LARGE LEARNING RATE TAMES HOMOGENEITY: CONVERGENCE AND BALANCING EFFECT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent empirical advances show that training deep models with large learning rate often improves generalization performance. However, theoretical justifications on the benefits of large learning rate are highly limited, due to challenges in analysis. In this paper, we consider using Gradient Descent (GD) with a large learning rate on a homogeneous matrix factorization problem, i.e., $\min_{X,Y} \|A - XY^\top\|_{\mathsf{F}}^2$. We prove a convergence theory for constant large learning rates well beyond $2/L$, where L is the largest eigenvalue of Hessian at the initialization. Moreover, we rigorously establish an implicit bias of GD induced by such a large learning rate, termed 'balancing', meaning that magnitudes of X and Y at the limit of GD iterations will be close even if their initialization is significantly unbalanced. Numerical experiments are provided to support our theory.

1 INTRODUCTION

Training machine learning models such as deep neural networks involves optimizing highly nonconvex functions. Empirical results indicate an intimate connection between training algorithms and the performance of trained models (Le et al., 2011; Bottou et al., 2018; Zhang et al., 2021; Soydaner, 2020; Zhou et al., 2020). Especially for widely used first-order training algorithms (e.g., GD and SGD), the learning rate is of essential importance and has received extensive focus from researchers (Smith, 2017; Jastrzebski et al., 2017; Smith, 2018; Gotmare et al., 2018; Liu et al., 2019; Li & Arora, 2019). A recent perspective is that large learning rates often lead to improved testing performance compared to the counterpart trained with small learning rates (Smith & Topin, 2019; Yue et al., 2020). Towards explaining the better performance, a common belief is that large learning rates encourage the algorithm to search for flat minima, which often generalize better and are more robust than sharp ones (Seong et al., 2018; Lewkowycz et al., 2020).

Despite abundant empirical observations, theoretical understandings of the benefits of large learning rate are still limited for non-convex functions, partly due to challenges in analysis. For example, the convergence (of GD or SGD) under large learning rate is not guaranteed. Even for globally smooth functions, very few general results exist if the learning rate exceeds certain threshold (Kong & Tao, 2020). Besides, popular regularity assumptions such as global smoothness for simplified analyses are often absent in homogeneous models, including commonly used ReLU neural networks.

This paper theoretically studies the benefits of large learning rate in a matrix factorization problem

$$\min_{X,Y} \frac{1}{2} \left\|A - XY^\top\right\|_{\mathsf{F}}^2, \quad \text{where } A \in \mathbb{R}^{n\times n},\ X, Y \in \mathbb{R}^{n\times d}. \tag{1}$$

We consider Gradient Descent (GD) for solving (1): at the k-th iteration, we have

$$X_{k+1} = X_k + h(A - X_k Y_k^\top)Y_k \quad \text{and} \quad Y_{k+1} = Y_k + h(A^\top - Y_k X_k^\top)X_k,$$

where h is the learning rate. Despite its simple formula, problem (1) serves as an important foundation of a variety of problems, including matrix sensing (Chen & Wainwright, 2015; Bhojanapalli et al., 2016; Tu et al., 2016), matrix completion (Keshavan et al., 2010; Hardt, 2014), and linear neural networks (Ji & Telgarsky, 2018; Gunasekar et al., 2018).

Problem (1) possesses several intriguing properties. Firstly, the objective function is non-convex, and critical points are either global minima or saddles (see e.g., Baldi & Hornik (1989); Li et al. (2019b);

Valavi et al. (2020a)). Secondly, problem (1) is homogeneous in X and Y, meaning that rescaling X, Y to $aX, a^{-1}Y$ for any $a \neq 0$ will not change the objective's value. This property is shared by commonly used ReLU neural networks. A direct consequence of homogeneity is that global minima of (1) are non-isolated and can be unbounded. The curvatures at these global minima are highly dependent on the magnitudes of X, Y. When X, Y have comparable magnitudes, the largest eigenvalue of Hessian is small, and this corresponds to a flat minimum; on the contrary, unbalanced X and Y give a sharp minimum. Last but not the least, the homogeneity impairs smoothness conditions of (1), rendering the gradient being not Lipschitz continuous unless X, Y are bounded. See a formal discussion in Section 2.

Existing approaches for solving (1) often uses explicit regularization (Ge et al., 2017; Tu et al., 2016; Cabral et al., 2013; Li et al., 2019a), or infinitesimal (or diminishing) learning rates for controlling the magnitudes of X, Y (Du et al., 2018; Ye & Du, 2021). In this paper, we go beyond the scope of aforementioned works, and analyze GD with a large learning rate for solving (1). In particular, we allow the learning rate h to be as large as approximately $4/L$, where L denotes the largest eigenvalue of Hessian at GD initialization. In connection to empirical observations, we provide positive answers to the following two questions:

Does GD with large learning rate converge at least for some cases of (1)*?*
Does larger learning rate biases toward flatter minima (i.e., X, Y with comparable magnitudes)?

We theoretically show the convergence of GD with large learning rate for the two situations $n = 1, d \in \mathbb{N}^+$ or $d = 1, n \in \mathbb{N}^+$ with isotropic A. We also observe a, perhaps surprising, "balancing effect" for general matrix factorization (i.e., any d, n, and A), meaning that when h is sufficiently large, the difference between X and Y shrinks significantly at the convergence of GD compared to its initial, even if the initial point is close to an unbalanced global minimum. In fact, with a proper large learning rate h, $\|X_k - Y_k\|_{\mathsf{F}}^2$ may decrease by an arbitrary factor at its limit. The following is a simple example of our theory for $n = 1$ (i.e. scalar factorization), and more general results will be presented later with a precise bound for h depending on the initial condition and A.

Theorem 1.1 (Informal version of Thm.3.1 & 3.2)**.** *Given scalar A and initial condition $X_0, Y_0 \in \mathbb{R}^{1\times d}$ chosen almost everywhere, with learning rate $h \lesssim 4/L$, GD converges to a global minimum (X, Y) satisfying $\|X\|_{\mathsf{F}}^2 + \|Y\|_{\mathsf{F}}^2 \leq \frac{2}{h}$. Consequently, its extent of balancing is quantified by $\|X - Y\|_{\mathsf{F}}^2 \leq \frac{2}{h} - 2A$.*

We remark that having a learning rate $h \approx 4/L$ is far beyond the commonly analyzed regime in optimization. Even for globally L-smooth objective, traditional theory requires $h < 2/L$ for GD convergence and $h = 1/L$ is optimal for convex functions (Boyd et al., 2004), not to mention that our problem (1) is never globally L-smooth due to homogeneity. Modified equation provides a tool for probing intermediate learning rates (see Hairer et al. (2006, Chapter 9) for a general review, and Kong & Tao (2020, Appendix A) for the specific setup of GD), but the learning rate here is too large for modified equation to work (see Appendix C). In fact, besides blowing up, GD with large learning rate may have a zoology of limiting behaviors (see e.g., Appendix B for convergence to periodic orbits under our setup, and Kong & Tao (2020) for convergence to chaotic attractors).

Our analyses (of convergence and balancing) leverage various mathematical tools, including a proper partition of state space and its dynamical transition (specifically invented for this problem), stability theory of discrete time dynamical systems (Alligood et al., 1996), and geometric measure theory (Federer, 2014).

The rest of the paper is organized as: Section 2 provides the background of studying (1) and discusses related works; Section 3 presents convergence and balancing results for scalar factorization problems; Section 4 generalizes the theory to rank-1 matrix approximation; Section 5 studies problem (1) with arbitrary A and its arbitrary-rank approximation; Section 6 summarizes the paper and discusses broadly related topics and future directions.

2 Background and Related Work

Notations. $\|v\|$ is ℓ_2 norm of a column or row vector v. $\|M\|_{\mathsf{F}}$ is the Frobenius norm of a matrix M.

Sharp and flat minima in (1) We discuss the curvatures at global minima of (1). To ease the presentation, consider simplified versions of (1) with either $n = 1$ or $d = 1$. In this case, X and Y become vectors and we denote them as x, y, respectively. We show the following proposition characterizing the spectrum of Hessian at a global minimum.

Proposition 2.1. *When $n = 1, d \in \mathbb{N}^+$ or $d = 1, n \in \mathbb{N}^+$ in (1), the largest eigenvalue of Hessian at a global minimum (x, y) is $\|x\|^2 + \|y\|^2$ and the smallest eigenvalue is 0.*

Homogenity implies global minimizers of (1) are not isolated, which is consistent with the 0 eigenvalue. On the other hand, if the largest eigenvalue $\|x\|^2 + \|y\|^2$ is large (or small), then the curvature at such a global minimum is sharp (or flat), in the direction of the leading eigenvector. Meanwhile, note this sharpness/flatness is an indication of the balancedness between magnitudes of x, y at a global minimum. To see this, singular value decomposition (SVD) yields that at a global minimum, (x, y) satisfies $\left\|xy^\top\right\|_{\mathsf{F}}^2 = \sigma_{\max}^2(A)$. Therefore, large $\|x\|^2 + \|y\|^2$ is obtained when $|\|x\| - \|y\||$ is large, i.e., x and y magnitudes are unbalanced, and small $\|x\|^2 + \|y\|^2$ is obtained when balanced.

Large learning rate We study smoothness properties of (1) and demonstrate that our learning rate is well beyond conventional optimization theory. We first define the smoothness of a function.

Definition 2.2 (L-smooth)**.** A function $f \in \mathcal{C}^1$ defined on $\mathbb{R}^N$ is *L-smooth* if for all $u_1, u_2 \in \mathbb{R}^N$,

$$\|\nabla f(u_1) - \nabla f(u_2)\| \leq L\|u_1 - u_2\|. \tag{2}$$

If further $f \in \mathcal{C}^2$, then $\nabla^2 f \preceq LI$. Moreover, if we have (2) for $u_1, u_2 \in \mathcal{X} \subseteq \mathbb{R}^N$, we call it locally L-smooth.

In traditional optimization (Nesterov, 2003; Polyak, 1987; Nesterov, 1983; Polyak, 1964; Beck & Teboulle, 2009), most analyzed objective functions often satisfy (i) (some relaxed form of) convexity or strong convexity, and (ii) L-smoothness. Choosing a step size $h < 2/L$ guarantees the convergence of GD to a minimum by the existing theory (reviewed in Appendix F.3). Our choice of learning rate $h \approx 4/L$ (more precisely, $4/L_0$; see below) goes beyond the classical analyses.

Besides, in our problem (1), the regularity is very different. Even simplified versions of (1), i.e., with either $n = 1$ or $d = 1$, suffer from (i) non-convexity and (ii) unbounded eigenvalues of Hessian, i.e., no global smoothness (see Appendix F.2 for more details). As shown in Du et al. (2018); Ye & Du (2021); Ma et al. (2021), decaying or infinitesimal learning rate ensures that the GD trajectory stays in a locally smooth region. However, the gap between the magnitudes of X, Y can only be maintained in that case. We show, however, that larger learning rate can shrink this gap.

Related work Matrix factorization problems in various forms have been extensively studied in the literature. The version of (1) is commonly known as the low-rank factorization, although here d can arbitrary and we consider both $d \leq \text{rank}(A)$ and $d > \text{rank}(A)$ cases. Baldi & Hornik (1989); Li et al. (2019b); Valavi et al. (2020b) provide landscape analysis of (1). Ge et al. (2017); Tu et al. (2016) propose to penalize the Frobenius norm of $\left\|X^\top X - Y^\top Y\right\|_{\mathsf{F}}^2$ to mitigate the homogeneity and establish global convergence guarantees of GD for solving (1). Cabral et al. (2013); Li et al. (2019a) instead penalize individual Frobenius norms of $\|X\|_{\mathsf{F}}^2 + \|Y\|_{\mathsf{F}}^2$. We remark that penalizing $\|X\|_{\mathsf{F}}^2 + \|Y\|_{\mathsf{F}}^2$ is closely related to nuclear norm regularization, since the variational formula for nuclear norm $\|Z\|_*^2 = \min_{Z=XY^\top} \|X\|_{\mathsf{F}}^2 + \|Y\|_{\mathsf{F}}^2$. More recently, Liu et al. (2021) consider using injected noise as regularization to GD and establish a global convergence. More specifically, by perturbing the GD iterate (X_k, Y_k) with Gaussian noise, GD will converge to a flat global optimum. On the other hand, Du et al. (2018); Ye & Du (2021); Ma et al. (2021) show that even without explicit regularization, when the learning rate of GD is infinitesimal, i.e., GD approximating gradient flow, X and Y maintain the gap in their magnitudes. Such an effect is more broadly recognized as implicit bias of learning algorithms (Neyshabur et al., 2014; Gunasekar et al., 2018; Soudry et al., 2018). Built upon this implicit bias, Du et al. (2018) further prove that GD with diminishing learning rates converges to a bounded global minimum of (1), and this conclusion is recently extended to the case of a constant small learning rate (Ye & Du, 2021). Our work goes beyond the scopes of these milestones and considers matrix factorization with much larger learning rates.

Additional results exist that demonstrate large learning rate can improve performance in various learning problems. Most of them involve non-constant learn rates. Specifically, Li et al. (2019c)

consider a two-layer neural network setting, where using learning rate annealing (initially large, followed by small ones) can improve classification accuracy compared to training with small learning rates. Nakkiran (2020) shows that the observation in Li et al. (2019c) even exists in convex problems. Lewkowycz et al. (2020) study constant large learning rates, and demonstrate distinct algorithmic behaviors of large and small learning rates, as well as empirically illustrate large learning rate yields better testing performance on neural networks. Their analysis is built upon the neural tangent kernel perspective (Jacot et al., 2018), with a focus on the kernel spectrum evolution under large learning rate. Worth noting is, Kong & Tao (2020) also study constant large learning rates, and show that large learning rate provides a mechanism for GD to escape local minima, alternative to noisy escapes due to stochastic gradients.

3 Overparameterized scalar factorization

In order to provide intuition before directly studying the most general problem, we begin with a simple special case, namely factorizing a scalar by two vectors. It corresponds to (1) with $n = 1$ and $d \in \mathbb{N}^+$, and this overparameterized problem is written as

$$\min_{x,y\in\mathbb{R}^{1\times d}} \frac{1}{2}(\mu - xy^\top)^2, \tag{3}$$

where μ is assumed without loss of generality to be a positive scalar. Problem (3) can be viewed as univariate regression using a linear two-layer neural network with the quadratic loss, which is studied in Lewkowycz et al. (2020) with atomic data distribution. Yet our analysis in the sequel can be used to study arbitrary univariate data distributions; see details in Section 6.

Although simplified, problem (3) is still nonconvex and exhibits the same homogeneity as (1). The convergence of its large learning rate GD optimization was previously not understood, let alone balancing. Many results that we will obtain for this problem will remain true for more general problems.

We first prove that GD converges despite of $> 2/L$ learning rate and for almost all initial conditions:

Theorem 3.1 (Convergence)**.** *Given $(x_0^\top, y_0^\top) \in (\mathbb{R}^d \times \mathbb{R}^d)\backslash\mathcal{B}$ where $\mathcal{B}$ is some Lebesgue measure-0 set, when the learning rate h satisfies*

$$h \le \min\left\{\frac{4}{\|x_0\|^2 + \|y_0\|^2 + 4\mu}, \frac{1}{3\mu}\right\},$$

GD converges to a global minimum.

Theorem 3.1 says that choosing a constant learning rate depending on GD initialization guarantees the convergence for almost every starting point in the whole space. This result is even stronger than the already nontrivial convergence under small learning rate with high probability over random initialization (Ye & Du, 2021). Furthermore, the upper bound on h is sufficiently large: on the one hand, suppose GD initialization (x_0, y_0) is close to an unbalanced global minimum. By Proposition 2.1, we can check that the largest eigenvalue $L(x_0, y_0)$ of Hessian $\nabla^2 f(x_0, y_0)$ is approximately $\|x_0\|^2 + \|y_0\|^2$. Consequently, our upper bound of h is almost $4/L$, which is beyond $2/L$ (see Section 2 for more details). On the other hand, we observe numerically that the $4/L$ upper bound is actually very close to the stability limit of GD when initialized away from the origin (see more details in Appendix E).

The convergence in Theorem 3.1 has an interesting searching-to-converging transition as depicted in Figure 1, where we observe two phases. In Phase 1, large learning rate drives GD to search for flat regions, escaping from the attraction of sharp minima. After some iterations, the algorithm enters the vicinity of a global minimum with more balanced magnitudes in x, y. Then in Phase 2, GD converges to the found balanced global minimum. We remark that the searching-to-converging transition also appears in Lewkowycz et al. (2020). However, the algorithmic behaviors are not the same. In fact, in our searching phase (phase 1), the objective function does not exhibit the blow-up phenomenon. In our convergence phase (phase 2), the analysis relies on a detailed state space partition (see Line 192) due to nonconvex nature of (3), while the analysis in Lewkowycz et al. (2020) is akin to monotone convergence in a convex problem.

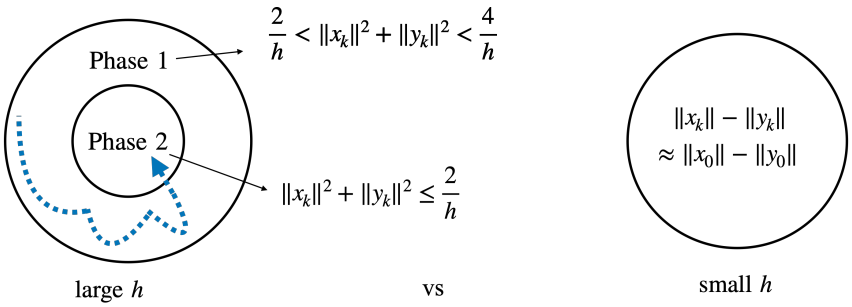

Figure 1: The dynamics of GD under different learning rate h

In comparison with the dynamics of small learning rate, we note that the searching phase (Phase 1) is vital to the convergence analysis. Meanwhile, the searching phase induces a balancing effect of large learning rate. The following theorem explicitly quantifies the extent of balancing.

Theorem 3.2 (Balancing)**.** *Under the same initial condition and learning rate h as Theorem 3.1, GD for (3) converges to a global minimizer (x, y) satisfying*

$$\|x\|^2 + \|y\|^2 \le \frac{2}{h}.$$

Consequently, its extent of balancing is quantified by

$$\|x - y\|^2 \le \frac{2}{h} - 2\mu.$$

One special case of Theorem 3.2 is the following theorem, which states that no matter how close to a global minimum does GD start, if this minimum does not correspond to well-balanced norms, a large learning rate will take the iteration to a more balanced limit. We also demonstrate a sharp shrinkage in the distance between x and y.

Corollary 3.3 (From 'unbalanced' to 'balanced')**.** *For any $\delta \in (0, \mu)$, let the GD initialization satisfy*

$$(x_0, y_0) \in \left\{(u, v) : |uv^\top - \mu| < \delta,\ \|u\|^2 + \|v\|^2 > 8\mu\right\}.$$

When the learning rate h satisfies $h = \frac{4}{\|x_0\|^2 + \|y_0\|^2 + 4\mu}$, the extent of balancing at the limiting point (x, y) of GD obeys

$$\|x - y\|^2 < \frac{1}{2}\|x_0 - y_0\|^2 + 2\mu.$$

Both Theorem 3.2 and Corollary 3.3 suggest that larger learning rate yields better balancing effect, as $\|x - y\|^2$ at the limit of GD may decrease a lot and is controlled by the learning rate. We remark that the balancing effect is a consequence of large learning rate, as small learning rate can only maintain the difference in magnitudes of x, y (Du et al., 2018).

In addition, the actual balancing effect can be quite strong with $\|x_k - y_k\|^2$ decreasing to be almost 0 at its limit under a proper choice of large learning rate. Figure 2 illustrates an almost perfect balancing case when $h = \frac{4}{\|x_0\|^2 + \|y_0\|^2 + 4} \approx 0.0122$ is chosen as the upper bound. The difference $\|x_k - y_k\|$ decreases from approximately 17.9986 to 0.0154 at its limit. Additional experiments with various learning rates and initializations can be found in Appendix A.

Technical Overview We sketch the main ideas behind Theorem 3.1, which lead to the balancing effect in Theorem 3.2. Full proof is deferred to Appendix G.

The convergence is proved by handling Phase 1 and 2 separately (see Fig.1). In Phase 1, we prove that $\|x_k\|^2 + \|y_k\|^2$ has a decreasing trend as GD searches for flat minimum. We show that $\|x_k\|^2 + \|y_k\|^2$ may not be monotone, i.e., it either decreases every iteration or decreases every other iteration.

In Phase 2, we carefully partition the state space and show GD at each partition will eventually enter a monotone convergence region. Note that the partition is based on detailed understanding of the dynamics and is highly nontrivial. Attentive readers may refer to Appendix G for more details. The combination of Phase 1 & 2 is briefly summarized as a proof flow chart in Figure 3.

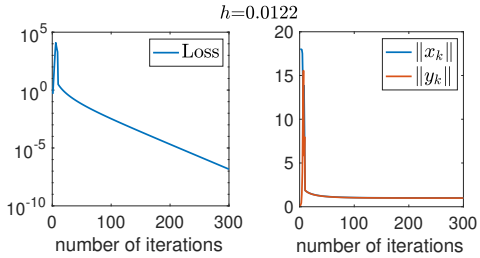

Figure 2: The objective function is $(1 - xy^\top)^2/2$, where $x^\top, y^\top \in \mathbb{R}^{10}$. Highly unbalanced initial condition is uniformly randomized, with the norms to be $\|x_0\| = 18$, $\|y_0\| = 0.09$.

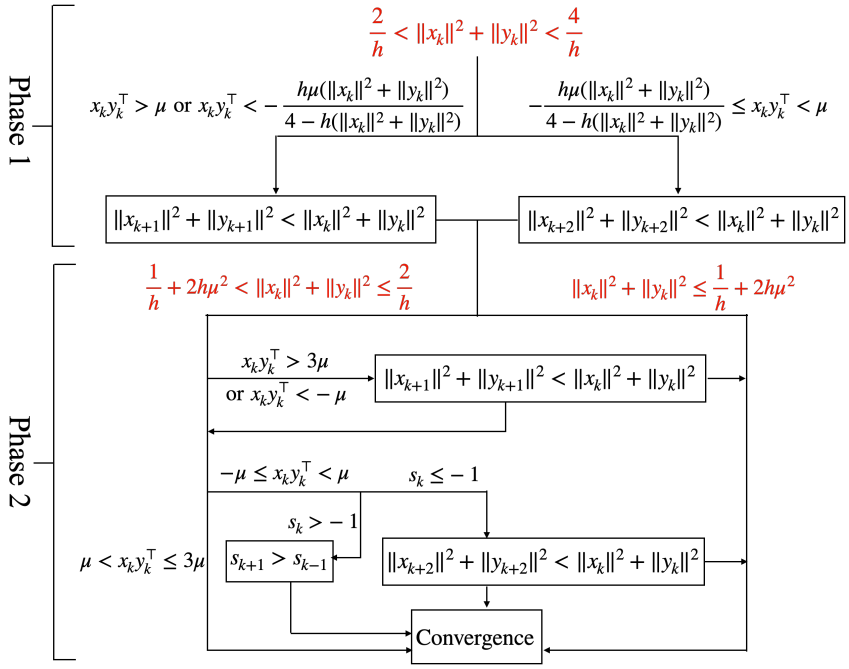

Figure 3: Proof overview of Theorem 3.1. At the k-th iteration, we denote (x_k, y_k) as the iterate and s_k is defined as $x_{k+1}y_{k+1}^\top - \mu = s_k(x_k y_k^\top - \mu)$.

4 Rank-1 approximation of isotropic A (an under-parameterized case)

Given the insights from scalar factorization, we consider rank-1 factorization of an isotropic matrix A, i.e., $A = \mu I_{n\times n}$ with $\mu > 0$, $d = 1$, and $n \in \mathbb{N}^+$. The corresponding optimization problem is written as

$$\min_{x,y\in\mathbb{R}^{n\times 1}} \frac{1}{2}\left\|\mu I_{n\times n} - xy^\top\right\|_{\mathsf{F}}^2. \tag{4}$$

Although similar at an uncareful glance, Problems (4) and (3) are rather different unless $n = d = 1$. First of all, Problem (4) is under-parameterized for $n > 1$, while (3) is overparameterized. More importantly, we'll show that, when (x, y) is a global minimum of (4), x, y must be aligned, i.e., $x = \ell y$ for some $\ell > 0$. In the scalar factorization problem, however, no such alignment is required. As a result, the set of global minima of (4) is an n-dimensional submanifold embedded in a $2n$-dimensional space, while in the scalar factorization problem the set of global minimum is a $(2d-1)$-dimensional submanifold — one rank deficient — in a $2d$-dimensional space. We expect the convergence in (4) is more complicated than that in (3), since searching in (4) is demanding.

To prove the convergence of large learning rate GD for (4), our theory consists of two steps: (i) show the convergence of the alignment between x and y (this is new); (ii) use that to prove the convergence of the full iterates (i.e., x & y). Step (i) first:

Theorem 4.1 (Alignment)**.** *Given $(x_0, y_0) \in (\mathbb{R}^n \times \mathbb{R}^n) \backslash \mathcal{B}$, where $\mathcal{B}$ is some Lebesgue measure-0 set, when learning rate $h \leq \min\left\{\frac{4}{\|x_0\|^2+\|y_0\|^2+4\sqrt{7}\mu}, \frac{1}{2\sqrt{7}\mu}\right\}$, the iterator (x_k, y_k) of GD at the k-th iteration satisfies $|\cos(\angle(x_k, y_k))| \to 1$ as $k \to \infty$.*

Proof sketch. A sufficient condition for the convergence of $|\cos(\angle(x_k, y_k))|$ is $\|x_k\|^2 \|y_k\|^2 - (x_k^\top y_k)^2 \to 0$. To ease the presentation, let $U_k = x_k^\top y_k$, $V_k = x_k^\top x_k$, and $W_k = y_k^\top y_k$. By the GD update and some algebraic manipulation, we derive

$$V_{k+1}W_{k+1} - U_{k+1}^2 = r_k \cdot (V_k W_k - U_k^2),$$

where $r_k = (1 - h(V_k + W_k) + h^2(V_k W_k - \mu^2))^2$. When k is sufficiently large, we can show a uniform upper bound on $r_k < 1 - c$ for some constant $c > 0$. In this way, we deduce that $V_k W_k - U_k^2$ will exponentially decay and converge to 0. More details are provided in Appendix H. □

Theorem 4.1 indicates that GD iterations will converge to the neighbourhood of $\{(x, y) : x = \ell y, \text{for some } \ell \in \mathbb{R}\backslash\{0\}\}$. This helps establish the global convergence as stated in Step (ii).

Theorem 4.2 (Convergence)**.** *Under the same initial conditions and learning rate h as Theorem 4.1, GD for (4) converges to a global minimum.*

Similar to the over-parametrized scalar case, this convergence can also be split into two phases where phase 1 motivates the balancing behavior with the decrease of $\|x_k\|^2 + \|y_k\|^2$, and phase 2 ensures the convergence. The following balancing theorem is thus obtained.

Theorem 4.3 (Balancing)**.** *Under the same initial conditions and learning rate h as Theorem 4.1, GD for (4) converges to a global minimizer that obeys*

$$\|x\|^2 + \|y\|^2 \leq \frac{2}{h},$$

and its extent of balancing is quantified by

$$\|x - y\|^2 \leq \frac{2}{h} - 2\mu.$$

This conclusion is the same as the one in Section 3. A quantitatively similar corollary like Corollary 3.3 can also be obtained from the above theorem, namely, if x_0 and y_0 start from an unbalanced point near a minimum, the limit will be a more balanced one.

5 General matrix factorization

In this section, we consider problem (1) with an arbitrary matrix $A \in \mathbb{R}^{n\times n}$. We replicate the problem formulation here for convenience,

$$\min_{X,Y\in\mathbb{R}^{n\times d}} \frac{1}{2}\left\|A - XY^\top\right\|_{\mathsf{F}}^2. \tag{5}$$

Note this is the most general case with $n, d \in \mathbb{N}^+$ and any square matrix A. Due to this generalization, we no longer utilize the convergence analysis and instead, establish the balancing theory via stability analysis of GD as a discrete time dynamical system.

Let $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq 0$ be the singular values of A. Assume for technical convenience $\|A\|_{\mathsf{F}}^2 = \sum_{i=1}^n \mu_i^2$ being independent of d and n. We denote the singular value decomposition of A as $A = UDV^\top$, where $U, D, V \in \mathbb{R}^{n\times n}$, U, V are orthogonal matrices and D is diagonal. Then we establish the following balancing effect.

Theorem 5.1. *Given almost all the initial conditions, for any learning rate h such that GD for (5) converges to a point (X, Y), there exists $c = c(n, d) > c_0$ with constant $c_0 > 0$ independent of h, n, and d, such that (X, Y) satisfies*

$$c(\|X\|_{\mathsf{F}}^2 + \|Y\|_{\mathsf{F}}^2) < \frac{2}{h},$$

and the extent of balancing is quantified by $\left\|X - (UV^\top)Y\right\|_{\mathsf{F}}^2 < \frac{2}{ch} - 2\sum_{i=1}^{\min\{d,n\}} \mu_i$, which means

$$\left| \|X\|_{\mathsf{F}} - \|Y\|_{\mathsf{F}} \right|^2 < \frac{2}{ch} - 2\sqrt{\sum_{i=1}^{\min\{d,n\}} \mu_i^2}.$$

In particular, when $d = 1$, i.e., rank-1 factorization of an arbitrary A, the constant c equals 1.

We observe that the extent of balancing can be quantified under some rotation of Y. This is necessary, since for factorizing a general matrix A (which can be asymmetric), at a global minimum, X, Y may only align after a rotation (which is however fixed by A, independent of initial or final conditions). Figure 4 illustrates an example of the balancing effect under different learning rates. Evidently, larger learning rate leads to a more balanced global minimizer. Additional experiments with various dimensions, learning rates, and initializations can be found in Appendix A; a similar balancing effect is also shown for additional problems including matrix sensing and matrix completion there.

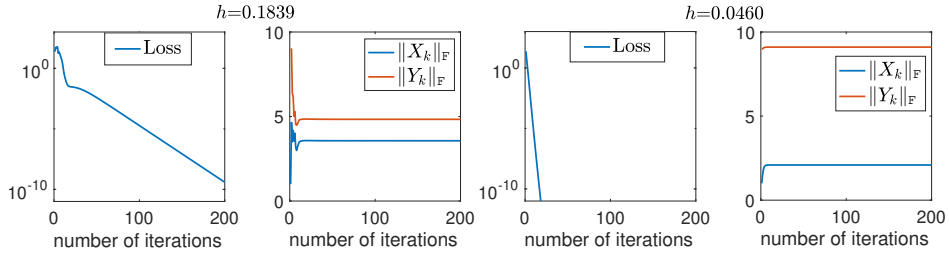

Figure 4: Balancing effect of general matrix factorization. We independently generate elements in $A \in \mathbb{R}^{6\times 6}$ from a Gaussian distribution. We choose $X, Y \in \mathbb{R}^{6\times 100}$ and randomly pick a pair of initial point (X_0, Y_0) with $\|X_0\|_{\mathsf{F}} = 1$ and $\|Y_0\|_{\mathsf{F}} = 9$.

Different from previous sections, Theorem 5.1 builds on stability analysis by viewing GD as a dynamical system in discrete time. More precisely, the proof of Theorem 5.1 (see Appendix I for details) consists of two parts: (i) the establishment of an easier but equivalent problem via the rotation of X and Y, (ii) stability analysis of the equivalent problem.

For (i), by singular value decomposition (SVD), $A = UDV^\top$, where $U, D, V \in \mathbb{R}^{n\times n}$, U and V are orthogonal matrices, and D is a non-negative diagonal matrix. Let $X_k = UR_k,\ Y_k = VS_k$. Then

$$\begin{cases} X_{k+1} = X_k + h(A - X_kY_k^\top)Y_k \\ Y_{k+1} = Y_k + h(A - X_kY_k^\top)^\top X_k \end{cases}$$

$$\Leftrightarrow \begin{cases} UR_{k+1} = UR_k + h(UDV^\top - UR_kS_k^\top V^\top)VS_k \\ VS_{k+1} = VS_k + h(UDV^\top - UR_kS_k^\top V^\top)^\top UR_k \end{cases}$$

$$\Leftrightarrow \begin{cases} R_{k+1} = R_k + h(D - R_kS_k^\top)S_k \\ S_{k+1} = S_k + h(D - R_kS_k^\top)^\top R_k \end{cases}.$$

Therefore, GD for problem (5) is equivalent to GD for the following problem

$$\min_{R,S\in\mathbb{R}^{n\times d}} \frac{1}{2}\left\|D - RS^\top\right\|_{\mathsf{F}}^2$$

and it thus suffices to work with diagonal non-negative A.

For (ii), here is a brief description of the idea of stability analysis: consider each iteration of GD as a mapping ψ from $u_k \sim (X_k, Y_K)$ to $u_{k+1} \sim (X_{k+1}, Y_{k+1})$, where matrices X and Y are flattened

and concatenated into a vector so that ψ is a closed map on vector space $\mathbb{R}^{2dn}$. GD iteration is thus a discrete time dynamical system on state space $\mathbb{R}^{2dn}$ given by

$$u_{k+1} = \psi(u_k) = u_k - h\nabla f(u_k),$$

where f is the objective function $f(u_k) = \frac{1}{2}\left\|A - X_k Y_k^\top\right\|_{\mathsf{F}}^2$, and gradient returns a vector that collects all component-wise partial derivatives.

It's easy to see that any stationary point of f, denoted by u^*, is a fixed point of ψ, i.e., $u^* = \psi(u^*)$. We are interested in what fixed point will the iterations of ψ converge to. For this, the following notions are helpful:

Proposition 5.2. *Consider a fixed point u^* of ψ. If all the eigenvalues of Jacobian matrix $\nabla\psi(u^*)$ are of complex modulus less than 1, it is a stable fixed point.*

Proposition 5.3. *Consider a fixed point u^* of ψ. If at least one eigenvalue of Jacobian matrix $\nabla\psi(u^*)$ is of complex modulus greater than 1, it is an unstable fixed point.*

Roughly put, the stable set of an unstable fixed point is of negligible size when compared to that of a stable fixed point, and thus what GD converges to is a stable fixed point for almost all initial conditions (Alligood et al., 1996). Thus, we investigate the stability of each global minimum of f (each saddle of f is an unstable fixed point of ψ and thus is irrelevant). By a detailed evaluation of $\nabla\psi$'s eigenvalues (Appendix I), we see that a global minimum (X, Y) of f corresponds to a stable fixed point of GD iteration if $\left|1 - ch\left(\|X\|_{\mathsf{F}}^2 + \|Y\|_{\mathsf{F}}^2\right)\right| < 1$, i.e., it is balanced as in Thm. 5.1.

6 Conclusion and Discussion

In this paper, we demonstrate an implicit regularization effect of large learning rate on the homogeneous matrix factorization problem solved by GD. More precisely, a phenomenon termed as "balancing" is theoretically illustrated, which says the difference between the two factors X and Y may decrease significantly at the limit of GD, and the extent of balancing can increase as learning rate increases. In addition, we provide theoretical analysis of the convergence of GD to the global minimum, and this is with large learning rate that can exceed the typical limit of $2/L$, where L is the largest eigenvalue of Hessian at GD initialization.

For the matrix factorization problem analyzed here, large learning rate avoids bad regularities induced by the homogeneity between X and Y. We feel it is possible that such balancing behavior can also be seen in problems with similar homogeneous properties, for example, in tensor decomposition (Kolda & Bader, 2009), matrix completion (Keshavan et al., 2010; Hardt, 2014), generalized phase retrieval (Candes et al., 2015; Sun et al., 2018), and neural networks with homogeneous activation functions (e.g., ReLU). Besides the balancing effect, the convergence analysis under large learning rate may be transplanted to other non-convex problems and help discover more implicit regularization effects.

In addition, factorization problems studied here are closely related to two-layer linear neural networks. For example, one-dimensional regression via a two-layer linear neural network can be formulated as the scalar factorization problem (3): Suppose we have a collection of data points $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$ for $i = 1, \ldots, n$. We aim to train a linear neural network $y = (u^\top v)x$ with $u, v \in \mathbb{R}^d$ for fitting the data. We optimize u, v by minimizing the quadratic loss,

$$(u^*, v^*) \in \arg\min_{u,v} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - (u^\top v)x_i\right)^2 = \arg\min_{u,v} \frac{1}{n}\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} - u^\top v\right)^2. \tag{6}$$

As can be seen, taking $\mu = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$ recovers (3). In this regard, our theory indicates that training of u, v by GD with large learning rate automatically balances u, v, and the obtained minimum is flat. This may provide some initial understanding of the improved performance brought by large learning rates in practice. Note that (6) generalizes to arbitrary data distribution of training a two-layer linear network with atomic data (i.e., $x = 1$ and $y = 0$) in Lewkowycz et al. (2020).

It is important to clarify, however, that there is a substantial gap between this demonstration and extensions to general neural networks, including deep linear and nonlinear networks. Although we suspect that large learning rate leads to similar balancing effect of weight matrices in the network, rigorous theoretical analysis is left as a future direction.

REFERENCES

Kathleen T Alligood, Tim D Sauer, and James A Yorke. *Chaos: an introduction to dynamical systems*. Springer, 1996.

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

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Srinadh Bhojanapalli, Anastasios Kyrillidis, and Sujay Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pp. 530–582. PMLR, 2016.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Ricardo Cabral, Fernando De la Torre, João P Costeira, and Alexandre Bernardino. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2488–2495, 2013.

Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Herbert Federer. *Geometric measure theory*. Springer, 2014.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pp. 1233–1242. PMLR, 2017.

Akhilesh Gotmare, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*, 2018.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10. IEEE, 2018.

Ernst Hairer, Marlis Hochbruck, Arieh Iserles, and Christian Lubich. Geometric numerical integration. *Oberwolfach Reports*, 3(1):805–882, 2006.

Moritz Hardt. Understanding alternating minimization for matrix completion. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 651–660. IEEE, 2014.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE transactions on information theory*, 56(6):2980–2998, 2010.

Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3): 455–500, 2009.

Lingkai Kong and Molei Tao. Stochasticity of deterministic gradient descent: Large learning rate for multiscale objective function. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2625–2638. Curran Associates, Inc., 2020.

Quoc V Le, Jiquan Ngiam, Adam Coates, Ahbik Lahiri, Bobby Prochnow, and Andrew Y Ng. On optimization methods for deep learning. In *ICML*, 2011.

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

Qiuwei Li, Zhihui Zhu, and Gongguo Tang. The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*, 8(1):51–96, 2019a.

Xingguo Li, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, Zhaoran Wang, and Tuo Zhao. Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization. *IEEE Transactions on Information Theory*, 65(6):3489–3514, 2019b.

Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *arXiv preprint arXiv:1907.04595*, 2019c.

Zhiyuan Li and Sanjeev Arora. An exponential learning rate schedule for deep learning. *arXiv preprint arXiv:1910.07454*, 2019.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

Tianyi Liu, Yan Li, Song Wei, Enlu Zhou, and Tuo Zhao. Noisy gradient descent converges to flat minima for nonconvex matrix factorization. 2021.

Cong Ma, Yuanxin Li, and Yuejie Chi. Beyond procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing. *IEEE Transactions on Signal Processing*, 69:867–877, 2021.

Preetum Nakkiran. Learning rate annealing can provably help generalization, even for convex problems. 2020.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pp. 543–547, 1983.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

Boris T Polyak. Introduction to optimization. optimization software. *Inc., Publications Division, New York*, 1, 1987.

Sihyeon Seong, Yegang Lee, Youngwook Kee, Dongyoon Han, and Junmo Kim. Towards flatter loss surface via nonmonotonic learning rate scheduling. In *UAI*, pp. 1020–1030, 2018.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pp. 464–472. IEEE, 2017.

Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pp. 1100612. International Society for Optics and Photonics, 2019.

Daniel Soudry, Elad Hoffer, 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.

Derya Soydaner. A comparison of optimization algorithms for deep learning. *International Journal of Pattern Recognition and Artificial Intelligence*, 34(13):2052013, 2020.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pp. 964–973. PMLR, 2016.

Hossein Valavi, Sulin Liu, and Peter Ramadge. Revisiting the landscape of matrix factorization. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1629–1638. PMLR, 26–28 Aug 2020a.

Hossein Valavi, Sulin Liu, and Peter J Ramadge. The landscape of matrix factorization revisited. *arXiv preprint arXiv:2002.12795*, 2020b.

Tian Ye and Simon S Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. *arXiv preprint arXiv:2106.14289*, 2021.

Xubo Yue, Maher Nouiehed, and Raed Al Kontar. Salr: Sharpness-aware learning rates for improved generalization. *arXiv preprint arXiv:2011.05348*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Hoi, et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *arXiv preprint arXiv:2010.05627*, 2020.