Convergence and Implicit Bias of Gradient Flow on Overparametrized Linear Networks*

Hancheng Min[†‡] hanchmin@jhu.edu
Salma Tarmoun[†§] starmou1@jhu.edu
René Vidal[†¶] rvidal@jhu.edu
Enrique Mallada[†‡] mallada@jhu.edu
[†]*Mathematical Institute for Data Science, Johns Hopkins University*
[‡]*Department of Electrical and Computer Engineering, Johns Hopkins University*
[§]*Department of Applied Mathematics and Statistics, Johns Hopkins University*
[¶]*Department of Biomedical Engineering, Johns Hopkins University*

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 to study how initialization and overparametrization affect convergence and implicit bias of training algorithms. In this paper, we present a novel analysis of single-hidden-layer linear networks trained under gradient flow, which connects initialization, optimization, and overparametrization. Firstly, we show that the squared loss converges exponentially to its optimum at a rate that depends on the level of imbalance and the margin of the initialization. Secondly, we show that proper initialization constrains the dynamics of the network parameters to lie within an invariant set. In turn, minimizing the loss over this set leads to the min-norm solution. Finally, we show that large hidden layer width, together with (properly scaled) random initialization, ensures proximity to such an invariant set during training, allowing us to derive a novel non-asymptotic upper-bound on the distance between the trained network and the min-norm solution.

Keywords: Linear Networks, Overparametrized Models, Gradient Flow, Convergence, Implicit Bias

1. Introduction

Neural networks have shown excellent empirical performance in many application domains such as vision (Krizhevsky et al., 2012; Rawat and 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 puzzles behind this success is the fact that gradient descent with random initialization, and without explicit regularization, enjoys good generalization performance despite being highly overparametrized.

One possible explanation of such phenomenon is the implicit bias or regularization that first order gradient algorithms induce under proper initialization assumptions. For example, in classification tasks, gradient descent on separable data can induce a bias towards the max-margin solution (Soudry et al., 2018; Ji and Telgarsky, 2019; Lyu and Li, 2019). Similarly, in regression tasks, it has been shown that (deep) matrix factorization models trained by

*. Preprint

first order methods yield solutions with low nuclear norm (Gunasekar et al., 2017) and low rank (Arora et al., 2019a). Along the same vein, Saxe et al. (2014); Gidel et al. (2019) have shown that deep linear networks sequentially learn dominant singular values of the input-output correlation matrix.

Another possible explanation is that, in the Neural Tangent Kernel (NTK) regime, the gradient flow of a randomly initialized infinitely wide neural network can be well approximated by the flow of its linearization at initialization Jacot et al. (2018); Chizat et al. (2019); Arora et al. (2019c,b). In this regime, training infinitely wide neural networks mimics kernel methods. In particular, the NTK flow is constrained to lie on a manifold, which improves generalization performance as discussed in (Arora et al., 2019b).

While the aforementioned analysis is quite insightful, it requires assumptions on the model and the initialization that are often disconnected. For example, the implicit bias characterized in (Gunasekar et al., 2017; Arora et al., 2019a) requires vanishing initialization, while the analysis of convergence of gradient algorithms for linear networks requires balanced (Arora et al., 2018b,a) or spectral (Saxe et al., 2014; Gidel et al., 2019) initialization. Similarly, the NTK regime (Jacot et al., 2018; Arora et al., 2019c), requires random initialization and infinitely wide networks, making the non-asymptotic analysis challenging (Arora et al., 2019c).

This paper aims to bridge some of these gaps. We present a novel analysis of the gradient flow dynamics of overparametrized single-hidden-layer linear networks, which provides a common set of conditions on initialization that lead to convergence and implicit bias. Specifically, we reveal the explicit role of weights imbalance and weights product on the convergence of linear networks, suggesting a broad set of initial parameter values that lead to exponential convergence. We further characterize a complementary condition, based on orthogonality, that enforces the learning trajectory to be constrained within an invariant set whose unique global optimum is the min-norm solution. While our analysis does not require infinite width, vanishing, spectral, or random initialization, we show that our exponential convergence and orthogonality conditions are probably approximately satisfied for wide networks with properly scaled random initialization, jointly leading to a bound on the distance to the min-norm solution. Hence, this paper formally connects initialization, exponential convergence of the optimization task, overparametrization and implicit bias.

This paper makes the following contributions:

1. In Section 3, we show that the convergence of linear networks explicitly depends on: 1) a weight imbalance matrix; and 2) the weights product (end-to-end function). With such observation, we provide two conditions, *sufficient imbalance* and *sufficient margin*, on the intialization, with either of them being sufficient for guaranteeing exponential convergence. Our convergence analysis unifies prior work's assumptions and expands them to a broader set of initial conditions, as discussed in Section 1.1.

2. In Section 4.1, we show the existence of a subset of the parameter space defined by an orthogonality condition, which is invariant under gradient flow. All trajectories within this invariant set lead to a unique minimizer (w.r.t. the end-to-end function), which corresponds to the min-norm solution. As a result, initializing the network within this invariant set always yields the min-norm solution upon convergence.

3. In Section 4.2, we further show that by randomly initializing the network weights using $\mathcal{N}(0; 1/h^2)$ (where h is the hidden layer width and $1/4 < \alpha \leq 1/2$), one can approximately satisfy both our sufficient imbalance and orthogonality conditions with high probability. Notably, initializations outside the invariant set require exponential convergence to control their deviation from the set. For linear networks our results also provide a novel non-asymptotic upper-bound on the operator norm distance between the trained network and the min-norm solution.

1.1 Other Related Work

Convergence of Linear Networks. Convergence in overparametrized linear networks has been studied for both gradient flow (Saxe et al., 2014; Tarmoun et al., 2021) and gradient descent (Gidel et al., 2019; Arora et al., 2018a,b). Saxe et al. (2014); Gidel et al. (2019); Tarmoun et al. (2021) analyze the trajectory of network parameters under spectral initialization. For non-spectral initialization, although the fact that the imbalance is conserved under gradient flow has been exploited in Arora et al. (2018a,b), the work studies balanced initialization and exploits the structure conveyed by it to study convergence of the learning dynamics. The analysis of convergence in the imbalanced case was recently studied in Tarmoun et al. (2021) for both spectral and non-spectral initializations. For non-spectral initialization, specifically, previous analyses largely rely on specific imbalance structure (For example, small imbalance (Arora et al., 2018a), and homogeneous imbalance (Tarmoun et al., 2021)). Our analysis improves upon prior works by studying general imbalance structures. Particularly, our analysis identifies three key parameters, that quantify gaps and the spread of the spectrum of an imbalance matrix, that affect the rate of convergence of gradient flow.

The summary of the convergence results for linear networks is shown in Table 1.1, and we also illustrate all aforementioned non-spectral initialization in Figure 1.

	Spectral	*Non-spectral*
Balanced	(Saxe et al., 2014) (Gidel et al., 2019)	Exactly balanced (Arora et al., 2018b) Sufficient margin + Approximately balanced (Arora et al., 2018a)
Imbalanced	(Tarmoun et al., 2021)	Homogeneous imbalance (Tarmoun et al., 2021) **Sufficient level of imbalance** (Our work) **Sufficient margin** (Our work)

Table 1: List of initialization types that have been studied for the convergence of gradient flow on the single-hidden-layer linear networks. All non-spectral initialization types listed here are illustrated in Figure 1

Wide Neural Networks. There has been a rich line of research that studies the convergence (Du et al., 2019b,a; Du and Hu, 2019; Allen-Zhu et al., 2019b) and generalization (Allen-Zhu et al., 2019a; Arora et al., 2019b,c; Li and Liang, 2018; Cao and Gu, 2019; Buchanan et al., 2020) of wide neural networks with random initialization. The behavior of such networks in their infinite width limit can be characterized by the *Neural Tangent Kernel* (NTK) (Jacot et al., 2018). Heuristically, training wide neural networks can be approximately viewed as kernel regression under gradient flow/descent (Arora et al., 2019c). Hence, 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; Allen-Zhu et al., 2019b; Arora et al., 2019b,c; 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 view on the effect of random initialization for linear networks when the hidden layer is sufficiently wide.

1.2 Notation

For a matrix A, we let A^T denote its transpose, $\mathrm{tr}(A)$ denote its trace, $\ _i(A)$ and $\ _i(A)$ denote its i-th eigenvalue and i-th singular value, respectively, in decreasing order (when adequate). For an $n \times m$ matrix A, we let $\ _{\min}(A) = \ _{\min\{n;m\}}(A)$, and we conventionally let $\ _i(A) = \ _i(A) = 0; \forall i > \min\{m; n\}$. 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 symmetric matrix A, we write $A \succ 0$ ($A \succeq 0$, $A \prec 0$, or $A \preceq 0$) when A is positive defnite (positive semi-definite, negative definite, or negative semi-definite), and $A \succ (\succeq) B$, $A \prec (\preceq) B$ are equivalent to $A - B \succ (\succeq) 0$, $A - B \prec (\preceq) 0$, 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 $\mathcal{N}(\ ;\ ^2)$ denote the normal distribution with mean $\ $ and variance $\ ^2$.

2. Problem Setup

We study the gradient flow on single-hidden-layer linear networks trained with squared l_2-loss. Given N training samples $\{x^{(i)}; y^{(i)}\}_{i=1}^N$, where $x^{(i)} \in \mathbb{R}^n$, $y^{(i)} \in \mathbb{R}^m$, we aim to solve the linear regression problem

$$\min_{\in \mathbb{R}^D \ m} \mathcal{L} = \frac{1}{2} \sum_{i=1}^{N} \|y^{(i)} - \Theta^T x^{(i)}\|_2^2 : \tag{1}$$

We do so by training a single-hidden-layer linear network $y = f(x; V; U) = V U^T x$, $V = \mathbb{R}^{m \times h}$, $U \in \mathbb{R}^{n \times h}$, where h is the hidden layer width, with gradient flow, i.e., gradient descent with "infinitesimal step size". In particular,

- we consider the under-determined case $n > \mathrm{rank}(X)$ for our regression problem, i.e., the input dimension is strictly larger than the rank of X. There are infinitely many solutions Θ^* that achieve optimal loss $\mathcal{L}^*$ of (1);

- we consider an *overparametrized* model such that $h \geq \min\{m; n\}$, i.e. there is no rank constraint on linear model Θ obtained from the linear network UV^T.

We rewrite the loss with respect to our parameters $V; U$ as

$$\mathcal{L}(V; U) = \frac{1}{2} \sum_{i=1}^{N} \|y^{(i)} - VU^T x^{(i)}\|_2^2 = \frac{1}{2}\|Y - XUV^T\|_F^2 ; \tag{2}$$

where $Y = [y^{(1)}; \cdots ; y^{(N)}]^T$ and $X = [x^{(1)}; \cdots ; x^{(N)}]^T$. The gradient flow dynamics are given by

$$\dot{V}(t) = -\frac{\partial \mathcal{L}}{\partial V}(V(t); U(t)) = (Y - XU(t)V^T(t))^T XU(t) ; \tag{3a}$$

$$\dot{U}(t) = -\frac{\partial \mathcal{L}}{\partial U}(V(t); U(t)) = X^T(Y - XU(t)V^T(t))V(t) : \tag{3b}$$

Since we study the under-determined case, it is necessary to reparametrize the gradient flow dynamics, as shown in the next section.

(Note: For the rest of this paper, we drop the explicit dependence on time t for scalar/matrix functions of time when such dependence is clear. For example, we will mostly write $U; \dot{U}$ instead of $U(t); \dot{U}(t)$.)

2.1 Reparametrization of Gradient Flow

Assuming that $n > r = \text{rank}(X)$, the singular value decomposition (SVD) of X can be written as

$$X = W \begin{bmatrix} \Sigma_X^{1=2} & 0 \end{bmatrix} \begin{bmatrix} \Phi_1^T \\ \Phi_2^T \end{bmatrix} ; \tag{4}$$

where $W \in \mathbb{R}^{N \times r}$, $\Phi_1 \in \mathbb{R}^{D \times r}$, and $\Phi_2 \in \mathbb{R}^{D \times (D-r)}$. Since $\Phi_1\Phi_1^T + \Phi_2\Phi_2^T = I_D$, we have

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

and hence we can reparametrize U as $(U_1; U_2)$ using the bijection $U = \Phi_1 U_1 + \Phi_2 U_2$, with inverse $(U_1; U_2) = (\Phi_1^T U; \Phi_2^T U)$.

We write the gradient flow in (3a)(3b) explicitly as

$$\dot{V} = \left(Y - XUV^T\right)^T XU = E^T \Sigma_X^{1=2} \Phi_1^T U ; \tag{5a}$$

$$\dot{U} = X^T \left(Y - XUV^T\right) V = \Phi_1 \Sigma_X^{1=2} EV ; \tag{5b}$$

where

$$E = E(V; U_1) := W^T Y - \Sigma_X^{1=2} U_1 V^T ; \tag{6}$$

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

$$\dot{V} = E^T \Sigma_X^{1=2} U_1 ; \; \dot{U}_1 = \Sigma_X^{1=2} EV ; \; \dot{U}_2 = 0 : \tag{7}$$

Notice that

$$\begin{aligned}\mathcal{L}(V;U) = \frac{1}{2}\|Y - XUV^T\|_F^2 = \ & \frac{1}{2}\|(I - WW^T)Y + WE\|_F^2 \\ = \ & \frac{1}{2}\|WE\|_F^2 + \frac{1}{2}\|(I - WW^T)Y\|_F^2 \\ = \ & \frac{1}{2}\|E\|_F^2 + \frac{1}{2}\|(I - WW^T)Y\|_F^2\,; \end{aligned} \tag{8}$$

where the last equality is because W has orthonormal columns. Here the last term in (8) does not dependson $V;U$, and it is the residual

$$\mathcal{L}^* = \frac{1}{2}\|(I - WW^T)Y\|_F^2\,;$$

which is also the optimal value of (1). Therefore, for convergence, it suffices to analyze the convergence of the error E under the dynamics of $V;U_1$ in (7). The role of U_2 is discussed when we study the implicit bias in Section 4.

3. Convergence Analysis for Gradient Flow on Single-Hidden-Layer Linear Networks

With the reparametrization of the gradient flow, we study, for convergence, the dynamics of $V;U_1$,

$$\dot{V} = E^T\Sigma_X^{1=2}U_1\,;\ \dot{U}_1 = \Sigma_X^{1=2}EV\,;$$

this is exactly the gradient flow dynamics on

$$\frac{1}{2}\|E\|_F^2 = \frac{1}{2}\|W^TY - \Sigma_X^{1=2}U_1V^T\|_F^2\,: \tag{9}$$

In particular, when $\Sigma_X^{1=2} = I_r$, (9) reduces to $\frac{1}{2}\|W^TY - U_1V^T\|_F^2$, the loss function for a matrix factorization problem. To motivate our main result, we start with the simplest scalar version of this factorization problem.

3.1 Warm-up: Scalar Dynamics

Consider the gradient flow dynamics on the loss function $\mathcal{L}_s(u;v) = \frac{1}{2}|y - uv|^2$, we have

$$\dot{u} = (y - uv)v;\ \dot{v} = (y - uv)u\,: \tag{10}$$

This dynamics appear when one studies the gradient flow on (9) under the spectral initialization (Saxe et al., 2014; Gidel et al., 2019; Tarmoun et al., 2021). One important feature of (10), is that the *imbalance* $d := u^2 - v^2$ is invariant under the gradient flow, namely

$$\dot{d} = 2u\dot{u} - 2v\dot{v} \equiv 0\,:$$

For the scalar dynamics, such invariance admits explicit solution $u(t);v(t)$ given a fixed imbalance at initialization (Saxe et al., 2014; Tarmoun et al., 2021), and the asymptotic convergence rate of $\mathcal{L}_s$ around the equilibrium explicitly depends on the imbalance (Tarmoun et al., 2021). In our analysis, the imbalance plays an critical role as well, though

in a more global sense. We show two types of initialization that guarantees exponential convergence of $\mathcal{L}_s$: 1) sufficient imbalance; 2) sufficient margin.

One sufficient condition for exponential convergence is a lower bound on the *instantaneous rate* $-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s}$, to see this, notice that $\forall t \geq 0$

$$-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s} \geq c > 0 \Rightarrow \int_0^t \frac{\dot{\mathcal{L}}_s(\tau)}{\mathcal{L}_s(\tau)} d\tau \leq \int_0^t -c\, d\tau \Rightarrow \log \mathcal{L}_s \Big|_0^t \leq -ct \Rightarrow \log \frac{\mathcal{L}_s(t)}{\mathcal{L}_s(0)} \leq -ct$$
$$\Rightarrow \mathcal{L}_s(t) \leq \exp(-ct)\mathcal{L}_s(0)\,;$$

i.e., a lower bound $c > 0$ on the instantaneous rate implies the loss converges to 0 exponentially at a rate at least c. Now under the scalar dynamics (10), one can easily verify that

$$-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s} = -\frac{-(y-uv)^2v^2 - (y-uv)^2u^2}{(y-uv)^2/2} = 2(u^2+v^2)\,. \tag{11}$$

From the definition of imbalance $d = u^2 - v^2$, we have

$$u^4 = u^2(u^2) = u^2(d+v^2) = du^2 + (uv)^2\,; \tag{12a}$$
$$v^4 = v^2(v^2) = v^2(-d+u^2) = -dv^2 + (uv)^2\,. \tag{12b}$$

Now if we regard the product uv as a known value, then (12a) and (12b) are quadratic equations with respect to u^2 and v^2, whose solutions are

$$u^2 = \frac{d + \sqrt{d^2 + 4(uv)^2}}{2}\,;\ v^2 = \frac{-d + \sqrt{d^2 + 4(uv)^2}}{2}\,. \tag{13}$$

Replacing u^2, v^2 in (11) with the solutions in (13), we have

$$-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s} = 2(u^2+v^2) = 2\sqrt{d^2 + 4(uv)^2}\,; \tag{14}$$

i.e. the instantaneous rate can be explicitly written as the function of the imbalance d and the product uv. More importantly, with proper initialization, we can control the value of d and uv throughout the entire trajectory. Specifically,

- Since the imbalance d is time-invariant, we have $d(t) = d(0)$. When $|d(0)| > 0$, there is *sufficient imbalance* at initialization, and

$$-\frac{\dot{\mathcal{L}}_s(t)}{\mathcal{L}_s(t)} = 2\sqrt{d^2(t) + 4(u(t)v(t))^2} \geq 2|d(t)| = 2|d(0)|\,.$$

- The product is tied to the loss function $\mathcal{L}_s = |y-uv|^2/2$, and the loss is non-decreasing. When $|y| - |y - u(0)v(0)| > 0$, there is *sufficient margin* at initialization, and from

$$|u(t)v(t)| \geq |y| - |y - u(t)v(t)| \geq |y| - |y - u(0)v(0)|\,;$$

we have

$$-\frac{\dot{\mathcal{L}}_s(t)}{\mathcal{L}_s(t)} = 2\sqrt{d^2(t) + 4(u(t)v(t))^2} \geq 4|u(t)v(t)| = 4(|y| - |y - u(0)v(0)|)\,.$$

Combining the two observations above, we have

$$-\frac{\dot{\mathcal{L}}_s}{\mathcal{L}_s} = 2\sqrt{d^2 + 4(uv)^2} \geq 2\sqrt{d^2(0) + 4(\max\{|y| - |y - u(0)v(0)|; 0\})^2}\,. \tag{15}$$

That is, $\mathcal{L}_s$ converges to zero exponentially when either $|d(0)| > 0$ (sufficient imbalance) or $|y| - |y - u(0)v(0)| > 0$ (sufficient margin). Our main results in the next section show that such observation can be completely generalized to the matrix factorization problem, allowing us to derive exponential convergence guarantees for gradient flow on single-hidden-layer linear networks.

3.2 Main results

Now we turn to study the gradient dynamics in (7). Similar to the scalar dynamics, we define the *imbalance* of the single-hidden-layer linear network under input data X as

$$\textit{Imbalance}: \; D = U_1^T U_1 - V^T V \in \mathbb{R}^{h\times h}\,. \tag{16}$$

This imbalance matrix, as expected, is time-invariant under gradient flow dynamics (7). To see this, we compute the time derivative of $U_1^T U_1$ and $V^T V$ as

$$\begin{aligned}\frac{d}{dt}U_1^T U_1 &= \dot{U}_1^T U_1 + U_1^T \dot{U}_1 = V^T E^T \Sigma_x^{1=2} U_1 + U_1^T \Sigma_x^{1=2} E V;\\ \frac{d}{dt}V^T V &= V^T \dot{V} + \dot{V}^T V = V^T E^T \Sigma_x^{1=2} U_1 + U_1^T \Sigma_x^{1=2} E V\,:\end{aligned}$$

Because $\frac{d}{dt}U_1^T U_1$ and $\frac{d}{dt}V^T V$ are identical, one have $\dot{D} = \frac{d}{dt}[U_1^T U_1 - V^T V] \equiv 0$.

Our first result is the lower bound on the instantaneous rate (Proof left to Section 3.3):

Proposition 1 (Bound on the instantaneous rate) *Consider the continuous time dynamics in* (7)*. Let* $\tilde{\mathcal{L}} := \mathcal{L} - \mathcal{L}^*$ *and* $D = U_1^T U_1 - V^T V$*, then we have*

$$-\frac{\dot{\tilde{\mathcal{L}}}}{\tilde{\mathcal{L}}} \geq \; {}_r(\Sigma_x) \; \left(-\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4\,{}^2_m(U_1V^T)}\right) \left(-\Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4\,{}^2_r(U_1V^T)}\right)\,; \tag{17}$$

where we define

$$\textit{(Positive imbalance spectrum spread)}\;\Delta_+ = \max\{\,{}_1(D); 0\} - \max\{\,{}_r(D); 0\}\,; \tag{18}$$

$$\textit{(Negative imbalance spectrum spread)}\;\Delta_- = \max\{\,{}_1(-D); 0\} - \max\{\,{}_m(-D); 0\}\,; \tag{19}$$

$$\textit{(Effective level of imbalance)}\;\underline{\Delta} = \max\{\,{}_r(D); 0\} + \max\{\,{}_m(-D); 0\}\,: \tag{20}$$

If we think the imbalance D and the product U_1V^T as two factors that contribute to the instantaneous rate, their "individual" contributions are (assuming ${}_r(\Sigma_x) = 1$): 1) When $D = 0$, the lower bound reduces to $2\,{}_{\min}(U_1V^T)$ (when $r \neq m$) or $4\,{}_{\min}(U_1V^T)$ (when $r = m$); 2) When $U_1V^T = 0$, the lower bound reduces to $2\underline{\Delta}$. That is, the product contributes to the rate through ${}_{\min}(U_1V^T)$, while the imbalance does so through level of imbalance $\underline{\Delta}$. As we see in (17), it is not as straightforward as in (14) to combine these two factors since extra terms $\Delta_+; \Delta_-$ enter the lower bound.

Remark 2 *Although in the scalar case the instantaneous rate can be exactly expressed by imbalance and product, the rate also depends on the target $\tilde{Y} = W^T Y$ for the general matrix dynamics. Our lower bound in* (17) *is considered optimal when regarding the target as being adversely chosen to minimize the rate. We refer the reader to Appendix C for detailed discussion.*

For the lower bound in (17), one can verify that

$$\begin{aligned} \sqrt{\underline{\Delta}^2 + 4\ \sigma_m^2(U_1 V^T)} &\geq -\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4\ \sigma_m^2(U_1 V^T)} \geq \underline{\Delta}; \\ \sqrt{\underline{\Delta}^2 + 4\ \sigma_r^2(U_1 V^T)} &\geq -\Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4\ \sigma_r^2(U_1 V^T)} \geq \underline{\Delta}: \end{aligned}$$

The right extreme is obtained when $\Delta_+; \Delta_- \to \infty$ or the singular values are zero, and the left extreme is obtained when $\Delta_+ = \Delta_- = 0$.

Therefore, when $\Delta_-; \Delta_+$ are much larger than $\underline{\Delta}$, the lower bound is approximately

$$2\ \sigma_r(\Sigma_X)\underline{\Delta}:$$

When $\Delta_-; \Delta_+$ are much smaller, the bound is approximately

$$\sigma_r(\Sigma_X) \left(\sqrt{\underline{\Delta}^2 + 4\ \sigma_m^2(U_1 V^T)} + \sqrt{\underline{\Delta}^2 + 4\ \sigma_r^2(U_1 V^T)} \right) :$$

The latter becomes $2\ \sigma_r(\Sigma_X) \sqrt{\underline{\Delta}^2 + 4\ \sigma_{\min}^2(U_1 V^T)}$ when $r = m$, which takes the similar form as in the scalar case. From the experiments in Section 5, we see that under random initialization, networks with small width falls into the first regime and ones with large width falls into the latter, and the loss curves behaves differently in these two regimes.

As we illustrated with the scalar dynamics, the lower bound in Proposition 1, which depends explicitly on imbalance and product, is useful because one can control the two factors for the entire trajectory with proper initialization. This allows us to derive exponential convergence guarantees for the gradient flow, as stated in our main theorem next (Proof left to Section 3.3).

Theorem 3 (Exponential Convergence Guarantee) *Consider the continuous dynamics in* (7)*. Let $\tilde{Y} := W^T Y$ and define*

$$\begin{aligned} c(t) = -\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4(\max\{\ \sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_X^{1=2} U_1(t) V(t)^T\|_F; 0\})^{2=}\ \sigma_1(\Sigma_X)} \\ - \Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4(\max\{\ \sigma_r(\tilde{Y}) - \|\tilde{Y} - \Sigma_X^{1=2} U_1(t) V(t)^T\|_F; 0\})^{2=}\ \sigma_1(\Sigma_X)}; \end{aligned} \tag{21}$$

where $\Delta_+; \Delta_-$, and $\underline{\Delta}$ are define as in (18),(19)*, and* (20)*. Then we have*

$$(\mathcal{L}(t) - \mathcal{L}^*) \leq \exp\left(-\ \sigma_r(\Sigma_X) c(0) t\right) (\mathcal{L}(0) - \mathcal{L}^*); \forall t \geq 0:$$

That is, if $c(0) > 0$, then the loss converges to its global minimum exponentially with a rate at least $\ \sigma_r(\Sigma_X) c(0)$.

Theorem 3 unifies several previously discovered sufficient conditions for exponential convergence of the gradient flow on two-layer linear networks:

Corollary 4 (Sufficient level of imbalance (Min et al., 2021)) *If $\underline{\Delta} > 0$ at initialization, then the loss converges to zero exponentially with a rate at least $2\ _r(\Sigma_x)\underline{\Delta}(0)$.*

Proof In (21), if we lower bound the margin term $(\max\{\ _m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1=2}U_1V^T\|_F; 0\})$ by 0, we have $c \geq 2\underline{\Delta}$. ■

Previous work (Min et al., 2021) identifies the role of *effective level of imbalance* $\underline{\Delta}$ and proves the convergence result in Corollary 4. Our result generalizes it by showing the combined contribution of level of imbalance and the margin to the convergence.

Corollary 5 (Sufficient margin) *If at initialization, $\ _{\min}(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1=2}U_1V^T\|_F > 0$, then $c(0) > 0$ and the loss converges to zero exponentially with a rate at least $\ _r(\Sigma_x)c(0)$.*

Previous work (Arora et al., 2018a) has shown that when the initialization has a positive margin, i.e., $\ _{\min}(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1=2}U_1V^T\|_F > 0$ and the imbalance has sufficiently small Frobenius norm (approximately balanced), then the gradient flow converges exponentially. Corollary 5 improves upon it by showing that a positive margin is sufficient, regardless of the imbalance.

Corollary 6 (Characterizing local convergence rate) *If at some $t_0 > 0$, we have $c(t_0) > 0$, then*

$$(\mathcal{L}(t) - \mathcal{L}^*) \leq \exp\left(-\ _r(\Sigma_x)c(t_0)t\right)(\mathcal{L}(t_0) - \mathcal{L}^*); \forall t \geq t_0:$$

That is, after t_0, the loss converges to zero exponentially with a rate of at least $\ _r(\Sigma_x)c(t_0)$. Notably, given any trajectory that eventually converges to a global minimum for $\mathcal{L}$, for sufficiently large t_0, we have

$$c(t_0) \simeq -\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4\ ^2_m(\tilde{Y})=\ _1(\Sigma_x)} - \Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4\ ^2_r(\tilde{Y})=\ _1(\Sigma_x)}: \tag{22}$$

For any trajectory that eventually converges, (22) is due to the fact that

$$\|W^TY - \Sigma_x^{1=2}U_1(t_0)V^T(t_0)\|_F \simeq 0;$$

at sufficiently large t_0. This corollary suggests that the asymptotic convergence rate around the equilibrium depends on the imbalance D and the target Y. Previous work (Tarmoun et al., 2021) has shown that when $\Sigma_x = I_r$, $h = r = m$ and $D = \ I_h$ for some $\ \neq 0$, the asymptotic convergence rate of the gradient flow is lower bounded by $2\sqrt{\ ^2 + 4\ ^2_{\min}(\tilde{Y})}$, and this can be exactly recovered from (22) with $\Delta_+ = \Delta_- = 0$; $\underline{\Delta} = \ $. Our result has no additional assumption on the dimension nor on the imbalance structure.

The major limitation of previous works on convergence is the requirement on the imbalance structure: exactly balanced (Arora et al., 2018b), or homogeneously imbalanced (Tarmoun et al., 2021) initialization admits explicit dynamics of the product U_1V^T (the end-to-end function), from which the convergence results are derived. Such analyses considers,

as illustrated in Figure 1, specific configurations in the parameter space, and only allow small variations (Arora et al., 2018a). Our analysis breaks such limitation by revealing fundamental relations between the convergence and the weight configuration (imbalance and product), as explicitly seen in the scalar dynamics, which provides convergence guarantees for a wide range of initialization.

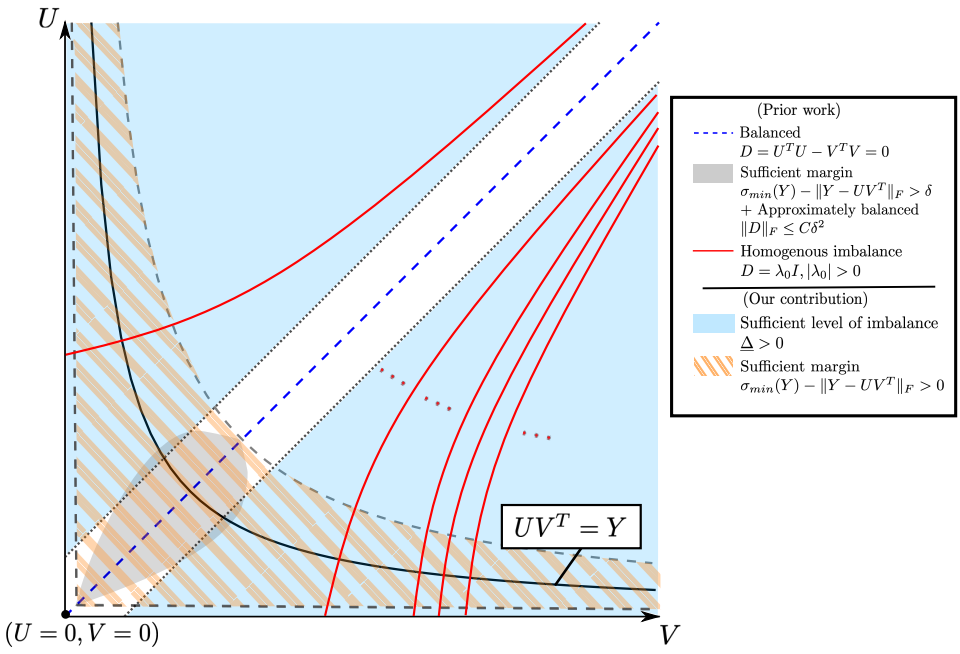

Figure 1: Illustration of non-spectral initialization studied for convergence of linear networks. Note: the conditions are presented for the gradient flow on $\frac{1}{2}\|Y - UV^T\|$, which is the special case of ours when $X = I_n$.

3.3 Proof Sketch for the Main Results

The proof of our main results Proposition 1 and Theorem 3 follows exactly the same procedure for the scalar dynamics in Section 3.1. We sketch the proof in this section and leave the proofs for all the stated Lemmas to Appendix B.

First of all, we lower bound the instantaneous rate with singular values of $U_1; V$, similar to (11).

Lemma 7 *Consider the continuous dynamics in* (7). *Let* $\tilde{\mathcal{L}} := \mathcal{L} - \mathcal{L}^*$, *then we have*

$$-\frac{\dot{\tilde{\mathcal{L}}}}{\tilde{\mathcal{L}}} \ge 2\ \ _r(\Sigma_X)(\ \ _r(U_1U_1^T) + \ \ _m(VV^T)):$$

Recall that for the scalar case, (12a)(12b) can be viewed as quadratic equations of u^2 and v^2 respectively. For the matrix case, one can derive quadratic inequalities of $\ \ _r(U_1U_1^T)$ and of $\ \ _m(VV^T)$, whose solutions give us lower bounds on $\ \ _r(U_1U_1^T)$ and $\ \ _m(VV^T)$), respectively. More generally, we have

Lemma 8 *Suppose* $h \geq \min\{r; m\}$. *Given any* $A \in \mathbb{R}^{r\times h}; B \in \mathbb{R}^{h\times m}$ *that satisfy* $A^TA - BB^T = D$ *for some* $D \in \mathbb{R}^{h\times h}$, *to get*

$$\sigma_m(B^TB) \geq \frac{-\bar{\lambda} + \lambda_- + \sqrt{(\bar{\lambda} + \lambda_-)^2 + 4\,\sigma_m^2(AB)}}{2}; \tag{23}$$

where $\bar{\lambda} = \max\{\sigma_1(D); 0\}$ *and* $\lambda_- = \max\{\sigma_m(-D); 0\}$.

Combining Lemma 7 and Lemma 8, we have the desired bound on the instantaneous rate

Proof [Proof of Proposition 1] From Lemma 8, let $A = U_1; B = V^T$, we have $A^TA - BB^T = D$, thus

$$\sigma_m(VV^T) \geq \frac{-\bar{\lambda}_+ + \lambda_- + \sqrt{(\bar{\lambda}_+ + \lambda_-)^2 + 4\,\sigma_m^2(U_1V^T)}}{2}; \tag{24}$$
$$\bar{\lambda}_+ = \max\{\sigma_1(D); 0\};\ \lambda_- = \max\{\sigma_m(-D); 0\};$$

then let $A = V; B = U_1^T$, we have $A^TA - BB^T = -D$, thus

$$\sigma_r(U_1U_1^T) \geq \frac{-\bar{\lambda}_- + \lambda_+ + \sqrt{(\bar{\lambda}_- + \lambda_+)^2 + 4\,\sigma_r^2(VU_1^T)}}{2}: \tag{25}$$
$$\bar{\lambda}_- = \max\{\sigma_1(-D); 0\};\ \lambda_+ = \max\{\sigma_r(D); 0\}$$

Now rewrite the lowerbounds (24)(25) in terms of

$$\Delta_+ := \bar{\lambda}_+ - \lambda_+;\ \Delta_- := \bar{\lambda}_- - \lambda_-;\ \underline{\Delta} := \lambda_+ + \lambda_-;$$

we have

$$\sigma_m(VV^T) \geq \frac{-\bar{\Delta}_+ + \lambda_- - \lambda_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4\,\sigma_m^2(U_1V^T)}}{2};$$
$$\sigma_r(U_1U_1^T) \geq \frac{-\bar{\Delta}_- + \lambda_+ - \lambda_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4\,\sigma_r^2(VU_1^T)}}{2}:$$

Then (17) follows immediately from Lemma 7. ■

Again, regarding the bound in Proposition 1, $\Delta_+; \Delta_-; \underline{\Delta}$ are time-invariant because the imbalance D is so, and the singular value $\sigma_m^2(U_1V^T)$ can be controlled via positive margin. This proves Theorem 3.

Proof [Proof of Theorem 3] When $m = r$, one have $\sigma_m(U_1V^T) = \sigma_r(U_1V^T) = \sigma_{\min}(U_1V^T)$. When $m > r$, we only need to lower bound $\sigma_m(U_1V^T)$ since $\sigma_r(U_1V^T) = 0$, and vise versa when $r > m$.

Therefore, without loss of generality, we assume $m \leq r$ and derive the lower bound on $\sigma_m(U_1V^T)$. By $\|A\|_F \geq \|A\|_2$ and Weyl's inequality (Horn and Johnson, 2012, 7.3.P16), one has

$$\|\tilde{Y} - \Sigma_x^{1=2}U_1V^T\|_F + \sigma_m(\Sigma_x^{1=2}U_1V^T) \geq \|\tilde{Y} - \Sigma_x^{1=2}U_1V^T\|_2 + \sigma_m(\Sigma_x^{1=2}U_1V^T) \geq \sigma_m(\tilde{Y});$$

from which one obtain the lower bound

$$\sigma_m(U_1 V^T) \geq \sigma_m(\Sigma_x^{1/2} U_1 V^T) = \lambda_1^{1/2}(\Sigma_x) \geq (\sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2} U_1 V^T\|_F) = \lambda_1^{1/2}(\Sigma_x).$$

The lower bound is trivial when $\sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2} U_1 V^T\|_F < 0$, thus we could write

$$\sigma_m(U_1 V^T) \geq \max\{\sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2} U_1 V^T\|_F, 0\} = \lambda_1^{1/2}(\Sigma_x). \tag{26}$$

Now because $\|\tilde{Y} - \Sigma_x^{1/2} U_1 V^T\|_F = \sqrt{2\tilde{\mathcal{L}}}$ is non-decreasing under gradient flow, we have $\forall t \geq 0$,

$$\begin{aligned}
\sigma_m^2(U_1(t) V^T(t)) &\geq (\max\{\sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2} U_1(t) V^T(t)\|_F, 0\})^2 = \lambda_1(\Sigma_x) \\
&\geq (\max\{\sigma_m(\tilde{Y}) - \|\tilde{Y} - \Sigma_x^{1/2} U_1(0) V^T(0)\|_F, 0\})^2 = \lambda_1(\Sigma_x).
\end{aligned} \tag{27}$$

Finally using (27) to further lower bound (17) in Proposition 1, we have our desired lower bound on the instantaneous rate

$$-\frac{\dot{\tilde{\mathcal{L}}}}{\tilde{\mathcal{L}}} \geq \lambda_r(\Sigma_x) c(0).$$

The result $\tilde{\mathcal{L}}(t) \leq \exp(-\lambda_r(\Sigma_x) c(0) t) \tilde{\mathcal{L}}(0)$ follows from Grönwall's inequality (Grönwall, 1919). ■

4. Implicit Bias of Gradient Flow on Single-Hidden-Layer Linear Network

In this section, we study a particular type of implicit bias of single-hidden-layer linear networks under gradient flow. We have assumed that $n > r = \operatorname{rank}(X)$, hence the regression problem (1) has infinitely many solutions Θ^* that achieve optimal loss. Among all these solutions, one that is of particular interest in high-dimensional linear regression is the *minimum norm solution* (min-norm solution)

$$\begin{aligned}
\hat{\Theta} &= \arg\min_{\Theta \in \mathbb{R}^{n \times m}} \{\|\Theta\|_F : \|Y - X\Theta\|_F^2 = \min \|Y - X\Theta\|_F^2\} \\
&= X^T (X X^T)^\dagger Y,
\end{aligned} \tag{28}$$

which has near-optimal generalization error for suitable data models (Bartlett et al., 2020; Mei and 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 invariant set. In turn, minimizing the loss over this set leads to the min-norm solution.

4.1 Decomposition of Trained Network

Notice that the end-to-end matrix $UV^T \in \mathbb{R}^{D \times m}$ associated with the single-hidden-layer linear network can be decomposed according to the SVD of data matrix X, (4), as

$$UV^T = (\Phi_1 \Phi_1^T + \Phi_2 \Phi_2^T) UV^T = \Phi_1 U_1 V^T + \Phi_2 U_2 V^T; \tag{29}$$

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

Convergence of Training Parameters. We have derived useful results regarding $U_1(t)V^T(t)$ for $t > 0$ in Section 3. When the condition in Theorem 3 is satisfied, exponential convergence of the loss implies $U_1(t)V^T(t)$ converges to some stationary point, as stated in the following proposition

Proposition 9 *Consider the continuous dynamics in* (7)*. If* $c(0)$*, defined in Theorem 3, is positive, then* $V(t); U_1(t); t > 0$ *converges to an equilibrium point* $V(\infty); U_1(\infty)$ *such that* $E(V(\infty); U_1(\infty)) = W^TY - \Sigma_x^{1=2}U_1V^T = 0$.

This Proposition is due to the fact that the states in gradient dynamics either converge to an equilibrium point or having its norm grow to infinity, and the exponential convergence excludes the later case. We left its proof to Appendix E.

Knowing that $V(t); U_1(t)$ converges, it is easy to check that

$$\Phi_1U_1(\infty)V^T(\infty) = \Phi_1\Sigma_x^{-1=2}W^TY = X^T(XX^T)^\dagger Y = \hat{\Theta}:$$

For $U_2(t)V^T(t)$, notice that $\dot{U}_2(t) = 0$ in dynamics (7), hence $U_2(t) = U_2(0); \forall t > 0$. Overall, given exponential convergence of the loss, $U(t)V^T(t)$ converges to some $U(\infty)V^T(\infty)$ and

$$U(\infty)V^T(\infty) = \Phi_1U_1(\infty)V^T(\infty) + \Phi_2U_2(0)V^T(\infty) = \hat{\Theta} + \Phi_2U_2(0)V^T(\infty): \tag{30}$$

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 (30). 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 $\Phi_2^TU_2(t)V^T(t) \equiv 0$ for all $t \geq 0$, thus the min-norm solution is obtained upon convergence, as suggested by the following proposition.

Proposition 10 *Let* $V(t); U_1(t); U_2(t); t > 0$ *be the solution of* (7) *starting from some* $V(0); U_1(0); U_2(0)$*. We assume* $V(t); U_1(t); t > 0$ *converges to some* $V(\infty); U_1(\infty)$ *with* $E(V(\infty); U_1(\infty)) = 0$*. If the initialization satisfies*

$$V(0)U_2^T(0) = 0;\ U_1(0)U_2^T(0) = 0; \tag{31}$$

then we have

$$U(\infty)V^T(\infty) = \hat{\Theta}:$$

Proof From (7) we have

$$\frac{d}{dt}\begin{bmatrix} VU_2^T \\ U_1U_2^T \end{bmatrix} = \begin{bmatrix} 0 & E^T\Sigma_x^{1=2} \\ \Sigma_x^{1=2}E & 0 \end{bmatrix}\begin{bmatrix} VU_2^T \\ U_1U_2^T \end{bmatrix}: \tag{32}$$

Since $VU_2^T = 0$; $U_1U_2^T = 0$ is an equilibrium point of (32), we have $V(t)U_2^T(0) = 0; \forall t \geq 0$ under the initialization in (31), hence $V(\infty)U_2^T(0) = 0$. From (30) we conclude that $U(\infty)V^T(\infty) = \hat{\Theta}$. ■

In the standard linear regression, where Θ follows the gradient flow on $\mathcal{L}(\Theta) = \frac{1}{2}\|Y - X\Theta\|_F^2$, it is well-known that if the columns of $\Theta(0)$ are initialized in $\text{span}(\Phi_1)$, namely $\Theta^T(0)\Phi_2 = 0$, then $\Theta(\infty) = \hat{\Theta}$. Proposition 10 is the extension of such results to the overparameterized setting. It is worth-noting that initializing the columns of $U(0)V^T(0)$ in $\text{span}(\Phi_1)$, namely $V(0)U_2^T(0) = 0$ is no longer sufficient for obtaining $\hat{\Theta}$ as the trained network, and additional condition $U_1(0)U_2^T(0) = 0$ is required.

Here the orthogonality constraints (31) defines an invariant subset of the parameter space $\{V; U : VU_2^T = 0; U_1U_2^T = 0\}$ under the gradient flow. Proposition 10 shows that given an initialization within the invariant set, the trained network (after convergence) is exactly the min-norm solution, which is the only minimizer in the invariant set.

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. Moreover, we note that while the zero initialization works for the standard linear regression case, such initialization $V(0) = 0; U(0) = 0$ is bad in the overparametrized case because it is an equilibrium point of the gradient flow, even though it satisfies the orthogonal condition $V(0)U_2^T(0) = 0$ and $U_1(0)U_2^T(0) = 0$.

In the next section, we show that under (properly scaled) random initialization and sufficiently large hidden layer width h, both conditions for convergence and implicit bias on initialization are probably approximately satisfied, i.e., with high probability the level of imbalance is sufficient for exponential convergence, and the parameters are initialized close to the invariant set, allowing us to obtain a non-asymptotic bound between the trained network and the min-norm solution.

4.2 Wide Single-Hidden-Layer Linear Network

In this section, we show how the previously mentioned conditions for convergence and implicit bias, i.e., high imbalance and orthogonality, are approximately satisfied with high probability under the following initialization

$$[U(0)]_{ij} \sim \mathcal{N}\quad 0; \frac{1}{h^2}\quad ;\ 1 \leq i \leq n; 1 \leq j \leq h\,;$$

$$[V(0)]_{ij} \sim \mathcal{N}\quad 0; \frac{1}{h^2}\quad ;\ 1 \leq i \leq m; 1 \leq j \leq h\,;$$

where all the entries are independent and $1=4 \leq \quad \leq 1=2$.

Both our parametrization and initialization are, at first sight, different from the one used in previous works (Jacot et al., 2018; Du and Hu, 2019; Arora et al., 2019c) on NTK analysis for wide neural networks. We note that with time-rescaling, however, we can relate our initialization to the one in Arora et al. (2019c). Please see Appendix D for a comparison.

Recall form 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 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 the following lemma. See Appendix E for the proof.

Lemma 11 *Let $\frac{1}{4} < \leq \frac{1}{2}$. Given data matrix X. $\forall \in (0;1)$, $\forall h > h_0 = \text{poly}(m;n;\frac{1}{})$, with probability at least $1-$ over random initialization with $[U(0)]_{ij};[V(0)]_{ij} \sim \mathcal{N}(0;h^{-2})$, the following conditions hold:*

1. *(Sufficient level of imbalance)*

$$\underline{\Delta}(0) > h^{1-2} \quad ; \tag{33}$$

where $\underline{\Delta}$ is the effective level of imbalance defined in (20).

2. *(Approximate orthogonality)*

$$\begin{matrix} V(0)U_2^T(0) \\ U_1(0)U_2^T(0) \end{matrix}_F \leq 2\sqrt{m+r}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{}}{h^{2\ -\frac{1}{2}}}; \tag{34}$$

$$U_1(0)V^T(0)\ _F \leq 2\sqrt{m}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{}}{h^{2\ -\frac{1}{2}}}: \tag{35}$$

From (34), we know that the parameters are initialized close to the invariant set of our interest, as measured by $\|VU_2^T\|_F + \|U_1U_2^T\|_F$. The dynamics (32) quantify at time t how fast this measure can maximally increase given that its current value is non-zero. It is clear that the smaller norm the current error E has, the lower is the rate at which this measure could increase. This suggests that as long as the error converges sufficiently fast, $\|VU_2^T\|_F + \|U_1U_2^T\|_F$ will not increase too much from its initial value. For our purpose, as the width h increases, we need at least a constant rate of exponential convergence of the error (given by (33)), and an initial error $E(0)$ that is bounded by some constant (derived from (35)). With these conditions satisfied with high probability, we have the following Theorem regarding the implicit bias of wide linear networks. We left its proof to Appendix E.

Theorem 12 *Let $\frac{1}{4} < \leq \frac{1}{2}$. Let $V(t);U(t);t > 0$ be the trajectory of the continuous dynamics (7) starting from some $V(0);U(0)$. Then, $\exists C > 0$, such that $\forall \in (0;1);\forall h > h_0^{1=(4\ -1)}$ with $h_0 = \text{poly}(m;n;\frac{1}{};\frac{1}{\frac{3}{r}(\ x)})$, with probability $1-$ over random initializations with $[U(0)]_{ij};[V(0)]_{ij} \sim \mathcal{N}(0;h^{-2})$, we have*

$$\|U(\infty)V^T(\infty) - \hat{\Theta}\|_2 \leq 2C^{1=h^{1\ \ 2}}\sqrt{m+r}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{}}{h^{2\ -\frac{1}{2}}}: \tag{36}$$

Here $C = \exp\left(1 + \frac{\frac{1=2}{1}(\ x)}{r(\ x)}\|Y\|_F\right)$, which depends on the data $X;Y$.

Previous works (Arora et al., 2019c) show non-asymptotic results on bounding the difference of predictions between the trained network and the kernel predictor of the NTK over a

finite number of testing point (non-global result) using more general network structure and activation functions. We work on a simpler model, we are able to study it without going through non-asymptotic NTK analysis, which is considerably more complicated than ours. We believe this theorem is a clear illustration of how overparametrization, in particular, in the hidden layer width, together with random initialization affects the convergence and implicit bias.

Notably, although our initialization is related to the NTK analysis (Jacot et al., 2018; Arora et al., 2019c) and the kernel regime (Chizat et al., 2019), we significantly simplify the non-asymptotic analysis with the exact charaterization of an invariant set tied to the regularized solution. Specifically, our analysis does not rely on approximating the training flow to one in the infinite width limit, or one from the linearized network at initialization. Instead, we have the exact characterization of the properties required to reach min-norm solution and show how such properties are approximately preserved during training.

5. Numerical Experiments

In this section, we first illustrate how the imbalance quantities $\Delta_+; \Delta_-; \underline{\Delta}$ are obtained from the spectrum of the imbalance matrix, as well as the role of width in shaping the imbalance quantities under random initialization. Then we run gradient descent (with small step size) on linear regression problem to validate our lower bounds for the convergence rate. We also refer readers to Appendix A.3 for numerical verification of our Theorem 12 on implicit bias of wide linear networks.

5.1 Imbalance Quantities

For simplicity, we consider the matrix factorization problem $\mathcal{L} = \frac{1}{2}\|Y - \frac{1}{\sqrt{mh}}UV^T\|_F^2$, $U \in \mathsf{R}^{r\times h}; V \in \mathsf{R}^{m\times h}$ under Xavier initialization (Glorot and Bengio, 2010). The scaling factor $\frac{1}{\sqrt{mh}}$ ensures that at initialization, the product UV^T keeps the same scale as we vary the hidden layer width h. Our convergence results Proposition 1 and Theorem 3 apply to this case and the imbalance quantities $\Delta_+; \Delta_-; \underline{\Delta}$ are defined from the imbalance matrix $D = U^TU - V^TV$ at initialization.

When $h \geq n + m$, then with probability 1 under random initialization, the imbalance matrix D has $rank(D) = n+m$ and it has n positive eigenvalues and m negative ones. Our experiment sets $n = 20; m = 5$ and consider the case of $h = 30$ (small width) and $h = 1000$ (large width). For initialization, we use $[U(0)]_{ij}; [V(0)]_{ij} \sim \mathcal{N}(0; 1)$.

Under Xavier initialization, the instantaneous rate is scaled by $\frac{1}{mh}$, hence we consider the scaled imbalance quantities, the details are given in Appendix A.1. We plot in Figure 4 all the non-zero eigenvalues of imbalance D and the imbalance quantities, scaled by $\frac{1}{mh}$. As illustrated by the plot, the imbalance quantities can be understood as the gaps between certain eigenvalues. It is clear that, compare to small width $h = 50$, large width $h = 1000$ has larger level of imbalance and smaller spectrum spread.

Moreover, as the width varies, the loss curve behaves differently:
(Small width): When $h = 30$, spectrum spreads $\Delta_-; \Delta_+$ are larger compared to the level of imbalance $\underline{\Delta}$. As we discussed in Section 3 after Proposition 1, the lower bound on the rate is approximately $2\underline{\Delta}$, which is not a good global bound for the convergence rate (see

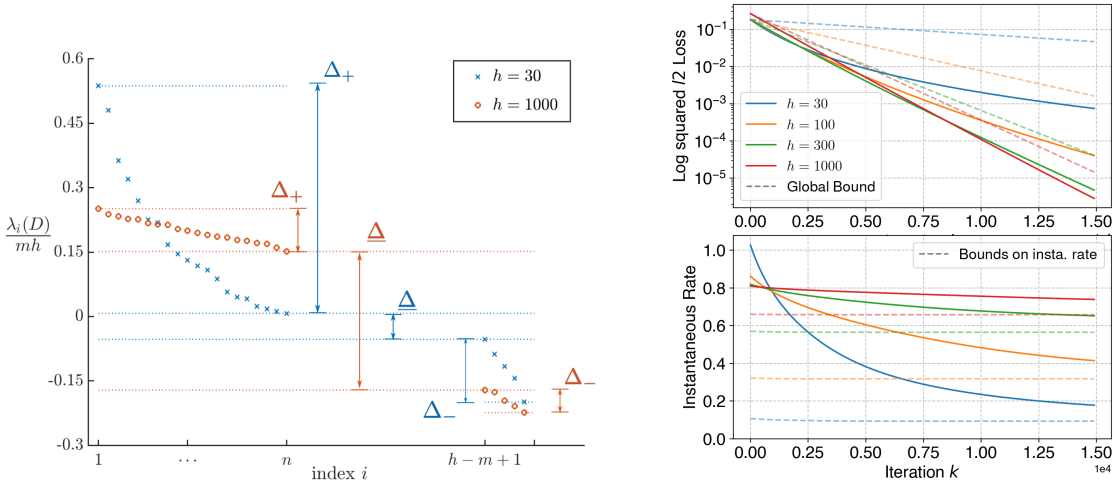

Figure 2: (Left): Scaled eigenvalues of the imbalance matrix D and the corresponding scaled imbalance quantities $\frac{1}{mh}\Delta_+; \frac{1}{mh}\Delta_-; \frac{1}{mh}\underline{\Delta}$ under random initialization, the scaling factor is omitted in the plot annotation for simplicity.
(Right): Gradient descent on $\mathcal{L} = \frac{1}{2}\left\| Y - \frac{1}{\sqrt{mh}} UV^T \right\|_F^2$ for different network width. The dashed lines represent the bound provided by our results (Proposition 1 and Theorem 3).

the top plot in Figure 5). However, interestingly, the instantaneous rate (see the bottom plot in Figure 5) starts off at large value and decreases as training proceeds. At late stage of the training, our lower bound for the instantaneous rate is reasonably good.

(Large width): When $h = 1000$, the level of imbalance $\underline{\Delta}$ is larger compared to spectrum spreads $\Delta_-; \Delta_+$. In this case $2\underline{\Delta}$ is a good global bound on the convergence rate (see the top plot in Figure 5). As for the instantaneous rate, there is no significant variation in the rate and our bound Proposition 1 is reasonably good during training.

Such observation hints more complicated relations between imbalance quantities and the training dynamics. We refer the readers to Appendix A.1 for more detailed discussion.

5.2 Convergence via Imbalanced Initialization

We train the linear network using gradient descent with a fixed small step size on the averaged loss $\mathcal{L}(U; V) = \|Y - XUV\|_F^2 / n$. We use the initialization $U(0) = \alpha_U U_0; V(0) = \alpha_V V_0$ for some randomly sampled $U_0; V_0$ with i.i.d. standard normal entries, and scalars α_U, α_V. Under this setting, we can change the relative scales of $\alpha_U; \alpha_V$ but keep their product fixed, so that we obtain initializations with different level of imbalance c while keeping the initial end-to-end matrix $U(0)V^T(0)$ fixed. To eliminate the effect of ill-conditioned Σ_X on the convergence, we have $\Sigma_X = I_r$ in this experiment.

For comparison, we also consider the balanced initialization that corresponds to the same end-to-end matrix. For a given $\Theta(0) = U(0)V^T(0)$, we choose an arbitrary $Q \in \mathbb{R}^{h \times m}$

with $Q^T Q = I_m$, then a balanced initialization is given by

$$U_{\text{balanced}}(0) = \Theta(0) \left(\Theta^T(0)\Phi_1\Phi_1^T\Theta(0)\right)^{-1/4} Q^T;$$
$$V_{\text{balanced}}(0) = \left(\Theta^T(0)\Phi_1\Phi_1^T\Theta(0)\right)^{1/4} Q:$$

Such initialization ensures the imbalanced is the zero matrix while keeping the end-to-end matrix as $\Theta(0)$. We note here the choice of Q does not affect the error trajectory $E(t)$, hence the loss $\mathcal{L}(t)$.

Figure 3: Convergence of gradient descent on linear networks with different initial imbalance matrices. We plot the loss function $\mathcal{L}$ on the left (Regular scale) and the middle(Log scale) figure. The instantaneous rate $-\dot{\mathcal{L}}/\mathcal{L}$ is shown on the right figure. The dashed line on the middle plot shows the bound on loss function by Theorem 3. Lastly, the dashed line on the right plot shows the lower bound by Proposition 1.

From Fig.3, we see that given fixed step size, the convergence rate is improved as we increase the level of the imbalance at initialization and the balanced initialization is the slowest among all cases. Notably, our lower bound on instantaneous rate is reasonably good for all cases except for case 2 at early training stage.

Moreover, the randomly initialized end-to-end function $(U\ V)(U_0 V_0^T)$ has zero margin, as there is no bound provided for the balance case (Middle plot in Figure 3). Therefore, the margin-based convergence analysis (Arora et al., 2018b) relies on carefully chosen initial end-to-end function and fail on the case of random initialization. On the contrary, random initialization almost surely yields a non-zero imbalance matrix, and our bound accounts for the effect of imbalance in convergence, resulting a much tighter bound on the rate.

Note that the goal of this experiment is to verify the improved convergence rate achieved by gradient flow initialized with a high level of imbalance. To this end, we approximate the continuous dynamics using gradient descent with a fixed small step size. However, this does not imply that one can always accelerate gradient descent by increasing the level of imbalance at initialization. This is because the step size for gradient descent is sometimes chosen to be close to the largest possible for convergence, but it is unknown how the level of imbalance affects such choice. Analyzing the effect of large step size on convergence is subject of current research.

6. Conclusion

In this paper, we study the explicit role of initialization on controlling the convergence and implicit bias of single-hidden-layer linear networks trained under gradient flow. We first provide a lower bound on the instantaneous rate based on the imbalance matrix and the product, from which convergence guarantees are derived based on sufficient imbalance or sufficient margin. We then show that proper initialization enforces the trajectory of network parameters to be exactly (or approximately) constrained in a low-dimensional invariant set, over which minimizing the loss yields the min-norm solution. Combining those results, we obtain a novel non-asymptotic bound regarding the implicit bias of wide linear networks under random initialization towards the min-norm solution. Our analysis, although on a simpler overparametrized model, connects overparametrization, initialization, and optimization. We think it is promising for future research 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.

Acknowledgments

The authors thank the support of the NSF-Simons Research Collaborations on the Mathematical and Scientific Foundations of Deep Learning (NSF grant 2031985), the NSF HDR TRIPODS Institute for the Foundations of Graph and Deep Learning (NSF grant 1934979), the NSF AMPS Program (NSF grant 1736448), and the NSF CAREER Program (NSF grant 1752362).

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*, pages 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*, pages 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, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019a.

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*, pages 322–332, 2019b.

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*, pages 8141–8150, 2019c.

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*, pages 10836–10846, 2019.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, pages 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*, pages 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*, pages 1675–1685, 2019a.

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.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pages 3202–3211. Curran Associates, Inc., 2019.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL `https://proceedings.mlr.press/v9/glorot10a.html`.

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*, pages 6645–6649. IEEE, 2013.

T. H. Gronwall. 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.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 6152–6160, 2017.

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. Matrix Analysis. Cambridge University Press, New York, NY, USA, 2nd edition, 2012.

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

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 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, pages 8157–8166, 2018.

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

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

Hancheng Min, Salma Tarmoun, René Vidal, and Enrique Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 7760–7768. PMLR, 18–24 Jul 2021.

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.

Kathrin Schacke. On the kronecker product. Master's thesis, University of Waterloo, 2004.

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.

Daniel Soudry, Elad Ho er, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data.The Journal of Machine Learning Research, 19(1):2822{2878, 2018.

Salma Tarmoun, Guilherme Franca, Benjamin D Hae ele, and Rene Vidal. Understanding the dynamics of gradient ow in overparameterized linear models. In Proceedings of the 38th International Conference on Machine Learning volume 139 of Proceedings of Machine Learning Research, pages 10153{10161. PMLR, 18{24 Jul 2021.

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

Appendix A. Numerical Verification

A.1 Effect of imbalance quantities on convergence

Experiment settings We consider the matrix factorization problem

$$L = \frac{1}{2}\left\| Y - \frac{1}{\sqrt{mh}} UV^T \right\|_F^2, \qquad U \in \mathbb{R}^{n \times h}, V \in \mathbb{R}^{m \times h},$$

under Xavier initialization (Glorot and Bengio, 2010), where $n = 20, m = 10$. The weights are initialized as $[U]_{ij}, [V]_{ij} \sim \mathcal{N}(0, 1)$. The scaling factor $\frac{1}{\sqrt{mh}}$ ensures that at initialization, the product UV^T keeps the same scale as we vary the hidden layer width h. Lastly, we set $[Y]_{ij} \sim \mathcal{N}(0, 0.1)$, which gives us a randomly chosen target Y with small norm.

Imbalance quantities at initialization When $h \geq n + m$, then with probability 1 under random initialization, the imbalance matrix D has $\text{rank}(D) = n + m$ and it has n positive eigenvalues and m negative ones. Our experiment considers the case of $h = 30$ (small width) and $h = 1000$ (large width).

Figure 4: Scaled eigenvalues of the imbalance matrix D and the corresponding scaled imbalance quantities $\frac{1}{mh}\lambda_+, \frac{1}{mh}\sigma, \frac{1}{mh}\lambda_-$ under random initialization, the scaling factor is omitted in the plot annotation for simplicity. When the network has small width $h = 30$, spectrum spreads λ_+, σ are larger compare to the level of imbalance λ_-. large width $h = 1000$ network shows the opposite.

Remark 13 Notice that the matrix factorization problem with scaling factor $\frac{1}{\sqrt{mh}}$ is equivalent to the regression problem (2) with $X = \frac{1}{\sqrt{mh}} I_n$, the lower bound for the rate is scaled by $\lambda_{\min}(\Sigma_X) = \frac{1}{mh}$. Therefore we analyze the scaled imbalance quantities $\frac{1}{mh}\lambda_+, \frac{1}{mh}\sigma, \frac{1}{mh}\lambda_-$.

We plot in Figure 4 all the non-zero eigenvalues of imbalance D and the imbalance quantities, scaled by $\frac{1}{mh}$. As illustrated by the plot, the imbalance quantities can be understood as

the gaps between certain eigenvalues. It is clear that, compare to small width $h = 50$, large width $h = 1000$ has larger level of imbalance and smaller spectrum spread.

Convergence of Gradient Descent Now under Xavier initialization, we run gradient descent with step size $= 0.05$ and plot

The loss function L in log scale, along with the bound given in Theorem 3;

The instantaneous rate $\frac{L_-}{L}$, along with the bound given in Proposition 1;

for each iteration. We run the experiment under different width $h = 30, 100, 300, 1000$.

Figure 5: Gradient descent on $L = \frac{1}{2} \left\| Y - \frac{1}{\sqrt{mh}} UV^T \right\|_F^2$.

As the width varies, the loss curve behaves differently:

(Small width): When $h = 30$, spectrum spreads $\ ; \ _+$ are larger compared to the level of imbalance $\_$. As we discussed in Section 3 after Proposition 1, the lower bound on the rate is approximately $2\_$, which is not a good global bound[1] for the convergence rate (see the top plot in Figure 5). However, interestingly, the instantaneous rate (see the bottom plot in Figure 5) starts off at large value and decreases as training proceeds. At late stage of the training, our lower bound for the instantaneous rate is reasonably good.

(Large width): When $h = 1000$, the level of imbalance $\_$ is larger compared to spectrum spreads $\ ; \ _+$. In this case $2\_$ is a good global bound on the convergence rate (see the

1. Under random initialization, the margin term in Theorem 3 is zero with high probability. Therefore the global bound generally depends on the imbalance quantities only.

top plot in Figure 5). As for the instantaneous rate, there is no signi cant variation in the rate and our bound Proposition 1 is reasonably good during training.

Our analysis provide some insights to these observations: Following the analysis in Appendix C, the dynamics of the error $E = Y - \frac{1}{mh} UV^T$ can be written as $\dot{E} = \frac{1}{mh} T_t E$, where T_t is a time-variant linear operator on $R^{n \times m}$. Moreover, the eigenvalues of T_t, which characterize the convergence rate of error in di erent directions, can be explicitly expressed as $_i(U(t)U(t)^T) + _j(V(t)V(t)^T); 1 \quad i \quad n; 1 \quad j \quad m$. When UV^T has small norm during training, which is the case in our experiment with target Y having small norm, positive eigenvalues of the imbalance serve as a good approximate to $_i(U(t)U(t)^T); 1 \quad i \quad n$ and negative eigenvalues serve as a good approximate to $_j(V(t)V(t)^T); 1 \quad i \quad m$.

When the width is small, there is large spectrum spread for the eigenvalues of the imbalance matrix, which implies the eigenvalues of T_t have large spread as well. The error E converges faster in some directions but much slower in others, and our lower bound only accounts for the slowest direction in which the error converges. Therefore, the lower bound in Proposition 1 is not tight at early stage of the training. The bound becomes better as training proceeds because at late stage, the main component of the error lies in the slow directions. On the contrary, when the width is large, small spectrum spread implies that the eigenvalues of T_t all concentrate at a certain value, and our lower bound accurately characterize the convergence rate of error in every directions.

In summary, for the convergence of linear networks, we observe two regimes, depending on the relative values between the spectrum spread and the level of imbalance, where the loss curve behaves di erently. Through our experiment, we show that random initialization could fall into one of the regime depending on the network width. Our analysis hints some relation between the imbalance quantities $_+; \quad ; \_$ and the behavior of the loss curve, and establishing such connection formally is left to future research.

A.2 Convergence of single-hidden-layer linear network via imbalanced initialization

The scale of the linear regression problem we consider in Section A.2 and A.3 is $D = 500$, $n = 100$, and $m = 1$.

Generating training data The synthetic training data is generated as following:

1) For data matrix X, rst we generate $X_0 \in R^{n \times D}$ with all the entries sampled from $N(0; 1)$, and take its SVD $X_0 = W \ ^{1=2} \ _1$. Then we let $X = W \ _1$, hence we have all the singular values of X being 1. Herer $= \text{rank}(X) = n = 100$.

2) For Y, we rst sample $N(0; D^{-1} I_D)$, and $N(0; 0.01^2 I_n)$, then we let $Y = X + \ $.

Initialization and Training We set the hidden layer width $h = 500$. We initialize $U(0); V(0)$ with

$$U(0) = \ _U U_0; \ V(0) = \ _V V_0; \ [U_0]_{ij}; [V_0]_{ij} \overset{i.i.d.}{\sim} N(0; 1);$$

and we consider three cases of such initialization: 1) $_U = 0.1; \ _V = 0.1$; 2) $_U = 0.5; \ _V = 0.02$; 3) $_U = 0.05; \ _V = 0.2$. Such setting ensures the initial end-to-end function are

identical for all cases but with di erent imbalance matrices. For these three cases, we run gradient descent on the averaged loss $\tilde{L} = \frac{1}{n}\|Y - XUV^T\|_F^2$ with step size $^2 = 5e-4$.

For comparison, we also consider the balanced initialization that corresponds to the same end-to-end matrix. For a given $(0) = U(0)V^T(0)$, we choose an arbitrary $Q \in R^{h \times m}$ with $Q^TQ = I_m$, then a balanced initialization is given by

$$U_{\text{balanced}}(0) = (0) \quad {}^T(0) \quad {}_1 \quad {}_1^T(0) \quad {}^{1=4}Q^T; \quad V_{\text{balanced}}(0) = \quad {}^T(0) \quad {}_1 \quad {}_1^T(0) \quad {}^{1=4}Q:$$

Such initialization ensures the imbalanced is the zero matrix while keeping the end-to-end matrix as (0). We note here the choice of Q does not a ect the error trajectory $E(t)$, hence the loss $L(t)$.

Figure 6: Convergence of gradient descent on linear networks with di erent initial imbalance matrices. We plot the loss function L on the left (Regular scale) and the middle(Log scale) gure. The instantaneous rate $L=L$ is shown on the right gure. The dashed line on the middle plot shows the bound on loss function by Theorem 3. Lastly, the dashed line on the right plot shows the lower bound by Proposition 1.

From Fig.6, we see that given xed step size, the convergence rate is improved as we increase the level of the imbalance at initialization and the balanced initialization is the slowest among all cases. Notably, our lower bound on instantaneous rate is reasonably good for all cases except for case 2 at early training stage.

Moreover, the randomly initialized end-to-end function $_U \ _V U_0V_0^T$ has zero margin, as there is no bound provided for the balance case (Middle plot in Figure 3). Therefore, the margin-based convergence analysis (Arora et al., 2018b) relies on carefully chosen initial end-to-end function and fail on the case of random initialization. On the contrary, random initialization almost surely yields a non-zero imbalance matrix, and our bound accounts for the e ect of imbalance in convergence, resulting a much tighter bound on the rate.

A.3 Implicit regularization on wide single-hidden-layer linear network

Generating training data The synthetic training data is generated as following:

2. To compute the bound from Theorem 1, the step size is scaled by $n=2$ to account for that the gradient descent uses rescaled loss function.

1) For data matrix X, first we generate $X \in \mathbb{R}^{n \times D}$ with all the entries sampled from $N(0; D^{-1})$;

2) For Y, we first sample $\sim N(0; D^{-1} I_D)$, and $\sim N(0; 0.01^2 I_n)$, then we let $Y = X + \ $.

Initialization and Training We initialize $U(0), V(0)$ with $[U(0)]_{ij} \sim N(0; h^{-1})$, $[V(0)]_{ij} \sim N(0; h^{-1})$ and run gradient descent on the averaged loss $L = \frac{1}{n}\|Y - XUV^T\|_F^2$ with step size $= 5e-3$. The training stops when the loss is below $1e-8$. We run the algorithm for various h from 500 to 10000, and we repeat 5 runs for each h.

Figure 7: Implicit bias of wide single-hidden-layer linear network under random initialization. The line is plotting the average over 5 runs for each h, and the error bar shows the standard deviation. The gradient descent stops at iteration t_f.

Fig.7 clearly shows that the distance between the trained network and the min-norm solution, $\|U(t_f)V^T(t_f) - \hat{\ }\|_F$, decreases as the width h increases and the middle plot verifies the asymptotic rate $O(h^{-1/2})$.

Appendix B. Proofs of Lemma 7 and 8

Proof [Proof of Lemma 7] Under (7), the time derivative of error is given by

$$\dot{E} = -\Sigma_X^{1/2} U_1 U_1^T \Sigma_X^{1/2} E + \ \Sigma_X E V V^T.$$

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

$$\frac{d}{dt}\|E\|_F^2 = \frac{d}{dt}\mathrm{tr}(E^T E) = -2\,\mathrm{tr}\left(E^T \Sigma_X^{1/2} U_1 U_1^T \Sigma_X^{1/2} E + E^T \ \Sigma_X E V V^T\right). \tag{B.1}$$

Use the trace inequality (Sheng-De Wang et al., 1986, Lemma 1) to get the lower bound the trace of two matrices respectively as

$$\begin{aligned}
\mathrm{tr}\left(E^T \Sigma_X^{1/2} U_1 U_1^T \Sigma_X^{1/2} E\right) &= \mathrm{tr}\left(\Sigma_X^{1/2} E E^T \Sigma_X^{1/2} U_1 U_1^T\right) \\
&\geq \lambda_r(U_1 U_1^T)\,\mathrm{tr}\left(\Sigma_X^{1/2} E E^T \Sigma_X^{1/2}\right) \\
&= \lambda_r(U_1 U_1^T)\,\mathrm{tr}\left(\Sigma_X E E^T\right) \\
&\geq \lambda_r(U_1 U_1^T)\,\lambda_r(\Sigma_X)\,\mathrm{tr}(E E^T) \\
&= \lambda_r(U_1 U_1^T)\,\lambda_r(\Sigma_X)\|E\|_F^2;
\end{aligned} \tag{B.2}$$

and

$$\begin{aligned} \mathrm{tr}\left(E^T \Sigma_x E V V^T\right) &\geq \lambda_m(VV^T)\,\mathrm{tr}\left(E^T \Sigma_x E\right) \\ &= \lambda_m(VV^T)\,\mathrm{tr}\left(\Sigma_x E E^T\right) \\ &\geq \lambda_m(VV^T)\,\lambda_r(\Sigma_x)\,\mathrm{tr}(EE^T) \\ &= \lambda_m(VV^T)\,\lambda_r(\Sigma_x)\|E\|_F^2 . \end{aligned} \tag{B.3}$$

Combine (B.1) with (B.2)(B.3), we have

$$\frac{d}{dt}\|E\|_F^2 \leq -2\lambda_r(\Sigma_x)\,\lambda_r(U_1 U_1^T) + \lambda_m(VV^T)\,\|E\|_F^2 \tag{B.4}$$

Notice that $\frac{1}{2}\|E\|_F^2$ is exactly $\tilde{L} = L - L^*$. It follows from (B.4) that

$$\frac{\dot{\tilde{L}}}{\tilde{L}} \leq -2\lambda_r(\Sigma_x)\,\lambda_r(U_1 U_1^T) + \lambda_m(VV^T) .$$

■

Proof [Proof of Lemma 8]

From the imbalance equation $A^T A - BB^T = D$, we have

$$(B^T B)^2 = B^T(BB^T)B = B^T(A^T A - D)B = B^T A^T AB - B^T DB .$$

Let $z_m \in S^{m-1}$ be the eigenvector of $(B^T B)^2$ (or $B^T B$) associated with eigenvalue $\sigma_m^2(B^T B)$ (or $\lambda_m(B^T B)$). The one have

$$\begin{aligned} \sigma_m^2(B^T B) = z_m^T (B^T B)^2 z_m &= z_m^T B^T A^T AB z_m - z_m^T B^T DB z_m \\ &\geq \lambda_m(B^T A^T AB) - z_m^T B^T DB z_m ; \\ &= \sigma_m^2(AB) - z_m^T B^T DB z_m \end{aligned} \tag{B.5}$$

and the rest of proof is to find a lower bound for $-z_m^T B^T DB z_m$.

First of all, we know that D has at most m negative eigenvalues: If D has more than m negative eigenvalues, then the subspace spanned by the all negative eigenvectors has dimension at least $m+1$, which must have non-trivial intersection with $\ker(-B^T)$, then there exists a nonzero vector $z \in \ker(B^T)$ such that $z^T Dz < 0$, which would imply $z^T A^T Az = z^T Dz < 0$, a contradiction.

When D has less than m negative eigenvalues, then $\lambda = 0$ and we simply lower bound $-z_m^T B^T DB z_m$ as

$$\begin{aligned} \sigma_m^2(B^T B) &\geq \sigma_m^2(AB) - z_m^T B^T DB z_m \\ &\geq \sigma_m^2(AB) - z_m^T B^T B z_m \\ &= \sigma_m^2(AB) - \lambda_m(B^T B) : \end{aligned}$$

This quadratic inequality w.r.t. $\lambda_m(B^T B)$ has nonnegative solutions

$$\lambda_m(B^T B) \geq \frac{-\lambda + \sqrt{\lambda^2 + 4\sigma_m^2(AB)}}{2} ;$$

which is exactly (23) when $\lambda = 0$.

When D has exactly m negative eigenvalues, the easy case is one with $h = m$, i.e. all eigenvalues of D are negative. We simply lower bound $z_m^T B^T D B z_m$ as

$$\begin{aligned}\sigma_m^2(B^T B) \quad & \sigma_m^2(AB) \quad z_m^T B^T D B z_m \\ & \sigma_m^2(AB) \quad (-\lambda z_m^T B^T B z_m) \\ = \quad & \sigma_m^2(AB) + \lambda\, \sigma_m(B^T B).\end{aligned}$$

This quadratic inequality w.r.t. $\sigma_m(B^T B)$ has nonnegative solutions

$$\sigma_m(B^T B) \quad \frac{-\lambda + \sqrt{\lambda^2 + 4\,\sigma_m^2(AB)}}{2};$$

which is exactly (23) when $\lambda = 0$.

Now we only left to prove the bound for the case $h > m$. We first consider any orthogonal matrix $Q \in O(h)$, we have $Q^T A^T A Q \quad Q^T B B^T Q = Q^T D Q$, $AQQ^T B = AB$, and $\sigma_m(B^T Q^T Q B) = \sigma_m(B^T B)$. Then it suffices to study the rotated matrices $\tilde{A} = AQ$, $\tilde{B} = Q^T B$, with $\tilde{A}^T \tilde{A} \quad \tilde{B}\tilde{B}^T = Q^T D Q$, $\tilde{A}\tilde{B} = AB$ and find a lower bound on $\sigma_m(\tilde{B}^T \tilde{B})$. We can pick Q that diagonalize D, thus without loss of generality, we assume D is diagonal and the eigenvalues are in decreasing order.

Since $h > m$, we write the diagonal D as a block matrix $D = \begin{matrix} D_+ & 0 \\ 0 & D_- \end{matrix}$; where

$$\begin{aligned}D_+ &= \operatorname{diag}\{\sigma_1(D), \quad , \sigma_{h-m}(D)\} \\ D_- &= \operatorname{diag}\{\sigma_{h-m+1}(D), \quad , \sigma_h(D)\} = \operatorname{diag}\{-\sigma_m(-D), \quad , -\sigma_1(-D)\}.\end{aligned}$$

Here, notice that D_+ is positive semi-definite and D_- positive definite with

$$D_+ \quad I_{h-m}, \qquad D_- \quad -\lambda I_m. \tag{B.6}$$

Now we write A, B as block matrices as well

$$\begin{aligned}A &= \begin{matrix} A_+ & A_- \end{matrix}, \quad B = \begin{matrix} B_+ \\ B_- \end{matrix}, \\ A_+ &\in \mathbb{R}^{r \times (h-m)}, A_- \in \mathbb{R}^{r \times m}, B_+ \in \mathbb{R}^{(h-m) \times m}, B_- \in \mathbb{R}^{m \times m},\end{aligned}$$

from which we can rewrite equations $A^T A \quad B B^T = D$ as

$$\begin{matrix} A_+^T \\ A_-^T \end{matrix} \begin{matrix} A_+ & A_- \end{matrix} \quad \begin{matrix} B_+ \\ B_- \end{matrix} \begin{matrix} B_+^T & B_-^T \end{matrix} = \begin{matrix} D_+ & 0 \\ 0 & D_- \end{matrix}.$$

By inspection, the equality for each block gives us

$$A_+^T A_+ = B_+ B_+^T + D_+; \tag{B.7}$$

$$A_-^T A_- = B_- B_-^T - D_-; \tag{B.8}$$

$$A_+^T A_- = B_+ B_-^T. \tag{B.9}$$

With these equalities, we know the following matrix is p.s.d., for any $\lambda > 0$,

$$\begin{bmatrix} B_+B_+^T + \lambda I_{h\,m} & B_+B^T \\ B\,B_+^T & B\,B^T - I_m \end{bmatrix} \overset{(B.6)}{=} \begin{bmatrix} B_+B_+^T + \lambda & B_+B^T \\ B\,B_+^T & B\,B^T \end{bmatrix} - \begin{bmatrix} A_+^T \\ A^T \end{bmatrix} \begin{bmatrix} A_+ & A \end{bmatrix} \succeq 0. \tag{B.10}$$

Since $B_+B_+^T + \lambda I_{h\,m} \succ 0$, positive semi-definiteness (B.10) is equivalent to

$$B\,B^T - I_m - B\,B_+^T(B_+B_+^T + \lambda I_{h\,m})^{-1}B_+B^T \succeq 0. \tag{B.11}$$

Now we use Woodbury's Identity (Horn and Johnson, 2012, 0.7.4), which says for matrices M, N, P with appropriate dimensions, we have

$$(M + P^TNP)^{-1} = M^{-1} - M^{-1}P^T(PM^{-1}P^T + N^{-1})^{-1}PM^{-1},$$

if all inverses exist. Let $M = I_m, N = \lambda^{-1}I_{h\,m}, P = B_+$, we have

$$(I_m + \lambda^{-1}B_+^TB_+)^{-1} = I_m - B_+^T(\lambda I_{h\,m} + B_+B_+^T)^{-1}B_+,$$

which leads to

$$B\,(I_m + \lambda^{-1}B_+^TB_+)^{-1}B^T = B\,B^T - B\,B_+^T(\lambda I_{h\,m} + B_+B_+^T)^{-1}B_+B^T. \tag{B.12}$$

Using (B.12), we can rewrite (B.11) as

$$I_m - B\,(I_m + \lambda^{-1}B_+^TB_+)^{-1}B^T \succeq 0. \tag{B.13}$$

Consider the following matrix congruence

$$\begin{bmatrix} I_m & B \\ B^T & I_m + \lambda^{-1}B_+^TB_+ \end{bmatrix}$$

$$= S_1 \begin{bmatrix} I_m - B\,(I_m + \lambda^{-1}B_+^TB_+)^{-1}B^T & 0 \\ 0 & I_m + \lambda^{-1}B_+^TB_+ \end{bmatrix} S_1^T \tag{B.14}$$

$$= S_2 \begin{bmatrix} I_m & 0 \\ 0 & I_m + \lambda^{-1}B_+^TB_+ - \lambda^{-1}B^TB \end{bmatrix} S_2^T \tag{B.15}$$

where

$$S_1 = \begin{bmatrix} I_m & -B\,(I_m + \lambda^{-1}B_+^TB_+)^{-1} \\ 0 & I_m \end{bmatrix}, \quad S_2 = \begin{bmatrix} I_m & 0 \\ -\lambda^{-1}B^T & I_m \end{bmatrix},$$

and S_1, S_2 are non-singular. By Sylvester's Intertia Theorem (Horn and Johnson, 2012, Theorem 4.5.8), the block diagonal matrix shown in (B.14) has exactly the same number of positive eigenvalues as the one shown in (B.15), and the number of positive eigenvalues is m, according to (B.13). Then for the block diagonal matrix in (B.15), we must have

$$I_m + \lambda^{-1}B_+^TB_+ - \lambda^{-1}B^TB \succeq 0,$$

hence

$$\begin{aligned}
&\begin{bmatrix} 0 & I_m \\ \hat{\lambda}B_+^TB_+ - \lambda B^TB & \hat{\lambda}I_m - \lambda B_+^TB_+ + \hat{\lambda}B^TB \end{bmatrix} \\
&\sim \begin{bmatrix} 0 & \hat{\lambda}I_m - \lambda B_+^TB_+ + \hat{\lambda}B^TB \\ \hat{\lambda}B_+^TB_+ - \lambda B^TB & \hat{\lambda}I_m - \lambda B_+^TB_+ + \hat{\lambda}B^TB \end{bmatrix} \\
&\sim \begin{bmatrix} \hat{\lambda}B_+^TB_+ - \lambda B^TB & \hat{\lambda}I_m - \lambda B_+^TB_+ + \hat{\lambda}B^TB \\ 0 & \hat{\lambda}I_m - \lambda B_+^TB_+ + \hat{\lambda}B^TB \end{bmatrix} \\
&\sim \begin{bmatrix} \hat{\lambda}B_+^TB_+ - \lambda B^TB & \hat{\lambda}I_m + (\hat{\lambda}-\lambda)(B_+^TB_+ + B^TB) \\ 0 & \hat{\lambda}I_m + (\hat{\lambda}-\lambda)B^TB \end{bmatrix}, \qquad \text{(B.16)}
\end{aligned}$$

where the last equivalence uses the fact $B^TB = B_+^TB_+ + B^TB$. This suggests that

$$\begin{aligned}
B^TDB &= B_+^T{}_+B_+ \quad B^T \quad B \qquad \hat{\lambda}B_+^TB_+ \quad \lambda B^TB \\
&\overset{\text{(B.16)}}{\sim} \hat{\lambda}I_m + (\hat{\lambda}-\lambda)B^TB \qquad \text{(B.17)}
\end{aligned}$$

Lastly, from (B.5) we have

$$\begin{aligned}
\sigma_m^2(B^TB) &= z_m^T(B^TB)^2z_m \qquad \sigma_m^2(AB) \quad z_m^TB^TDBz_m \\
&\overset{\text{(B.17)}}{\sim} \sigma_m^2(AB) + \hat{\lambda} \quad (\hat{\lambda} \quad \lambda)z_m^TB^TBz_m \\
&= \sigma_m^2(AB) + \hat{\lambda} \quad (\hat{\lambda} \quad \lambda)\ \sigma_m(B^TB).
\end{aligned}$$

This quadratic inequality w.r.t. $\sigma_m(B^TB)$ has nonnegative solutions

$$\sigma_m(B^TB) \quad \frac{-\hat{\lambda} + \lambda + \sqrt{(\lambda - \hat{\lambda})^2 + 4\hat{\lambda}\lambda + 4\ \sigma_m^2(AB)}}{2} = \frac{\hat{\lambda} + \lambda + \sqrt{(\hat{\lambda} + \lambda)^2 + 4\ \sigma_m^2(AB)}}{2}.$$

Since we can choose any $\hat{\lambda} > 0$, we have

$$\sigma_m(B^TB) \quad \lim_{\hat{\lambda}\to 0} \frac{\hat{\lambda} + \lambda + \sqrt{(\hat{\lambda} + \lambda)^2 + 4\ \sigma_m^2(AB)}}{2} = \frac{\lambda + \sqrt{(\lambda)^2 + 4\ \sigma_m^2(AB)}}{2}.$$

This is exactly (23).

(Note that when $\lambda > 0$, one can pick $\hat{\lambda} = \lambda$ and obtain the desired bound directly. Taking the limit $\hat{\lambda} \to 0$ is necessary only when $\lambda = 0$). ■

Appendix C. Detailed analysis for the matrix factorization problem

Consider the gradient flow on $L(U,V) = \frac{1}{2}\|Y - UV^T\|_F^2$,[3] where $Y \in \mathbb{R}^{r \times m}, U \in \mathbb{R}^{r \times h}, V \in \mathbb{R}^{h \times n}$. Still we define $E := Y - UV^T$.

3. When $\Sigma_x = I_r$, $L(U,V) = \frac{1}{2}\|W^TY - \Sigma_x^{1/2}U_1V^T\|_F^2 = \frac{1}{2}\|\Upsilon - U_1V^T\|_F^2$ is exactly of this form.

We start with the exact expression for the instantaneous rate

$$\frac{\dot{L}}{L} = -2\frac{\mathrm{tr}(E^T\dot{E})}{\|E\|_F^2} = 2\frac{\mathrm{tr}(E^T(\dot{U}V^T + U\dot{V}^T))}{\|E\|_F^2} = 2\frac{\mathrm{tr}(E^T(UU^TE + EVV^T))}{\|E\|_F^2}.$$

If we define the Hermitian linear operator $T_{U,V}$ on $\mathbb{R}^{m\times n}$ as $T_{U,V}E = UU^TE + EVV^T$. Then the instantaneous rate is actually a Rayleigh quotient

$$\frac{\dot{L}}{L} = -2\frac{\langle E, T_{U,V}E\rangle_F}{\langle E, E\rangle_F},$$

where $\langle\cdot,\cdot\rangle_F$ is the Frobenius inner product on $\mathbb{R}^{m\times n}$. Notice that both E and $T_{U,V}$ depend on U, V here.

Now our goal is find the best lower bound on the instantaneous rate, provided that the imbalance $D = U^TU - V^TV$ and product $W = UV^T$ is known to us (Recall that we can express the instantaneous rate exactly by the imbalance and product in the scalar case (14)). That is, the following problem

Problem 1 Suppose $h \geq \min\{m, n\}$. Given $Y \in \mathbb{R}^{m\times n}$, $D \in \mathbb{R}^{h\times h}$ and $W \in \mathbb{R}^{m\times n}$, find

$$c(Y, D, W) = \min -2\frac{\langle E, T_{U,V}E\rangle_F}{\langle E, E\rangle_F} : U^TU - V^TV = D, UV^T = W.$$

c is the best bound we can obtain by knowing the imbalance D and the product W, but it also depends on Y because we have defined $E = Y - UV^T$. Problem 1 is generally hard to solve except for very special cases, thus we consider a lower bound for it:

$$\begin{aligned} c(Y, D, W) &= \min -2\frac{\langle E, T_{U,V}E\rangle_F}{\langle E, E\rangle_F} : U^TU - V^TV = D, UV^T = W \\ &\geq \min -2\min_Y \frac{\langle E, T_{U,V}E\rangle_F}{\langle E, E\rangle_F} : U^TU - V^TV = D, UV^T = W \\ &= \min -2\lambda_{\min}(T_{U,V}) : U^TU - V^TV = D, UV^T = W := c(D, W), \end{aligned}$$

Here the second equality is obtained by choosing $Y = E_{\min} + UV^T$ where $E_{\min}$ is the least eigenmatrix of $T_{U,V}$. Moreover, one can show that (Schacke, 2004)

$$\lambda_{\min}(T_{U,V}) = \lambda_r(UU^T) + \lambda_m(VV^T).$$

This left us to consider the following problem

Problem 2 Suppose $h \geq \min\{m, n\}$. Given $Y \in \mathbb{R}^{m\times n}$, $D \in \mathbb{R}^{h\times h}$ and $W \in \mathbb{R}^{m\times n}$, find

$$c(D, W) = \min -2(\lambda_r(UU^T) + \lambda_m(VV^T)) : U^TU - V^TV = D, UV^T = W.$$

Now following the results in Section 3.3, one obtain, by Lemma 8,

$$c(D, W) \geq -\alpha_+ + \sqrt{(\alpha_+ + \beta_-)^2 + 4\sigma_m^2(W)} \geq -\alpha_+ + \sqrt{(\alpha_+ - \beta_-)^2 + 4\sigma_r^2(W)}.$$

It turns out that this lower bound for $C(D, W)$ is tight in most cases. Formally speaking, we have

Proposition 14 Suppose $h \leq \min\{r, m\}$. When $r \neq m$, we have

$$c(D;W) = \frac{\alpha_+ + \sqrt{(\alpha_+ + \beta_-)^2 + 4\sigma_m^2(W)}}{2} + \frac{\alpha_+ + \sqrt{(\alpha_+ + \beta_-)^2 + 4\sigma_r^2(W)}}{2}.$$

In summary, we have shown

$$\frac{\tilde{L}}{\tilde{L}}\, c(Y;D;W) \geq c(D;W) \geq \frac{\alpha_+ + \sqrt{(\alpha_+ + \beta_-)^2 + 4\sigma_m^2(W)}}{2} + \frac{\alpha_+ + \sqrt{(\alpha_+ + \beta_-)^2 + 4\sigma_r^2(W)}}{2}.$$

Therefore, the bound in Proposition 1, in some sense, is not the best lower bound we can obtain. Potential improvement for the current results would be studying $c(Y;D;W)$ directly.

We end this section with the proof for Proposition 14.

Proof [Proof of Proposition 14] Assume $r > m$, the proof for the opposite case is identical.

The statement in Proposition 14 is equivalent to that given $W \in \mathbb{R}^{r \times m}$ and a symmetric $D \in \mathbb{R}^{h \times h}$ with $\text{rank}(D) \leq r + m$, there exist $U \in \mathbb{R}^{r \times h}, V \in \mathbb{R}^{m \times h}$ such that

$$U^T U - V^T V = D, \quad UV^T = W,$$

and

$$2(\sigma_r(UU^T) + \sigma_m(VV^T)) = \alpha_+ + \sqrt{(\alpha_+ + \beta_-)^2 + 4\sigma_m^2(W)} + \alpha_+ + \sqrt{(\alpha_+ + \beta_-)^2 + 4\sigma_r^2(W)}.$$

Now consider the SVD of W, $W = P\bar{W}Q^T$, where $P \in O(r), Q \in O(m)$ and $\bar{W} \in \mathbb{R}^{r \times m}$ has singular values of W on its main diagonal. Additionally, we have the eigendecomposition of D, $D = R\bar{D}R^T$, where $R \in O(h)$ and $\bar{D}$ has eigenvalues of D on its diagonal. Notice that there are at most $r + m$ non-zero eigenvalues of D.

With the decomposition of W and D. We only need to show that there exist $\bar{U} \in \mathbb{R}^{r \times h}, \bar{V} \in \mathbb{R}^{m \times h}$ such that

$$\bar{U}^T\bar{U} - \bar{V}^T\bar{V} = \bar{D}, \quad \bar{U}\bar{V}^T = \bar{W},$$

and

$$2(\sigma_r(\bar{U}\bar{U}^T) + \sigma_m(\bar{V}\bar{V}^T)) = \alpha_+ + \sqrt{(\alpha_+ + \beta_-)^2 + 4\sigma_m^2(W)} + \alpha_+ + \sqrt{(\alpha_+ + \beta_-)^2 + 4\sigma_r^2(W)}.$$

Then the statement on the existence of U, V is true by setting $U = P\bar{U}R^T$ and $V = Q\bar{V}R^T$. Therefore, without loss of generality, we assume W is main diagonal and D diagonal. Furthermore, all main diagonal entries of W are non-negative, denoted by $\sigma_i(W), i = 1, \ldots, m$, and we reorder the diagonal entries of D such that

$$D = \begin{bmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_m & & & \\ & & & \lambda_{m+1}(D) & & \\ & & & & \ddots & \\ & & & & & \lambda_r(D) \\ & & & & & & 0 \end{bmatrix},$$

where

$$\Sigma_i = \begin{bmatrix} \sigma_i(D) & 0 \\ 0 & \sigma_{h-m+i}(D) \end{bmatrix} = \begin{bmatrix} \sigma_i(D) & 0 \\ 0 & -\sigma_{m-i+1}(-D) \end{bmatrix}.$$

One can verify that this is indeed a valid reordering of the eigenvalues of D. Notice that under this reordering, we are pairing some eigenvalues of D, as seen in the definition of the 2×2 blocks $\Sigma_i, i = 1, \ldots, m$. Moreover, for any of these pairs Σ_i, we have one non-negative eigenvalue $\sigma_i(D)$ and one non-positive eigenvalue $\sigma_{h-m+i}(D)$.

We claim that given such W and D, one of the solution U, V to

$$U^T U - V^T V = D, \; UV^T = W,$$

is of the form

$$U = \begin{bmatrix} U_{\text{nonzero}} & 0_{r \times (h-r-m)} \end{bmatrix}, U_{\text{nonzero}} = \begin{bmatrix} u_1^T & & & \\ & \ddots & & \\ & & u_m^T & \\ & & & \sqrt{\sigma_{m+1}(D)} \\ & & & & \ddots \\ & & & & & \sqrt{\sigma_r(D)} \end{bmatrix},$$

$$V = \begin{bmatrix} V_{\text{nonzero}} & 0_{m \times (h-2m)} \end{bmatrix}, V_{\text{nonzero}} = \begin{bmatrix} v_1^T & & \\ & \ddots & \\ & & v_m^T \end{bmatrix},$$

where $u_i, v_i, i = 1, \ldots, m$ are 2×1 vectors, and they necessarily satisfies for $i = 1, \ldots, m$,

$$u_i u_i^T - v_i v_i^T = \Sigma_i, \; u_i^T v_i = \sigma_i(W). \tag{C.18}$$

Let $u_i = \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix}, v_i = \begin{bmatrix} v_{i1} \\ v_{i2} \end{bmatrix}$, (C.18) yields 4 quadratic equations in terms of 4 unknowns $u_{i1}, u_{i2}, v_{i1}, v_{i2}$. (C.18) has at least one real solution whose exact expression is too complicated to write here, but for that real solution, we have

$$u_i^T u_i = \frac{\sigma_i(D) - \sigma_{m-i+1}(-D) + \sqrt{(\sigma_i(D) + \sigma_{m-i+1}(-D))^2 + 4\sigma_i^2(W)}}{2},$$

and

$$v_i^T v_i = \frac{\sigma_i(D) + \sigma_{m-i+1}(-D) + \sqrt{(\sigma_i(D) + \sigma_{m-i+1}(-D))^2 + 4\sigma_i^2(W)}}{2}.$$

Notice that under this construction $\sigma_i(UU^T) = u_i^T u_i$; $\sigma_i(VV^T) = v_i^T v_i, i = 1, \ldots, m$ and $\sigma_j(UU^T) = \sigma_j(D), j = m+1, \ldots, r$. Therefore,

$$\sigma_r(UU^T) + \sigma_m(VV^T) = \sigma_r(D) + \frac{\sigma_1(D) + \sigma_m(-D) + \sqrt{(\sigma_1(D) + \sigma_m(-D))^2 + 4\sigma_m^2(W)}}{2},$$

which is exactly the lower bound we get using Lemma 8. Then the statement

$$2(\ _r(UU^T) + \ _m(VV^T)) = \ _+ + \sqrt{(\ _+ + \_)^2 + 4\ _m^2(W)} \qquad + \sqrt{(\quad + \_)^2 + 4\ _r^2(W)}:$$

follows from the definition of $\ _+; \quad ; \_$. ■

Appendix D. Comparison with the NTK Initialization for wide single-hidden-layer linear networks

In Section 4.2, we analyzed implicit bias of wide single-hidden-layer linear networks under properly scaled random initialization. Our initialization for network weights $U; V$ is different from the typical setting in previous works (Jacot et al., 2018; Du and Hu, 2019; Arora et al., 2019c). In this section, we show that under our setting, the gradient flow is related to the NTK flow by 1) reparametrization and rescaling in time $\ $; 2) proper scaling of the network output. The use of output scaling is also used in Arora et al. (2019c).

In this paper we work with a single-hidden-layer linear network defined as $f : R^n \to R^m; f(x; V; U) = VU^T x$, which is parametrized by $U; V$. Then we analyze the gradient flow on the loss function $L(V; U) = \frac{1}{2}\ Y - XUV^T\ _F^2$, given the data and output matrix $X; Y$. Lastly, in Section 4.2, we initialize $U(0); V(0)$ such that all the entries are randomly drawn from $N(0; h^{\ 2\ (1=4 < \ \ 1=2)})$, where h is the hidden layer width.

Now we define $\tilde{U} := h^{\ }U; \tilde{V} := h^{\ }V$, then the loss function can be written as

$$\begin{aligned} L(V; U) = \tilde{L}(\tilde{V}; \tilde{U}) &= \frac{1}{2}\ Y - \frac{1}{h^2}X\tilde{U}\tilde{V}^T\ _F^2 = \frac{1}{2}\ Y - \frac{\sqrt{m}}{h^{2\ \frac{1}{2}}}\sqrt{\frac{1}{mh}}X\tilde{U}\tilde{V}^T\ _F^2 \\ &= \frac{1}{2}\sum_{i=1}^n \ y^{(i)} - \frac{\sqrt{m}}{h^{2\ \frac{1}{2}}}\sqrt{\frac{1}{mh}}\tilde{V}\tilde{U}^T x^{(i)}\ _2^2 \\ &:= \sum_{i=1}^n \ y^{(i)} - \frac{\sqrt{m}}{h^{2\ \frac{1}{2}}}\tilde{f}(x; \tilde{V}; \tilde{U})\ _2^2 \end{aligned}$$

Notice that $\tilde{f}(x; \tilde{V}; \tilde{U}) = \sqrt{\frac{1}{mh}}\tilde{V}\tilde{U}^T x$ is the typical network discussed in previous works (Jacot et al., 2018; Du and Hu, 2019; Arora et al., 2019c). When all the entries of $U(0); V(0)$ are initialized randomly as $N(0; h^{\ 2})$, the entries of $\tilde{U}(0); \tilde{V}(0)$ are random samples from $N(0; 1)$, which is the typical choice of initialization for NTK analysis.

However, the difference is that $\tilde{f}(x; \tilde{V}; \tilde{U})$ is scaled by $\frac{\sqrt{m}}{h^{2\ \frac{1}{2}}}$. In previous work showing non-asymptotic bound between wide neural networks and its infinite width limit (Arora et al., 2019c, Theorem 3.2), the wide neural network is scaled by a small constant such that the prediction by the trained network is within $\ $-distance to the one by the kernel predictor of its NTK. Moreover, Arora et al. (2019c) suggests $\frac{1}{\ }$ should scale as $\mathrm{poly}(\frac{1}{\ })$, i.e., to make sure the trained network is arbitrarily close to the kernel predictor, $\ $ should be vanishingly small. In our setting, the random initialization implicitly enforces such a vanishing scaling $\frac{\sqrt{m}}{h^{2\ \frac{1}{2}}}$, as the width of network increases.

Lastly, we show that the gradient flow on $L(V;U)$ only differs from the flow on $\tilde{L}(\tilde{V};\tilde{U})$ by the time scale.

Suppose U, V follows the gradient flow on $L(V;U)$, we have

$$\begin{aligned}\frac{1}{h}\frac{\partial}{\partial U}L(V;U) &= \frac{1}{h}X^T(Y - XUV^T)V \\ &= \frac{1}{h^2}X^TY - \frac{1}{h^2}X^TXU\tilde{V}^T\tilde{V} = \frac{\partial}{\partial\tilde{U}}\tilde{L}(\tilde{V};\tilde{U}); \end{aligned} \tag{D.1}$$

and

$$\begin{aligned}\frac{1}{h}\frac{\partial}{\partial V}L(V;U) &= \frac{1}{h}(Y - XUV^T)^TXU \\ &= \frac{1}{h^2}Y^TXU - \frac{1}{h^2}\left(X\tilde{U}\tilde{V}^T\right)^TX\tilde{U} = \frac{\partial}{\partial\tilde{V}}\tilde{L}(\tilde{V};\tilde{U}): \end{aligned} \tag{D.2}$$

From (D.1), we have

$$\begin{aligned}\dot{U} = -\frac{\partial}{\partial U}L(V;U), \quad \frac{1}{h}\dot{\tilde{U}} &= -\frac{\partial}{\partial U}L(V;U) \\ , \quad \frac{1}{h}\dot{\tilde{U}} &= -h\frac{\partial}{\partial\tilde{U}}\tilde{L}(\tilde{V};\tilde{U}) \\ , \quad \dot{\tilde{U}} &= -h^2\frac{\partial}{\partial\tilde{U}}\tilde{L}(\tilde{V};\tilde{U}); \end{aligned} \tag{D.3}$$

Similarly from (D.2) we have

$$\dot{V} = -\frac{\partial}{\partial V}L(V;U), \quad \dot{\tilde{V}} = -h^2\frac{\partial}{\partial\tilde{V}}\tilde{L}(\tilde{V};\tilde{U}): \tag{D.4}$$

From (D.3) and (D.4) we know that the gradient flow on $L(V;U)$ w.r.t. time t essentially runs the gradient flow on $\tilde{L}(\tilde{V};\tilde{U})$ with an scaled-up rate by h^2.

Appendix E. Proofs of Proposition 9, Lemma 11 and Theorem 12

We start with the proof of Proposition 9.

Proof [Proof of Proposition 9] Since $c(0) > 0$, for the gradient system (7), the states (parameters) $(U_1;V)$ converge either to an equilibrium point which minimizes the potential $\frac{1}{2}\|E\|_F^2 = L - L^*$ or have its l_2-norm grow to infinity (Hirsch et al., 1974).

Consider the following dynamics

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

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

From (E.1), we have

$$\begin{aligned}\frac{d}{dt}\|Z\|_F^2 &= 2\,\mathrm{tr}\left(Z^T A_Z Z\right)\\ &= 2\,\mathrm{tr}\left(ZZ^T A_Z\right)\\ &\quad 2\|A_Z\|_2\,\mathrm{tr}\left(ZZ^T\right)\\ &= 2\left(\lambda_x^{1/2}\|E\|_2\|Z\|_F^2\right)\\ &\quad 2\left(\lambda_1^{1/2}(\lambda_x)\|E\|_2\|Z\|_F^2\right)\\ &\quad 2\left(\lambda_1^{1/2}(\lambda_x)\|E\|_F\|Z\|_F^2\right).\end{aligned}$$

By Grönwall's inequality (Grönwall, 1919), we have

$$\|Z(t)\|_F^2 \quad \exp\left(\int_0^t 2\left(\lambda_1^{1/2}(\lambda_x)\|E(\tau)\|_F\right)d\tau\right)\|Z(0)\|_F^2.$$

Finally, by Theorem 3, we have $\|E(t)\|_F \quad \exp\left(-\lambda_n(\lambda_x)c(0)t/2\right)\|E(0)\|_F,\ \forall t > 0$, since $\|E\|_F = \sqrt{2(L - L^\star)}$, which leads to

$$\begin{aligned}&\exp\left(\int_0^t 2\left(\lambda_1^{1/2}(\lambda_x)\|E(\tau)\|_F\right)d\tau\right)\\ &\exp\left(2\left(\lambda_1^{1/2}(\lambda_x)\|E(0)\|_F\right)\int_0^t \exp\left(-\lambda_n(\lambda_x)c(0)\tau/2\right)d\tau\right)\\ &\exp\left(2\left(\lambda_1^{1/2}(\lambda_x)\|E(0)\|_F\right)\int_0^{\infty} \exp\left(-\lambda_n(\lambda_x)c(0)\tau/2\right)d\tau\right)\\ &= \exp\left(\frac{4\left(\lambda_1^{1/2}(\lambda_x)\right)}{c(0)\,\lambda_n(\lambda_x)}\|E(0)\|_F\right).\end{aligned}$$

Therefore we have

$$\|Z(t)\|_F^2 \quad \exp\left(\frac{4\left(\lambda_1^{1/2}(\lambda_x)\right)}{c(0)\,\lambda_n(\lambda_x)}\|E(0)\|_F\right)\|Z(0)\|_F^2,$$

which implies that the trajectory $(V(t), U_1(t)), t > 0$ is bounded, i.e. its l_2-norm can not grow to infinity, then it has to converge to some equilibrium point $(V(\infty), U_1(\infty))$ such that its potential is zero, i.e., $E(V(\infty), U_1(\infty)) = 0$. ■

Now we turn to prove Lemma 11 and Theorem 12. We need a basic result in random matrix theory

Lemma E.1 Given $m, n \in \mathbb{N}$ with $m \quad n$. Let A be an $n \quad m$ random matrix with i.i.d. standard normal entries $A_{ij} \sim N(0,1)$. For > 0, with probability at least $1 \quad 2\exp(-^2)$, we have

$$\sqrt{n} \quad (\sqrt{m} +) \quad \sigma_m(A) \quad \sigma_1(A) \quad \sqrt{n} + (\sqrt{m} +).$$

The proof can be found in Davidson and Szarek (2001, Theorem 2.13). We also need the following inequality.

Lemma E.2 Let $A \in \mathbb{R}^{k \times n}, B \in \mathbb{R}^{n \times m}$. Suppose $n \geq m$, then

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

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

Proof We start with the case where $k = n$. When $\sigma_n(B^T) = 0$, the result is trivial. When $\sigma_n(B^T) \neq 0$, we have $BB^y = I$, where B^y is the Moore–Penrose inverse of B. By Weyl's inequality (Horn and Johnson, 2012, 7.3.P16), it follows that

$$\sigma_i(A) \geq \sigma_i(AB)\,\sigma_1(B^y), \quad 1 \leq i \leq n.$$

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

When $k > n$, we have $1 \leq i \leq n$,

$$\sigma_i(A) = \sigma_i\begin{bmatrix} A & 0_{k\times(k-n)} \end{bmatrix} \geq \sigma_i(AB)\,\sigma_1\begin{pmatrix} B^y & 0_{m\times(k-n)} \end{pmatrix} = \sigma_i(AB)\,\sigma_1(B^y),$$

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 $1 \leq i \leq k$,

$$\sigma_i(A)\,\sigma_n(B) = \sigma_i\begin{bmatrix} A \\ 0_{(n-k)\times n} \end{bmatrix}\sigma_n(B) \leq \sigma_i\begin{bmatrix} AB \\ 0_{(n-k)\times m} \end{bmatrix} = \sigma_i(AB).$$

■

Now we are ready to prove Lemma 11.

Lemma 10 (restated) Let $\frac{1}{4} < \alpha < \frac{1}{2}$. Given data matrix X. $\forall\, \delta \in (0,1)$, $\forall h > h_0 = \mathrm{poly}(m, n, \frac{1}{\delta})$, with probability at least $1-\delta$ over random initialization with $[U(0)]_{ij}, [V(0)]_{ij} \sim N(0, h^2)$, the following conditions hold:

1. (Sufficient level of imbalance)

$$\lambda_+(0) + \lambda_-(0) > h^{1/2-\alpha}; \tag{E.2}$$

where λ_+, λ_- are defined in (19).

2. (Approximate orthogonality)

$$\left\| \begin{matrix} V(0)U_2^T(0) \\ U_1(0)U_2^T(0) \end{matrix} \right\|_F \leq 2\sqrt{m+r}\,\frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}}; \tag{E.3}$$

$$\left\| U_1(0)V^T(0) \right\|_F \leq 2\sqrt{m}\,\frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\delta}}{h^{2\alpha-\frac{1}{2}}}. \tag{E.4}$$

Proof [Proof of Lemma 10] For readability we simply write $U(0), U_1(0), U_2(0), V(0), D(0)$ as U, U_1, U_2, V, D.

Consider the matrix $V^T \quad U^T$ which is $h \quad (m+n)$. Apply Lemma E.1 to matrix $A = h \quad V^T \quad U^T$, with probability at least $1 \quad$, we have

$$_{m+n}(h \quad V^T \quad U^T) \quad \frac{p}{h} \quad p\overline{m+n} + \quad ;$$

which leads to

$$_{m+n}(V^T \quad U^T) \quad h^{\frac{1}{2}} \quad \frac{p\overline{m+n} + \frac{1}{2}\log\frac{2}{}}{h}: \tag{E.5}$$

Regarding the first inequality, we write the imbalance as

$$U_1^T U_1 \quad V^T V = V^T \quad U_1^T \quad \begin{matrix} V \\ U_1 \end{matrix} = V^T \quad U^T \quad \begin{matrix} I_m & 0 \\ 0 & {}_1{}^T_1 \end{matrix} \quad \begin{matrix} V \\ U \end{matrix}:$$

For $h > \quad p\overline{m+n} + \frac{1}{2}\log\frac{2}{} \quad ^2$, assume event (E.5) happens, then

$$_{m+n} \quad V^T \quad U^T \quad h^{\frac{1}{2}} \quad \frac{p\overline{m+n} + \frac{1}{2}\log\frac{2}{}}{h} > 0;$$

hence we have

$$\begin{aligned}
_{r+m}(D) &= \ _{r+m}(U_1^T U_1 \quad V^T V) \\
&= \ _{r+m} \quad V^T \quad U^T \quad \begin{matrix} I_m & 0 \\ 0 & {}_1{}^T_1 \end{matrix} \quad \begin{matrix} V \\ U \end{matrix} \\
&\text{(Lemma E:2)} \quad _{r+m} \quad V^T \quad U^T \quad \begin{matrix} I_m & 0 \\ 0 & {}_1{}^T_1 \end{matrix} \quad _{m+n} \quad \begin{matrix} V \\ U \end{matrix} \\
&= \ _{r+m} \quad \begin{matrix} I_m & 0 \\ 0 & {}_1{}^T_1 \end{matrix} \quad \begin{matrix} V \\ U \end{matrix} \quad _{m+n} \quad \begin{matrix} V \\ U \end{matrix} \\
&\text{(Lemma E:2)} \quad _{r+m} \quad \begin{matrix} I_m & 0 \\ 0 & {}_1{}^T_1 \end{matrix} \quad {}^2_{m+n} \quad \begin{matrix} V \\ U \end{matrix} \\
&= \ _{r+m} \quad \begin{matrix} I_m & 0 \\ 0 & {}_1{}^T_1 \end{matrix} \quad {}^2_{m+n} \quad V^T \quad U^T \\
&= \ {}^2_{m+n} \quad V^T \quad U^T \quad ;
\end{aligned}$$

where the last equality is due to the fact that $\begin{matrix} I_m & 0 \\ 0 & {}_1{}^T_1 \end{matrix}$ has exactly $r+m$ non-zero singular value and all of them are 1.

We further assume $h > 16 \quad p\overline{m+n} + \frac{1}{2}\log\frac{2}{} \quad ^2$, conditioned on event (E.5), with probability 1 we have

$$\begin{aligned}
&_{r+m}(D) \quad {}^2_{m+n} \quad V^T \quad U^T \\
&\qquad h^{\frac{1}{2}} \quad \frac{p\overline{m+n} + \frac{1}{2}\log\frac{2}{}}{h}^{!} \ ^2 \\
&= h^{1 \ 2} \quad 2\frac{p\overline{m+n} + \frac{1}{2}\log\frac{2}{}}{h^2 \quad ^{\frac{1}{2}}} + \quad \frac{p\overline{m+n} + \frac{1}{2}\log\frac{2}{}}{h}^{!} \ ^2 \\
&> h^{1 \ 2} \quad 2\frac{p\overline{m+n} + \frac{1}{2}\log\frac{2}{}}{h^2 \quad ^{\frac{1}{2}}} \quad \frac{1}{2}h^{1 \ 2}:
\end{aligned} \tag{E.6}$$

Lastly, due to the minimax property of symmetric matrix (Horn and Johnson, 2012, Theorem 4.2.6), we have

$$\begin{aligned}
\sigma_{r+1}(D) &= \min_{\dim(S)=h-r} \max_{0\neq x\in S} \frac{x^T D x}{x^T x} \\
(\dim(\ker(U_1)) \geq h-r) \quad &\min_{S\supseteq \ker(U_1)\atop \dim(S)=h-r} \max_{0\neq x\in S} \frac{x^T D x}{x^T x} \\
&= \min_{S\supseteq \ker(U_1)\atop \dim(S)=r} \max_{0\neq x\in S} \frac{x^T(I-V^TV)x}{x^T x} \geq 0;
\end{aligned}$$

and

$$\begin{aligned}
\sigma_r(D) &= \max_{\dim(S)=r} \min_{0\neq x\in S} \frac{x^T D x}{x^T x} \\
(\dim(\ker(V(0))) \geq h-m-r) \quad &\max_{S\subseteq \ker(V(0))\atop \dim(S)=r} \min_{0\neq x\in S} \frac{x^T D x}{x^T x} \\
&= \max_{S\subseteq \ker(V(0))\atop \dim(S)=r} \min_{0\neq x\in S} \frac{x^T U_1^T U_1 x}{x^T x} \geq 0.
\end{aligned}$$

Similarly, we have

$$\sigma_{m+1}(I-D) \geq \min_{S\supseteq \ker(V)\atop \dim(S)=h-m} \max_{0\neq x\in S} \frac{x^T(I-U_1^TU_1)x}{x^T x} \geq 0;$$

and

$$\sigma_m(I-D) \leq \max_{S\subseteq \ker(U_1(0))\atop \dim(S)=m} \min_{0\neq x\in S} \frac{x^T V^T V x}{x^T x} \geq 0.$$

These inequalities together imply

$$\min\{\sigma_r(D), \sigma_m(I-D)\} = \sigma_{r+m}(D).$$

Here we also use the fact that D is symmetric. Now by (E.6), we immediately obtain that conditioned on event (E.5), with probability 1, the following holds,

$$\lambda_+ + \lambda_- = \sigma_r(D) + \sigma_m(I-D) \geq 2\sigma_{r+m}(D) \geq h^{1-2\alpha},$$

which is exactly (E.2).

Regarding the second and third inequality, using the fact that

$$\|A\|_F \geq \sqrt{\min\{n,m\}}\|A\|_2, \ \forall A\in\mathbb{R}^{n\times m},$$

we have

$$\frac{1}{\sqrt{m}}\left\|U_1V^T\right\|_F - \left\|U_1V^T\right\|_2 = \left\|\begin{bmatrix}0 & \\ & U_1^T\end{bmatrix}\begin{bmatrix}V & \\ & U\end{bmatrix}\begin{bmatrix}V^T & \\ & U^T\end{bmatrix}\begin{bmatrix}I_m \\ 0\end{bmatrix}\right\|_2$$

$$= \left\|\begin{bmatrix}0 & \\ & U_1^T\end{bmatrix}\begin{bmatrix}V & \\ & U\end{bmatrix}\begin{bmatrix}V^T & \\ & U^T\end{bmatrix}I_{m+n}\begin{bmatrix}I_m \\ 0\end{bmatrix}\right\|_2$$

$$\left\|\begin{bmatrix}V & \\ & U\end{bmatrix}\begin{bmatrix}V^T & \\ & U^T\end{bmatrix}I_{m+n}\right\|_2; \text{for any } \in R;$$

where the second equality is due to the fact that $\left\|\begin{bmatrix}0 & \\ & U_1^T\end{bmatrix}\begin{bmatrix}I_m \\ 0\end{bmatrix}\right\|_2 = 0$. And

$$\frac{1}{\sqrt{m+r}}\left\|\begin{bmatrix}V & \\ & U_2^T\end{bmatrix}\begin{bmatrix}V^T & \\ & U_2^T\end{bmatrix}\right\|_F - \left\|\begin{bmatrix}V & \\ & U_1U_2^T\end{bmatrix}\begin{bmatrix}V^T & \\ & U_2^T\end{bmatrix}\right\|_2 = \left\|\begin{bmatrix}I_m & \\ 0 & \end{bmatrix}\begin{bmatrix}0 & \\ & U_1^T\end{bmatrix}\begin{bmatrix}V & \\ & U\end{bmatrix}\begin{bmatrix}V^T & \\ & U^T\end{bmatrix}\begin{bmatrix}0 \\ 0\end{bmatrix}\right\|_2\left\|_2\right.$$

$$= \left\|\begin{bmatrix}I_m & \\ 0 & \end{bmatrix}\begin{bmatrix}0 & \\ & U_1^T\end{bmatrix}\begin{bmatrix}V & \\ & U\end{bmatrix}\begin{bmatrix}V^T & \\ & U^T\end{bmatrix}I_{m+n}\begin{bmatrix}0 \\ 0\end{bmatrix}\right\|_2\left\|_2\right.$$

$$\left\|\begin{bmatrix}V & \\ & U\end{bmatrix}\begin{bmatrix}V^T & \\ & U^T\end{bmatrix}I_{m+n}\right\|_2; \text{for any } \in R;$$

where the second equality is due to the fact that $\left\|\begin{bmatrix}I_m & \\ 0 & \end{bmatrix}\begin{bmatrix}0 & \\ & U_1^T\end{bmatrix}\begin{bmatrix}0 \\ 0\end{bmatrix}\right\|_2 = 0$. Notice that

$$\left\|\begin{bmatrix}V & \\ & U\end{bmatrix}\begin{bmatrix}V^T & \\ & U^T\end{bmatrix}I_{m+n}\right\|_2 = \max_i {}_i^2(V^T \quad U^T) \quad :$$

Again we let $h > \sqrt{m+n} + \frac{1}{2}\log\frac{2}{}\ ^2$. When event (E.5) happens, all ${}_i^2(V^T \quad U^T)$ are within the interval $h^{\frac{1}{2}} \quad \frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{}\ ^2}{h}\ ;\ h^{\frac{1}{2}} \quad \frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{}\ ^2}{h}$. Since the choice of is arbitrary, we pick

$$= h^{1\ 2} + \frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{}\ ^{!\ 2}}{h}\ ; \tag{E.7}$$

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

$$\max_i {}_i^2(V^T \quad U^T)$$

$$\max\left\{h^{\frac{1}{2}} \quad \frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{}\ ^{!\ 2}}{h}\ ;\ h^{\frac{1}{2}} + \frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{}\ ^{!\ 2}}{h}\right\};$$

(is the mid-point)

$$h^{\frac{1}{2}} \quad \frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{}\ ^{!\ 2}}{h} \quad h^{1\ 2} \quad \frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{}\ ^{!\ 2}}{h}$$

$$= 2\frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{}}{h^{2\ \frac{1}{2}}}$$

Therefore, when $h > \sqrt{m+n} + \frac{1}{2}\log\frac{2}{} \ ^2$, conditioned on event (E.5), with probability 1, we have

$$\begin{aligned} & U_1V^T \quad {}_F \quad \sqrt{m} \quad \begin{matrix} V \\ U \end{matrix} \quad V^T \quad U^T \quad I_{m+n} \quad {}_2 \quad 2\sqrt{m}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{}}{h^2 \quad \frac{1}{2}}; \\ \text{and} \quad & \begin{matrix} V U_2^T \\ U_1 U_2^T \end{matrix} \quad {}_F \quad \sqrt{m+r} \quad \begin{matrix} V \\ U \end{matrix} \quad V^T \quad U^T \quad I_{m+n} \quad {}_2 \quad 2\sqrt{m+r}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{}}{h^2 \quad \frac{1}{2}}; \end{aligned} \tag{E.8}$$

where we choose as in (E.7).

When $h > h_0 = 16\sqrt{m+n} + \frac{1}{2}\log\frac{2}{} \ ^2$ and conditioned on event (E.5), events (E.6) and (E.8) happen with probability 1, hence the probability that both (E.6) and (E.8) happen is at least the probability of event (E.5), which is at least $1 \quad$. ■

With Lemma 10, we can prove Theorem 12.

Theorem 2 (restated) Let $\frac{1}{4} < \quad \frac{1}{2}$. Let $V(t); U(t); t > 0$ be the trajectory of the continuous dynamics (7) starting from some $V(0); U(0)$. Then, $9C > 0$, such that $8 \ 2 (0;1); 8h > h_0^{1=(4 \quad 1)}$ with $h_0 = \text{poly} \ m; n; \frac{1}{}; \frac{{}_1(\ x)}{{}_r^3(\ x)}$, with probability $1 \quad$ over random initializations with $[U(0)]_{ij}; [V(0)]_{ij} \quad N \ (0; h^{\ 2})$, we have

$$kU(1 \ )V^T(1 \ ) \quad \hat{} k_2 \quad 2C^{1=h^{1 \ 2}}\sqrt{m+r}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{}}{h^2 \quad \frac{1}{2}}: \tag{E.9}$$

Here $C = \exp \ 1 + \frac{{}_1^{1=2}(\ x)}{{}_r(\ x)} kY k_F$, which depends on the data $X; Y$.

Proof [Proof of Theorem 2] From Corollary 4 and Proposition 9, the stationary point $U(1 \ ); V(1 \ )$ satisfy

$$U_1(1 \ )V^T(1 \ ) = \quad {}_1^T \hat{}; \quad U_2(1 \ ) = U_2(0);$$

provided that level of imbalance $_+ + \ _$ is non-zero, which is guaranteed with high probability by Lemma 10. Hence we have

$$\begin{aligned} kU(1 \ )V^T(1 \ ) \quad \hat{} k_2 &= k \ {}_1U_1(1 \ )V^T(1 \ ) + \quad {}_2U_2(1 \ )V^T(1 \ ) \quad \hat{} k_2 \\ &= k \ {}_1 \ {}_1^T \hat{} + \quad {}_2U_2(1 \ )V^T(1 \ ) \quad \hat{} k_2 \\ &= k \ {}_2U_2(1 \ )V^T(1 \ )k_F \\ &= k \ {}_2U_2(0)V^T(1 \ )k_F = kU_2(0)V^T(1 \ )k_2 \quad k \ U_2(0)V^T(1 \ )k_F : \end{aligned}$$

Consider the following dynamics

$$\frac{d}{dt} \begin{matrix} V U_2^T \\ U_1 U_2^T \end{matrix} = \underbrace{\begin{matrix} 0 & E^T \ {}_x^{1=2} \\ {}_x^{1=2}E & 0 \end{matrix}}_{:= A_Z} \underbrace{\begin{matrix} V U_2^T \\ U_1 U_2^T \end{matrix}}_{:= Z}; \tag{E.10}$$

which can be viewed as a time-variant linear system, and in particular, by Horn and Johnson (2012, Theorem 7.3.3), we have $\|A_Z\|_2 = \|\Sigma_x^{1/2} E\|_2$. Notice that here the Z is different from the one in the proof for Proposition 9.

From (E.10), we have

$$\begin{aligned}\frac{d}{dt}\|Z\|_F^2 &= 2\,\mathrm{tr}\left(Z^T A_Z Z\right)\\ &= 2\,\mathrm{tr}\left(ZZ^T A_Z\right)\\ &\quad 2\|A_Z\|_2\,\mathrm{tr}\left(ZZ^T\right)\\ &= 2\|\Sigma_x^{1/2} E\|_2\|Z\|_F^2\\ &\quad 2\left(\lambda_1^{1/2}(\Sigma_x)\|E\|_2\|Z\|_F^2 \quad 2\left(\lambda_1^{1/2}(\Sigma_x)\|E\|_F\|Z\|_F^2\right.\right.:\end{aligned}$$

By Grönwall's inequality (Grönwall, 1919), we have $\forall t \quad 0$,

$$\begin{aligned}\|Z(t)\|_F^2 \quad \exp\left(\int_0^t 2\left(\lambda_1^{1/2}(\Sigma_x)\|E(\tau)\|_F d\tau\right)\right) \|Z(0)\|_F^2\\ \Rightarrow \|Z(t)\|_F \quad \exp\left(\int_0^t \left(\lambda_1^{1/2}(\Sigma_x)\|E(\tau)\|_F d\tau\right)\right) \|Z(0)\|_F \qquad \text{(E.11)}\end{aligned}$$

Using Lemma 10, for $h > h_0^0 := 16\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\quad}^2$, with probability at least 1 we have all the following.

$$\sigma_+(0) + \sigma_-(0) > h^{1/2}: \qquad \text{(E.12)}$$

$$U_1(0)V^T(0) \qquad 2\sqrt{m}\frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\quad}}{h^2 \quad \frac{1}{2}}; \qquad \text{(E.13)}$$

$$\|Z(0)\|_F = \left\|\begin{matrix}V(0)U_2^T(0)\\ U_1(0)U_2^T(0)\end{matrix}\right\|_F \qquad 2\sqrt{m+r}\frac{\sqrt{m+n} + \frac{1}{2}\log\frac{2}{\quad}}{h^2 \quad \frac{1}{2}} \qquad \text{(E.14)}$$

By Corollary 4, we have

$$\|E(t)\|_F^2 \quad \exp\left(\quad \lambda_r(\Sigma_x)c'(0)t\right)\|E(0)\|_F^2;$$

where $c'(0) = 2(\sigma_+(0) + \sigma_-(0))$, then by (E.12), we have

$$\begin{aligned}\|E(t)\|_F^2 \quad \exp\left(2h^{1/2} \quad \lambda_r(\Sigma_x)t\right)\|E(0)\|_F^2\\ \Rightarrow \|E(t)\|_F \quad \exp\left(h^{1/2} \quad \lambda_r(\Sigma_x)t\right)\|E(0)\|_F:\end{aligned}$$

Finally, from (E.11), we have

$$\begin{aligned}\|Z(t)\|_F \quad &\exp\left(\int_0^t \lambda_1^{1/2}(\Sigma_x)\|E(\tau)\|_F d\tau\right)\|Z(0)\|_F\\ &\exp\left(\lambda_1^{1/2}(\Sigma_x)\|E(0)\|_F \int_0^t \exp\left(h^{1/2} \quad \lambda_r(\Sigma_x)\right) d\right)\|Z(0)\|_F\\ &\exp\left(\lambda_1^{1/2}(\Sigma_x)\|E(0)\|_F \int_0^1 \exp\left(h^{1/2} \quad \lambda_r(\Sigma_x)\right) d\right)\|Z(0)\|_F\\ &= \exp\left(\frac{\lambda_1^{1/2}(\Sigma_x)}{h^{1/2} \quad \lambda_r(\Sigma_x)}\|E(0)\|_F \quad \|Z(0)\|_F:\right.\end{aligned} \qquad \text{(E.15)}$$

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

$$\begin{aligned}\|E(0)\|_F &= \|W^T Y - \Sigma_x^{-1=2} U_1(0) V^T(0)\|_F \\ &\le \|W^T Y\|_F + \|\Sigma_x^{-1=2} U_1(0) V^T(0)\|_F \\ &\le \|Y\|_F + \bar{r}^{-1=2}(\Sigma_x)\|U_1(0)V^T(0)\|_F\end{aligned}$$

then we can write (E.15) as

$$\begin{aligned}\|Z(t)\|_F &\le \exp\left(\frac{\frac{1}{r}^{1=2}(\Sigma_x)}{h^{1-2}\ _r(\Sigma_x)}\|Y\|_F\right)\exp\left(\frac{\frac{1}{r}^{1=2}(\Sigma_x)}{h^{1-2}\ \frac{3=2}{r}(\Sigma_x)}\|U_1(0)V^T(0)\|_F\right)\|Z(0)\|_F \\ &= \exp\left(\frac{\frac{1}{r}^{1=2}(\Sigma_x)}{{}_r(\Sigma_x)}\|Y\|_F\right)\exp\left(\frac{\frac{1}{r}^{1=2}(\Sigma_x)}{\frac{3=2}{r}(\Sigma_x)}\|U_1(0)V^T(0)\|_F\right)^{1=h^{1\ 2}}\|Z(0)\|_F :\end{aligned} \tag{E.16}$$

For the second exponential, we let $h_0 := \max\left\{h_0'; 4\frac{1}{\frac{3}{r}(\ x)}m\ \sqrt{m+n}+\frac{1}{2}\log\frac{2}{\ }^{\ 2}\right\}$, then $\forall h > h_0^{1=(4\ -1)}$, by (E.13) we have

$$\exp\left(\frac{\frac{1}{r}^{1=2}(\Sigma_x)}{\frac{3=2}{r}(\Sigma_x)}\|U_1(0)V^T(0)\|_F\right) \le \exp\left(2\frac{\frac{1}{r}^{1=2}(\Sigma_x)}{\frac{3=2}{r}(\Sigma_x)}\sqrt{m}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\ }}{h^{2\ -\frac{1}{2}}}\right) \le e: \tag{E.17}$$

Notice that $h > h_0^{1=(4\ -1)}$ also ensures $h > h_0^{1=(4\ -1)} \ge h_0 \ge h_0'$, hence the width condition for (E.12)(E.14)(E.13) to hold is satisfied.

Finally by (E.14)(E.17), we write (E.16) as

$$\begin{aligned}\|Z(t)\|_F &\le \exp\left(1+\frac{\frac{1}{r}^{1=2}(\Sigma_x)}{{}_r(\Sigma_x)}\|Y\|_F\right)^{1=h^{1\ 2}}\|Z(0)\|_F \\ &\le \underbrace{\exp\left(1+\frac{\frac{1}{r}^{1=2}(\Sigma_x)}{{}_r(\Sigma_x)}\|Y\|_F\right)^{1=h^{1\ 2}}}_{:=C^{1=h^{1\ 2}}} 2\sqrt{m+r}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\ }}{h^{2\ -\frac{1}{2}}} \\ &= 2C^{1=h^{1\ 2}}\sqrt{m+r}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\ }}{h^{2\ -\frac{1}{2}}} :\end{aligned}$$

Therefore for some $C > 0$ that depends on the data $(X; Y)$, given any $0 < \ < 1$, when $h > h_0^{1=(4\ -1)}$ as defined above, with at least probability $1 - \ $, we have

$$\begin{aligned}\|U(\infty)V^T(\infty) - \hat{\Theta}\|_2 &\le \|U_2(0)V^T(\infty)\|_F \\ &\le \sup_{t>0}\|U_2(0)V^T(t)\|_F \\ &\le \sup_{t>0}\|Z(t)\|_F \le 2C^{1=h^{1\ 2}}\sqrt{m+r}\frac{\sqrt{m+n}+\frac{1}{2}\log\frac{2}{\ }}{h^{2\ -\frac{1}{2}}} :\end{aligned}$$

■