Linear Convergence of Gradient Descent for Finite Width Over-parametrized Linear Networks with General Initialization

Ziqing Xu
Johns Hopkins University

Hancheng Min
Johns Hopkins University

Salma Tarmoun
Johns Hopkins University

Enrique Mallada
Johns Hopkins University

René Vidal
University of Pennsylvania

Abstract

Recent theoretical analyses of the convergence of gradient descent (GD) to a global minimum for over-parametrized neural networks make strong assumptions on the step size (infinitesimal), the hidden-layer width (infinite), or the initialization (spectral, balanced). In this work, we relax these assumptions and derive a linear convergence rate for two-layer linear networks trained using GD on the squared loss in the case of finite step size, finite width and general initialization. Despite the generality of our analysis, our rate estimates are significantly tighter than those of prior work. Moreover, we provide a time-varying step size rule that monotonically improves the convergence rate as the loss function decreases to zero. Numerical experiments validate our findings.

1 INTRODUCTION

The empirical success of neural networks on a wide variety of applications, such as natural language processing [Vaswani et al., 2017, Vaswani et al., 2018], computer vision [He et al., 2015, Minaee et al., 2021] and decision making [Silver et al., 2016, Vo et al., 2019], has motivated significant research on understanding theoretically why neural networks work so well in practice. One interesting and puzzling phenomenon is that over-parametrized neural networks trained with gradient descent (GD) enjoy fast convergence even if their loss landscape is non-convex. Much of the recent work in this area has focused on deriving convergence rates for over-parametrized networks.

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Table 1: Summary of prior work and our contributions.

	step size	width	initialization
[Jacot et al., 2018, Du et al., 2018b, Lee et al., 2019, Liu et al., 2022, Oymak and Soltanolkotabi, 2020]	**finite**	very large	sufficiently large
[Mei et al., 2018, Chizat and Bach, 2018, Ding et al., 2022, Sirignano and Spiliopoulos, 2020]	infinite-simal	infinite	**general**
[Saxe et al., 2013, Gidel et al., 2019, Tarmoun et al., 2021]	infinite-simal	**finite**	spectral
[Tarmoun et al., 2021, Min et al., 2022]	infinite-simal	**finite**	**general**
[Arora et al., 2018, Du et al., 2018a, Nguegnang et al., 2021]	**finite**	**finite**	large margin and small imbalance
This work	**finite**	**finite**	**general**

However, existing results require stringent assumptions on the step size (infinitesimally small), the hidden-layer width (infinitely large), or the initialization (spectral, balanced).

Prior work One line of work [Jacot et al., 2018, Du et al., 2018b, Lee et al., 2019, Liu et al., 2022] studies the convergence of GD when the scale of the initialization and the network width are sufficiently large. Under these assumptions, the network weights remain close to their initialization during training, and one can show that GD converges linearly to a global minimum. However, [Chizat et al., 2019, Chen et al., 2022] show that this "lazy training" regime is unrealistic in practice as it limits feature learning. A convergence analysis beyond the so-called lazy regime can be undertaken in the (mean-field) limit of infinitely wide networks [Mei et al., 2018, Rotskoff and Vanden-Eijnden, 2018b, Rotskoff and Vanden-Eijnden, 2018a, Chizat and Bach, 2018, Sirignano and Spiliopoulos, 2020, Ding et al., 2022], where suitable assumptions on the

initialization and step size make GD become a Wasserstein flow; a partial differential equation commonly appearing in optimal transport theory. However, while such analysis can guarantee convergence to the global optimum for a wider range of initializations, it still imposes strong assumptions on the network width (infinite) and step size (infinitesimal).

Another line of work studies the convergence of gradient algorithms for over-parametrized networks with finite width. In this finite-width setting, the vast majority of existing results consider linear networks trained using gradient flow (GF). GF can be seen as GD with infinitesimal step size, but its dynamics in this setting are generally easier to analyze. For example, [Saxe et al., 2013, Gidel et al., 2019, Tarmoun et al., 2021] show that under spectral initialization the dynamics of GF decouple into several scalar dynamics, which allows them to derive a linear convergence rate. For non-spectral initialization, [Tarmoun et al., 2021, Min et al., 2022] show that a large *imbalance* or large *margin* of the initialization can lead to faster convergence of GF, significantly extending the range of initializations from which linear convergence of GF is guaranteed. However, such results require infinitesimal step size. For finite step size, [Arora et al., 2018, Du et al., 2018a, Nguegnang et al., 2021] prove linear convergence of GD when there is sufficient margin at initialization and the imbalance is small. However, such assumptions rarely hold in practice since commonly used random initializations have a large imbalance.

Paper contributions In this work, we derive a linear convergence rate for GD in the case of over-parametrized, finite-width, two-layer linear networks with general initialization. Our analysis can be seen as a natural extension of recent results for GF, which cover finite width and imbalanced initializations. However, a key challenge in the case of GD is that quantities such as *imbalance*, which are preserved by GF, are no longer preserved by GD. To address this challenge, we derive quantities that effectively bound the deviation of the discrete dynamics from the continuous dynamics as a function of the step size, thus ensuring sufficient control (via upper and lower bounds) of the level of imbalance throughout training. This leads to a convergence rate that naturally depends on the step size, as well as other quantities, such as the current loss value. Moreover, the dependency of the rate on the step size is a low-degree polynomial, which allows us to easily compute an optimal step size at each iteration of training. Furthermore, we prove that the resulting time-varying step size is *lower-bounded by* the optimal rate of GD for the non-overparametrized problem. Finally, our numerical results show that, despite the generality of our analysis, the step size we derive leads to faster convergence and Theorem 3.1 admits a wider range of step size than in [Du et al., 2018a, Arora et al., 2018].

Notation We use lower case letters a to denote a scalar, and capital letters A and $A^\top$ to denote a matrix and its transpose. We use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote the largest and smallest eigenvalues of A, $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ to denote its largest and smallest singular values, $\|A\|_F$ and $\|A\|_2$ to denote its Frobenius and spectral norms, and $A[i,j]$ to denote its (i,j)-th element. Given two matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{k\times l}$, it will be convenient to use either $\begin{pmatrix} A \\ B \end{pmatrix}$ or (A,B) to represent an element in the product space $\mathbb{R}^{n\times m} \times \mathbb{R}^{k\times l}$, irrespectively of the dimensions. For a function $f(Z)$, we use $\nabla f(Z) := \frac{\partial}{\partial Z} f(Z)$ to denote its gradient, and whenever Z depends on an independent variable t, we use $f(t) := f(Z(t))$ and $\dot{Z}(t) = \frac{d}{dt} Z(t)$, dropping the dependence on t when it is implicit from the context, e.g., $\dot{Z} = \frac{d}{dt} Z$. Finally, we use $\mathcal{N}(\mu, \sigma^2)$ to denote a normal distribution with mean μ and variance σ^2.

2 CONVERGENCE OF GRADIENT FLOW FOR TWO-LAYER LINEAR NETWORKS

In this section, we first consider a linear regression problem and its over-parametrized version, which is equivalent to training a two-layer linear neural network. We then summarize the convergence results for GF in [Min et al., 2022], which constitute the starting point of our work. Throughout this section, we thus consider a continuous time $t \in \mathbb{R}$.

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 consider the following linear regression problem

$$\min_W \ell(W) = \frac{1}{2}\|Y - XW\|_F^2, \tag{1}$$

where $W \in \mathbb{R}^{n\times m}$, $X = [x_1, \cdots, x_N]^\top \in \mathbb{R}^{N\times n}$ and $Y = [y_1, \cdots, y_N]^\top \in \mathbb{R}^{N\times m}$. We are interested in solving the optimization problem in equation 1 by solving the following over-parametrized problem

$$\min_{W_1, W_2} L(W_1, W_2) = \frac{1}{2}\|Y - XW_1W_2\|_F^2, \tag{2}$$

where $W_1 \in \mathbb{R}^{n\times h}$, $W_2 \in \mathbb{R}^{h\times m}$. This over-parametrized problem corresponds to training a two-layer linear neural network with n inputs, h hidden neurons, m outputs, and weight matrices W_1 and W_2.

To simplify exposition, we consider the above problems in the under-determined case, i.e., $N \leq n$. We assume that the input data matrix X is full rank, i.e., $\operatorname{rank}(X) = N$.[1] We also assume that $h \geq \min\{n, m\}$. These assumptions imply that the minimum of both problems is zero, i.e., $\min_W \ell(W) = 0$ and $L^* := \min_{W_1,W_2} L(W_1, W_2) = 0$.

[1] When X is rank deficient, one can reformulate the problem into one with full-rank input data matrix (see Appendix A for details).

We note, however, that our results generalize the case $N > n$, by properly accounting for a non-zero L^*.

Convergence under GF Let us consider solving equation 2 via GF

$$\begin{pmatrix}\dot{W}_1\\ \dot{W}_2\end{pmatrix} = -\nabla L(W_1, W_2) = -\begin{pmatrix}\nabla \ell(W) W_2^\top\\ W_1^\top \nabla \ell(W)\end{pmatrix}, \quad (3)$$

where $\nabla \ell(W) = X^\top (Y - XW)$. Notice that there exists a linear operator $\gamma(\cdot; W_1, W_2) : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times h} \times \mathbb{R}^{h\times m}$,

$$\gamma(\nabla \ell(W); W_1, W_2) := \begin{pmatrix}\nabla \ell(W) W_2^\top\\ W_1^\top \nabla \ell(W)\end{pmatrix}, \quad (4)$$

which depends on W_1, W_2, that maps the gradient of the loss $\nabla \ell(W) \in \mathbb{R}^{n\times m}$ to the gradient of the over-parametrized loss $\nabla L(W_1, W_2) \in \mathbb{R}^{n\times h} \times \mathbb{R}^{h\times m}$.

Then, one can show that the evolution of L under GF is

$$\begin{aligned}&\dot{L}(W_1, W_2)\\ &= \left\langle \frac{\partial L}{\partial W_1}(W_1, W_2), \dot{W}_1 \right\rangle + \left\langle \frac{\partial L}{\partial W_2}(W_1, W_2), \dot{W}_2 \right\rangle \quad (5)\\ &= -\ \langle \gamma(\nabla \ell(W); W_1, W_2), \gamma(\nabla \ell(W); W_1, W_2)\rangle\\ &= -\ \langle \nabla \ell(W), \gamma^* \circ \gamma(\nabla \ell(W); W_1, W_2)\rangle\,,\end{aligned}$$

where $\gamma^*(\cdot; W_1, W_2)$ is the adjoint of $\gamma(\cdot; W_1, W_2)$. Therefore, the dynamics of L are defined by the following positive semi-definite Hermitian linear operator on $\nabla \ell(W)$:

$$\begin{aligned}\tau(\nabla \ell(W); W_1, W_2) &:= \gamma^* \circ \gamma(\nabla \ell(W); W_1, W_2) \quad (6)\\ &= \nabla \ell(W)\ W_2^\top W_2 + W_1 W_1^\top \nabla \ell(W).\end{aligned}$$

Then, from equation 5 and the min-max principle of Hermitian operators, we have

$$\dot{L}(t) = -\langle \nabla \ell(t), \tau_t(\nabla \ell(t))\rangle \le -\lambda_{\min}(\tau_t) \|\nabla \ell(t)\|_F^2, \quad (7)$$

where for simplicity we use $\ell(t)$, $L(t)$ and $\tau_t(\nabla \ell(t))$, resp., as a shorthand for $\ell(W(t))$, $L(W_1(t), W_2(t))$ and $\tau(\nabla \ell(W(t)); W_1(t), W_2(t))$. Similarly, we use $\lambda_{\min}(\tau_t)$ and $\lambda_{\max}(\tau_t)$ as a shorthand for $\lambda_{\min}(\tau(\,\cdot\,; W_1(t), W_2(t)))$ and $\lambda_{\max}(\tau(\,\cdot\,; W_1(t), W_2(t)))$, respectively.

The core contribution of [Min et al., 2022] is to provide a lower bound on $\lambda_{\min}(\tau_t)$ using two quantities: *imbalance*

$$D(t) = W_1^\top(t) W_1(t) - W_2(t) W_2(t)^\top, \quad (8)$$

and *product* $W(t) = W_1(t) W_2(t)$. Specifically, they show there exists an non-negative function $\alpha(D, \sigma_{\min}(W))$ that depends on imbalance and product, such that for all $t \ge 0$,

$$\lambda_{\min}(\tau_t) \ge \alpha(D(t), \sigma_{\min}(W(t))). \quad (9)$$

To find a uniform lower bound on $\alpha(D(t), \sigma_{\min}(W(t)))$ for all $t \ge 0$, they exploit the fact that the imbalance matrix remains constant along the trajectories of GF [Arora et al., 2018, Du et al., 2018a], i.e., $\dot{D} \equiv 0$ so that $D(t) = D(0)$. As for the product, [Min et al., 2022] show (from the fact that the loss $L(t)$ is non-increasing) that

$$\sigma_{\min}(W(t)) \ge p_1 (:= margin)\,. \quad (10)$$

Therefore, we can replace the imbalance $D(t)$ in equation 9 by its initial value $D(0)$. Moreover, it can be shown that $\alpha(D, \sigma)$ is a non-decreasing function of the second argument σ, allowing us to use equations 9 and 10 to show that

$$\lambda_{\min}(\tau_t) \ge \alpha(D(t), p_1) = \alpha(D(0), p_1) := \alpha_0, \quad (11)$$

where the expression for α_0 is shown in Table 2. Observe that equation 11 yields a uniform lower bound on $\lambda_{\min}(\tau_t)$. Combining equation 11 with the fact that $\ell(t)$ satisfies the PL condition $\frac{1}{2}\|\nabla \ell(t)\|_F^2 \ge \mu \ell(t)$ with $\mu = \sigma_{\min}^2(X) > 0$, we show that equation 7 can be further upper-bounded by:

$$\begin{aligned}\dot{L}(t) &\le -\lambda_{\min}(\tau_t) \|\nabla \ell(t)\|_F^2 \le -\alpha_0 \|\nabla \ell(t)\|_F^2\\ &\le -2\mu\alpha_0 \ell(t) = -2\mu\alpha_0 L(t),\end{aligned} \quad (12)$$

where the third inequality follows from the PL condition. Moreover, if $\alpha_0 > 0$, it follows from Grönwall's inequality that $L(t) \le \exp(-2\mu\alpha_0 t) L(0)$, showing that GF converges exponentially with a rate $2\mu\alpha_0$.

As discussed in the introduction, the imbalance matrix $D(t)$ measures the difference of the weights in the two layers, while the margin p_1 depends on the initial error $\|Y - XW_1(0)W_2(0)\|_F$ (the smaller the error, the larger the margin). [Min et al., 2022] show that $\alpha_0 > 0$ when there is either 1) sufficient imbalance $\underline{\Delta} > 0$ or 2) sufficient margin $p_1 > 0$, where $\underline{\Delta}$ is defined in Table 2. Moreover, a larger imbalance (as measured by $\underline{\Delta}$) or a larger margin p_1 improves the rate of convergence α_0. In summary, the convergence of GF is completely determined by the initialization $W_1(0), W_2(0)$, and convergence is guaranteed when the initialization satisfies $\alpha_0 > 0$, which is achieved by either being imbalanced or having sufficient margin.

3 CONVERGENCE OF GRADIENT DESCENT FOR TWO-LAYER LINEAR NETWORKS

In this section, we analyze the convergence of GD for over-parametrized two-layer linear networks. We start in §3.1 by highlighting the challenges of analyzing over-parametrized GD when compared to (1) the standard GD algorithm applied to $\ell(W)$ and (2) the GF algorithm applied to $L(W_1, W_2)$ described in the previous section. Alongside, we provide a high-level overview of the overall strategy we use to overcome these challenges. Based on the proposed approach, we then derive in §3.2 a rigorous convergence rate that depends on not only the imbalance and margin at the initialization but also the step size and condition number of the data. Finally, in §3.3 we propose an adaptive

step size scheme that accelerates convergence. Due to the discrete nature of our updates, we thus consider t to be discrete, i.e., $t \in \mathbb{N}$.

3.1 Challenges in the Analysis of Over-parametrized Gradient Descent

Standard GD We start by deriving the convergence rate of the non-overparametrized regime described in equation 1. Notice that $\ell(t)$ is K-smooth and satisfies μ-PL condition, where $K = \sigma^2_{\max}(X), \mu = \sigma^2_{\min}(X)$. Then, the following smoothness inequality holds for any $W(t), W(t+1)$:

$$\begin{aligned}\ell(t+1) \leq \ell(t) + \langle \nabla \ell(t), W(t+1) - W(t) \rangle \\ + \frac{K}{2}\|W(t+1) - W(t)\|_F^2\end{aligned} \tag{13}$$

After substituting the GD update with fixed step size η

$$W(t+1) = W(t) - \eta \nabla \ell(t). \tag{14}$$

into the smoothness inequality in equation 13 we obtain

$$\begin{aligned}\ell(t+1) &\leq \ell(t) - \eta\|\nabla \ell(t)\|_F^2 + \frac{K}{2}\eta^2\|\nabla \ell(t)\|_F^2 \\ &= \ell(t) - \eta\left(1 - K\frac{\eta}{2}\right)\|\nabla \ell(t)\|_F^2\end{aligned} \tag{15}$$

Then, if the step size satisfies $\eta < \frac{2}{K}$, then the loss is non-increasing. Moreover, if we apply the PL condition $\frac{1}{2}\|\nabla \ell(t)\|_F^2 \geq \mu \ell(t)$ to equation 15, we obtain

$$\ell(t+1) \leq (1 - 2\eta\mu + K\mu\eta^2)\ell(t), \tag{16}$$

which suffices to show the linear convergence of GD, for properly chosen η.

Over-parametrized GD In the over-parametrized case, we use the chain rule to write the gradient of L with respect to W_1, W_2 in terms of $\nabla \ell(W), W_1, W_2$. The update of weights in GD is

$$\begin{pmatrix} W_1(t+1) \\ W_2(t+1) \end{pmatrix} = \begin{pmatrix} W_1(t) \\ W_2(t) \end{pmatrix} - \eta \gamma_t(\nabla \ell(t)) \tag{17}$$

Thus, the update of the product is

$$\begin{aligned}&W(t+1) = W_1(t+1)W_2(t+1) \\ &= \big(W_1(t) - \eta \nabla \ell(t) W_2(t)^\top\big)\big(W_2(t) - \eta W_1(t)^\top \nabla \ell(t)\big) \\ &= W(t) - \eta \tau_t(\nabla \ell(t)) + \eta^2 \nabla \ell(t) W(t)^\top \nabla \ell(t).\end{aligned} \tag{18}$$

In other words, the update of the product is a polynomial of degree two on the step size η, unlike the update in equation 14, which is a polynomial of degree one. Substituting equation 18 into the smoothness inequality 13, and using the PL condition, we can connect the loss at iteration $t+1$ with the loss at iteration t. The following lemma characterizes this property.

Lemma 3.1. *If at the t-th iteration of GD applied to the over-parametrized loss L, the step size η satisfies*

$$\begin{aligned}&\lambda_{\min}(\tau_t) - \eta\|\nabla \ell(t)\|_F\|W(t)\|_F \\ &- \frac{K\eta}{2}\big[\lambda_{\max}(\tau_t) + \eta\|\nabla \ell(t)\|_F\|W(t)\|_F\big]^2 \geq 0\,,\end{aligned} \tag{19}$$

then the following inequality holds

$$L(t+1) \leq \rho(\eta, t) L(t)\,, \tag{20}$$

where

$$\begin{aligned}\rho(\eta, t) = 1 &- 2\eta\mu\lambda_{\min}(\tau_t) + K\mu\eta^2\lambda^2_{\max}(\tau_t) \\ &+ 2\eta^2\mu\sigma_{\max}(W(t))\|\nabla \ell(t)\|_F \\ &+ 2\eta^3\mu K\lambda_{\max}(\tau_t)\sigma_{\max}(W(t))\|\nabla \ell(t)\|_F \\ &+ \eta^4\mu K\sigma^2_{\max}(W(t))\|\nabla \ell(t)\|_F^2.\end{aligned} \tag{21}$$

The proof of the above lemma can be found in the Appendix B.

Comparison with non-overparametrized GD The difference between the inequality we derive in Lemma 3.1 and the one in equation 16 is twofold. Firstly, $\rho(\eta, t)$ in equation 50 includes a quadratic polynomial of η:

$$1 - 2\eta\mu\lambda_{\min}(\tau_t) + K\mu\eta^2\lambda^2_{\max}(\tau_t) \tag{22}$$

that resembles the one in equation 16. The only difference is that the second coefficient is now scaled by $\lambda_{\min}(\tau_t)$ and the third coefficient by $\lambda^2_{\max}(\tau_t)$. Equation 22 comes from the term $\eta\tau_t(\nabla \ell(t))$ in the product update in equation 18, which corresponds to moving the weight $W(t)$ along the "skewed gradient direction" $\tau_t(\nabla \ell(t))$ instead of $\nabla \ell(t)$. Secondly, equation 50 has extra second- and higher-order terms in η which come from the other term $\eta^2 \nabla \ell(t) W^\top(t) \nabla \ell(t)$ in equation 18. Overall, compared to equation 16, the over-parametrized GD introduces a more complicated update on the product $W(t)$, leading to the inequality in equation 49 that not only is a polynomial of degree four in η, but also depends on the weights $W_1(t), W_2(t)$ at the current iteration. These differences pose additional challenges in deriving a linear convergence rate for over-parametrized GD.

Towards linear convergence Lemma 3.1 provides an upper bound on $L(t+1)$, $\rho(\eta, t)L(t)$, which implicitly depends on $W_1(t)$ and $W_2(t)$ via $\lambda_{\min}(\tau_t), \sigma_{\max}(W(t)), \ell(t)$ and $\lambda_{\max}(\tau_t)$. However, it is unclear whether one can find some step size η that can simultaneously satisfy equation 48 and uniformly bound $\rho(\eta, t) \leq \bar{\rho} < 1$, for all t. Only under such conditions Lemma 3.1 would lead to

$$L(t+1) < \bar{\rho} L(t) < (\bar{\rho})^{t+1} L(0).$$

We approach this challenge in a similar spirit as it was done in GF [Min et al., 2022].

Step 1. Spectral bounds for τ_t and $W(t)$: First, we seek to find bounds for $\lambda_{\min}(\tau_t)$ and $\lambda_{\max}(\tau_t)$ based on the imbalance $D(t)$ and the singular values of the product, i.e.,

$$\begin{aligned}\alpha(D(t), \sigma_{\min}(W(t))) \le \lambda_{\min}(\tau_t)\\ \lambda_{\max}(\tau_t) \le \beta(D(t), \sigma_{\max}(W(t))),\end{aligned} \tag{23}$$

where both functions $\alpha(D, \sigma)$ and $\beta(D, \sigma)$ are increasing on the second argument, σ. As a result, if one is able to control $D(t)$ and the singular values of $W(t)$, one can attempt to upper-bound $\rho(\eta, t)$ in equation 50.

For the case of $\sigma_{\min}(W(t))$ and $\sigma_{\max}(W(t))$, a similar monotonicity argument as in GF can be done to obtain

$$p_1 \le \sigma_{\min}(W(t)) \le \sigma_{\max}(W(t)) \le p_2. \tag{24}$$

The additional, non-trivial challenge present in GD is the fact that the imbalance $D(t)$ is no longer preserved, i.e., $D(t) \neq D(0)$, which makes it still difficult control $\lambda_{\min}(\tau_t), \lambda_{\max}(\tau_t)$ by equation 23. Nevertheless, we show in *Theorem 3.1* that if η is sufficiently small, but not infinitesimal, it is possible to control how much the imbalance changes by bounding $\|D(t) - D(0)\|$ for all t, which leads to a uniform bound of the form

$$\alpha_0 c_1 \le \lambda_{\min}(\tau_t) \le \lambda_{\max}(\tau_t) \le \beta_0 c_2\,, \tag{25}$$

where $\beta_0 := \beta(D(0), p_1)$, and the parameters $0 < c_1 < 1$, $c_2 > 1$ represent an additional level of conservativeness in the bound that is necessary to accommodate the time varying nature of $D(t)$ in GD; see discussion after Theorem 3.1 for more details.

Stage 2. Uniform upper-bound on rate $\rho(\eta, t)$: Once bounds for the spectrum of $W(t)$ and τ_t have been established, one can then proceed to bound $\rho(\eta, t)$ in equation 50. In particular, we will show that $\rho(\eta, t) \le f(\eta, t)$, where

$$f(\eta, t) := 1 - a_1\eta + a_2(t)\eta^2 + a_3(t)\eta^3 + a_4(t)\eta^4\,, \tag{26}$$

and the dependency on time is only through $L(t)$, i.e.,

$$\begin{aligned}a_1 &= 2(c_1\alpha_0)\sigma^2_{\min}(X),\\ a_2(t) &= 2\sqrt{2\kappa L(t)\sigma^6_{\min}(X)}p_2 + \kappa\sigma^4_{\min}(X)(c_2\beta_0)^2,\\ a_3(t) &= 2\sqrt{2\kappa^3 L(t)\sigma^{10}_{\min}(X)}c_2\beta_0 p_2,\\ a_4(t) &= 2\kappa^2\sigma^6_{\min}(X)p_2^2 L(t).\end{aligned} \tag{27}$$

The above bound for $\rho(\eta, t)$ in equation 26, whose derivation is provided in *Theorem 3.2*, can be then leverage in multiple ways.

- **Uniform linear rate.** Under mild conditions no the step size, here exists η independent of t such that $f(\eta, t) \le f(\eta, 0)$ (also in *Theorem 3.2*), leading to

$$L(t) \le \Pi_{k=0}^t f(\eta, k)\, L(0) \le (f(\eta, 0))^t L(0). \tag{28}$$

- **Time-varying step size.** A natural consequence of equations 26 and 28 is the possibility to adaptively choose η_t, using only knowledge of the current loss $L(t)$, so as to improve the convergence rate. This is explored in §3.3; call for *Algorithm 1*.

3.2 General bound on linear convergence rate

In this subsection, we derive conditions under which Lemma 3.1 is a descent lemma. Based on this result, we can prove that GD converges linearly to a global minimum of equation 2. We refer the reader to Table 2 for the definition of various quantities appearing in this section.

Before stating our main result, we note that prior work [Arora et al., 2018, Du et al., 2018a] studied optimizing equation 2 via GD, but their results require the initial imbalance to have small Frobenius norm and the initial margin to be sufficiently large. The NTK initialization [Du and Hu, 2019] does not require small imbalance, but it does require a large hidden-layer width h, and the weights needs to be randomly initialized. To the best of our knowledge, Theorem 3.2 is the first convergence result for GD which provides an explicit convergence rate without making the assumption that the initial imbalance is small or that the width of the network is large.

Table 2: Table of Notation

SYMBOL	DEFINITION
$\ell(t)$	$\ell(W(t))$
$L(t)$	$L(W_1(t), W_2(t))$
$\tau_t(\nabla\ell(t))$	$\tau(\nabla\ell(W(t)); W_1(t), W_2(t))$
$\lambda_{\min}(\tau_t)$	$\lambda_{\min}(\tau(\cdot; W_1(t), W_2(t)))$
$\lambda_{\max}(\tau_t)$	$\lambda_{\max}(\tau(\cdot; W_1(t), W_2(t)))$
$D(t)$	$W_1^\top(t)W_1(t) - W_2(t)W_2(t)^\top$
$W(t)$	$W_1(t)W_2(t)$
$E(t)$	$Y - XW_1(t)W_2(t)$
κ	$\frac{\sigma^2_{\max}(X)}{\sigma^2_{\min}(X)}$
p_1	$\frac{\max\{\sigma_{\min}(Y) - \|E(0)\|_F, 0\}}{\sigma_{\max}(X)}$
p_2	$\frac{\|Y\|_F + \|E(0)\|_F}{\sigma_{\min}(X)}$
Δ_+	$\max(\lambda_{\max}(D(0)), 0) - \max(\lambda_n(D(0)), 0)$
Δ_-	$\max(\lambda_{\max}(-D(0), 0)) - \max(\lambda_m(-D(0)), 0)$
$\underline{\Delta}$	$\max(\lambda_n(D(0)), 0) + \max(\lambda_m(-D(0)), 0)$
λ_+	$\max(\lambda_{\max}(D(0)), 0)$
λ_-	$\max(\lambda_{\max}(-D(0)), 0)$
α_0	$\frac{-\Delta_+ + \sqrt{(\Delta_+ + \underline{\Delta})^2 + 4p_1^2}}{2} + \frac{-\Delta_- + \sqrt{(\Delta_- + \underline{\Delta})^2 + 4p_1^2}}{2}$
β_0	$\frac{\lambda_+ + \sqrt{\lambda_+^2 + 4p_2^2}}{2} + \frac{\lambda_- + \sqrt{\lambda_-^2 + 4p_2^2}}{2}$

Theorem 3.1 (Uniform bounds on eigenvalues of τ_t and singular values of $W(t)$)**.** *Assume $\alpha_0 > 0$, and choose $0 < c_1 < 1$, and $c_2 > 1$. Let $\eta_1^{\max}$ and $\eta_2^{\max}$ be, respectively, the unique positive roots of the following two polynomials*

in η

$$a_4(0)\eta^3+a_3(0)\eta^2+\big(a_2(0)+\frac{4c_2L(0)\sigma^2_{\max}(X)}{c_2-1}\big)\eta=a_1,$$
$$a_4(0)\eta^3+a_3(0)\eta^2+\big(a_2(0)+\frac{8c_2\beta_0L(0)\sigma^2_{\max}(X)}{(1-c_1)\alpha_0}\big)\eta=a_1. \tag{29}$$

Then, for any $0<\eta\le\eta_{\max}:=\min\{\eta_1^{\max},\eta_2^{\max}\}$*, the following holds for all* $t=0,1,\dots$

$$\begin{aligned} c_1\alpha_0\le\lambda_{\min}(\tau_t)\le\lambda_{\max}(\tau_t)\le c_2\beta_0\\ p_1\le\sigma_{\min}(W(t))\le\sigma_{\max}(W(t))\le p_2. \end{aligned} \tag{30}$$

The above theorem says that when the step size is small, we can bound the eigenvalues of τ_t and the singular values of $W(t)$ using the initial *imbalance* and *margin*. When $\alpha_0>0$, we have $a_1>0$, and the LHS of equation 29 is a monotonically increasing function of η, when $\eta>0$, and is equal to zero, when $\eta=0$. Therefore, each polynomial has a unique positive root. The condition $\eta\le\eta_{\max}$ is used to control $\|D(t)-D(0)\|_F$. We use $\lambda_{\min}(\tau_t)$ as an example to illustrate why we need to control $\|D(t)-D(0)\|_F$. In GD, equation 9 still holds. However, since the imbalance is no longer constant, i.e. since $D(t)\neq D(0)$, we no longer have $\alpha(D(t),p_1)=\alpha(D(0),p_1)$. Nonetheless, after careful analysis, we observe that the change of imbalance at each iteration is of order η^2. Moreover, as long as the loss decreases linearly and η is small (see equation 29), we can prove that $\|D(t)-D(0)\|_F\le O(\eta)$. Thus, we first introduce c_1 to control the change of the eigenvalues of the imbalance matrix. Then, if the step size is bounded, i.e. $\eta\le\eta_{\max}$, we can show that $\alpha(D(t),p_1)\ge c_1\alpha(D(0),p_1)$. A similar analysis yields the upper bound for $\lambda_{\max}(\tau_t)$. When c_1,c_2 are chosen to be close to one, the change in eigenvalues of imbalance is guaranteed to be small, but it requires a smaller step as $\eta_{\max}$ is small.

Then, based on Theorem 3.1, we can prove the linear convergence of GD.

Theorem 3.2 (Convergence rate of gradient descent on two-layer linear networks)**.** *Under the assumptions in Theorem 3.1, for any* $0<\eta\le\eta_{\max}:=\min\{\eta_1^{\max},\eta_2^{\max}\}$*, the loss function under GD satisfies*

$$L(t+1)\le f(\eta,t)L(t),$$

for $f(\eta,t)$ *as defined in equation 26, and with*

$$0<f(\eta,t)\le f(\eta,0)<1,\quad\forall t\ge0. \tag{31}$$

Thus, the loss converges linearly, i.e.,

$$L(t)\le\Pi_{k=0}^t f(\eta,k)\,L(0)\le f(\eta,0)^tL(0). \tag{32}$$

with rate given by $f(\eta,0)$.

In $f(\eta,t)$, $-a_1\eta$ is an important term that facilitates convergence because it is the only term that is associated with a negative coefficient. a_1 depends on $p_1,D(0)$ via α_0, and when $\alpha_0>0$, i.e., there is either sufficient margin or imbalance, we have $a_1>0$. The proof Theorem 3.1 and Theorem 3.2 is presented in Appendix C.

Detailed comparison with SOTA We compare our results with other works studying the same problem [Du et al., 2018a, Arora et al., 2018]. In both works, the authors make assumptions that the initial imbalance is small. In our work, Theorem 3.2 holds if there is either a sufficient imbalance or sufficient margin at initialization, which is a more general setting. In [Du et al., 2018a], they prove the loss decreases, and the imbalance remains small during training, but the paper does not provide an explicit convergence rate. More importantly, a decay in step size is needed to control the difference between $D(t)$ and $D(0)$. In our work, we provide an explicit convergence rate without the need to decrease step size. In [Arora et al., 2018], the authors provide an explicit convergence rate. However, their result depends on the property that when step size is small, $\|D(t)\|_F\le2\|D(0)\|_F$. We think the *two* used in their proof is an artifact and improve it by introducing c_1 and c_2 and characterize the dependence between step size and c_1,c_2, which is a more general case.

Comparison with non-overparametrized regime In the GF regime, [Min et al., 2022, Tarmoun et al., 2021] show that if α_0 is sufficiently large, the over-parametrized model can have a faster convergence rate than the non-overparametrized model. However, as shown in the next proposition, such a result does not extend to the GD regime.

Proposition 3.1. *If* $\alpha_0>0$*, for all* $0<\eta\le\eta_{\max}$ *and for all* $t=0,1,\cdots$*, the following inequality holds*

$$f(\eta,t)\ge1-\frac{1}{\kappa} \tag{33}$$

where $\kappa=\frac{K}{\mu}$ *is the condition number of the non-overparametrized Problem 1*

In Proposition 3.1, $1-\frac{1}{\kappa}$ is the theoretical optimal convergence rate of solving Problem 1 via GD (see §3.1 for a derivation of it). As a result, Proposition 3.1 states that the convergence rate derived in Theorem 3.2, i.e., $f(\eta,t)$, for solving the over-parametrized Problem 2 via GD, is always larger. Nevertheless, we point out that Theorem 3.2 only provides an upper bound on the rate, and further study is needed to characterize its tightness.

3.3 Adaptive Step Size Scheme

In Theorem 3.2, we show that a fixed step size $\eta\le\eta_{\max}$ guarantees linear rate of convergence of the loss $L(t)$. However, our analysis also suggests that faster convergence

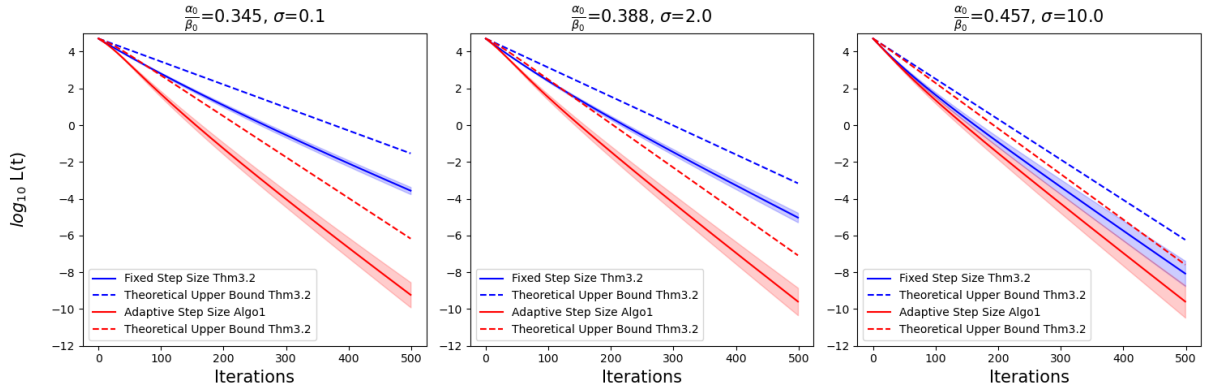

Figure 1: Tightness check for upper bound in Theorem 3.2 versus reconstruction error. We run the simulations three times. The dashed lines represent the upper bound on the training loss we derive in the Theorem!3.2 and the solid lines represent the $\log_{10}$ of the mean reconstruction error $L(t)$ (blue lines: GD under a fixed step size, red: GD under our proposed adaptive step size). The blue dashed lines are calculated using the LHS of equation (32), while the red dashed lines are calculated using $\Pi_{k=0}^{t} f(\eta, k)\, L(0)$, where η_k is the optimal step size derived in equation (35).

can be achieved if we use a time-varying step size η_t at every iteration t. Specifically, in Theorem 3.2, we show that when $\eta \leq \eta_{\max}$ the following result holds

$$L(t+1) \leq f(\eta, t) L(t), \forall t = 0, 1, 2, \dots \tag{34}$$

At every iteration t, the best choice for the step size, suggested by our theoretical result in equation 34 is the one that minimizes $f(\eta, t)$, subject to our constraint for convergence $\eta_t \leq \eta_{\max}$:

$$\eta_t = \arg \min_{\eta \leq \eta_{\max}} f(\eta, t)\,. \tag{35}$$

Finding the solution to equation 35 only requires solving a third-order polynomial:

Claim 3.1. *Suppose $\alpha_0 > 0$. Let η_t' be the unique positive root of the following equation*

$$-a_1 + 2a_2(t)\eta + 3a_3(t)\eta^2 + 4a_4(t)\eta^3 = 0. \tag{36}$$

Then the solution to Problem 35 is $\eta_t = \min(\eta_t', \eta_{\max})$.

The proof is in Appendix E. This suggests that one can find η_t very efficiently at each iteration. We present the GD algorithm with adaptive step size scheduling below:

Convergence rate under adaptive step size Notice that $f(\eta, t)$ depends on the iteration t via the loss function $L(t)$. As the training proceeds, the adaptive step size scheme ensures $f(\eta, t) < 1$ such that the loss $L(t)$ converges to zero. This in turn affects the asymptotic expression for $f(\eta, t)$. Specifically, when t is sufficiently large (so that $L(t) \simeq 0$),

Algorithm 1: GD with Adaptive Step Size Scheme

Data: X, Y, and initial $W_1(0), W_2(0)$
Result: W_1^*, W_2^* which minimize $\frac{1}{2}\|Y - XW_1W_2\|_F^2$.
for $t = 0, 1, 2 \cdots$ **do**
  /* adaptive step size */
  $\eta_t \leftarrow \arg\min_{\eta \leq \eta_{\max}} f(\eta, t)$ /* GD update with η_t */
  $\begin{pmatrix} W_1(t+1) \\ W_2(t+1) \end{pmatrix} = \begin{pmatrix} W_1(t) \\ W_2(t) \end{pmatrix} - \eta_t \gamma_t (\nabla \ell(t))$.
end

we have

$$f(\eta, t) \simeq 1 - 2(c_1\alpha_0)\sigma_{\min}^2(X)\eta + \kappa\sigma_{\min}^4(X)(c_2\beta_0)^2\eta^2\,. \tag{37}$$

Under a proper choice of c_1, c_2 such that

$$\eta_{\max} \geq \frac{c_1\alpha_0}{c_2^2\beta_0^2\kappa\sigma_{\min}^2(X)} \tag{38}$$

the adaptive step size scheduler yields a rate

$$f(\eta_t, t) \simeq 1 - \frac{(c_1\alpha_0)^2}{(c_2\beta_0)^2}\frac{1}{\kappa}\,. \tag{39}$$

In this case, the asymptotic convergence rate of GD with adaptive size depends on both $\frac{\alpha_0}{\beta_0}$ and $\frac{1}{\kappa}$. In Appendix E, we show that there always exists such choice of c_1, c_2 such that equation 38 holds. Our numerical simulations show that GD with the adaptive step size strategy achieves faster

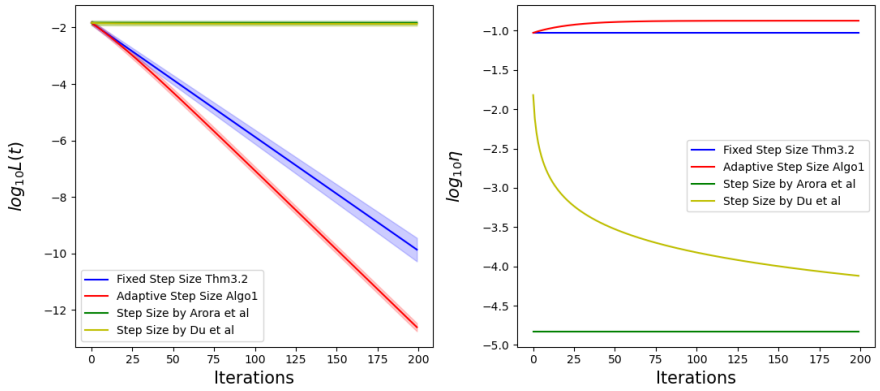

Figure 2: Comparison between different step sizes is presented here. We run the simulations three times. On the left plot, the solid line is the $\log_{10}$ of the mean reconstruction error $L(t)$, and the shaded area is the mean plus and minus one standard deviation. We select step sizes in [Arora et al., 2018, Du et al., 2018a] and the other two step sizes are the same as previous experiment. On the right plot, how each choice of step size changes during training is presented.

convergence than GD with a fixed step size. Please refer to Section 4.1 for details.

4 SIMULATIONS

In this section, we first present empirical evidence that Theorem 3.2 provides a good characterization of the actual convergence rate of the loss under several different initializations. Moreover, we compared the convergence rate of GD using the step sizes presented in Theorem 3.2 and Algorithm 1 with those using step size proposed in previous work [Arora et al., 2018, Du et al., 2018a], and our step sizes achieve considerably faster convergence rate.

Throughout the experiments, we fix $c_1 = 0.5$ and $c_2 = 1.5$ in our choice of step sizes. The details of the simulations are presented in Appendix F.

4.1 Evaluation of the tightness of the theoretical bound on the convergence rate

We train a two-layer linear network using GD on the squared loss in equation 2. We generate the data matrix as follows: $X \in \mathbb{R}^{20\times 20}, X[i,j] \sim \mathcal{N}(0,1)$, and $Y = X\Theta$ where $\Theta \in \mathbb{R}^{20\times 20}, \Theta[i,j] \sim \mathcal{N}(0,1)$. The initial weight matrices are generated as $W_1(0) = \sigma U_0, W_2(0) = \frac{1}{\sigma} V_0$, where $U_0 \in \mathbb{R}^{20\times 1000}, V_0 \in \mathbb{R}^{1000\times 20}$ and have entry-wise i.i.d. samples drawn from a standard gaussian $\mathcal{N}(0,1)$. We choose different values of σ to test our convergence rate in different regimes.

Figure 1 compares the convergence rate predicted by Theorem 3.2 with the actual convergence rate of $L(t)$, under different values of σ and approximately similar values of $\frac{\alpha_0}{\beta_0}$. In all scenarios, our theoretical bounds follow the empirical results relatively well. Moreover, we see, on the one hand, that the rate of convergence of the adaptive step size regime is relatively insensitive to the value of σ in the initialization. This is not surprising, given the discussion in §3.3 on the asymptotic rate of the adaptive step size regime, which mostly depends on $\frac{\alpha_0}{\beta_0}$. On the other hand, the rate of the fixed step size value in Theorem 3.2 varies significantly with σ. As a result, this experiment suggests a certain level of robustness provided by the adaptive step size scheme. Finally, in this experiment, the initial margin is 0 and the there is large initial imbalance. Those initial conditions violate the assumptions in [Arora et al., 2018, Du et al., 2018a], but still enjoys linear convergence.

4.2 Comparison between different learning rates presented in previous work

In this section, we compare the step sizes proposed in Theorem 2 of [Arora et al., 2018] and Theorem 3.1 of [Du et al., 2018a] to the step size of Theorem 3.2 and the adaptive step size proposed in Section 3.3. We note that the analyses in [Arora et al., 2018, Du et al., 2018a] assume that the initialization is approximately balanced ($\|D(0)\|_F$ is small). In addition, [Arora et al., 2018] requires the initialization to have sufficient margin ($\|Y - XW_1(0)W_2(0)\|_F$ is small). Therefore, we compare our results with [Arora et al., 2018, Du et al., 2018a] under initialization that is balanced ($D(0) = 0$) and has a sufficiently large margin. In order to do so, we generate the training data using the following:

$$\begin{aligned} &X = I_{20}, Y = XW(0) + 0.01\varepsilon\,, \\ &W(0) \in \mathbb{R}^{20\times 1}, W(0)[i,j] \sim \mathcal{N}(0, 1/4)\,, \\ &\varepsilon \in \mathbb{R}^{20\times 1}, \varepsilon[i,j] \sim \mathcal{N}(0,1)\,. \end{aligned} \tag{40}$$

Here we first randomly initialize the product $W(0)$ and the construction ensures that $\|Y - XW(0)\|_F$ is small so

that there is a sufficiently large margin. Then, we initialize weights $W_1(0), W_2(0)$ of the linear networks such that $W_1(0)W_2(0) = W(0)$. The width of linear networks is 1000. To construct a balanced initialization, we compute the SVD of the initial product $W(0) = U\Sigma V^\top$. Then, we initialize the weights as $W_1(0) = U\Sigma^{1/2}, W_2(0) = \Sigma^{1/2}V^\top$.

Figure 2 shows the step size proposed in our paper achieves the fastest convergence compared with other SOTA methods [Arora et al., 2018, Du et al., 2018a]. On the right plot, step sizes proposed in this work is larger than the one proposed in [Arora et al., 2018, Du et al., 2018a].

5 CONCLUSIONS

This paper studied the convergence of GD for optimizing two-layer linear networks. In particular, we derived a convergence rate for networks of finite width that are initialized in a non-NTK regime. Our results build upon recent work for GF, which derived convergence rates that depend on the imbalance and margin of the initialization. However, a key challenge in the GD regime is that the imbalance of the weights changes with the iterations of GD. In this paper, we show that when the step size is small, the imbalance at iteration t is close to its value at initialization. Moreover, we show that under this constraint on the step size, the loss is decreasing. In addition, we derive an explicit convergence rate that depends on the margin, imbalance, and condition number of the data matrix. Finally, based on the convergence rate, we propose an adaptive step size scheme that accelerates convergence compared with a constant step size. Empirically, we show the convergence rate derived in our work is tighter than in previous work.

Acknowledgements

The authors acknowledge the support of the Office of Naval Research (grant 503405-78051), the National Science Foundation (grants 203198, 1934979, 1752362,2136324), and the Simons Foundation (grant 814201).

References

[Arora et al., 2018] Arora, S., Cohen, N., Golowich, N., and Hu, W. (2018). A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*.

[Chen et al., 2022] Chen, Z., Vanden-Eijnden, E., and Bruna, J. (2022). On feature learning in neural networks with global convergence guarantees. *arXiv preprint arXiv:2204.10782*.

[Chizat and Bach, 2018] Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31.

[Chizat et al., 2019] Chizat, L., Oyallon, E., and Bach, F. (2019). On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32.

[Ding et al., 2022] Ding, Z., Chen, S., Li, Q., and Wright, S. J. (2022). Overparameterization of deep resnet: Zero loss and mean-field analysis. *J. Mach. Learn. Res.*, 23:48–1.

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

[Du et al., 2018a] Du, S. S., Hu, W., and Lee, J. D. (2018a). Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in Neural Information Processing Systems*, 31.

[Du et al., 2018b] Du, S. S., Zhai, X., Poczos, B., and Singh, A. (2018b). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*.

[Gidel et al., 2019] Gidel, G., Bach, F., and Lacoste-Julien, S. (2019). Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

[Jacot et al., 2018] Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

[Lee et al., 2019] Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32.

[Liu et al., 2022] Liu, C., Zhu, L., and Belkin, M. (2022). Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116.

[Mei et al., 2018] Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671.

[Min et al., 2022] Min, H., Tarmoun, S., Vidal, R., and Mallada, E. (2022). Convergence and implicit bias of gradient flow on overparametrized linear networks. *arXiv preprint arXiv:2105.06351*.

[Minaee et al., 2021] Minaee, S., Boykov, Y. Y., Porikli, F., Plaza, A. J., Kehtarnavaz, N., and Terzopoulos, D. (2021). Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*.

[Nguegnang et al., 2021] Nguegnang, G. M., Rauhut, H., and Terstiege, U. (2021). Convergence of gradient descent for learning linear neural networks. *arXiv preprint arXiv:2108.02040*.

[Oymak and Soltanolkotabi, 2020] Oymak, S. and Soltanolkotabi, M. (2020). Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105.

[Rotskoff and Vanden-Eijnden, 2018a] Rotskoff, G. and Vanden-Eijnden, E. (2018a). Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. *Advances in neural information processing systems*, 31.

[Rotskoff and Vanden-Eijnden, 2018b] Rotskoff, G. M. and Vanden-Eijnden, E. (2018b). Trainability and accuracy of neural networks: An interacting particle system approach. *arXiv preprint arXiv:1805.00915*.

[Saxe et al., 2013] Saxe, A. M., McClelland, J. L., and Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

[Silver et al., 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503.

[Sirignano and Spiliopoulos, 2020] Sirignano, J. and Spiliopoulos, K. (2020). Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752.

[Tarmoun et al., 2021] Tarmoun, S., Franca, G., Haeffele, B. D., and Vidal, R. (2021). Understanding the dynamics of gradient flow in overparameterized linear models. In *International Conference on Machine Learning*, pages 10153–10161. PMLR.

[Vaswani et al., 2018] Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., et al. (2018). Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416*.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

[Vo et al., 2019] Vo, N. N., He, X., Liu, S., and Xu, G. (2019). Deep learning for decision making and the optimization of socially responsible investments and portfolio. *Decision Support Systems*, 124:113097.

A CASE WHEN DATA MATRIX IS RANK DEFICIENT

Here, we show for any data matrix X of arbitrary dimensions and rank, the over-parametrized problem

$$\min_{W_1,W_2} L(W_1,W_2)=\frac{1}{2}\|Y-XW_1W_2\|_F^2, \tag{41}$$

can be reparametrized into the following problem

$$\min_{\tilde{W}_1,\tilde{W}_2} L(\tilde{W}_1,\tilde{W}_2)=\frac{1}{2}\|\tilde{Y}-\tilde{X}\tilde{W}_1\tilde{W}_2\|_F^2, \tag{42}$$

where $\tilde{X}$ is a square matrix of full rank.

Let singular value decomposition of X be

$$X=[U_1,U_2]\begin{bmatrix}\Sigma_X & 0\\ 0 & 0\end{bmatrix}\begin{bmatrix}V_1^\top\\ V_2^\top\end{bmatrix}, \tag{43}$$

where Σ_X contains all non-zero singular values of X. Then, we have $X=U_1\Sigma_X V_1^\top$. The GD update of W_1,W_2 is

$$\begin{aligned}W_1(t+1)&=W_1(t)+\eta X^\top E(t)W_2^\top(t)=W_1(t)+\eta V_1\Sigma_X U_1^\top E(t)W_2^\top(t),\\ W_2(t+1)&=W_2(t)+\eta W_1^\top(t)X^\top E(t)=W_2(t)+\eta W_1^\top(t)V_1\Sigma_X U_1^\top E(t).\end{aligned} \tag{44}$$

We project W_1 onto the space spanned by V_1,V_2,

$$\begin{aligned}W_{11}&=V_1^\top W_1,\\ W_{12}&=V_2^\top W_1.\end{aligned} \tag{45}$$

Furthermore, we define $\tilde{E}(t)=U_1^\top E(t)$. Based on above, one has

$$\begin{aligned}W_{11}(t+1)&=W_{11}(t)+\eta\Sigma_X\tilde{E}(t)W_2(t),\\ W_{12}(t+1)&=W_{12}(t),\\ W_2(t+1)&=W_2(t)+\eta W_{11}^\top(t)\Sigma_X\tilde{E}(t).\end{aligned} \tag{46}$$

The update of W_{11},W_2 is the same to the following problem

$$\min_{W_{11},W_2} L(W_{11},W_2)=\frac{1}{2}\|U_1^\top Y-\Sigma_X W_{11}W_2\|_F^2, \tag{47}$$

where Σ_X is a sqaure matrix of full rank. The above problem takes the same form as equation 42 where $\tilde{Y}=U_1^\top Y,\tilde{X}=\Sigma_X,\tilde{W}_1=W_{11},\tilde{W}_2=W_2$.

B PROOF OF LEMMA 3.1

In this section, we present detailed proof of Lemma 3.1.

Lemma 3.1. *If at the t-th iteration of GD applied to the over-parametrized loss L, the step size η satisfies*

$$\begin{aligned}&\lambda_{\min}(\tau_t)-\eta\|\nabla\ell(t)\|_F\sigma_{\max}(W(t))\\ &-\frac{K\eta}{2}\big[\lambda_{\max}(\tau_t)+\eta\|\nabla\ell(t)\|_F\sigma_{\max}(W(t))\big]^2\geq 0\,,\end{aligned} \tag{48}$$

then the following inequality holds

$$L(t+1)\leq\rho(\eta,t)L(t)\,, \tag{49}$$

where

$$\begin{aligned}\rho(\eta,t)=1&-2\eta\mu\lambda_{\min}(\tau_t)+K\mu\eta^2\lambda_{\max}^2(\tau_t)\\ &+2\eta^2\mu\sigma_{\max}(W(t))\|\nabla\ell(t)\|_F\\ &+2\eta^3\mu K\lambda_{\max}(\tau_t)\sigma_{\max}(W(t))\|\nabla\ell(t)\|_F\\ &+\eta^4\mu K\sigma_{\max}^2(W(t))\|\nabla\ell(t)\|_F^2.\end{aligned} \tag{50}$$

Proof. Applying smoothness equation 13to the update of the product in equation 18, we get

$$\begin{aligned} L(t+1) &\leq L(t) - \eta\langle\nabla\ell(t), \tau_t(\nabla\ell(t)) - \eta\nabla\ell(t)W(t)^\top\nabla\ell(t)\rangle \\ &\quad + \frac{K}{2}\eta^2\|\tau_t(\nabla\ell(t)) - \eta\nabla\ell(t)W(t)^\top\nabla\ell(t)\|_F^2 \\ &= L(t) - \eta\langle\nabla\ell(t), \tau_t(\nabla\ell(t))\rangle \\ &\quad + \eta^2\big(\langle\nabla\ell(t), \nabla\ell(t)W(t)^\top\nabla\ell(t)\rangle + \tfrac{K}{2}\|\tau_t(\nabla\ell(t))\|_F^2\big) \\ &\quad - \eta^3 K\langle\tau_t(\nabla\ell(t)), \nabla\ell(t)W(t)^\top\nabla\ell(t)\rangle \\ &\quad + \frac{K}{2}\eta^4\|\nabla\ell(t)W(t)^\top\nabla\ell(t)\|_F^2 \end{aligned} \tag{51}$$

Then, we upper bound each term in the above inequality separately. First, since τ_t is a positive semi-definite operator, we have

$$\begin{aligned} \langle\nabla\ell(t), \tau_t(\nabla\ell(t))\rangle &\geq \lambda_{\min}(\tau_t)\|\nabla\ell(t)\|_F^2 \\ \|\tau_t(\nabla\ell(t))\|_F^2 &\leq \lambda_{\max}^2(\tau_t)\|\nabla\ell(t)\|_F^2 \end{aligned} \tag{52}$$

Then, using the sub-multiplicative property of Frobenius norm and Cauchy Schwartz inequality, we can bound the rest terms in equation 51

$$\begin{aligned} |\langle\nabla\ell(t), \nabla\ell(t)W(t)^\top\nabla\ell(t)\rangle| &\leq \|\nabla\ell(t)\|_F\|\nabla\ell(t)W(t)^\top\nabla\ell(t)\|_F \leq \|\nabla\ell(t)\|_F^3\sigma_{\max}(W(t)) \\ |\langle\tau_t(\nabla\ell(t)), \nabla\ell(t)W(t)^\top\nabla\ell(t)\rangle| &\leq \|\nabla\ell(t)\|_F\sigma_{\max}(W(t))\langle\nabla\ell(t), \tau_t(\nabla\ell(t))\rangle \leq \lambda_{\max}(\tau_t)\sigma_{\max}(W(t))\|\nabla\ell(t)\|_F^3 \\ \|\nabla\ell(t)W(t)^\top\nabla\ell(t)\|_F^2 &\leq \sigma_{\max}^2(W(t))\|\nabla\ell(t)\|_F^4. \end{aligned} \tag{53}$$

Based on above results, we can further upper bound equation 51

$$L(t+1) \leq L(t) - \eta\langle\nabla\ell(t), \tau_t(\nabla\ell(t))\rangle \tag{54}$$
$$+ \eta^2\big(\langle\nabla\ell(t), \nabla\ell(t)W(t)^\top\nabla\ell(t)\rangle + \tfrac{K}{2}\|\tau_t(\nabla\ell(t))\|_F^2\big) \tag{55}$$
$$- \eta^3 K\langle\tau_t(\nabla\ell(t)), \nabla\ell(t)W(t)^\top\nabla\ell(t)\rangle \tag{56}$$
$$+ \frac{K}{2}\eta^4\|\nabla\ell(t)W(t)^\top\nabla\ell(t)\|_F^2 \tag{57}$$
$$\leq L(t) - \eta\lambda_{\min}(\tau_t)\|\nabla\ell(t)\|_F^2 \tag{58}$$
$$+ \eta^2(\sigma_{\max}(W(t))\|\nabla\ell(t)\|_F^3 + \frac{K}{2}\lambda_{\max}^2(\tau_t)\|\nabla\ell(t)\|_F^2) \tag{59}$$
$$+ \eta^3 K\lambda_{\max}(\tau_t)\sigma_{\max}(W(t))\|\nabla\ell(t)\|_F^3 \tag{60}$$
$$+ \eta^4\frac{K}{2}\sigma_{\max}^2(W(t))\|\nabla\ell(t)\|_F^4 \tag{61}$$
$$= L(t) - \eta\|\nabla\ell(t)\|_F^2 g(\eta) \tag{62}$$

where

$$\begin{aligned} g(\eta) = \lambda_{\min}(\tau_t) &- \eta(\sigma_{\max}(W(t))\|\nabla\ell(t)\|_F + \frac{K}{2}\lambda_{\max}^2(\tau_t)) \\ &- \eta^2 K\lambda_{\max}(\tau_t)\sigma_{\max}(W(t))\|\nabla\ell(t)\|_F - \eta^3\frac{K}{2}\sigma_{\max}^2(W(t))\|\nabla\ell(t)\|_F^2. \end{aligned} \tag{63}$$

When $g(\eta) > 0$, which is assumed in equation 48, we apply PL condition $\frac{1}{2}\|\nabla\ell(t)\|_F^2 \geq \mu\ell(t)$ to the above equation to get

$$\begin{aligned} L(t+1) &\leq L(t) \times \Big\{1 - 2\eta\mu\lambda_{min}(\tau_t) \\ &\quad + 2\eta^2\mu\big(\sigma_{\max}(W(t))\|\nabla\ell(t)\|_F + \frac{K}{2}\lambda_{\max}^2(\tau_t)\big) \\ &\quad + 2\eta^3\mu K\lambda_{\max}(\tau_t)\sigma_{\max}(W(t))\|\nabla\ell(t)\|_F \\ &\quad + \eta^4\mu K\sigma_{\max}^2(W(t))\|\nabla\ell(t)\|_F^2\Big\} \\ &= \rho(\eta, t)L(t). \end{aligned} \tag{64}$$

□

C PROOF OF THEOREM 3.1 AND THEOREM 3.2

Here we prove a new Theorem which implies Theorem 3.1 and Theorem 3.2.

Theorem C.1. *Under the assumptions in Theorem 3.1, for any* $0 < \eta \leq \eta_{\max} := \min\{\eta_1^{\max}, \eta_2^{\max}\}$*, the following four properties hold for all* $t = 0, 1, 2, \cdots$.

- $A_1(t) : L(t) \leq f(\eta, t)L(t-1)$*, where* $f(\eta, t) = 1 - a_1\eta + a_2(t)\eta^2 + a_3(t)\eta^3 + a_4(t)\eta^4 < 1$.
- $A_2(t) : p_1 \leq \sigma_{\min}(W(t)) \leq \sigma_{\max}(W(t)) \leq p_2$.
- $A_3(t) : \|D(t) - D(0)\|_F \leq \frac{2c_2\beta_0\sigma_{\max}^2(X)L(0)\eta^2}{1-f(\eta,0)}$ *when* $\eta < \eta_{\max}$.
- $A_4(t) : c_1\alpha_0 \leq \sigma_{\min}^2(W_1(t)) + \sigma_{\min}^2(W_2(t)) \leq \lambda_{\min}(\tau_t) \leq \lambda_{\max}(\tau_t) \leq \sigma_{\max}^2(W_1(t)) + \sigma_{\max}^2(W_2(t) \leq c_2\beta_0$.

Notice Theorem 3.1 is Property $A_2(t), A_4(t)$, and Theorem 3.2 is implied by Property $A_1(t)$ because when $L(k) \leq L(0)$ hold for all $k = 0, 1, \cdots, t$, we have $a_2(k) \leq a_2(0), a_3(k) \leq a_3(0), a_4(k) \leq a_4(0)$. Thus, $f(\eta, k) \leq f(\eta, 0)$. As a result, the following inequality holds

$$L(t) \leq f(\eta, t)L(t-1) \leq L(0)\prod_{k=0}^{t-1} f(\eta, k) \leq f(\eta, 0)^t L(0). \tag{65}$$

Before proving Theorem C.1, we first present several **preliminary lemmas**.

Lemma C.1. *For matrix* A, B*, we have*

$$\begin{aligned}\sigma_{\min}^2(A)\|B\|_F^2 \leq \|AB\|_F^2 \leq \sigma_{\max}^2(A)\|B\|_F^2\\ \sigma_{\min}^2(B)\|A\|_F^2 \leq \|AB\|_F^2 \leq \sigma_{\max}^2(B)\|A\|_F^2.\end{aligned} \tag{66}$$

Proof.

$$\begin{aligned}\|AB\|_F^2 &= \operatorname{tr}\left(ABB^\top A^\top\right)\\ &= \operatorname{tr}\left(A^\top ABB^\top\right) \qquad \text{use cyclic property of trace}\\ &\leq \lambda_{\max}\left(A^\top A\right)\|B\|_F^2 \qquad \text{use trace inequality}\\ &= \sigma_{\max}^2(A)\|B\|_F^2.\end{aligned} \tag{67}$$

For the other way

$$\begin{aligned}\|AB\|_F^2 &= \operatorname{tr}\left(ABB^\top A^\top\right)\\ &= \operatorname{tr}\left(A^\top ABB^\top\right)\\ &\leq \lambda_{\max}\left(BB^\top\right)\|A\|_F^2\\ &= \sigma_{\max}^2(B)\|A\|_F^2.\end{aligned} \tag{68}$$

The lower bound is similar. □

Lemma C.2. *Let* $X \in \mathbb{R}^{N\times n}, Y \in \mathbb{R}^{N\times m}$*. Assume* $N \leq n$ *and* $rank(X) = N$*. For arbitrary* $W \in \mathbb{R}^{n\times m}$*, the following holds for* $\ell(W) = \frac{1}{2}\|Y - XW\|_F^2$

$$2\sigma_{\min}^2(X)\ell(W) \leq \|\nabla\ell(W)\|_F^2 \leq 2\sigma_{\max}^2(X)\ell(W). \tag{69}$$

Proof. The first inequality is PL inequality. We then prove the second

$$\begin{aligned}\|\nabla\ell(W)\|_F^2 &= \|X^\top(Y - XW)\|_F^2 \qquad \text{gradient calculation}\\ &\leq \sigma_{\max}^2(X)\|Y - XW\|_F^2 \qquad \text{use Lemma C.1}\\ &= 2\sigma_{\max}^2(X)\ell(W).\end{aligned} \tag{70}$$

□

Lemma C.3. *The difference of the imbalance between iteration $t+1$ and t can be upper bounded by*

$$\|D(t+1) - D(t)\|_F \leq 2\eta^2 \sigma_{\max}^2(X)\big(\sigma_{\max}^2(W_1(t)) + \sigma_{\max}^2(W_2(t))\big)L(t). \tag{71}$$

Proof. Notice the definition of imbalance is $D(t) := W_1^\top(t)W_1(t) - W_2(t)W_2^\top(t)$ and the update of GD is given in equation 4. Thus, using both results, one has

$$\begin{aligned} D(t+1) &= \big(W_1(t) - \eta\nabla\ell(t)W_2(t)^\top\big)^\top\big(W_1(t) - \eta\nabla\ell(t)W_2(t)^\top\big) \qquad \text{plug in GD update} \\ &\quad - \big(W_2(t) - \eta W_1(t)^\top\nabla\ell(t)\big)\big(W_2(t) - \eta W_1(t)^\top\nabla\ell(t)\big)^\top \\ &= D(t) + \eta^2\big(W_2(t)\nabla\ell(t)^\top\nabla\ell(t)W_2(t)^\top - W_1(t)^\top\nabla\ell(t)\nabla\ell(t)^\top W_1(t)\big). \end{aligned} \tag{72}$$

Then, we can upper bound $\|D(t+1) - D(t)\|_F$ using Lemma C.1 and Lemma C.2

$$\begin{aligned} \|D(t+1) - D(t)\|_F &= \eta^2\|W_2(t)\nabla\ell(t)^\top\nabla\ell(t)W_2(t)^\top - W_1(t)^\top\nabla\ell(t)\nabla\ell(t)^\top W_1(t)\|_F \\ &\leq \eta^2\big(\|W_2(t)\nabla\ell(t)^\top\nabla\ell(t)W_2(t)^\top\|_F + \|W_1(t)^\top\nabla\ell(t)\nabla\ell(t)^\top W_1(t)\|_F\big) \\ &\leq \eta^2\big(\|W_2(t)\nabla\ell(t)^\top\|_F^2 + \|W_1(t)^\top\nabla\ell(t)\|_F^2\big) \qquad \text{by Lemma C.1} \\ &\leq \eta^2\big(\sigma_{\max}^2(W_1(t)) + \sigma_{\max}^2(W_2(t))\big)\|\nabla\ell(t)\|_F^2 \qquad \text{by Lemma C.2} \\ &\leq 2\eta^2\sigma_{\max}^2(X)\big(\sigma_{\max}^2(W_1(t)) + \sigma_{\max}^2(W_2(t))\big)L(t). \end{aligned} \tag{73}$$

□

Lemma C.4. *Suppose $h > \min\{r, m\}$. Given any $A \in \mathbb{R}^{r\times h}, B \in \mathbb{R}^{h\times m}$ that satisfy $A^\top A - BB^\top = D$, we have*

$$\lambda_m(B^\top B) \geq \frac{-\bar{\lambda} + \underline{\lambda} + \sqrt{(\bar{\lambda} + \underline{\lambda})^2 + 4\sigma_m^2(AB)}}{2} \tag{74}$$

where $\bar{\lambda} = \max\{\lambda_1(D), 0\}$ and $\underline{\lambda} = \max\{\lambda_m(-D), 0\}$.

Lemma C.4 is cited from [Min et al., 2022] and the proof can be found in [Min et al., 2022] Lemma 8.

Lemma C.5. *Suppose $h > \min\{r, m\}$. Given any $A \in \mathbb{R}^{r\times h}, B \in \mathbb{R}^{h\times m}$ that satisfy $A^\top A - BB^\top = D$, we have*

$$\lambda_{\max}(B^\top B) \leq \frac{\max(\lambda_{\max}(-D), 0) + \sqrt{\max(\lambda_{\max}(-D), 0)^2 + 4\sigma_{\max}^2(AB)}}{2} \tag{75}$$

Proof. We first choose $z \in \mathbb{R}^m$ with $\|z\|_2 = 1$ s.t.

$$z^\top B^\top Bz = \lambda_{\max}(B^\top B). \tag{76}$$

Then, we have

$$\begin{aligned} \lambda_{\max}^2(B^\top B) - z^\top B^\top A^\top ABz &= z^\top B^\top BB^\top Bz - z^\top B^\top A^\top ABz \\ &= z^\top(B^\top BB^\top B - B^\top A^\top AB)z \\ &= z^\top B^\top(BB^\top - A^\top A)Bz \\ &= z^\top B^\top(-D)Bz. \end{aligned} \tag{77}$$

Notice

$$\begin{aligned} &\lambda_{\max}^2(B^\top B) - z^\top B^\top A^\top ABz \geq \lambda_{\max}^2(B^\top B) - \sigma_{\max}^2(AB) \\ &z^\top B^\top(-D)Bz \leq \max(\lambda_{\max}(-D), 0)\|Bz\|_2^2 \leq \max(\lambda_{\max}(-D), 0)\lambda_{\max}(B^\top B). \end{aligned} \tag{78}$$

Thus, we have

$$\lambda_{\max}(B^\top B)^2 - \sigma_{\max}^2(AB) \leq \max(\lambda_{\max}(-D), 0)\lambda_{\max}(B^\top B). \tag{79}$$

The solution to the above inequality gives us the results. □

Then, we begin the proof of Theorem C.1.

Proof. Assume $A_1(k), A_2(k), A_3(k), A_4(k)$ hold at iteration $k = 1, 2, \cdots, t$, then we prove they all hold at iteration $t+1$. First, **we prove** $A_1(t+1)$ **hold**. According to Lemma 3.1, we have

$$\begin{aligned} L(t+1) \leq L(t) \times \Big\{ & 1 - 2\eta\mu\lambda_{min}(\tau_t) \\ & + 2\eta^2\mu\big(\sigma_{\max}(W(t))\|\nabla\ell(t)\|_F + \frac{K}{2}\lambda^2_{\max}(\tau_t)\big) \\ & + 2\eta^3\mu K\lambda_{\max}(\tau_t)\sigma_{\max}(W(t))\|\nabla\ell(t)\|_F \\ & + \eta^4\mu K\sigma^2_{\max}(W(t))\|\nabla\ell(t)\|^2_F \Big\} \end{aligned} \tag{80}$$

Since $A_2(t), A_4(t)$ hold, we can further upper bound the above inequality

$$\begin{aligned} L(t+1) \leq L(t) \times \Big\{ & 1 - 2\eta\mu c_1\alpha_0 + 2\eta^2\mu\big(p_2\|\nabla\ell(t)\|_F + \frac{K}{2}(c_1\beta_0)^2\big) \\ & + 2\eta^3\mu K c_1\beta_0 p_2\|\nabla\ell(t)\|_F + \eta^4\mu K p_2^2\|\nabla\ell(t)\|^2_F \Big\} \end{aligned} \tag{81}$$

Apply Lemma C.2

$$\begin{aligned} L(t+1) &\leq L(t) \times \Big\{ 1 - 2\eta\mu c_1\alpha_0 + 2\eta^2\mu\big(p_2\sqrt{2\sigma^2_{\max}(X)L(t)} + \frac{\sigma^2_{\max}(X)}{2}(c_2\beta_0)^2\big) \\ &\quad + 2\eta^3\mu\sigma^2_{\max}(X)c_2\beta_0 p_2\sqrt{2\sigma^2_{\max}(X)L(t)} + 2\eta^4\mu\sigma^4_{\max}(X)p_2^2L(t) \Big\} \\ &= L(t) \times \Big\{ 1 - 2\eta\sigma^2_{\min}(X)c_1\alpha_0 + 2\eta^2\big(p_2\sqrt{2\kappa\sigma^6_{\min}(X)L(t)} + \frac{\kappa\sigma^4_{\min}(X)}{2}(c_2\beta_0)^2\big) \\ &\quad + 2\eta^3 c_2\beta_0 p_2\sqrt{2\kappa^3\sigma^{10}_{\min}(X)L(t)} + 2\eta^4 p_2^2\kappa^2\sigma^6_{\min}(X)L(t) \Big\} \\ &= L(t) \times [1 - a_1\eta + a_2(t)\eta^2 + a_3(t)\eta^3 + a_4(t)\eta^4] \end{aligned} \tag{82}$$

Finally, we show when $0 < \eta \leq \eta_{\max}, f(\eta, t) < 1$. Notice $f(\eta, t)$ is a decreasing functions in t, it suffices to show $f(\eta, 0) < 1$

$$f(\eta, 0) < 1 \iff a_4(0)\eta^3 + a_3(0)\eta^2 + a_2(0)\eta < a_1. \tag{83}$$

Compare the above inequality with equation 29, one has

$$\begin{aligned} a_4(0)\eta^3 + a_3(0)\eta^2 + a_2(0)\eta &< a_4(0)\eta^3 + a_3(0)\eta^2 + \big(a_2(0) + \frac{4c_2L(0)\sigma^2_{\max}(X)}{c_2 - 1}\big)\eta \\ a_4(0)\eta^3 + a_3(0)\eta^2 + a_2(0)\eta &< a_4(0)\eta^3 + a_3(0)\eta^2 + \big(a_2(0) + \frac{8c_2\beta_0L(0)\sigma^2_{\max}(X)}{(1 - c_1)\alpha_0}\big)\eta. \end{aligned} \tag{84}$$

Thus, when $0 < \eta \leq \eta_{\max}$, we have

$$\begin{aligned} a_4(0)\eta^3 + a_3(0)\eta^2 + a_2(0)\eta &< a_4(0)\eta^3 + a_3(0)\eta^2 + \big(a_2(0) + \frac{4c_2L(0)\sigma^2_{\max}(X)}{c_2 - 1}\big)\eta \leq a_1 \\ a_4(0)\eta^3 + a_3(0)\eta^2 + a_2(0)\eta &< a_4(0)\eta^3 + a_3(0)\eta^2 + \big(a_2(0) + \frac{8c_2\beta_0L(0)\sigma^2_{\max}(X)}{(1 - c_1)\alpha_0}\big)\eta \leq a_1. \end{aligned} \tag{85}$$

which is equivalent to $f(\eta, 0) < 1$. Thus, $A_1(t+1)$ is proved.

Then, **we prove** $A_2(t+1)$ **hold**. Since loss is decreasing, i.e. $L(t+1) \leq L(t) \leq L(0)$, we have

$$\|Y - XW(t+1)\|_F \leq \|E(0)\|_F. \tag{86}$$

equation 86 is equivalent to

$$\|Y\|_F - \|E(0)\|_F \leq \|XW(t+1)\|_F \leq \|Y\|_F + \|E(0)\|_F. \tag{87}$$

In [Min et al., 2022], Theorem 3, the lower bound is proved. For the upper bound,

$$\sigma_{\max}(W(t+1))\sigma_{\min}(X) \leq \|W(t+1)\|_F \sigma_{\min}(X) \leq \|XW(t+1)\|_F \leq \|Y\|_F + \|E(0)\|_F, \tag{88}$$

Thus,

$$\sigma_{\max}(W(t+1)) \leq \frac{\|Y\|_F + \|E(0)\|_F}{\sigma_{\min}(X)} =: p_2. \tag{89}$$

Then, **we prove** $A_3(t+1)$ **hold**.

$$\begin{aligned}
\|D(t+1) - D(0)\|_F &\leq \sum_{k=0}^{t} \|D(K+1) - D(k)\|_F \qquad \text{use Lemma C.3} \\
&\leq \sum_{k=0}^{t} 2\eta^2 \sigma_{\max}^2(X)\big(\sigma_{\max}^2(W_1(k)) + \sigma_{\max}^2(W_2(k))\big)L(k) \qquad \text{use } A_4(k) \\
&\leq 2\eta^2 \sigma_{\max}^2(X) c_2 \beta_0 \sum_{k=0}^{t} L(k) \qquad \text{use } A_1(k) \\
&\leq 2\eta^2 c_2 \beta_0 \sigma_{\max}^2(X) \sum_{k=0}^{t} (1 - a_1\eta + a_2(k)\eta^2 + a_3(k)\eta^3 + a_4(k)\eta^4)^k L(0) \qquad (90) \\
&\leq 2\eta^2 c_2 \beta_0 \sigma_{\max}^2(X) \sum_{k=0}^{t} (1 - a_1\eta + a_2(0)\eta^2 + a_3(0)\eta^3 + a_4(0)\eta^4)^k L(0) \\
&\leq \frac{2\eta^2 c_2 \beta_0 \sigma_{\max}^2(X) L(0)}{1 - f(\eta, 0)}. \qquad (91)
\end{aligned}$$

where we upper bound $a_i(k)$ by $a_i(0)$ in equation 90 for $i = 1, 2, 3, 4$.

Finally, **we prove** $A_4(t+1)$ **hold**. $\lambda_{\min}(\tau) \leq \lambda_{\max}(\tau)$ is obvious. We begin with the second inequality

$$\begin{aligned}
\lambda_{\min}(\tau) &= \min_{\|W\|_F=1} \langle W, WW_2^\top W_2 + W_1 W_1^\top W \rangle \qquad \text{definition of operator norm} \\
&\geq \min_{\|W\|_F=1} \langle W, WW_2^\top W_2 \rangle + \min_{\|W\|_F=1} \langle W, W_1 W_1^\top W \rangle \\
&= \sigma_{\min}^2(W_1) + \sigma_{\min}^2(W_2). \qquad (92)
\end{aligned}$$

The fourth inequality can be proved similarly

$$\begin{aligned}
\lambda_{\max}(\tau) &= \max_{\|W\|_F=1} \langle W, WW_2^\top W_2 + W_1 W_1^\top W \rangle \\
&\leq \max_{\|W\|_F=1} \langle W, WW_2^\top W_2 \rangle + \max_{\|W\|_F=1} \langle W, W_1 W_1^\top W \rangle \\
&= \sigma_{\max}^2(W_1) + \sigma_{\max}^2(W_2) \qquad (93)
\end{aligned}$$

Then, we prove the first inequality and last inequality holds. According to Lemma C.4, we have

$$\begin{aligned}
\sigma_{\min}^2(W_1(t+1)) &\geq \frac{-\bar{\lambda}_{\text{-}}(t+1) + \underline{\lambda}_{+}(t+1) + \sqrt{\big(\bar{\lambda}_{\text{-}}(t+1) + \underline{\lambda}_{+}(t+1)\big)^2 + 4\sigma_n^2\left(W(t+1)\right)}}{2}. \\
\sigma_{\min}^2(W_2(t+1)) &\geq \frac{-\bar{\lambda}_{+}(t+1) + \underline{\lambda}_{-}(t+1) + \sqrt{\big(\bar{\lambda}_{+}(t+1) + \underline{\lambda}_{-}(t+1)\big)^2 + 4\sigma_m^2\left(W(t+1)\right)}}{2}. \qquad (94)
\end{aligned}$$

where

$$\begin{aligned}\bar{\lambda}_+(t) &= \max\left(\lambda_1(D(t)), 0\right)\\ \underline{\lambda}_-(t) &= \max\left(\lambda_m(-D(t)), 0\right)\\ \bar{\lambda}_-(t) &= \max\left(\lambda_1(-D(t)), 0\right)\\ \underline{\lambda}_+(t) &= \max\left(\lambda_n(D(t)), 0\right)\end{aligned} \tag{95}$$

We define

$$\begin{aligned}h_1(\Delta_1, \Delta_2) &:= \frac{-\bar{\lambda}_-(0) + \Delta_1 + \underline{\lambda}_+(0) + \Delta_2 + \sqrt{\left(\bar{\lambda}_-(0) + \underline{\lambda}_+(0) + \Delta_1 + \Delta_2\right)^2 + 4p_1^2}}{2}\\ h_2(\Delta_3, \Delta_4) &= \frac{-\bar{\lambda}_+(0) + \Delta_1 + \underline{\lambda}_-(0) + \Delta_2 + \sqrt{\left(\bar{\lambda}_+(0) + \underline{\lambda}_-(0) + \Delta_1 + \Delta_2\right)^2 + 4p_1^2}}{2}\end{aligned} \tag{96}$$

where

$$\begin{aligned}\Delta_1 &= \bar{\lambda}_-(t+1) - \bar{\lambda}_{\text{-}}(0)\\ \Delta_2 &= \underline{\lambda}_+(t+1) - \underline{\lambda}_+(0)\\ \Delta_3 &= \bar{\lambda}_+(t+1) - \bar{\lambda}_+(0)\\ \Delta_4 &= \underline{\lambda}_-(t+1) - \underline{\lambda}_-(0).\end{aligned} \tag{97}$$

Then, we use $\sigma_{\min}(W(t+1)) \geq p_1$ to lower bound equation 94

$$\begin{aligned}\sigma_{\min}^2(W_1(t+1)) &\geq \frac{-\bar{\lambda}_{\text{-}}(t+1) + \underline{\lambda}_+(t+1) + \sqrt{\left(\bar{\lambda}_{\text{-}}(t+1) + \underline{\lambda}_+(t+1)\right)^2 + 4\sigma_n^2\left(W(t+1)\right)}}{2}.\\ &\geq \frac{-\bar{\lambda}_{\text{-}}(t+1) + \underline{\lambda}_+(t+1) + \sqrt{\left(\bar{\lambda}_{\text{-}}(t+1) + \underline{\lambda}_+(t+1)\right)^2 + 4p_1^2}}{2}.\\ &:= h_1(\Delta_1, \Delta_2).\end{aligned} \tag{98}$$

Similarly, we have

$$\sigma_{\min}^2(W_2(t+1)) \geq h_2(\Delta_3, \Delta_4). \tag{99}$$

Notice $h_1(0,0) + h_2(0,0) = \alpha_0$ which is independent of t. Our goal is to lower bound $h_1(\Delta_1, \Delta_2) + h_2(\Delta_3, \Delta_4)$ using $h_1(0,0) + h_2(0,0)$. A natural solution is that if we can quantify how large $|\Delta_k|, k = 1, 2, 3, 4$ is, i.e.$|\Delta_k| \leq \Delta_h$, and if we can show $h_1(\cdot,\cdot), h_2(\cdot,\cdot)$ are both L_h-Lipschitz continuous. Using these two ingredients, one can show

$$\begin{aligned}&|h_1(\Delta_1, \Delta_2) - h_1(0,0)| \leq L_h\sqrt{\Delta_1^2 + \Delta_2^2}\\ &\Rightarrow h_1(\Delta_1, \Delta_2) \geq h_1(0,0) - L_h\sqrt{\Delta_1^2 + \Delta_2^2} \geq h_1(0,0) - \sqrt{2}L_h\Delta_h.\end{aligned} \tag{100}$$

Similarly, we have

$$h_2(\Delta_3, \Delta_4) \geq h_2(0,0) - \sqrt{2}L_h\Delta_h. \tag{101}$$

Based on above two equations, one has

$$h_1(\Delta_1, \Delta_2) + h_2(\Delta_3, \Delta_4) \geq h_1(0,0) + h_2(0,0) - 2\sqrt{2}L_h\Delta_h. \tag{102}$$

Next, we show the above two assumptions hold

1. $h_1(\cdot,\cdot), h_2(\cdot,\cdot)$ are both L_h-Lipschitz continuous.
2. $|\Delta_k| \leq \Delta_h$ hold for all $k = 1, 2, 3, 4$.

For the first one, using Weyl's inequality and Property $A_3(t+1)$, we can upper bound $|\Delta_k|$

$$\begin{aligned}
|\Delta_1| &= \big|\max(\lambda_1(-D(t+1)),0) - \max(\lambda_1(-D(0)),0)\big| \\
&\le \big|\lambda_1(-D(t+1)) - \lambda_1(-D(0)),0)\big| \qquad \text{use Weyl's inequality} \\
&\le \|D(t+1) - D(0)\|_F \qquad \text{use Lemma C.3} \\
&\le \frac{2\eta^2 c_2\beta_0\sigma_{\max}^2(X)L(0)}{1-f(\eta,0)}.
\end{aligned} \tag{103}$$

Similarly, we have

$$|\Delta_2|, |\Delta_3|, |\Delta_4| \le \|D(t+1) - D(0)\|_F \le \frac{2\eta^2 c_2\beta_0\sigma_{\max}^2(X)L(0)}{1-f(\eta,0)}. \tag{104}$$

What's more,

$$\begin{aligned}
\left|\frac{dh_1(x,y)}{dx}\right| &= \left|-\frac{1}{2} + \frac{x+y+\bar{\lambda}_-(0)+\underline{\lambda}_+(0)}{2\sqrt{\big(\bar{\lambda}_-(0)+\underline{\lambda}_+(0)+x+y\big)^2+4p_1^2}}\right| \\
&\le \frac{1}{2} + \left|\frac{x+y+\bar{\lambda}_-(0)+\underline{\lambda}_+(0)}{2\sqrt{\big(\bar{\lambda}_-(0)+\underline{\lambda}_+(0)+x+y\big)^2+4p_1^2}}\right| \\
&\le \frac{1}{2} + \frac{1}{2} \\
&\le 1.
\end{aligned} \tag{105}$$

Similarly, we have $\big|\frac{dh_1(x,y)}{dy}\big|, \big|\frac{dh_2(x,y)}{dx}\big|, \big|\frac{dh_2(x,y)}{dy}\big| \le 1$. Combine with equation 105, we have $h_1(\cdot,\cdot), h_2(\cdot,\cdot)$ are $\sqrt{2}$-Lipschitz continuous. Thus, we have

$$\begin{aligned}
\sigma_{\min}^2(W_1(t)) + \sigma_{\min}^2(W_2(t)) &\ge h_1(\Delta_1,\Delta_2) + h_2(\Delta_3,\Delta_4) \\
&\ge \alpha_0 - 2L_h\sqrt{2}\|D(t+1) - D(0)\|_F \qquad L_h = \sqrt{2} \\
&\ge \alpha_0 - \frac{8\eta^2 c_2\beta_0\sigma_{\max}^2(X)L(0)}{1-f(\eta,0)}.
\end{aligned} \tag{106}$$

Although the above lower bound is smaller than α_0, it is close to α_0 when η is small. This motivates us to introduce $0 < c_1 < 1$ so that when η is small, the above inequality is lower bounded by $c_1\alpha_0$. To derive the upper bound on η, it is equivalent to ensure

$$\begin{aligned}
&\alpha_0 - \frac{8\eta^2 c_2\beta_0\sigma_{\max}^2(X)L(0)}{1-f(\eta,0)} \ge c_1\alpha_0 \\
\iff\ &(1-c_1)\alpha_0 \ge \frac{8\eta^2 c_2\beta_0\sigma_{\max}^2(X)L(0)}{1-f(\eta,0)} \\
\iff\ &(1-c_1)\alpha_0 \ge \frac{8\eta c_2\beta_0\sigma_{\max}^2(X)L(0)}{a_1 - a_2(0)\eta - a_3(0)\eta^2 - a_4(0)\eta^3} \\
\iff\ &a_4(0)\eta^3 + a_3(0)\eta^2 + \big(a_2(0) + \frac{8c_2\beta_0 L(0)\sigma_{\max}^2(X)}{(1-c_1)\alpha_0}\big)\eta \le a_1
\end{aligned} \tag{107}$$

which is ensured when $0 < \eta < \eta_{\max}$.

The proof for the fourth inequality $\sigma_{\max}^2(W_1(t+1)) + \sigma_{\max}^2(W_2(t+1)) \le c_2\beta_0$ in $A_4(t+1)$ is similar. According to Lemma C.5, we have

$$\begin{aligned}
&\sigma_{\max}^2(W_1(t+1)) + \sigma_{\max}^2(W_2(t+1)) \\
&\le \frac{\max(\lambda_{\max}(D(0)),0) + \Delta_3 + \sqrt{4\sigma_{\max}^2(W(t+1)) + [\max(\lambda_{\max}(D(0)),0) + \Delta_3]^2}}{2} \\
&\quad + \frac{\max(\lambda_{\max}(-D(0)),0) + \Delta_4 + \sqrt{4\sigma_{\max}^2(W(t+1)) + [\max(\lambda_{\max}(-D(0)),0) + \Delta_4]^2}}{2}
\end{aligned}$$

$$\leq \frac{\max(\lambda_{\max}(D(0)),0)+\Delta_3+\sqrt{4p_2^2+[\max(\lambda_{\max}(D(0)),0)+\Delta_3]^2}}{2}$$
$$+\frac{\max(\lambda_{\max}(-D(0)),0)+\Delta_4+\sqrt{4p_2^2+[\max(\lambda_{\max}(-D(0)),0)+\Delta_4]^2}}{2}$$
$$:= h_3(\Delta_5,\Delta_6), \tag{108}$$

where

$$\begin{aligned}\Delta_5 &= \max(\lambda_{\max}(D(t+1)),0)-\max(\lambda_{\max}(D(0)),0)\\ \Delta_6 &= \max(\lambda_{\max}(-D(t+1)),0)-\max(\lambda_{\max}(-D(0)),0).\end{aligned} \tag{109}$$

Since

$$\left|\frac{dh_3(x,y)}{dx}\right| = \left|\frac{1}{2}+\frac{x+\max(\lambda_{\max}(D(t+1)),0)}{2\sqrt{4p_2^2+[\max(\lambda_{\max}(D(t+1)),0)+\Delta_5]^2}}\right| \leq 1. \tag{110}$$

Similarly, $|\frac{dh_3(x,y)}{dy}| \leq 1$. What's more, Weyl's inequality gives us

$$\begin{aligned}|\Delta_5| &= |\max(\lambda_{\max}(D(t+1)),0)-\max(\lambda_{\max}(D(0)),0)|\\ &\leq |\lambda_{\max}(D(t+1))-\lambda_{\max}(D(0))|\\ &\leq \|D(t+1)-D(0)\|_F\end{aligned} \tag{111}$$

Similarly, we have $|\Delta_6| \leq \|D(t+1)-D(0)\|_F$. Thus, we have

$$\begin{aligned}\sigma^2_{\max}(W_1(t+1))+\sigma^2_{\max}(W_2(t+1)) &= h_3(\Delta_5,\Delta_6)\\ &\leq h_3(0,0)+\sqrt{2}\sqrt{\Delta_5^2+\Delta_6^2}\\ &\leq \beta_0+\frac{4\eta^2c_2\beta_0\sigma^2_{\max}(X)L(0)}{1-f(\eta,0)}\\ &\leq \beta_0 c_2\end{aligned} \tag{112}$$

where the last inequality holds if and only if

$$a_4\eta^3+a_3\eta^2+\big(a_2+\frac{4c_2L(0)\sigma^2_{\max}(X)}{c_2-1}\big)\eta \leq a_1. \tag{113}$$

□

D PROOF OF PROPOSITION 3.1

Proposition 3.1. *If $\alpha_0 > 0$, for all $0 < \eta \leq \eta_{\max}$ and for all $t = 0, 1, \cdots$, the following inequality holds*

$$f(\eta,t) \geq 1-\frac{1}{\kappa} \tag{114}$$

where $\kappa = \frac{K}{\mu}$ is the condition number of the non-overparametrized Problem 1

Proof. The theoretical optimal convergence rate for non-overparametrized regime is $1-\frac{1}{\kappa}$. Then

$$\begin{aligned}f(\eta,t)-(1-\frac{1}{\kappa}) &= \frac{1}{\kappa}-a_1\eta+a_2(t)\eta^2+a_3(t)\eta^3+a_4(t)\eta^4 \qquad \text{drop last two terms which are non-negative}\\ &\geq \frac{1}{\kappa}-2c_1\alpha_0\sigma^2_{\min}(X)\eta+\big(2\sqrt{2\kappa L(t)\sigma^6_{\min}(X)}p_2+\kappa\mu^2c_2^2\beta_0^2\big)\eta^2\\ &\geq \frac{1}{\kappa}-2c_1\alpha_0\sigma^2_{\min}(X)\eta+\kappa\sigma^4_{\min}(X)c_2^2\beta_0^2\eta^2 \qquad \text{use } \beta_0\geq\alpha_0 \text{ to lower bound last term}\\ &\geq \frac{1}{\kappa}-2c_1\alpha_0\sigma^2_{\min}(X)\eta+\kappa\sigma^4_{\min}(X)c_2^2\alpha_0^2\eta^2\\ &= (\frac{1}{\sqrt{\kappa}}-\sqrt{\kappa}\sigma^2_{\min}(X)c_2\alpha_0\eta)^2\\ &\geq 0.\end{aligned} \tag{115}$$

Thus, the results are proved. □

E PROOF OF CLAIM 3.1

Claim 3.1. *Suppose $\alpha_0 > 0$. Let η_t' be the unique positive root of the following equation*

$$-a_1 + 2a_2(t)\eta + 3a_3(t)\eta^2 + 4a_4(t)\eta^3 = 0. \tag{116}$$

Then the solution to Problem 35 is $\eta_t = \min(\eta_t', \eta_{\max})$.

Proof. We first observe the derivative of $f(\eta, t)$ with respect to η is monotonically increasing when $\eta > 0$

$$\frac{df(\eta,t)}{d\eta} = -a_1 + 2a_2(t)\eta + 3a_3(t)\eta^2 + 4a_4(t)\eta^3, \tag{117}$$

and $\frac{d^2 f(\eta,t)}{d\eta^2} > 0$. Thus, if $\eta_t' \le \eta_{\max}$, the minimizer of Problem 35 is $\eta_{\max}$. If $\eta_t' \ge \eta_{\max}$, since $\frac{df(\eta,t)}{d\eta}$ is negative when $0 < \eta \le \eta_{\max} \le \eta_t'$, $f(\eta, t)$ is decreasing in the same range. Thus, the minimizer is $\eta_{\max}$. Combing the above two cases, the minimizer of Problem 35 is

$$\eta_t = \min(\eta_t', \eta_{\max}). \tag{118}$$

□

Claim E.1. *Given some $0 < c_1 < \frac{2}{3}$, pick any*

$$c_2 \ge \max\left\{ \frac{M + \frac{16L(0)}{\beta_0}}{c_1\alpha_0\sigma_{\min}^2(X)}, \sqrt{\frac{M + \frac{8\alpha_0 L(0)}{\beta_0^2}}{\alpha_0\sigma_{\min}^2(X)}}, 2 \right\}, \tag{119}$$

where $M = \frac{2\alpha_0^3 p_2^2 L(0)}{\beta_0^6\kappa} + \frac{2\sqrt{2\sigma_{\min}^2(X)L(0)}p_2\alpha_0^2}{\sqrt{\kappa}\beta_0^3} + \frac{2\sqrt{2L(0)\sigma_{\min}^2(X)}p_2\alpha_0}{\beta_0^2\sqrt{\kappa}}$.

Such choice of c_1, c_2 ensures $\eta_{\max} \ge \eta_t'$ for all $t = 0, 1, 2, \cdots$.

Remark E.1. *Claim E.1 implies for proper choice of c_1, c_2, one has $\eta_{\max} \ge \eta_t'$ for all $t = 0, 1, 2, \cdots$. In the limiting case when $t \to \infty$, one has*

$$\lim_{t\to\infty} f(\eta, t) = 1 - 2(c_1\alpha_0)\sigma_{\min}^2(X)\eta + \kappa\sigma_{\min}^4(X)(c_2\beta_0)^2\eta^2 \tag{120}$$

With the choice of c_1, c_2 specified, we have $\eta_\infty' \le \eta_{\max}$. Thus, the asymptotic convergence rate is

$$f(\eta_\infty', \infty) = 1 - \frac{(c_1\alpha_0)^2}{(c_2\beta_0)^2}\frac{1}{\kappa} \tag{121}$$

The asymptotic convergence rate is determined by $\frac{c_1\alpha_0}{c_2\beta_0}$ and condition number κ. The smaller κ is, the faster convergence rate is. What's more, since $\frac{\lambda_{\min}(\tau_t)}{\lambda_{\max}(\tau_t)} \ge \frac{c_1\alpha_0}{c_2\beta_0}$, we can view $\frac{c_1\alpha_0}{c_2\beta_0}$ as a lower bound on the condition number of the operator τ_t. The more ill-conditioned τ_t is, i.e. $\frac{c_1\alpha_0}{c_2\beta_0}$ is small, the slower the convergence rate is.

Proof. Notice $a_2(t), a_3(t), a_4(t)$ depends on $L(t)$ and $L(t)$ decreases as t increases, so $a_2(t), a_3(t), a_4(t)$ decrease as t increase. From equation 117, we can see η_t' increases as t increases. Thus, to prove $\eta_t' \le \eta_{\max}$, it suffices to show

$$\lim_{t\to\infty} \eta_t' = \frac{c_1\alpha_0}{c_2^2\beta_0^2\kappa\sigma_{\min}^2(X)} \le \eta_{\max}. \tag{122}$$

which is equivalent to the following inequalities

$$a_4(0)\left(\frac{c_1\alpha_0}{c_2^2\beta_0^2\kappa\sigma_{\min}^2(X)}\right)^3 + a_3(0)\left(\frac{c_1\alpha_0}{c_2^2\beta_0^2\kappa\sigma_{\min}^2(X)}\right)^2 + \left(a_2(0) + \frac{4c_2L(0)\sigma_{\max}^2(X)}{c_2 - 1}\right)\frac{c_1\alpha_0}{c_2^2\beta_0^2\kappa\sigma_{\min}^2(X)} \le a_1, \tag{123}$$

$$a_4(0)\left(\frac{c_1\alpha_0}{c_2^2\beta_0^2\kappa\sigma_{\min}^2(X)}\right)^3 + a_3(0)\left(\frac{c_1\alpha_0}{c_2^2\beta_0^2\kappa\sigma_{\min}^2(X)}\right)^2 + \left(a_2(0) + \frac{8c_2\beta_0L(0)\sigma_{\max}^2(X)}{(1 - c_1)\alpha_0}\right)\frac{c_1\alpha_0}{c_2^2\beta_0^2\kappa\sigma_{\min}^2(X)} \le a_1. \tag{124}$$

For equation 123 to hold, we study its LHS

$$\begin{aligned}\text{LHS of equation 123} &= \frac{2c_1^3\alpha_0^3p_2^2L(0)}{c_2^6\beta_0^6\kappa} + \frac{2\sqrt{2L(0)\sigma_{\min}^2(X)}p_2c_1^2\alpha_0^2}{\sqrt{\kappa}c_2^3\beta_0^3} \\ &\quad + \frac{4c_1\alpha_0L(0)}{(c_2-1)c_2\beta_0^2} + \frac{2\sqrt{2L(0)\sigma_{\min}^2(X)}p_2c_1\alpha_0}{c_2^2\beta_0^2\sqrt{\kappa}} + c_1\alpha_0\sigma_{\min}^2(X) \\ &= \frac{2c_1^3\alpha_0^3p_2^2L(0)}{c_2^6\beta_0^6\kappa} + \frac{Pc_1^2\alpha_0}{c_2^3\beta_0} + \frac{4c_1\alpha_0L(0)}{(c_2-1)c_2\beta_0^2} + \frac{Pc_1}{c_2^2} + c_1\alpha_0\sigma_{\min}^2(X)\end{aligned}$$

where

$$P = \frac{2\sqrt{2L(0)\sigma_{\min}^2(X)}p_2\alpha_0}{\sqrt{\kappa}\beta_0^2}. \tag{125}$$

Since $c_2 \geq 2$, so $c_2 - 1 \geq \frac{c_2}{2}$. Then, we upper bound the above equality by substituting higher order terms of $c_1^k, k \geq 2$ with c_1 in the numerator by one except for the last term and replace higher order terms of $c_2^k, k \geq 3$ with c_2^2,

$$\begin{aligned}\text{LHS of equation 123} &= \frac{2c_1^3\alpha_0^3p_2^2L(0)}{c_2^6\beta_0^6\kappa} + \frac{Pc_1^2\alpha_0}{c_2^3\beta_0} + \frac{4c_1\alpha_0L(0)}{(c_2-1)c_2\beta_0^2} + \frac{Pc_1}{c_2^2} + c_1\alpha_0\sigma_{\min}^2(X) \\ &\leq \frac{2c_1^3\alpha_0^3p_2^2L(0)}{c_2^6\beta_0^6\kappa} + \frac{Pc_1^2\alpha_0}{c_2^3\beta_0} + \frac{8c_1\alpha_0L(0)}{c_2^2\beta_0^2} + \frac{Pc_1}{c_2^2} + c_1\alpha_0\sigma_{\min}^2(X) \qquad \text{use } c_2 - 1 \geq \frac{c_2}{2} \text{ in the first term} \\ &\leq \frac{2c_1\alpha_0^3p_2^2L(0)}{c_2^2\beta_0^6\kappa} + \frac{Pc_1\alpha_0}{c_2^2\beta_0} + \frac{8c_1\alpha_0L(0)}{c_2^2\beta_0^2} + \frac{Pc_1}{c_2^2} + c_1\alpha_0\sigma_{\min}^2(X) \\ &\qquad \text{use } c_1 \geq c_1^k, k \geq 2 \text{ in the numerator and } c_2^2 \leq c_2^k, k \geq 3 \text{ in denominator} \\ &= \frac{c_1}{c_2^2}\left[\frac{2\alpha_0^3p_2^2L(0)}{\beta_0^6\kappa} + \frac{P\alpha_0}{\beta_0} + \frac{8\alpha_0L(0)}{\beta_0^2} + P\right] + c_1\alpha_0\sigma_{\min}^2(X) \\ &= \frac{c_1}{c_2^2}(M + \frac{8\alpha_0L(0)}{\beta_0^2}) + c_1\alpha_0\sigma_{\min}^2(X) \qquad \text{use second condition in equation 119} \\ &\leq c_1\alpha_0\sigma_{\min}^2(X) + c_1\alpha_0\sigma_{\min}^2(X) = a_1.\end{aligned} \tag{126}$$

For equation 124 to hold, we study its LHS

$$\text{LHS of equation 124} = \frac{2c_1^3\alpha_0^3p_2^2L(0)}{c_2^6\beta_0^6\kappa} + \frac{Pc_1^2\alpha_0}{c_2^3\beta_0} + \frac{8L(0)c_1}{(1-c_1)c_2\beta_0} + \frac{Pc_1}{c_2^2} + c_1\alpha_0\sigma_{\min}^2(X). \tag{127}$$

Since $0 < c_1 < \frac{2}{3}$, we have $1 - c_1 \geq \frac{c_1}{2}$. Then, we upper bound the above equality by substituting c_1 with 1 in the numerator by one except for the last term and replace higher order terms of $c_2^k, k \geq 2$ with c_2,

$$\begin{aligned}\text{LHS of equation 124} &= \frac{2c_1^3\alpha_0^3p_2^2L(0)}{c_2^6\beta_0^6\kappa} + \frac{Pc_1^2\alpha_0}{c_2^3\beta_0} + \frac{8L(0)c_1}{(1-c_1)c_2\beta_0} + \frac{Pc_1}{c_2^2} + c_1\alpha_0\sigma_{\min}^2(X) \\ &\leq \frac{2c_1^3\alpha_0^3p_2^2L(0)}{c_2^6\beta_0^6\kappa} + \frac{Pc_1^2\alpha_0}{c_2^3\beta_0} + \frac{16L(0)}{c_2\beta_0} + \frac{Pc_1}{c_2^2} + c_1\alpha_0\sigma_{\min}^2(X) \qquad \text{use } 1 - c_1 \geq \frac{c_1}{2} \\ &\leq \frac{2\alpha_0^3p_2^2L(0)}{c_2\beta_0^6\kappa} + \frac{P\alpha_0}{c_2\beta_0} + \frac{16L(0)}{c_2\beta_0} + \frac{P}{c_2} + c_1\alpha_0\sigma_{\min}^2(X) \\ &\qquad \text{use } c_1 \leq 1 \text{ in the numerator and } c_2 \geq c_2^k, k \geq 1 \text{ in the numerator} \\ &= \frac{1}{c_2}\left[\frac{2\alpha_0^3p_2^2L(0)}{\beta_0^6\kappa} + \frac{P\alpha_0}{\beta_0} + \frac{16L(0)}{\beta_0} + P\right] + c_1\alpha_0\sigma_{\min}^2 \\ &= \frac{1}{c_2}[M + \frac{16L(0)}{\beta_0}] + c_1\alpha_0\sigma_{\min}^2 \qquad \text{use first condition in equation 119} \\ &\leq c_1\alpha_0\sigma_{\min}^2(X) + c_1\alpha_0\sigma_{\min}^2(X) = a_1.\end{aligned} \tag{128}$$

□

F SIMULATIONS

In Section 4.2, we compare the step sizes proposed in [Arora et al., 2018, Du et al., 2018a], Theorem 3.2 and Algorithm 1. In [Du et al., 2018a], they choose an adaptive step size

$$\eta_t = \frac{\sqrt{\epsilon/r}}{100(t+1)\|Y\|_F^{\frac{3}{2}}}, \tag{129}$$

where $0 < \epsilon < \|Y\|_F$ is the final precision we want to achieve, r is the rank of Y. When comparing, we set $\epsilon = \|Y\|_F$ to select the largest step size possible in their work.

In [Arora et al., 2018], they choose constant step size which satisfies

$$\eta \leq \frac{p_1^3}{6144 \times 2^3 \times \|Y\|_F^4}, \tag{130}$$

When comparing, we select the largest step size possible, i.e. $\eta = \frac{p_1^3}{6144 \times 2^3 \times \|Y\|_F^4}$.

In [Arora et al., 2018, Du et al., 2018a], the authors make assumptions that there is sufficient margin and zero imbalance at initialization. What's more, they both choose the setting of matrix factorization and claim it's equivalent to linear networks. To make fair comparison, we generate X using identity matrix. For initialization of the network, we follow Proposition F.1 in [Arora et al., 2018] to create a balanced initialization. The magnitude 0.05 of noise added to Y is a hyperparameter which ensures there is sufficient margin at initialization. The procedure to ensure there is zero imbalance at initialization is given below

Proposition F.1 (Balanced Initialization)**.** *Given $d_0, d_1, \ldots, d_N \in \mathbb{N}$ such that $\min\{d_1, \ldots, d_{N-1}\} \geq \min\{d_0, d_N\}$ and a distribution $\mathcal{D}$ over $d_N \times d_0$ matrices, a* balanced initialization *of $W_j \in \mathbb{R}^{d_j \times d_{j-1}}$, j=1, . . . , N, assigns these weights as follows:*

1. *Sample $A \in \mathbb{R}^{d_N \times d_0}$ according to $\mathcal{D}$.*
2. *Take singular value decomposition $A = U\Sigma V^\top$, where $U \in \mathbb{R}^{d_N \times \min\{d_0, d_N\}}$, $V \in \mathbb{R}^{d_0 \times \min\{d_0, d_N\}}$ have orthonormal columns, and $\Sigma \in \mathbb{R}^{\min\{d_0, d_N\} \times \min\{d_0, d_N\}}$ is diagonal and holds the singular values of A.*
3. *Set $W_N \simeq U\Sigma^{1/N}, W_{N-1} \simeq \Sigma^{1/N}, \ldots, W_2 \simeq \Sigma^{1/N}, W_1 \simeq \Sigma^{1/N} V^\top$, where the symbol "$\simeq$" stands for equality up to zero-valued padding.*