From Gradient Flow on Population Loss to Learning with Stochastic Gradient Descent

Anonymous Author(s)
Affiliation
Address
email

Abstract

Stochastic Gradient Descent (SGD) has been the method of choice for learning large-scale non-convex models. While a general analysis of when SGD works has been elusive, there has been a lot of recent progress in understanding the convergence of Gradient Flow (GF) on the population loss, partly due to the simplicity that a continuous-time analysis buys us. An overarching theme of our paper is providing general conditions under which SGD converges, assuming that GF on the population loss converges. Our main tool to establish this connection is a general *converse Lyapunov* like theorem, which implies the existence of a Lyapunov potential under mild assumptions on the rates of convergence of GF. In fact, using these potentials, we show a one-to-one correspondence between rates of convergence of GF and geometrical properties of the underlying objective. When these potentials further satisfy certain self-bounding properties, we show that they can be used to provide a convergence guarantee for Gradient Descent (GD) and SGD (even when the paths of GF and GD/SGD are quite far apart). It turns out that these self-bounding assumptions are in a sense also necessary for GD/SGD to work. Using our framework, we provide a unified analysis for GD/SGD not only for classical settings like convex losses, or objectives that satisfy PL/ KL properties, but also for more complex problems including Phase Retrieval and Matrix sq-root, and extending the results in the recent work of Chatterjee [8].

1 Introduction

Stochastic Gradient Descent (SGD) has been a method of choice to train complex, large scale machine learning models. While understanding of SGD for convex objectives is comprehensive, a general understanding of when SGD works for non-convex models has been somewhat elusive. A large slew of properties like, convexity [36], one-point-convexity [27], linearizability [24], KŁ [3, 31] and PŁ [25, 37, 32] properties, and more problem specific, tailored analysis of SGD and Gradient Descent (GD) for specific problem instances like matrix square-root problem, matrix completion [23], phase retrieval [7, 10, 41] and Dictionary learning [2] have been proposed. Recent success of SGD in over-parameterized deep learning models have lead to the idea that SGD perhaps optimizes training objective with an implicit bias given by some implicit regularizer [16, 38, 21, 17, 18]. However, in [24] it is argued that there are over-parameterized models for which SGD works but no method that minimizes an implicit regularized training objective can learn successfully, thus showing that in general, the success of SGD cannot be explained by implicit regularization.

The goal of our paper is to provide a unifying analysis for when SGD/GD works. More specifically, we do this via first showing that Gradient Flow (GF) works and then extending this analysis to SGD and GD. Gradient Flow (GF) can be seen as a continuous time analogue of GD. In an idealized world, if one had access to the population loss, it turns out that convergence analysis for running

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

gradient flow on population loss is somewhat simpler due to tools from continuous time analysis and PDEs. There has been several recent works [4, 9, 11] that have provided convergence analysis for GF even on non-convex objectives. The high level theme of this paper is to show that, under some mild/appropriate assumptions of population loss/objective and on the noise of gradient estimates, "if, GF converges on population loss, then SGD that uses one fresh example per iteration is successful at learning". Notice, that GF converging on population loss is a purely deterministic optimization problem. However, the fact that SGD works is a learning result that implies a sample complexity bound.

There have been past works that have aimed at providing convergence analysis for Gradient Descent (GD) starting from Gradient Flow (GF). Typical route to obtain a convergence analysis of GD starting from GF tries to think of GD updates as approximating GF path. Even with more sophisticated discretization schemes like Euler discretization, obtaining convergence for GD, starting from GF can be quite complex. In this paper, to show that when GF converges, SGD/GD also converges, we take a different approach. A key tool for proving convergence results for GF is by constructing so called Lyapunov potentials. In the literature of Ordinary Differential Equations (ODEs), when ODEs have regular enough convergence rates, one can show, so called converse Lyapunov theorems (see [26] for a nice survey of classic results) that state that when an ODE converges to stable solutions, there has to exist a corresponding Lyapunov potential. While convergence of GF in terms of sub-optimality is quite different from convergence in the ODE sense, in this paper, we first prove a converse Lyapunov style theorem for GF. Specifically, we show that when GF converges in terms of sub-optimality to a global minimum, then there has to exist a corresponding Lyapunov potential and using such potential, the rates can be recovered. This result becomes a starting point for our analysis. We show that if this Lyapunov potential (obtained from the converse Lyapunov style theorem) satisfies certain extra self-bounding regularity conditions, then one can show that GD and SGD algorithms converge in terms of sub-optimality when appropriate step sizes are used. Such convergence for SGD/GD happens even when the GF path and GD/SGD paths can be quite different.

We now summarize our main contributions below:

- We prove a converse Lyapunov style theorem that shows that if gradient flow converges with rate specified by with an appropriate rate function, then there exists a corresponding Lyapunov potential that recovers this rate.
- We provide a geometric characterization for a given rate of convergence of gradient flow (ie. GF converges at a particular rate if and only if a specific geometric condition on objective holds.)
- There are problems for which GF converges at a specific rate but GD can be arbitrarily slow to converge.
- This motivates the necessity of additional conditions to ensure GD/SGD converges even when GF converges. We provide certain self-bounding regularity conditions on the Lyapunov potential, under which we show that GD converges. We also provide conditions on gradient estimate noise under which we show that SGD using these gradient estimates also converges.
- We instantiate our results for problems such as matrix square-root and phase retrieval, amongst other applications.

2 Setup

Given a continuously differentiable function and non-negative function $F : \mathbb{R}^d \mapsto \mathbb{R}$, our goal is to minimize $F(w)$. Without any loss of generality, we assume that $\min_w F(w) = 0$. First-order algorithms are popular for such optimization tasks. In the following, we formally describe the Gradient Descent and Stochastic Gradient Descent algorithm, and their continuous time counterpart called gradient flow.

Gradient Descent (GD). Gradient descent is the most popular iterative algorithm to minimize differentiable functions. Starting from an initial point w_0, GD on the function $F(w)$ performs the following update on every iteration:

$$w_{t+1} \leftarrow w_t - \eta \nabla F(w_t), \tag{1}$$

where η denotes the step size. After T rounds, GD algorithm returns the point $\widehat{w}_T := \operatorname{argmin}_{s \le T} F(w_s)$.

Stochastic Gradient Descent (SGD). Stochastic gradient descent (SGD) has been the method of choice for optimizing complex convex and non-convex learning problems in practice. In the learning setting, $F(w)$ corresponds to the unknown population loss and can be written as $F(w) = \mathbb{E}_{z \sim \mathcal{D}}[f(w; z)]$ where the expectation is taken with respect to samples z drawn from an unknown distribution $\mathcal{D}$. SGD algorithm (mini-batch size 1) is an iterative algorithm that at every round $t \ge 0$, draws a fresh sample z_t from $\mathcal{D}$ to compute a stochastic unbiased estimate $\nabla f(w_i; z_i)$ of the gradient $\nabla F(w_i)$, and performs the update

$$w_{t+1} \leftarrow w_t - \eta \nabla f(w_t; z_t) \tag{2}$$

where η is the step size and w_0 denotes the initial point. After T rounds, SGD algorithm returns $\widehat{w}_T$ by sampling a point uniformly at random from the set $\{w_1, \ldots, w_T\}$.

Gradient Flow (GF). Gradient flow from a point w_0 is continuous time process $(w(t))_{t \ge 0}$ that starts at $w(0) = w_0$ and evolves as

$$\frac{\mathrm{d}w(t)}{\mathrm{d}t} = -\nabla F(w(t)). \tag{3}$$

GF has been thought of as a continuous time analogue of GD and is popularly used to understand behavior of gradient based optimization algorithms in the limit, primarily due to its simplicity and lack of step size.

3 Gradient Flow, Potentials and Geometry

Lyapunov potentials are a popular tool for understanding convergence of GF [43, 45, 29]. At an intuitive level, a Lyapunov potential is any non-negative function Φ that satisfies $\langle \nabla \Phi(w), -\nabla F(w) \rangle \le 0$, i.e. Φ decreases along the GF paths of F. This monotonicity property helps to show asymptotic convergence of GF to stable points of the underlying objective. In our work, we consider potential functions for which the rate of change (decrease) of the potential along the GF path is related to the suboptimality of the objective at that point.

Definition 1 (Admissible potentials)**.** *A differentiable potential function $\Phi_g : \mathbb{R}^d \mapsto \mathbb{R}^+$ is admissible w.r.t. F on a set $\mathcal{W}$ if there exists a monotonically increasing function $g : \mathbb{R}^+ \mapsto \mathbb{R}^+$ with $g(0) = 0$ such that for any $w \in \mathcal{W}$,*[1]

$$\langle \nabla \Phi_g(w), \nabla F(w) \rangle \ge g(F(w)). \tag{4}$$

Existence of potential functions of the above form can be used to provide rates of convergence for GF as show in the following theorem.

Theorem 1 (From potentials to gradient flow)**.** *Let $\mathcal{W}$ be a set of initial points that we want to consider, and let Φ_g be an admissible potential w.r.t. F on the set $\operatorname{clo}(\mathcal{W})$. Then, for any initialization $w_0 \in \mathcal{W}$, the point $w(t)$ on the GF path with $w(0) = w_0$ satisfies for any $t \ge 0$,*

$$g(F(w(t))) \le \frac{\Phi_g(w_0)}{t}.$$

The idea that admissible potential functions imply convergence rates for GF has appeared in various forms in the prior literature [5, 28, 43, 45]. As an example, consider the potential function $\Phi(w) = \|w - w^*\|/2$. Notice that Φ is an admissible potential for any F that is convex with $g(z) = z$. This is because convexity implies that (4) is true for any w. Hence, using Theorem 1 we get a $\|w - w^*\|^2/2t$ rate of convergence for GF on any convex objective.

Our main result in this section is to establish a converse Lyapunov style theorem−that given a rate, finds a potential function corresponding to that rate. We start by defining admissible rate functions.

Definition 2 (Admissible rate functions)**.** *A function $R : \mathbb{R}^d \times \mathbb{R} \mapsto \mathbb{R}^+$ is an admissible rate function w.r.t. F if for any $w \in \mathbb{R}^d$,*

[1] Whenever not specified, we assume that the function $g(z) = z$.

(a) *$R(w,t)$ is a non-increasing function of t such that* $\lim_{t\to\infty} R(w,t) = 0$.

(b) *R satisfies the relation:* $\int_{t=0}^{\infty}\left(\frac{\partial R(w,t)}{\partial t} + \langle \nabla R(w,t), \nabla F(w)\rangle\right) \mathrm{d}t \geq 0$.

Remark 1. *In order to simplify the task of checking whether a given rate is admissible, note that Definition 2-(b) is satisfied whenever the condition $\frac{\partial R(w,t)}{\partial t} + \langle \nabla R(w,t), \nabla F(w)\rangle \geq 0$ holds for every $w, t \geq 0$. Many rate functions, e.g. $R(w,t) = F(w)e^{-t}$ and F being KŁ, in fact satisfy this condition for every $w, t \geq 0$.*

Furthermore, also note that Definition 2-(b) is satisfied whenever the rate function is such that $R(w(s),t) \leq R(w, s+t)$ for all $s, t \geq 0$ and $w \in \mathbb{R}^d$, which may be an easier to check condition, e.g. when $R(w,t) = F(w)e^{-t}$.

We utilize admissible rate functions to characterize behavior of GF on F. Before we proceed, let us motivate the two properties above. Property (a) is natural for any rate function and captures the fact that running GF for more time leads to better guarantees. The property (b), while seeming a bit mysterious, captures the fact that having additional information about the GF path should only improve the rate. Note that $R(w(0), s+t)$ corresponds to an upper bound on the sub-optimality at $w(s+t)$ and $R(w(s),t)$ corresponds to an upper bound on the same quantity but with the additional information that $w(s)$ is a point on the GF path. Furthermore, given any rate function R, it is easy to construct a new rate function $\bar{R}$ that always satisfies property (b) by defining $\bar{R}(w,t) = \min_{s\geq 0} R(w', t+s)$ where w' is any point such that the point w lies on the GF path from w' at time s. We also remark that the function $R(w,t) = F(w(t))$ is always an admissible rate function. Furthermore, all the rate functions appearing in this paper satisfy the above properties.

Our next result shows how existence of rate functions for GF that satisfy the above properties can be used to construction admissible potentials w.r.t. F.

Theorem 2 (From gradient flow to potentials)**.** *Let $\mathcal{W} \subseteq \mathbb{R}^d$ be any set of initial points that we want to consider, and R be an admissible rate function w.r.t. all GF paths originating from any point in $\mathcal{W}$. Further, suppose that for any $w_0 \in \mathcal{W}$, the point $w(t)$ on the GF path satisfies $F(w(t)) \leq R(w_0, t)$, then the function Φ_g defined as*

$$\Phi_g(w) = \int_{t=0}^{\infty} g(R(w,t))\,\mathrm{d}t \tag{5}$$

is an admissible potential w.r.t. F on the set $\mathrm{clo}(\mathcal{W})$, for any differentiable and monotonically increasing function $g : \mathbb{R}^+ \mapsto \mathbb{R}^+$ that satisfies $\int_{t=0}^{\infty} g(R(w,t))\,\mathrm{d}t < \infty$ and $\int_{t=0}^{\infty} g'(R(w,t))\|\nabla R(w,t)\|\,\mathrm{d}t < \infty$ for every $w \in \mathrm{clo}(\mathcal{W})$.

As an illustration on how to apply Theorem 2, assume that for F the rate for GF is $R(w,t) = \|w - w^*\|^2/2t$. For instance, we already know that such a rate holds when F is convex. For this rate, by choosing $g(z) = (1 + 1/z)^{-2}$, we get that the function $\Phi_g(w) = \int_{t=0}^{\infty}\left(1 + 2t/\|w - w^*\|^2\right)^{-2}\mathrm{d}t = \|w - w^*\|^2/2$ is an admissible potential w.r.t. F. We provide more examples in Section 5.

Theorem 1 and Theorem 2 are, in a sense, converse of each other. Theorem 1 shows that the existence of an admissible potential function implies a rate of convergence for GF. On the other hand, Theorem 2 shows how to construct admissible potentials starting from the fact that GF has a rate. One might wonder whether there always exist a Lyapunov function, more specifically a g function above, such that the rate implied by the constructed potential in Theorem 1 matches the rate that we started with for Theorem 2, i.e. $R(w,t) \approx g^{-1}(\Phi_g(w)/t)$. We answer this in the positive for rate functions that are of the product form.

Corollary 1. *Let $\mathcal{W} \subseteq \mathbb{R}^d$ be any set of initial points that we want to consider, and R be an admissible rate function w.r.t. all GF paths originating from points in $\mathcal{W}$. Additionally, suppose R has the product form $R(w,t) = h(w)r(t)$ where h is differentiable and r is a non-increasing function that satisfies $r(t) \leq \lambda|r'(t)|\max\{1,t\}$ for any $t \in \mathbb{R}$ (where λ is a universal constant). Furthermore, suppose that for any $w_0 \in \mathcal{W}$, the point $w(t)$ on the GF path satisfies $F(w(t)) \leq R(w_0, t)$. Then, there exists a monotonically increasing function $g : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that the potential $\Phi(w)$ constructed in Theorem 2 using g, when plugged in Theorem 1, implies that GF has the rate*

$$F(w(t)) \leq \max_{w\in\mathcal{W}} R(w, t/\log^2(t))$$

for any initialization $w(0) \in \mathcal{W}$.

3.1 Geometric Interpretation

The definition of an admissible potential comes with a geometric condition on the function F given in (4). Since Theorem 2 constructs admissible potentials, when a rate $R(w,t)$ holds for GF it suggests that the geometric property in (4) holds for the objective function F. As an example, say GF on F satisfies the rate $R(w,t) = F(w)e^{-t}$. From Theorem 2, we note that $\Phi_g(w) = F(w)$ is an admissible potential w.r.t F with $g(z) = z$. This implies the geometric property

$$\langle \nabla F(w), \nabla F(w) \rangle \geq F(w) \tag{6}$$

holds for F whenever GF has rate $R(w,t) = F(w)e^{-t}$. On the other, we know that whenever (6) holds the function $\Phi(w) = F(w)$ satisfies (4) and is thus an admissible potential for F (with $g(z) = z$), and hence Theorem 1 implies the rate of $F(w)/t$, which is equivalent to the rate $F(w)e^{-t}$ (c.f. Lemma 7). This implies an equivalence between the rates $R(w,t) = F(w)e^{-t}$ and the geometric property (6). We formalize this in the following.

Proposition 1. *The following two properties are equivalent:*

(a) For any $w(0) \in \mathbb{R}^d$ and $t \geq 0$, GF has the rate $F(w(t)) \leq F(w(0))e^{-\lambda t}$,

(b) $F(w)$ satisfies the Polyak-Łojasiewicz (PL) property i.e. $\lambda F(w) \leq \|\nabla F(w)\|^2$,

for any $\lambda \geq 0$. Theorem 2 implies (b) and yields the potential function $\Phi(w) = F(w)$.

In the following, we show a correspondence between the rate $R(w,t) = \frac{\|w(0)-w^*\|^2 - \|w(t)-w^*\|^2}{2t}$, and linearizability—a condition that is weaker than convexity but is sufficient for the corresponding rate of convergence for GF.

Proposition 2. *The following two properties are equivalent: For any $\lambda \geq 0$,*

(a) For any $w(0) \in \mathbb{R}^d$ and $t \geq 0$, GF has the admissible rate $F(w(t)) \leq \lambda \frac{\|w(0)-w^*\|^2 - \|w(t)-w^*\|^2}{2t}$,

(b) $F(w)$ is linearizable w.r.t. w^* i.e. $F(w) \leq \lambda \langle \nabla F(w), w - w^* \rangle$.

More generally, the equivalence between GF rates and the corresponding geometry on F can be characterized as follows.

Remark 2. *GF on F enjoys the admissible rate $R(w,t) = g^{-1}(\Phi_g(w)/t)$ if and only if F has the geometric property $\langle \nabla \Phi_g(w), \nabla F(w) \rangle \geq g(F(w))$.*

4 Stochastic Gradient Descent and Gradient Descent

GD can be thought of as an approximate discretization of gradient flow. Thus, for problems where GF converges with a given rate R, one may try to get convergence guarantees for GD from an initial point w_0 by bounding the distance between the GD and GF trajectories starting from w_0. This is exactly the approach taken in prior works [19, 30, 44, 40, 48, 14]. However, coming up with non-vacuous bounds on the distance between corresponding GF and GD iterated is often quite challenging and requires much stronger assumptions on the underlying objective. In fact there are cases where both GF and GD converge to the same global minimum but their paths can be quite far away from each other. We take a different approach for proving convergence of GD/SGD which directly relies on the properties of corresponding potential for F. In the following theorem, we note that further assumption on top of the premise that GF has a rate are required, to even hope that GD succeeds.

Theorem 3. *For any integer $T_0 > 0$, there exists a continuously differentiable convex function F for which $\min_w F(w) = 0$ and $w^* = 0$ is the unique minimizer, such that:*

(a) $\Phi(w) = \|w\|^2/2$ is an admissible potential for F. Thus, Theorem 1 implies that for any initial point w_0, the point $w(t)$ on its GF path satisfies $F(w(t)) \leq \frac{\|w_0\|^2}{2t}$.

(b) There exists an initial point w_0 with $\|w_0\| \leq 1$ and $F(w_0) \leq 2$ such that GD fails to find an $1/10$-suboptimal solution for any step size η within $t \leq T_0$ steps.

Below we present two regularity conditions that are sufficient to show convergence of GD, given a prescribed rate for GF.

Assumption 1. *There exists a monotonically increasing function $\psi : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that $\|\nabla F(w)\|^2 \leq \psi(F(w))$ for any point $w \in \mathcal{W}$.*

Assumption 2. *The potential function Φ is second-order differentiable, and there exists a monotonically increasing function $\rho : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that $\|\nabla^2 \Phi(w)\| \leq \rho(\Phi(w))$ at any point $w \in \mathcal{W}$.*

We will refer to the above conditions on F and Φ as self-bounding regularity conditions. The following theorem provides convergence guarantees for GD when an admissible potential exists and the above assumptions are satisfies.

Theorem 4 (GD convergence guarantee)**.** *Let Φ_g be an admissible potential w.r.t. F. Assume that F satisfies Assumption 1 and Φ_g satisfies Assumption 2. Then, for any $T \geq 0$ and setting η appropriately, the point $\widehat{w}_T$ returned by GD algorithm has the convergence guarantee*[2]

$$g(F(\widehat{w}_T)) = O\big(\tfrac{1}{\sqrt{T}}\big). \tag{7}$$

Furthermore, if the function $\frac{\psi}{g}$ is monotonically increasing, then for a different appropriate choice of η,

$$g(F(\widehat{w}_T)) = O\big(\tfrac{1}{T}\big). \tag{8}$$

Let us consider an example. Suppose GF on F achieves the rate $R(w,t) = \|w - w^*\|^2/2t$. Then, $\Phi_g(w) = \|w - w^*\|^2/2$ is an admissible potential for F from Theorem 2 with $g(z) = (\epsilon^{-1/\epsilon} + 1/z)^{-1-\epsilon}$ for any $\epsilon > 0$. However, as we saw in Theorem 3 just existence of such a rate function does not imply the GD will succeed and we need to make further assumptions. Notice that in this case $\Phi_g(w)$ satisfies Assumption 2 with $\rho(z) = 1$. If we further assume that F is L-Lipschitz, then Assumption 1 is satisfied with $\psi(z) = L^2$. Hence, applying Theorem 4 and taking the limit of $\epsilon \to 0$, we get that GD has convergence rate $F(\widehat{w}_T) = O(\|w - w^*\| L/\sqrt{T})$. Instead if F was H-smooth, Assumption 1 is satisfied with $\psi(z) = 4Hz$ and $\psi(z)/g(z) = 4H$ is a monotonically increasing function and thus using (8), we get that GD has the convergence rate $F(\widehat{w}_T) = O(H\|w - w^*\|^2/T)$. Notice that both of these rates are optimal for GD under the Lipschitz/Smoothness assumptions on F and given the fact that GF converges at the rate R starting from w [36]. On similar lines, using the rates for GF convergence on PL/KŁ functions, we can also recover optimal convergence rates for GD under appropriate smoothness assumptions on F. Details are deferred to the appendix.

We next consider the convergence of SGD algorithm. Recall that at the iterate w_t, SGD performs the updated using $\nabla f(w_t, z_t)$, a stochastic and unbiased estimate of $\nabla F(w_t)$. Of course, unless one has some form of control over the distribution of $\nabla f(w_t, z_t)$, one cannot hope to prove any convergence guarantees of SGD. To this end, we make the following regularity assumption on the noise in $\nabla f(w, z_t)$ while estimating $\nabla F(w)$.

Assumption 3 (Noise regularity)**.** *There exists a monotonically increasing function $\chi : \mathbb{R}^+ \mapsto \mathbb{R}^+$ such that for any point w, the gradient estimate $\nabla f(w, z)$ satisfies*

$$\Pr\big(\|\nabla f(w; z) - \nabla F(w)\|^2 \geq t \cdot \chi(F(w))\big) \leq e^{-t}.$$

Assumption 3 is quite general, and can be specialized by appropriately setting the function χ to model various stochastic optimization problem settings observed in practice. For example, the classical stochastic optimization setting in which $\nabla f(w; z) = \nabla F(w) + \varepsilon_t$ where ε_t is a sub-Gaussian random variable with mean 0 and variance σ^2 is captured by the above assumption when $\chi(z) = \sigma^2$ [35]. However, it turns out that for many interesting ML problems, the noise typically scales with the function value [46, 47].

Theorem 5 (SGD convergence guarantee)**.** *Let Φ_g be an admissible potential w.r.t. F. Assume that F satisfies Assumption 1, Φ_g satisfies Assumption 2 and the stochastic gradient estimates $\nabla f(w; z)$ satisfy Assumption 3. Then, for any $T \geq 0$ and setting η appropriately, the point $\widehat{w}_T$ returned by SGD algorithm has the convergence guarantee*[1]

$$g(F(\widehat{w}_T)) = \widetilde{O}\big(\tfrac{1}{\sqrt{T}}\big).$$

with probability at least 0.7 over the randomization of the algorithm and stochastic gradients.

[2] The $O(\cdot)$ notation here hides initialization and problem dependent constants fully specified in the Appendix.

Remark 3. *In most classic settings, one expects a $1/\sqrt{T}$ rate for SGD [6]. However, in cases where Φ_g is an admissible potential and $g(z) = o(z)$, Theorem 5 seems to suggest a $g^{-1}(1/\sqrt{T})$ rate of convergence which is faster than $1/\sqrt{T}$. This is where the self-bounding regularity conditions play an important role. As an example for PŁ style rates, one can show that $F(w)^p$ is an admissible potential with $g(z) = z^p$ for any p. However, the self-regularity conditions are not satisfied unless $p \geq 1$. Setting $p = 1$ recovers the $1/\sqrt{T}$ rate of SGD for PŁ functions which is optimal [1].*

5 Examples: From Gradient Flow to Gradient Descent

So far, we have already discussed classical examples like PŁ functions, convex functions, etc. In this section, we show the versatility of our framework by discussing more complex non-convex losses like phase retrieval and matrix square root.

5.1 Phase retrieval

In the phase retrieval problem [7, 10, 41], we wish to reconstruct a hidden vector $w^* \in \mathbb{R}^d$ with $\|w^*\| = 1$ using phaseless observations $\mathcal{S} = \{(a_j, y_j)\}_{j \leq T}$ of the form $y_j = \langle a_j, w^* \rangle^2$ where $a_j \sim \mathcal{N}(0, \mathrm{I}_d)$. The classical approach to recover w^* is by using the per-sample loss function $f_{\mathrm{pr}}(w; (a_j, y_j)) = ((a_j^\top w)^2 - y_j)^2$ for which the corresponding population loss is given by

$$F_{\mathrm{pr}}(w) = \mathbb{E}[f_{\mathrm{pr}}(w; (a, y))] = \mathbb{E}_{a \sim \mathcal{N}(0, \mathrm{I}_d)}\Big[\big((a^\top w)^2 - (a^\top w^*)^2\big)^2\Big]. \tag{9}$$

F_{pr} is non-convex, and has stationary points (and local minima) that do not correspond to the global minima. In the following, we provide convergence guarantees for GD algorithm on F_{pr}, and SGD algorithms that computes stochastic gradient estimates using $\mathcal{S}$. We first note that F_{pr} satisfies self-bounding regularity conditions, and GF on F_{pr} converges to the global minimizer for any initial point w_0.

Lemma 1. *F_{pr} satisfies Assumption 1. Furthermore, for any initial point w_0, the point $w(t)$ on its gradient flow path satisfies*

$$F_{\mathrm{pr}}(w(t)) \leq \min\Big\{F_{\mathrm{pr}}(w_0), F_{\mathrm{pr}}(w_0) e^{-t + \frac{1}{\langle w_0, w^* \rangle^2}}\Big\} =: R_{\mathrm{pr}}(w(0), t).$$

Furthermore, the function R_{pr} above is an admissible rate of convergence w.r.t. F_{pr}.

The rate above follows from independently analysis the terms parallel and perpendicular components $\langle w, w^* \rangle$ and $\|w\|^2 - \langle w, w^* \rangle$ respectively. Our main tool for getting the convergence guarantee for GD / SGD is to utilize Theorem 2 to get an admissible potential w.r.t. F_{pr}, which can be plugged in Theorem 4 and 5 to get the corresponding rates.

Theorem 6. *Consider the phase retrieval objective F_{pr} given in (9). For any initial point w_0 and $T \geq 1$, setting η appropriately,*

(a) The point $\widehat{W}_T$ returned by GD starting from w_0 satisfies $F_{\mathrm{pr}}(\widehat{W}_T) = O\big(\min\{\frac{1}{T}, e^{-O(T - t_0)}\}\big)$ for all $T \geq t_0$, where t_0 is a w_0 dependent constant.

(b) The point $\widehat{w}_T$ returned by SGD starting from w_0 and using stochastic gradient estimates for which Assumption 3 holds, satisfies $F_{\mathrm{pr}}(\widehat{w}_T) = \widetilde{O}\big(\frac{1}{\sqrt{T}}\big)$ with probability at least 0.7.

The $O(\cdot)$ notation above hides w_0 dependent constants which we specify in the Appendix. Our rate for GD above matches the best known result in the literature in terms of the dependence on T [10]. To the best of our knowledge, ours is also the first convergence analysis of SGD under arbitrary noise conditions satisfying Assumption 3. While this rate is optimal under certain noise conditions, e.g. when $\chi(z) = \sigma^2$, further improvements are possible when χ is favorable. For example, suppose the stochastic gradient estimates were computed using samples from $\mathcal{S}$ by taking a fresh sample for each estimate, i.e. $\nabla f(w; (a, y)) = 4((a^\top w)^2 - y)(a^\top w)w$. In this case, the stochastic gradient satisfy Assumption 3 with $\chi(z) = \min\{\sqrt{z}, c\}$ where c is a universal constant (c.f. Candes et al. [7, Lemma 7.4, 7.7]). While, our framework implies that this SGD algorithm (computing estimates using samples) converges at the rate of $1/\sqrt{T}$, this rate can be improved further [10], and we defer the refined analysis for future research.

5.2 Initialization specific rates

In many applications, GF is only known to converge from nice enough initial points that satisfy certain properties. In this section, we show how to use our tools to show convergence of GD/SGD for such problems, and consider matrix square root as an example. We first provide the following general utility lemma that shows how to construct admissible potentials when the rate for GF from w_0 holds only when w_0 satisfies a certain property characterized by $h(w_0) \geq 0$.

Lemma 2. *Let $h : \mathbb{R}^d \mapsto [0,1]$ be a continuously differentiable function, and suppose that for any point w for which $h(w) > 0$, GF with $w(0) = w$ has rate $F(w(t)) \leq R(w,t)$ where $R(w,\cdot)$ is a monotonically decreasing function in t. Furthermore, suppose that $F(w) \leq R(w,0)$, F satisfies Assumption 1, $R(w, h(w)t)$ is an admissible rate function w.r.t. F, and for any w,*

- *(a) the function $\Gamma(w) := \int_{t=0}^{\infty} R(w,t)\,\mathrm{d}t$ is continuously differentiable, and $\max\{\|\nabla\Gamma(w)\|, \|\nabla^2\Gamma(w)\|\} \leq \lambda(\Gamma(w))$ where λ is a positive, non-decreasing.*
- *(b) $\max\{\|\nabla h(w)\|, \|\nabla^2 h(w)\|\} \leq \pi(\Gamma(w))$ where π is a positive, non-decreasing.*
- *(c) $(h(w) - h(w^*))^2 \leq \mu(\Gamma(w))$ where μ is a positive, monotonically increasing function with the property that $k\mu(z) \leq \mu(kz)$ for any $k \geq 1$.*

Then, the function $\Phi_g(w) = \Gamma(w)/h(w)$ is an admissible potential w.r.t. F with $g(z) = z$, and satisfies the self-bounding regularity condition in Assumption 2.

5.2.1 Matrix square root

In the matrix square root problem [12, 20], we are given a positive definite and symmetric matrix $M \in \mathbb{R}^{d\times d}$ with $\sigma_d(M) > 0$, and wish to find a symmetric $W \in \mathbb{R}^{d\times d}$ that minimizes the objective

$$F_{\mathrm{ms}}(W) = \|M - W^2\|_F^2. \tag{10}$$

F_{ms} is non-convex in W, and has spurious stationary points. In the following, we provide convergence guarantees for GD/SGD algorithm on F_{ms}. We first note that F_{ms} satisfies self-bounding regularity conditions, and GF on F_{ms} converges to the global minimizer when the initial point w_0 satisfies additional assumptions. We capture these initial conditions using the function h_{ms} defined as

$$h_{\mathrm{ms}}(W) = \sigma\big(\phi(W^2) - \alpha\big), \tag{11}$$

where the function $\phi(Z) := \frac{-1}{\gamma}\log(\mathrm{tr}(e^{-\gamma Z}) + e^{-16\alpha\gamma})$, $\alpha = \sigma_{\min}(M)/1600$, $\gamma = \log(d+1)/\alpha$, and σ denotes a smoothened version of the indicator function given by $\sigma(z) = \{0 \text{ if } z \leq 0,\ \frac{2}{\alpha^2}z^2 \text{ if } 0 \leq z \leq \frac{\alpha}{2},\ -\frac{2}{\alpha^2}z^2 + \frac{4}{\alpha}z - 1 \text{ if } \frac{\alpha}{2} \leq z \leq \alpha,\ \text{and } 1 \text{ if } \alpha \leq z\}$.

Lemma 3. *F_{ms} satisfies Assumption 1. Furthermore, for any initial point W_0 for which $h_{\mathrm{ms}}(W_0) > 0$, the point $W(t)$ on its GF path satisfies*

$$F_{\mathrm{ms}}(W(t)) \leq F_{\mathrm{ms}}(W_0)\exp(-16\alpha t) =: R_{\mathrm{ms}}(W(0), t),$$

where $\alpha = \sigma_{\min}(M)/1600$, $\gamma = \log(d+1)/\alpha$ and the function h_{ms} is defined in (11).

The above rate follows from directly solving the PDE associated with the gradient flow on the underlying objective. Lemma 3 provides conditions on W_0 under which the GF path converges with the rate function R_{ms}. Our main tool for showing the convergence of GD / SGD is by using Lemma 2 to get admissible potentials. Note that the function h_{ms} takes values in $[0,1]$, is continuously differentiable, and as we show in the appendix satisfies all the required self-bounding regularity conditions in Lemma 2. Thus, Lemma 2 provides an admissible potential w.r.t. F_{pr} which can be used to get the following rates.

Theorem 7. *Consider the matrix square root objective F_{ms} given in (10). For any $\kappa > 0$, initial point W_0 for which $h_{\mathrm{ms}}(W_0) > 0$ and setting η appropriately,*

- *(a) The point $\widehat{W}_T$ returned by GD starting from w_0 satisfies $F_{\mathrm{ms}}(\widehat{W}_T) = O\big(\min\{\frac{1}{T}, e^{-O(T-t_0)}\}\big)$ for all $T \geq t_0$, where t_0 is a w_0 dependent constant.*
- *(b) The point $\widehat{W}_T$ returned by SGD starting from w_0 and using stochastic gradient estimates for which Assumption 3 holds, satisfies $F_{\mathrm{ms}}(\widehat{W}_T) = \widetilde{O}\big(\frac{1}{\sqrt{T}}\big)$ with probability at least 0.7.*

The $O(\cdot)$ notation above hides W_0 dependent constants which we specify in the Appendix. Our rate for GD above matches the best known result in the literature in terms of the dependence on T [20]. Ours is also the first convergence analysis of SGD under arbitrary noise conditions satisfying Assumption 3. Note that the classical stochastic optimization setting in which $\nabla f_{\mathrm{ms}}(w; z) = 2(W^2 - M)W + 2W(W^2 - M) + \varepsilon_t$ where ε_t is a sub-Gaussian random variable with mean 0 and variance σ^2 satisfies Assumption 3 with $\chi(z) = \sigma^2$, and as a result of Theorem 7, we get that SGD converges at the rate of $1/\sqrt{T}$. To the best of our knowledge, convergence of SGD in the stochastic optimization setting for matrix square root problem was not known before.

5.3 Extending Chatterjee 2022 [8]

If the objective F is such that some potential Φ_g satisfies the geometric condition in (4) for every w, then we have a rate of convergence for GF (Theorem 1). As we saw earlier, for instance, using this machinery one can obtain rates for GF/GD/SGD when F has PŁ property everywhere. However, such global properties, that (4) holds for every w are often too stringent to hold in practice. In order to go beyond global assumption, in Lemma 2 we showed how to extend our tools (by defining corresponding admissible potentials) when such properties (and thus rates for GF) only hold in some region. Convergence under such local properties has also been considered before in other works [10, 13, 22, 34, 42, 20, 33]. However, all of these results usually rely on being able to choose an initialization w_0 in the good region, where the corresponding local property holds, and is close enough to the global minima that we wish to converge to. This is not always practical, and to circumvent this issue in a recent work of Chatterjee [8], an assumption that is "local" w.r.t. initial point is provided under which one can show that GF and GD starting from this initialization is guaranteed to converge (at an exponential rate). The interesting property of this condition is that it is local to initial point w_0 considered and does not make any global assumption on the objective.

Using the tools in this paper, this type of local property can be easily extended to more general properties than what was considered in Chatterjee [8]. For ease of presentation, we present below the result for H-smooth objective F and for GF convergence, the corresponding techniques can be easily extended show GD/SGD convergence when Assumption 1 holds. Given a function $r : \mathbb{R}^d \mapsto \mathbb{R}^+$ and a monotonically increasing positive function g, define

$$\alpha_{r,g}(w_0, \kappa) = \inf_{w:\|w-w_0\|_2 \le \kappa, F(w) \ne 0} \frac{\nabla r(w)^\top \nabla F(w)}{g(F(w))} \tag{12}$$

Our main assumption on the initial point w_0 is that for some $\kappa > 0$ and some functions R and g,

$$\int_0^\infty \sqrt{g^{-1}\left(\frac{r(w_0)}{t\alpha_{r,g}(w_0, \kappa)}\right)} dt \le \frac{\kappa}{H} \tag{13}$$

The next lemma shows that for any initial point w_0 that satisfies the local condition above, one has a rate of convergence for GF starting from w_0.

Lemma 4. *Suppose w_0 satisfies (13) for some functions R and g, and radius $\kappa = \kappa_0 > 0$. Then, gradient flow starting from $w(0) = w_0$ satisfies for any $t \ge 0$,*

$$F(w(t)) \le g^{-1}\left(\frac{r(w_0)}{T\alpha(w_0, \kappa_0)}\right).$$

To obtain nearly matching rates for the type of condition in Chatterjee [8], one can choose $r(w) = p \cdot F(w)^{1/p}$ and $g(z) = z^{1/p}$. Since p is arbitrary, setting $p = T\alpha(w_0, \kappa_0)/e$ we obtain nearly the same rate and the local condition as Chatterjee [8] (upto constants). The interesting part though, is that this is for only one choice of g and r, whereas we can get the convergence for GF when the condition holds for any g, r. In Chatterjee [8], examples of overparmeterized deep neural nets are shown to satisfy the assumption (for the specific r and g above). With a wider choice of g and r we can extend these to more general models (eg. neural networks with milder assumptions on the activation function).

References

[1] A. Agarwal, M. J. Wainwright, P. Bartlett, and P. Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

[2] S. Arora, R. Ge, T. Ma, and A. Moitra. Simple, efficient, and neural algorithms for sparse coding. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, 2015.

[3] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010. ISSN 0364765X, 15265471.

[4] S. Azulay, E. Moroshko, M. S. Nacson, B. E. Woodworth, N. Srebro, A. Globerson, and D. Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 468–477. PMLR, 2021.

[5] N. Bansal and A. Gupta. Potential-function proofs for first-order methods. *arXiv preprint arXiv:1712.04581*, 2017.

[6] S. Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[7] E. J. Candes, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[8] S. Chatterjee. Convergence of gradient descent for deep neural networks. *arXiv preprint arXiv:2203.16462*, 2022.

[9] N. S. Chatterji, P. M. Long, and P. L. Bartlett. The interplay between implicit bias and benign overfitting in two-layer linear networks. *CoRR*, abs/2108.11489, 2021.

[10] Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1):5–37, 2019.

[11] L. Chizat and F. Bach. A note on lazy training in supervised differentiable programming. 2018. URL http://arxiv.org/abs/1812.07956. cite arxiv:1812.07956.

[12] C. De Sa, C. Re, and K. Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *International conference on machine learning*, pages 2332–2341. PMLR, 2015.

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

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

[15] Y. Fang, K. A. Loparo, and X. Feng. Inequalities for the trace of matrix product. *IEEE Transactions on Automatic Control*, 39(12):2489–2490, 1994.

[16] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[17] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Implicit bias of gradient descent on linear convolutional networks. *arXiv preprint arXiv:1806.00468*, 2018.

[18] S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

[19] S. Gunasekar, B. Woodworth, and N. Srebro. Mirrorless mirror descent: A natural derivation of mirror descent. In *International Conference on Artificial Intelligence and Statistics*, pages

2305–2313. PMLR, 2021.

[20] P. Jain, C. Jin, S. Kakade, and P. Netrapalli. Global convergence of non-convex gradient descent for computing matrix squareroot. In *Artificial Intelligence and Statistics*, pages 479–488. PMLR, 2017.

[21] Z. Ji and M. Telgarsky. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

[22] C. Jin, S. M. Kakade, and P. Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. *Advances in Neural Information Processing Systems*, 29, 2016.

[23] C. Jin, S. M. Kakade, and P. Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. In *NIPS*, 2016.

[24] S. Kale, A. Sekhari, and K. Sridharan. SGD: the role of implicit regularization, batch-size and multiple-epochs. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27422–27433, 2021.

[25] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9851*, ECML PKDD 2016, page 795–811, Berlin, Heidelberg, 2016. Springer-Verlag. ISBN 9783319461274. doi: 10.1007/978-3-319-46128-1_50.

[26] C. M. Kellett. Classical converse theorems in lyapunov's second method. *Discrete and Continuous Dynamical Systems - B*, 20(8):2333–2360, 2015.

[27] R. Kleinberg, Y. Li, and Y. Yuan. An alternative view: When does sgd escape local minima? In J. G. Dy and A. Krause, editors, *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2703–2712. PMLR, 2018.

[28] W. Krichene. Continuous and discrete dynamics for online learning and convex optimization. *Ph. D. Dissertation*, 2016.

[29] W. Krichene. *A Lyapunov Approach to Accelerated First-Order Optimization In Continuous and Discrete Time*. PhD thesis, University of California, Berkeley, 2016.

[30] W. Krichene, A. Bayen, and P. L. Bartlett. Accelerated mirror descent in continuous and discrete time. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[31] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783, 1998.

[32] S. Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

[33] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354. PMLR, 2018.

[34] H. Mohammadi, A. Zare, M. Soltanolkotabi, and M. R. Jovanović. Global exponential convergence of gradient methods over the nonconvex landscape of the linear quadratic regulator. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 7474–7479. IEEE, 2019.

[35] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Stochastic approximation approach to stochastic programming. In *SIAM J. Optim*. Citeseer.

[36] A. S. Nemirovskij and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983.

[37] B. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. ISSN 0041-5553.

[38] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[39] N. Srebro, K. Sridharan, and A. Tewari. Smoothness, low noise and fast rates. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/file/76cf99d3614e23eabab16fb27e944bf9-Paper.pdf.

[40] W. Su, S. Boyd, and E. Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *Advances in neural information processing systems*, 27, 2014.

[41] Y. S. Tan and R. Vershynin. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *arXiv preprint arXiv:1910.12837*, 2019.

[42] G. Vardi and O. Shamir. Implicit regularization in relu networks with the square loss. In *Conference on Learning Theory*, pages 4224–4258. PMLR, 2021.

[43] A. Wilson. *Lyapunov arguments in optimization*. University of California, Berkeley, 2018.

[44] A. C. Wilson, B. Recht, and M. I. Jordan. A lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021.

[45] A. C. Wilson, B. Recht, and M. I. Jordan. A lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021.

[46] S. Wojtowytsch. Stochastic gradient descent with noise of machine learning type. part i: Discrete time analysis. *arXiv preprint arXiv:2105.01650*, 2021.

[47] S. Wojtowytsch. Stochastic gradient descent with noise of machine learning type. part ii: Continuous time analysis. *arXiv preprint arXiv:2106.02588*, 2021.

[48] P. Zhang, A. Orvieto, H. Daneshmand, T. Hofmann, and R. S. Smith. Revisiting the role of euler numerical integration on acceleration and stability in convex optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3979–3987. PMLR, 2021.