t)\|_2\,\mathrm{tr}\left(Z(t) Z^T(t)\right) \\
=&\ 2\|\Sigma_x^{1/2} E(t)\|_2 \|Z(t)\|_F^2 \\
\leq&\ 2\sigma_1^{1/2}(\Sigma_x)\|E(t)\|_2 \|Z(t)\|_F^2 \leq 2\sigma_1^{1/2}(\Sigma_x)\|E(t)\|_F \|Z(t)\|_F^2\,.
\end{aligned}
$$

By Grönwall's inequality, Lemma B.1, we have

$$
\|Z(t)\|_F^2 \leq \exp\left(\int_0^t 2\sigma_1^{1/2}(\Sigma_x)\|E(\tau)\|_F d\tau\right) \|Z(0)\|_F^2\,.
$$

Finally, from (23), we have $\|E(t)\|_F \leq \exp\left(-\sigma_n(\Sigma_x) ct\right) \|E(0)\|_F,\ \forall t > 0$, which leads to

$$
\begin{aligned}
\|Z(t)\|_F^2 \leq&\ \exp\left(\int_0^t 2\sigma_1^{1/2}(\Sigma_x)\|E(\tau)\|_F d\tau\right) \|Z(0)\|_F^2 \\
\leq&\ \exp\left(2\sigma_1^{1/2}(\Sigma_x)\|E(0)\|_F \left(\int_0^t \exp\left(-\sigma_n(\Sigma_x) c\tau\right) d\tau\right)\right) \|Z(0)\|_F^2 \\
\leq&\ \exp\left(2\sigma_1^{1/2}(\Sigma_x)\|E(0)\|_F \left(\int_0^\infty \exp\left(-\sigma_n(\Sigma_x) c\tau\right) d\tau\right)\right) \|Z(0)\|_F^2 \\
=&\ \exp\left(\frac{2\sigma_1^{1/2}(\Sigma_x)}{c\sigma_n(\Sigma_x)}\|E(0)\|_F\right) \|Z(0)\|_F^2\,.
\end{aligned} \tag{25}
$$

Therefore the trajectory $V(t), U_1(t), t > 0$ is bounded, i.e. it can not converge to infinity, then it has to converges to some equilibrium point $(V(\infty), U_1(\infty))$ such that $E(V(\infty), U_1(\infty)) = 0$. □

D PROOF OF THEOREM 2

To prove Theorem 2, we use a basic result in random matrix theory

Lemma D.1. *Given $m, n \in \mathbb{N}$ with $m \leq n$. Let A be an $n \times m$ random matrix with i.i.d. standard normal entries $A_{ij} \sim \mathcal{N}\left(0, \frac{1}{n}\right)$. With probability at least $1 - 2\exp(-\delta^2)$, we have*

$$1 - \frac{\sqrt{m} + \delta}{\sqrt{n}} \leq \sigma_m(A) \leq \sigma_1(A) \leq 1 + \frac{\sqrt{m} + \delta}{\sqrt{n}} .$$

The proof can be found in Davidson & Szarek (2001, Theorem 2.13)

As we mentioned in Section.4, under random initialization, with high probability we have sufficient rank of imbalance and $\|V(0)U_2^T(0)\|_F, \|U_1(0)U_2^T(0)\|_F \sim \mathcal{O}(h^{-1/2})$. Additionally, the fact that $\left\|U_1(0)V^T(0)\right\|_F \sim \mathcal{O}(h^{-1/2})$ will be also useful. Here we state them formally as a lemma.

Lemma D.2. *Given data matrix X. $\forall \delta \in (0,1)$, $\forall h > h_0 = poly\left(m, D, \frac{1}{\delta}\right)$, with probability at least $1 - \delta$ over random initializations, we have all the following hold.*

1. *(Sufficient rank of imbalance)*

$$\sigma_{n+m}\left(U_1^T(0)U_1(0) - V^T(0)V(0)\right) > 1 - 2\frac{\sqrt{m+D} + \frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}} ,$$

2. *(Approximate orthogonality condition)*

$$\left\|\begin{bmatrix} V(0)U_2^T(0) \\ U_1(0)U_2^T(0) \end{bmatrix}\right\|_F \leq 2\sqrt{m+n}\frac{\sqrt{m+D} + \frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}} ,$$

$$\left\|U_1(0)V^T(0)\right\|_F \leq 2\sqrt{m}\frac{\sqrt{m+D} + \frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}} .$$

Proof. For readability we simply write $U(0), U_1(0), U_2(0), V(0)$ as U, U_1, U_2, V.

Consider the matrix $\begin{bmatrix} V^T & U^T \end{bmatrix}$ which is $h \times (m + D)$. Apply Theorem D.1 to matrix $A = \begin{bmatrix} V^T & U^T \end{bmatrix}$, with probability at least $1 - \delta$, we have

$$1 - \frac{\sqrt{m+D} + \frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}} \leq \sigma_{m+D}(\begin{bmatrix} V^T & U^T \end{bmatrix}) \leq \sigma_1(\begin{bmatrix} V^T & U^T \end{bmatrix}) \leq 1 + \frac{\sqrt{m+D} + \frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}} . \tag{26}$$

Regarding the first inequality, write the imbalance as

$$U_1^T U_1 - V^T V = \begin{bmatrix} V^T & U_1^T \end{bmatrix} \begin{bmatrix} -V \\ U_1 \end{bmatrix} = \begin{bmatrix} V^T & U^T \end{bmatrix} \begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^T \end{bmatrix} \begin{bmatrix} V \\ U \end{bmatrix} .$$

For $h > \left(\sqrt{m+D} + \frac{1}{2}\log\frac{2}{\delta}\right)^2$, assume event (26) happens, then $\sigma_{m+D}\left(\begin{bmatrix} V^T & U^T \end{bmatrix}\right) \geq 1 - \frac{\sqrt{m+D} + \frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}} > 0$, then we have

$$\begin{aligned}
& \sigma_{n+m}(U_1^T U_1 - V^T V) \\
&= \sigma_{n+m}\left(\begin{bmatrix} V^T & U^T \end{bmatrix} \begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^T \end{bmatrix} \begin{bmatrix} V \\ U \end{bmatrix}\right) \\
(\text{Lemma } B.3) &\geq \sigma_{n+m}\left(\begin{bmatrix} V^T & U^T \end{bmatrix} \begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^T \end{bmatrix}\right) \sigma_{m+D}\left(\begin{bmatrix} V \\ U \end{bmatrix}\right) \\
&= \sigma_{n+m}\left(\begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^T \end{bmatrix} \begin{bmatrix} V \\ U \end{bmatrix}\right) \sigma_{m+D}\left(\begin{bmatrix} V \\ U \end{bmatrix}\right) \\
(\text{Lemma } B.3) &\geq \sigma_{n+m}\left(\begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^T \end{bmatrix}\right) \sigma_{m+D}^2\left(\begin{bmatrix} V \\ U \end{bmatrix}\right) \\
&= \sigma_{n+m}\left(\begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^T \end{bmatrix}\right) \sigma_{m+D}^2\left(\begin{bmatrix} V^T & U^T \end{bmatrix}\right)
\end{aligned}$$

$$= \sigma^2_{m+D}\left(\begin{bmatrix} V^T & U^T \end{bmatrix}\right) ,$$

where the last equality is due to the fact that $\begin{bmatrix} -I_m & 0 \\ 0 & \Phi_1\Phi_1^T \end{bmatrix}$ has exactly $n+m$ non-zero singular value and all of them are 1.

Therefore when $h > \left(\sqrt{m+D} + \frac{1}{2}\log\frac{2}{\delta}\right)^2$, conditioned on event (26), with probability 1 we have

$$\begin{aligned} \sigma_{n+m}(U_1^T U_1 - V^T V) \geq \sigma^2_{m+D}\left(\begin{bmatrix} V^T & U^T \end{bmatrix}\right) &\geq \left(1 - \frac{\sqrt{m+D} + \frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}}\right)^2 \\ &> 1 - 2\frac{\sqrt{m+D} + \frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}} . \end{aligned} \tag{27}$$

Regarding the second and third inequality, using the fact that $\|A\|_F \leq \sqrt{\min\{n,m\}}\|A\|_2, A \in \mathbb{R}^{n\times m}$, we have

$$\begin{aligned} \frac{1}{\sqrt{m+n}}\left\|\begin{bmatrix} VU_2^T \\ U_1U_2^T \end{bmatrix}\right\|_F &\leq \left\|\begin{bmatrix} VU_2^T \\ U_1U_2^T \end{bmatrix}\right\|_2 \\ &= \left\|\begin{bmatrix} I_m & 0 \\ 0 & \Phi_1^T \end{bmatrix}\begin{bmatrix} V \\ U \end{bmatrix}\begin{bmatrix} V^T & U^T \end{bmatrix}\begin{bmatrix} 0 \\ \Phi_2 \end{bmatrix}\right\|_2 \\ &= \left\|\begin{bmatrix} I_m & 0 \\ 0 & \Phi_1^T \end{bmatrix}\left(\begin{bmatrix} V \\ U \end{bmatrix}\begin{bmatrix} V^T & U^T \end{bmatrix} - \eta I_{m+D}\right)\begin{bmatrix} 0 \\ \Phi_2 \end{bmatrix}\right\|_2 \\ &\leq \left\|\begin{bmatrix} V \\ U \end{bmatrix}\begin{bmatrix} V^T & U^T \end{bmatrix} - \eta I_{m+D}\right\|_2 , \text{for any } \eta \in \mathbb{R} , \end{aligned}$$

where the second equality is by the fact that $\begin{bmatrix} I_m & 0 \\ 0 & \Phi_1^T \end{bmatrix}\begin{bmatrix} 0 \\ \Phi_2 \end{bmatrix} = 0$, and

$$\begin{aligned} \frac{1}{\sqrt{m}}\left\|U_1V^T\right\|_F &\leq \left\|U_1V^T\right\|_2 \\ &= \left\|\begin{bmatrix} 0 & \Phi_1^T \end{bmatrix}\begin{bmatrix} V \\ U \end{bmatrix}\begin{bmatrix} V^T & U^T \end{bmatrix}\begin{bmatrix} I_m \\ 0 \end{bmatrix}\right\|_2 \\ &= \left\|\begin{bmatrix} 0 & \Phi_1^T \end{bmatrix}\left(\begin{bmatrix} V \\ U \end{bmatrix}\begin{bmatrix} V^T & U^T \end{bmatrix} - \eta I_{m+D}\right)\begin{bmatrix} I_m \\ 0 \end{bmatrix}\right\|_2 \\ &\leq \left\|\begin{bmatrix} V \\ U \end{bmatrix}\begin{bmatrix} V^T & U^T \end{bmatrix} - \eta I_{m+D}\right\|_2 , \text{for any } \eta \in \mathbb{R} , \end{aligned}$$

where the second equality is by the fact that $\begin{bmatrix} 0 & \Phi_1^T \end{bmatrix}\begin{bmatrix} I_m \\ 0 \end{bmatrix} = 0$.

Notice that

$$\left\|\begin{bmatrix} V \\ U \end{bmatrix}\begin{bmatrix} V^T & U^T \end{bmatrix} - \eta I_{m+D}\right\|_2 = \max_i \left|\sigma_i^2(\begin{bmatrix} V^T & U^T \end{bmatrix}) - \eta\right| .$$

Again we let $h > \left(\sqrt{m+D} + \frac{1}{2}\log\frac{2}{\delta}\right)^2$. When event (26) happens, all $\sigma_i^2(\begin{bmatrix} V^T & U^T \end{bmatrix})$ are within the interval $\left[\left(1 - \frac{\sqrt{m+D}+\frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}}\right)^2, \left(1 + \frac{\sqrt{m+D}+\frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}}\right)^2\right]$. Since the choice of η is arbitrary, we pick

$$\eta = 1 + \left(\frac{\sqrt{m+D} + \frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}}\right)^2 , \tag{28}$$

which is the mid-point of this interval, then we have

$$\max_i \left|\sigma_i^2(\begin{bmatrix} V^T & U^T \end{bmatrix}) - \eta\right|$$

$$
\begin{aligned}
&\leq \max\left\{\left|\left(1-\frac{\sqrt{m+D}+\delta}{\sqrt{h}}\right)^2-\eta\right|, \left|\left(1+\frac{\sqrt{m+D}+\delta}{\sqrt{h}}\right)^2-\eta\right|\right\} \\
(\eta \text{ is the mid-point}) &\leq \left|\left(1-\frac{\sqrt{m+D}+\delta}{\sqrt{h}}\right)^2-1-\left(\frac{\sqrt{m+D}+\delta}{\sqrt{h}}\right)^2\right| \\
&= 2\frac{\sqrt{m+D}+\frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}}
\end{aligned}
$$

Therefore, when $h > h_0 = \left(\sqrt{m+D}+\frac{1}{2}\log\frac{2}{\delta}\right)^2$, conditioned on event (26), with probability 1, we have

$$
\begin{aligned}
&\left\|\begin{bmatrix} VU_2^T \\ U_1U_2^T\end{bmatrix}\right\|_F \leq \sqrt{m+n}\left\|\begin{bmatrix} V \\ U\end{bmatrix}\begin{bmatrix} V^T & U^T\end{bmatrix}-\eta I_{m+D}\right\|_2 \leq 2\sqrt{m+n}\frac{\sqrt{m+D}+\frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}}, \\
\text{and } &\left\|U_1V^T\right\|_F \leq \sqrt{m}\left\|\begin{bmatrix} V \\ U\end{bmatrix}\begin{bmatrix} V^T & U^T\end{bmatrix}-\eta I_{m+D}\right\|_2 \leq 2\sqrt{m}\frac{\sqrt{m+D}+\frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}},
\end{aligned} \tag{29}
$$

where we choose η as in (28). Conditioned on event (26), events (27) and (29) happen with probability 1, hence the probability that both (27) and (29) happen is at least the probability of event (26), which is at least $1-\delta$. □

Theorem 2 (Generalization of wide single-hidden-layer linear network, restated)**.** *Let $(V(t), U(t), t > 0)$ be a trajectory of continuous dynamics* (7)*. Then, $\exists C > 0$, such that $\forall \delta \in (0,1), \forall h > h_0 = poly\left(m, D, \frac{1}{\delta}, \frac{\sigma_1(\Sigma_x)}{\sigma_n^3(\Sigma_x)}\right)$, with probability $1-\delta$ over random initializations, we have*

$$
\|U(\infty)V^T(\infty)-\hat{\Theta}\|_2 \leq C\sqrt{m+n}\frac{\sqrt{m+D}+\frac{1}{2}\log\frac{2}{\delta}}{\sqrt{h}},
$$

where C depends on the data X, Y.

Proof. From the continuous dynamics (7) and Theorem 1, the stationary point $U(\infty), V(\infty)$ satisfy

$$
U_1(\infty)V^T(\infty) = \Phi_1^T\hat{\Theta}, \quad U_2(\infty) = U_2(0).
$$

Hence we have

$$
\begin{aligned}
\|U(\infty)V^T(\infty)-\hat{\Theta}\|_2 &= \|\Phi_1U_1(\infty)V^T(\infty)+\Phi_2U_2(\infty)V^T(\infty)-\hat{\Theta}\|_2 \\
&= \|\Phi_1\Phi_1^T\hat{\Theta}+\Phi_2U_2(\infty)V^T(\infty)-\hat{\Theta}\|_2 \\
&= \|\Phi_2U_2(\infty)V^T(\infty)\|_F \\
&= \|\Phi_2U_2(0)V^T(\infty)\|_F = \|U_2(0)V^T(\infty)\|_2 \leq \|U_2(0)V^T(\infty)\|_F.
\end{aligned}
$$

Consider the following dynamics

$$
\frac{d}{dt}\begin{bmatrix} V(t)U_2^T(0) \\ U_1(t)U_2^T(0)\end{bmatrix} = \underbrace{\begin{bmatrix} 0 & E^T(t)\Sigma_x^{1/2} \\ \Sigma_x^{1/2}E(t) & 0\end{bmatrix}}_{:=A_Z(t)}\underbrace{\begin{bmatrix} V(t)U_2^T(0) \\ U_1(t)U_2^T(0)\end{bmatrix}}_{:=Z(t)}, \tag{30}
$$

which is a time-variant linear system, and in particular, by Horn & Johnson (2012, Theorem 7.3.3), we have $\|A_Z(t)\|_2 = \|\Sigma_x^{1/2}E(t)\|_2$. Notice that here the $Z(t)$ is different from the one in the proof for Theorem 1.

From (30), we have

$$
\begin{aligned}
\frac{d}{dt}\|Z(t)\|_F^2 &= 2\operatorname{tr}\left(Z^T(t)A_Z(t)Z(t)\right) \\
&= 2\operatorname{tr}\left(Z(t)Z^T(t)A_Z(t)\right) \\
&\leq 2\|A_Z(t)\|_2\operatorname{tr}\left(Z(t)Z^T(t)\right) \\
&= 2\|\Sigma_x^{1/2}E(t)\|_2\|Z(t)\|_F^2
\end{aligned}
$$

$$\leq\ 2\sigma_1^{1/2}(\Sigma_x)\|E(t)\|_2\|Z(t)\|_F^2 \leq 2\sigma_1^{1/2}(\Sigma_x)\|E(t)\|_F\|Z(t)\|_F^2\,.$$

By Grönwall's inequality, Lemma B.1, we have

$$\begin{aligned}\|Z(t)\|_F^2 &\leq \exp\left(\int_0^t 2\sigma_1^{1/2}(\Sigma_x)\|E(\tau)\|_F d\tau\right)\|Z(0)\|_F^2\\ \Rightarrow \|Z(t)\|_F &\leq \exp\left(\int_0^t \sigma_1^{1/2}(\Sigma_x)\|E(\tau)\|_F d\tau\right)\|Z(0)\|_F \end{aligned}\tag{31}$$

Using Lemma D.2, for $h > h_0' := 16\left(\sqrt{m+D}+\frac{1}{2}\ln\frac{2}{\delta}\right)^2$, with probability at least $1-\delta$ we have all the following.

$$\sigma_{n+m}\left(U_1^T(0)U_1(0)-V^T(0)V(0)\right) >\ 1-2\frac{\sqrt{m+D}+\frac{1}{2}\ln\frac{2}{\delta}}{\sqrt{h}} \geq \frac{1}{2}\,. \tag{32}$$

$$\|Z(0)\|_F = \left\|\begin{bmatrix}V(0)U_2^T(0)\\ U_1(0)U_2^T(0)\end{bmatrix}\right\|_F \leq\ 2\sqrt{m+n}\frac{\sqrt{m+D}+\frac{1}{2}\ln\frac{2}{\delta}}{\sqrt{h}}\,, \tag{33}$$

$$\left\|U_1(0)V^T(0)\right\| \leq\ 2\sqrt{m}\frac{\sqrt{m+D}+\frac{1}{2}\ln\frac{2}{\delta}}{\sqrt{h}} \tag{34}$$

Here h_0' is larger than one from Lemma D.2 because in (32) we want the least non-zero singular value of the imbalance to be further bounded by $\frac{1}{2}$.

From Theorem 1, we have

$$\|E(t)\|_F^2 \leq \exp\left(-2\sigma_n(\Sigma_x)ct\right)\|E(0)\|_F^2\,,$$

where $c = \sigma_{n+m-1}\left(U_1^T(0)U_1(0)-V^T(0)V(0)\right)$, then by (32), we have

$$\begin{aligned}\|E(t)\|_F^2 &\leq \exp\left(-\sigma_n(\Sigma_x)t\right)\|E(0)\|_F^2\\ \Rightarrow \|E(t)\|_F &\leq \exp\left(-\sigma_n(\Sigma_x)t/2\right)\|E(0)\|_F\,.\end{aligned}$$

Finally, from (31), we have

$$\begin{aligned}\|Z(t)\|_F \leq\ & \exp\left(\int_0^t \sigma_1^{1/2}(\Sigma_x)\|E(\tau)\|_F d\tau\right)\|Z(0)\|_F\\ \leq\ & \exp\left(\sigma_1^{1/2}(\Sigma_x)\|E(0)\|_F\left(\int_0^t \exp\left(-\sigma_n(\Sigma_x)\tau/2\right)d\tau\right)\right)\|Z(0)\|_F\\ \leq\ & \exp\left(\sigma_1^{1/2}(\Sigma_x)\|E(0)\|_F\left(\int_0^\infty \exp\left(-\sigma_n(\Sigma_x)\tau/2\right)d\tau\right)\right)\|Z(0)\|_F\\ =\ & \exp\left(2\frac{\sigma_1^{1/2}(\Sigma_x)}{\sigma_n(\Sigma_x)}\|E(0)\|_F\right)\|Z(0)\|_F\,.\end{aligned}\tag{35}$$

The initial error depends on the initialization but can be upper bounded as

$$\begin{aligned}\|E(0)\|_F =\ & \|W^TY-\Sigma_x^{-1/2}U_1(0)V^T(0)\|_F\\ \leq\ & \|W^TY\|_F+\|\Sigma_x^{-1/2}U_1(0)V^T(0)\|_F\\ \leq\ & \|Y\|_F+\sigma_n^{-1/2}(\Sigma_x)\|U_1(0)V^T(0)\|_F\end{aligned}$$

then we can write (35) as

$$\|Z(t)\|_F \leq\ \exp\left(2\frac{\sigma_1^{1/2}(\Sigma_x)}{\sigma_n(\Sigma_x)}\|Y\|_F\right)\exp\left(2\frac{\sigma_1^{1/2}(\Sigma_x)}{\sigma_n^{3/2}(\Sigma_x)}\|U_1(0)V^T(0)\|_F\right)\|Z(0)\|_F\,. \tag{36}$$

For the second exponential, for $h > h_0 := \max\left\{h_0', 16\frac{\sigma_1(\Sigma_x)}{\sigma_n^3(\Sigma_x)}m\left(\sqrt{m+D}+\frac{1}{2}\ln\frac{2}{\delta}\right)^2\right\}$, by (34) we have

$$\exp\left(2\frac{\sigma_1^{1/2}(\Sigma_x)}{\sigma_n^{3/2}(\Sigma_x)}\|U_1(0)V^T(0)\|_F\right) \leq \exp\left(4\frac{\sigma_1^{1/2}(\Sigma_x)}{\sigma_n^{3/2}(\Sigma_x)}\sqrt{m}\frac{\sqrt{m+D}+\frac{1}{2}\ln\frac{2}{\delta}}{\sqrt{h}}\right) \leq e \tag{37}$$

Finally by (33)(37), we write (36) as

$$\begin{aligned}
\|Z(t)\|_F \leq\ & \exp\left(1 + 2\frac{\sigma_1^{1/2}(\Sigma_x)}{\sigma_n(\Sigma_x)}\|Y\|_F\right)\|Z(0)\|_F \\
\leq\ & \underbrace{\exp\left(1 + 2\frac{\sigma_1^{1/2}(\Sigma_x)}{\sigma_n(\Sigma_x)}\|Y\|_F\right)2}_{:=C}\sqrt{m+n}\frac{\sqrt{m+D} + \frac{1}{2}\ln\frac{2}{\delta}}{\sqrt{h}} \\
=\ & C\sqrt{m+n}\frac{\sqrt{m+D} + \frac{1}{2}\ln\frac{2}{\delta}}{\sqrt{h}}\,.
\end{aligned}$$

Therefore for some $C > 0$ that depends on the data (X, Y), given any $0 < \delta < 1$, when $h > h_0$ as defined above, with at least probability $1 - \delta$, we have

$$\begin{aligned}
\|U(\infty)V^T(\infty) - \hat{\Theta}\|_2 \leq\ & \|U_2(0)V^T(\infty)\|_F \\
\leq\ & \sup_{t>0}\|U_2(0)V^T(t)\|_F \\
\leq\ & \sup_{t>0}\|Z(t)\|_F \leq C\sqrt{m+n}\frac{\sqrt{m+d} + \frac{1}{2}\ln\frac{2}{\delta}}{\sqrt{h}}\,.
\end{aligned}$$

□