Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

Convergence of Stochastic Gradient Descent under a Local Łojasiewicz Condition for Deep Neural Networks

Jing An * [1] and Jianfeng Lu [1]

[1] Department of Mathematics, Duke University, Durham, NC 27710, USA.

Abstract. We study the convergence of stochastic gradient descent (SGD) for non-convex objective functions. We establish the local convergence with positive probability under the local Łojasiewicz condition introduced by Chatterjee [arXiv:2203.16462, 2022] and an additional local structural assumption of the loss function landscape. A key component of our proof is to ensure that the whole trajectories of SGD stay inside the local region with a positive probability. We also provide examples of neural networks with finite widths such that our assumptions hold.

Keywords:
Non-convex optimization,
Stochastic gradient descent,
Convergence analysis.

Article Info.:
Volume: 4
Number: 2
Pages: 89 - 107
Date: June/2025
doi.org/10.4208/jml.240724

Article History:
Received: 07/07/2024
Accepted: 11/02/2025

Communicated by:
Qianxiao Li

1 Introduction

The stochastic gradient descent and its variants are widely applied in machine learning problems due to its computational efficiency and generalization performance. A typical empirical loss function for the training writes as

$$F(\theta) = \mathbb{E}_{\xi\sim\mathcal{D}}[f(\theta;\xi)], \tag{1.1}$$

where ξ denotes the random sampling from the training data set following the distribution $\mathcal{D}$. A standard SGD iteration to train parameters $\theta \in \mathbb{R}^n$ is of the form

$$\theta_{k+1} = \theta_k - \eta_k \nabla f(\theta_k;\xi_k). \tag{1.2}$$

Here, the step size η_k can be either a fixed constant or iteration-adapted, and $\nabla f(\theta_k;\xi_k)$ is a unbiased stochastic estimate of the gradient $\nabla F(\theta_k)$, induced by the sampling of the dataset.

The convergence of SGD for convex objective functions has been well established, and we give an incomplete list of works [4, 6, 7, 18, 23, 29] here for reference. Since SGD algorithms in practice are often applied to non-convex problems in machine learning such as complex neural networks and demonstrate great empirical success, much attention has been drawn to study the SGD in non-convex optimization [12, 19, 31]. Compared with

*Corresponding author. jing.an@duke.edu

convex optimization, the behavior of stochastic gradient algorithms over the non-convex landscape is unfortunately much less understood. It is natural to investigate whether stochastic gradient algorithms converge through the training, and what minimum they converge to in non-convex problems. However, these questions are noticeably challenging since the trajectory of stochastic iterates is more difficult to track due to the noise. Most available results are limited. For example, works such as [1,2,16,38] provide convergence guarantees to a critical point in terms of quantifying the vanishing of ∇F, but little information is given on what critical points that SGD converges to. Many convergence results are based on global assumptions on the objective function, including the global Poylak-Łojasiewicz condition [24, 25], the global quasar-convexity [17], or assumptions of weak convexity and global boundedness of iterates [14, 37]. Those global assumptions are often not realistic, at least they cannot cover general multi-modal landscapes.

More specifically for optimization problems for deep neural network architectures, most convergence results are obtained in the overparametrized regime, which means that the number of neurons grow at least polynomially with respect to the sample size. For example, works including [10,20,36,40] consider wide neural networks, which essentially linearize the problem by extremely large widths. Particularly in such settings, Poylak-Łojasiewicz type conditions are shown to be satisfied, and they thus prove convergence with linear rates [3, 25]. Let us also mention convergence results of shallow neural networks in the mean field regime [9,27,32], while the convergence has not been fully established for deep neural networks.

Convergence results are very limited for neural networks with finite widths and depths, and we refer to [8, 21, 26] for recent progresses in terms of the convergence of gradient descent in such scenario. In particular, [8] constructs feedforward neural networks with smooth and strictly increasing activation functions, with the input dimension being greater than or equal to the number of data points. Such neural networks satisfy a local version of the Łojasiewicz inequality, and the convergence of gradient descent to a global minimum given appropriate initialization are fully analyzed. In this work, our goal is to extend the convergence result in [8] to stochastic gradient descent, with minimal additional assumptions added to the loss function $F(\theta)$.

In this work, we extend Chatterjee's convergence result to SGD for non-convex objective functions with minimal additional assumptions applicable to finitely wide neural networks. Our main result Theorem 3.1 asserts that, with a positive probability, SGD converges to a zero minimum within a locally initialized region satisfying the Łojasiewicz condition (Assumption 1). In particular, our proof relies on assuming that the noise scales with the objective function (Assumption 4), and in the end we provide an negative argument showing that convergence with the bounded noise and Robbins-Monro type step sizes can fail in specific scenarios (Theorem 4.1).

Notation

Throughout the note, $|\cdot|$ denotes the Euclidean norm, $B(\theta, r)$ denotes the Euclidean ball of radius r centered at θ. Unless otherwise specified, the expectation $\mathbb{E} = \mathbb{E}_{\xi\sim\mathcal{D}}$, and the gradients $\nabla = \nabla_\theta$.

2 Preliminaries

We will refer the major assumption used in [8] as the local Łojasiewicz condition. Let us first recall the definition as in [8]: Given the dimension d, let $F : \mathbb{R}^d \to [0, \infty)$ be a non-negative objective function. For any $\theta_0 \in \mathbb{R}^d$ and $r > 0$, we define

$$\alpha(\theta_0, r) := \inf_{B(\theta_0, r),\, F(\theta) \neq 0} \frac{|\nabla F(\theta)|^2}{F(\theta)}. \tag{2.1}$$

If $F(\theta) = 0$ for all $\theta \in B(\theta_0, r)$, then we set $\alpha(\theta_0, r) = \infty$. With the same initialization θ_0 and (2.1), the main assumption is

Assumption 1 (Local Łojasiewicz Condition)**.** We assume that for some $r > 0$,

$$4F(\theta_0) < r^2 \alpha(\theta_0, r). \tag{2.2}$$

Detailed discussions on this local Łojasiewicz condition can be found in [8]. Let us mention that the local Łojasiewicz condition does have limitations as standard Polyak-Łojasiewicz conditions have. However, aiming at finding the zero minimum, the choice of our initialization (2.2) has excluded the possibilities of saddles or sub-optimal local minima that exist in neural networks.

Assumption 2 (C_L-Smoothness)**.** Furthermore, we assume that for a compact set $\mathcal{K}$, there exist a constant $C_L > 0$ such that

$$|\nabla F(\theta_1) - \nabla F(\theta_2)| \leq C_L |\theta_1 - \theta_2| \tag{2.3}$$

for any $\theta_1, \theta_2 \in \mathcal{K}$.

Based on that, we can obtain a local growth control of $|\nabla F|$. The following result is a local version of [39, Lemma B.1]. The proof is very similar except that some modification is needed for the localization with the local boundedness of $|\nabla F|$. We provide the proof in the appendix.

Lemma 2.1. *Suppose that $F : \mathbb{R}^d \to \mathbb{R}$ is a non-negative function. Assume ∇F is Lipschitz continuous in a compact set $\mathcal{K}$ with the constant C_L, and there exists $\bar{C} > 0$ such that $\max_{\theta \in \mathcal{K}} |\nabla F(\theta)| = \bar{C}$. Then there exists a compact set $\tilde{\mathcal{K}} \subset \mathcal{K}$, where $\operatorname{dist}(\partial\mathcal{K}, \partial\tilde{\mathcal{K}})$ depends on C_L and $\bar{C}$ such that for any $\theta \in \tilde{\mathcal{K}}$,*

$$|\nabla F(\theta)|^2 \leq 2C_L F(\theta). \tag{2.4}$$

Throughout the note, we assume that $B(\theta_0, r) \subset \tilde{\mathcal{K}}$ for a given radius $r > 0$.

2.1 Additional assumptions

Besides the local Łojasiewicz condition, we need an additional structural assumption on F. Given a large radius R with which (2.2) is satisfied, it is natural to assume that the $F(\theta)$ is bounded away from zero near the domain boundary.

Assumption 3 (Confinement Near Boundary)**.** For some constant $M_0 > 0$,

$$F(\theta) \geq M_0, \quad \forall\, \theta \in B(\theta_0, R) \setminus B(\theta_0, R-1). \tag{2.5}$$

Here we set the annulus radius to be 1 for simplicity. The assumption (2.5) can include more general F by applying geometric transformations to F, or linear transformations to the dataset to change the annulus radius.

The assumption (2.5) is in order to ensure that the support of the zero minimum $F(\theta^*) = 0$ is away from the boundary $\partial B(\theta_0, R)$. In [8, Theorem 2.1], Chatterjee constructed a feedforward neural network with finite width and depth that satisfies the local Łojasiewicz condition (2.2). We will verify in the appendix that such a subgroup of feedforward neural networks can satisfy (2.5) as well.

Stochastic gradients

To analyze the convergence of stochastic gradient algorithms, it is unavoidable to assume some structure of the gradient noise. We may write the stochastic gradient in the decomposition form

$$\nabla f(\theta; \xi) = \nabla F(\theta) + Z(\theta; \xi), \tag{2.6}$$

where the noise term $Z(\theta; \xi)$ is unbiased

$$\mathbb{E}[Z(\theta; \xi)] = 0. \tag{2.7}$$

In most papers, usually two kinds of structural assumptions are imposed on the noise. We take a brief review.

- If one plans to analyze SGD with non-constant step sizes η_k, i.e. in the Robbins-Monro flavor [35]

$$\sum_{k=1}^{\infty} \eta_k = \infty, \quad \sum_{k=1}^{\infty} \eta_k^{1+\frac{m}{2}} < \infty, \quad m \geq 2, \tag{2.8}$$

  then it is typical to assume the bounded moments of stochastic gradients, that is, for some constant $c > 0$ and $m \geq 2$,

$$\mathbb{E}[|Z(\theta; \xi)|^m] \leq c^m < \infty. \tag{2.9}$$

  The bounded variance ($m = 2$) assumption is standard in stochastic optimization, and we refer to classical books, lecture notes and papers [5, 22, 30, 33] on this setup. The noise boundedness and adaptive step sizes have been a popular combination to establish convergence, for example, [28] provides a SGD convergence based on the local convexity assumption, [15] gives the convergence of SGD to the local manifold of minima by estimating $\mathbb{P}(F(\theta_k) - \inf_{\theta \in \mathbb{R}^d} F(\theta) \geq \epsilon)$. We also mention works such as [34] that introduces the stochastic variance reduction method, which is motivated by improving the convergence rate for $\min_{0 \leq k \leq N-1} \mathbb{E}[|\nabla F(\theta_k)|^2]$.

- On the other hand, if one just wants to establish global convergence for a fixed step size η, additional assumptions on how the stochastic gradient being related to the loss function can help. One option, according to [39], is that

Assumption 4. For some constant $\sigma > 0$,

$$\nabla f(\theta;\xi) = \nabla F(\theta) + \sqrt{\sigma F(\theta)} Z_{\theta,\xi} \tag{2.10}$$

with

$$\mathbb{E}[Z_{\theta,\xi}] = 0, \quad \mathbb{E}[|Z_{\theta,\xi}|^2] \leq 1. \tag{2.11}$$

Here, $\sqrt{\sigma F(\theta_k)} Z_{\theta,\xi}$ is named as the machine learning noise, which scales with the function value. We refer to [39, Section 2.5] for many types of machine learning problems satisfying the assumptions (2.10)-(2.11). A more relaxed and general version, considered in [13], is that there exists a monotonically increasing function $\varrho : \mathbb{R}_+ \to \mathbb{R}_+$ so that

$$\Pr\left(|\nabla f(\theta;\xi) - \nabla F(\theta)|^2 \geq t\varrho\big(F(\theta)\big)\right) \leq e^{-t}. \tag{2.12}$$

In this note, we will present the convergence of SGD under the second type assumptions (2.10)-(2.11). Moreover, we will discuss why convergence under assumptions (2.8)-(2.9) can fail for the local Łojasiewicz condition.

Comparison to previous convergence results

Our motivation is in the same vein as [8]: We assume the local Łojasiewicz condition for the initialization so that the convergence result may apply to feedforward neural networks of bounded widths and depths.

Several convergence results for SGD in non-convex optimization have been developed in recent years, and we list a few here and highlight their differences:

- [28]: Convergence to a local minima θ^* that are Hurwicz regular, i.e. $\nabla^2 F(\theta^*) \succ 0$. The key difference is that they assume that there exists $a > 0$ such that $\langle \nabla F(\theta), \theta - \theta^* \rangle \geq a|\theta - \theta^*|^2$ for all θ in a convex compact neighborhood of θ^*.
- [39]: It assumes that the Łojasiewicz condition holds in an ϵ-sublevel set of F. As a consequence, a convergence rate for $F(\theta) \to 0$ is obtained, but the proof cannot track where the global minimizer x^* locates.

Similar to the result in [8] for gradient descents, we will show that under our Assumptions 1-4 as in the setup, with some computable probability, stochastic iterates will convergence to the zero minimum almost surely. We need to choose the initialization in a ball where (2.2) holds, and as a consequence, the minimizer also lies in the same ball without assuming its existence a priori, similar to [8]. Compared to [39], our results is Euclidean as we can track the stochastic trajectories in the training.

3 Main result

Given the assumptions we describe in the previous section, we present the convergence of SGD with a quantitative rate and a step-size bound. We point out that (2.10)-(2.11) assuming noise in SGD being scaled with the objective function play an important role here.

Theorem 3.1. *We choose an initialization θ_0 with a radius $R > 1$ such that (2.2) holds, and write $\alpha := \alpha(\theta_0, R)$ as defined in (2.1). Let F be a loss function such that its gradient is Lipschitz continuous (2.3), and (2.5) holds for $B(\theta_0, R)$. Then for every $\delta > 0$, there exists $\varepsilon > 0$ such that if $F(\theta_0) \leq \varepsilon$, with probability at least $1 - \delta$, we have $\theta_k \in B(\theta_0, R-1)$ for all $k \in \mathbb{N}$.*

Conditioned on the event that $\theta_k \in B(\theta_0, R-1)$ for all $k \in \mathbb{N}$, we have

$$\lim_{k\to\infty} \beta^k F(\theta_k) = 0 \tag{3.1}$$

almost surely for every $\beta \in [1, \rho^{-1})$, where

$$\rho = 1 - \eta\alpha + \frac{\eta^2}{2} C_L(2C_L + \sigma) \in (0, 1), \tag{3.2}$$

if the step size η satisfies

$$0 < \eta \leq \min\left\{\frac{1}{\alpha}, \frac{\alpha}{4C_L(2C_L+\sigma)}\right\}.$$

Moreover, conditioned on the same event, θ_k converges almost surely to a point θ_* where $F(\theta_*) = 0$.

Proof. Let us define the event where all the iterates up to k-th step stay in a ball $B(\theta_0, r)$

$$E_k(r) := \bigcap_{j=0}^{k} \{|\theta_j - \theta_0| \leq r\} \tag{3.3}$$

for some $r \in (0, R]$ to be determined later. We denote $\mathcal{F}_k$ as the filtration generated by $\xi_1, \cdots, \xi_{k-1}$.

Our proof strategy is roughly outlined as follows:

(1) Conditioned on the event $E_k(R-1)$, it can be shown that the expectation of loss $F(\theta_k)$ exhibits a contraction by a factor $\rho \in (0, 1)$.

(2) Given that, the expectation of the travel distance $(\theta_{k+1} - \theta_k)\mathbb{1}_{E_k(R-1)}$ can be bounded by $C\rho^k/2$ with some constant $C > 0$. Therefore, the total distance $\sum_{k=0}^{\infty}(\theta_{k+1} - \theta_k)\mathbb{1}_{E_k(R-1)}$ is finite with high probability.

(3) What remains is to show the event $E_\infty(R-1)$ contained in all $E_k(R-1)$ will happen with a positive probability. This can be done by the estimates of (3.14) and (3.15). Conditioned on $E_\infty(R-1)$, $\{\theta_k\}$ form a Cauchy sequence and $F(\theta_k)$ converges to the zero minimum.

Taking the interpolation $\theta_{k+s} = \theta_k - s\eta\nabla f(\theta_k, \xi_k)$ for $s \in [0, 1]$, we have that

$$\begin{aligned}
&F(\theta_{k+1}) - F(\theta_k) \\
&= \int_0^1 \frac{d}{ds} F\big(\theta_k - s\eta\nabla f(\theta_k; \xi_k)\big) ds \\
&= -\eta \int_0^1 \nabla F(\theta_{k+s}) \cdot \nabla f(\theta_k; \xi_k) ds
\end{aligned}$$

$$
\begin{aligned}
&= -\eta \int_0^1 \nabla F(\theta_k) \cdot \nabla f(\theta_k; \xi_k) ds - \eta \int_0^1 \left(\nabla F(\theta_{k+s}) - \nabla F(\theta_k) \right) \cdot \nabla f(\theta_k; \xi_k) ds \\
&\leq -\eta \nabla F(\theta_k) \cdot \nabla f(\theta_k; \xi_k) + \eta C_L \int_0^1 |\theta_{k+s} - \theta_k| \, |\nabla f(\theta_k; \xi_k)| ds \\
&= -\eta \nabla F(\theta_k) \cdot \nabla f(\theta_k; \xi_k) + \eta^2 C_L \int_0^1 s |\nabla f(\theta_k; \xi_k)|^2 ds \\
&= -\eta \nabla F(\theta_k) \cdot \nabla f(\theta_k; \xi_k) + \frac{\eta^2}{2} C_L |\nabla f(\theta_k; \xi_k)|^2, \qquad (3.4)
\end{aligned}
$$

and we apply the smoothness assumption (2.3) in the middle inequality. Rearrange terms and multiply with the indicate function on both sides, we have

$$
F(\theta_{k+1}) \mathbb{1}_{E_{k+1}(r)} \leq \left(F(\theta_k) - \eta \nabla F(\theta_k) \cdot \nabla f(\theta_k; \xi_k) + \frac{\eta^2}{2} C_L |\nabla f(\theta_k; \xi_k)|^2 \right) \mathbb{1}_{E_{k+1}(r)}. \qquad (3.5)
$$

Now we want to replace $\mathbb{1}_{E_{k+1}(r)}$ by $\mathbb{1}_{E_k(r)}$ on the right side. Note that $\mathbb{1}_{E_k(r)} \geq \mathbb{1}_{E_{k+1}(r)}$, and moreover,

$$
F(\theta_k) - \eta \nabla F(\theta_k) \cdot \nabla f(\theta_k; \xi_k) + \frac{\eta^2}{2} C_L |\nabla f(\theta_k; \xi_k)|^2 \geq 0, \qquad (3.6)
$$

since the discriminant of this quadratic equation is

$$
|\nabla F(\theta_k)|^2 - 2 C_L F(\theta_k) \leq 0
$$

by (2.4). We thus get

$$
F(\theta_{k+1}) \mathbb{1}_{E_{k+1}(r)} \leq \left(F(\theta_k) - \eta \nabla F(\theta_k) \cdot \nabla f(\theta_k; \xi_k) + \frac{\eta^2}{2} C_L |\nabla f(\theta_k; \xi_k)|^2 \right) \mathbb{1}_{E_k(r)}. \qquad (3.7)
$$

We replace ∇f above by the decomposition (2.10), that is

$$
\nabla f(\theta_k; \xi_k) = \nabla F(\theta_k) + \sqrt{\sigma F(\theta_k)} Z_{\theta, \xi},
$$

and use the bound (2.4) as well as the local Łojasiewicz condition (2.2). By taking the expectation we have that

$$
\mathbb{E}[F(\theta_{k+1}) \mathbb{1}_{E_{k+1}(r)} | \mathcal{F}_k] \leq \left(1 - \eta \alpha + \frac{\eta^2}{2} C_L (2 C_L + \sigma) \right) F(\theta_k) \mathbb{1}_{E_k(r)}. \qquad (3.8)
$$

We may choose a small step-size

$$
0 < \eta^* \leq \min \left\{ \frac{1}{\alpha}, \frac{\alpha}{4 C_L (2 C_L + \sigma)} \right\},
$$

so that

$$
\rho = 1 - \eta^* \alpha + \frac{\eta^{*2}}{2} C_L (2 C_L + \sigma) \leq 1 - \frac{\alpha \eta^*}{2} \in (0, 1).
$$

With this, we obtain a contraction

$$\mathbb{E}[F(\theta_{k+1})\mathbb{1}_{E_{k+1}(r)}|\theta_0] \leq \rho\mathbb{E}[F(\theta_k)\mathbb{1}_{E_k(r)}|\theta_0] \leq \rho^{k+1}F(\theta_0). \tag{3.9}$$

Based on the contraction of F, we can nicely control the Euclidean distance between θ_k and θ_0. Note that

$$\begin{aligned}
&\mathbb{E}[|(\theta_{k+1}-\theta_k)\mathbb{1}_{E_k(r)}|] \\
&= \eta^*\mathbb{E}[|\nabla f(\theta_k;\xi_k)\mathbb{1}_{E_k(r)}|] \\
&= \eta^*\mathbb{E}\left[\left|\left(\nabla F(\theta_k) + \sqrt{\sigma F(\theta_k)}Z_{\theta_k,\xi_k}\right)\mathbb{1}_{E_k(r)}\right|\right] \\
&\leq \eta^*\sqrt{2C_L}\sqrt{\mathbb{E}[F(\theta_k)\mathbb{1}_{E_k(r)}]} \\
&\quad + \eta^*\sqrt{\mathbb{E}[\sigma F(\theta_k)\mathbb{1}_{E_k(r)}]}\sqrt{\mathbb{E}\left[|Z_{\theta_k,\xi_k}|^2\right]} \\
&= \left(\sqrt{2C_L}+\sqrt{\sigma}\right)\eta^*\rho^{\frac{k}{2}}F(\theta_0),
\end{aligned} \tag{3.10}$$

where we apply the Cauchy-Schwarz inequality, the growth bound (2.4), and (2.10)-(2.11). The total distance is thus finite

$$\mathbb{E}\left[\sum_{k=0}^{\infty}|(\theta_{k+1}-\theta_k)\mathbb{1}_{E_k(r)}|\right] \leq \frac{\left(\sqrt{2C_L}+\sqrt{\sigma}\right)\eta^*}{1-\sqrt{\rho}}F(\theta_0) < \infty. \tag{3.11}$$

By the Markov's inequality, for any $\tilde{\delta} \in (0,1)$, we have the length of trajectory bounded by

$$\sum_{k=0}^{\infty}|(\theta_{k+1}-\theta_k)\mathbb{1}_{E_k(r)}| \leq \frac{\left(\sqrt{2C_L}+\sqrt{\sigma}\right)\eta^*}{(1-\sqrt{\rho})\tilde{\delta}}F(\theta_0) \tag{3.12}$$

with probability at least $1-\tilde{\delta}$. This means that as long as the local Łojasiewicz condition holds, the stochastic iterates will converge to a point with high probability. What remains to show that there exists a positive probability for

$$E_\infty(r) := \bigcap_{j=0}^{\infty}\{|\theta_j-\theta_0| \leq r\} \tag{3.13}$$

to occur. In fact, we can set $r = R-1$ and argue that there exists a positive probability that all stochastic iterates are trapped in the ball $B(\theta_0, R-1)$. Let us consider the following two cases:

- Utilizing (3.10), the chance for the next iterate to escape the ball $B(\theta_0, R)$ is

$$\begin{aligned}
&\Pr\left(E_k(R-1) \text{ occurs but } \theta_{k+1} \in B^c(\theta_0,R)\right) \\
&\leq \Pr(|\theta_{k+1}-\theta_k|\mathbb{1}_{E_k(R-1)} \geq 1) \\
&\leq \mathbb{E}[|\theta_{k+1}-\theta_k|\mathbb{1}_{E_k(R-1)}] \leq \left(\sqrt{2C_L}+\sqrt{\sigma}\right)\eta^*\rho^{\frac{k}{2}}F(\theta_0).
\end{aligned} \tag{3.14}$$

- Utilizing (2.5), the chance for the next iterate to enter the annulus $B(\theta_0,R)\backslash B(\theta_0,R-1)$ is

$$\begin{aligned}&\Pr\left(E_k(R-1) \text{ occurs but } \theta_{k+1}\in B(\theta_0,R)\setminus B(\theta_0,R-1)\right)\\ \leq\ &\Pr\left(F(\theta_{k+1})\mathbb{1}_{E_{k+1}(R)}\geq M_0\right)\leq \frac{\mathbb{E}[F(\theta_{k+1})\mathbb{1}_{E_{k+1}(R)}]}{M_0}\leq \frac{\rho^{k+1}F(\theta_0)}{M_0}.\end{aligned} \tag{3.15}$$

We are ready to conclude, note that the event defined in (3.3) over radius $r = R-1$ is monotonically decreasing,

$$E_{k+1}(R-1)\subseteq E_k(R-1). \tag{3.16}$$

Let us define

$$\tilde{E}_{k+1}(R-1) := E_k(R-1)\setminus E_{k+1}(R-1), \tag{3.17}$$

whose probability is indeed the summation of (3.14) and (3.15)

$$\Pr\left(\tilde{E}_{k+1}(R-1)\right)\leq F(\theta_0)\left(\left(\sqrt{2C_L}+\sqrt{\sigma}\right)\eta^*\rho^{\frac{k}{2}}+\frac{\rho^{k+1}}{M_0}\right). \tag{3.18}$$

The probability of the complementary event can be bounded,

$$\begin{aligned}\Pr\left(E_k^c(R-1)\right)&=\sum_{i=1}^{k}\Pr\left(\tilde{E}_i(R-1)\right)\\ &<F(\theta_0)\left(\left(\sqrt{2C_L}+\sqrt{\sigma}\right)\frac{\eta^*\sqrt{\rho}}{1-\sqrt{\rho}}+\frac{\rho^2}{M_0(1-\rho)}\right)\end{aligned} \tag{3.19}$$

for any $k\in\mathbb{N}$. Thus for every $0<\delta<1$, there exists $\varepsilon>0$ depending on C_L,σ,η^*,ρ and M_0 such that if $F(\theta_0)\leq\varepsilon$, then

$$\Pr\left(E_k(R-1)\right)\geq 1-\delta \tag{3.20}$$

for all $k\in\mathbb{N}$. Taking $k\to\infty$, we then have

$$\Pr\left(E_\infty(R-1)\right)\geq 1-\delta.$$

Conditioned on the event $E_\infty(R-1)$, we can conclude from (3.9) that for every $\beta\in[1,\rho^{-1})$,

$$\lim_{k\to\infty}\beta^k F(\theta_k)=0 \tag{3.21}$$

almost surely. Moreover, for all $1\leq j<k$, by (3.10),

$$\begin{aligned}&\mathbb{E}[|\theta_k-\theta_j|\mathbb{1}_{E_\infty(R-1)}]\\ \leq&\sum_{i=j}^{k-1}\mathbb{E}[|\theta_{i+1}-\theta_i|\mathbb{1}_{E_\infty(R-1)}]\\ \leq&\sum_{i=j}^{k-1}\mathbb{E}[|\theta_{i+1}-\theta_i|\mathbb{1}_{E_i(R-1)}]\end{aligned}$$

$$\leq \sum_{i=j}^{k-1} \left(\sqrt{2C_L} + \sqrt{\sigma}\right)\eta^* \rho^{\frac{i}{2}} F(\theta_0)$$

$$< \left(\sqrt{2C_L} + \sqrt{\sigma}\right)\eta^* F(\theta_0) \frac{\rho^{j/2}}{1-\sqrt{\rho}}. \tag{3.22}$$

As $j \to \infty$, this bound goes to zero. Thus $\{\theta_k\}_{k\geq 0}$ from a Cauchy sequence conditioned on $E_\infty(R-1)$, and

$$\theta_k \;\; \to \;\; \theta_* \in B(\theta_0, R-1) \quad \text{as} \quad k \;\; \to \;\; \infty.$$

By (3.9), we see that $F(\theta_*) = 0$ conditioned on $E_\infty(R-1)$. We also know that with probability at least $1-\delta$, $\theta_k \in B(\theta_0, R-1)$ for all $k \in \mathbb{N}$. Therefore, we can conclude that with probability at least $1-\delta$, θ_k converges to the minimizer $\theta_* \in B(\theta_0, R-1)$. □

4 Non-convergence with bounded noises

One may wonder if the machine learning noise is relaxed by only assuming the boundedness (2.9), whether it can be shown that SGD iterates still converge inside $B(\theta_0, r)$ for some radius $r > 0$. As a companion of bounded noises, we should consider to take a Robbins-Monro type step sizes

$$\eta_k = \frac{\gamma}{(k+n_0)^q}, \quad q \in \left(\frac{1}{2}, 1\right] \tag{4.1}$$

for SGD (1.2) with constants $\gamma, n_0 > 0$ to be chosen.

If all the stochastic iterates θ_k stay inside the ball where the local Łojasiewicz condition (2.2) holds, then we show in Lemma 4.2 that the convergence happens with an algebraic rate. Lemma 4.2 relies on classical results in numerical sequences, which can be traced back to [11].

Lemma 4.1 ([11, Lemmas 1 and 4])**.** *Let $\{b_k\}_{k\geq 1}$ be a non-negative sequence such that*

$$b_{k+1} \leq \left(1 - \frac{C_1}{(k+n_0)^q}\right) b_k + \frac{C_2}{(k+n_0)^{q+p}}, \tag{4.2}$$

where $q \in (0,1]$, $p > 0$ and $C_1, C_2 > 0$, then

1. *If $q = 1$ and $C_1 > p$, we have*

$$b_k \leq \frac{C_2}{C_1 - p}\frac{1}{k} + o\left(\frac{1}{k}\right). \tag{4.3}$$

2. *If $q < 1$, we have*

$$b_k \leq \frac{C_2}{C_1}\frac{1}{k^p} + o\left(\frac{1}{k^p}\right). \tag{4.4}$$

In the proof of Lemma 4.2, an iteration inequality of the form (4.2) will show up after we apply the local Łojasiewicz condition (2.2), set up suitable n_0, and use the boundedness of the noise (2.9). The power q comes from (4.1), and the extra power p comes from η_k^2.

Lemma 4.2. *Given a radius $r > 0$ where (2.2) holds. Assume that the noise term in (2.6) satisfies (2.9), and the SGD iterates are updated with step sizes $\eta_k = \gamma/(k+n_0)^q, 1/2 < q \leq 1$. Conditioned on the event that $\theta_k \in B(\theta_0, r)$ for all $k \in \mathbb{N}$, then for $n_0 \geq (2C_L^2\gamma/\alpha)^{1/q}$, we have the convergence*

$$\mathbb{E}[F(\theta_k)] \leq \begin{cases} \dfrac{\gamma^2 c^2 C_L}{\alpha\gamma - 2}\dfrac{1}{k} + o\left(\dfrac{1}{k}\right), & q = 1, \quad \gamma > \dfrac{2}{\alpha}, \\ \dfrac{\gamma^2 c^2 C_L}{\alpha\gamma}\dfrac{1}{k^q} + o\left(\dfrac{1}{k^q}\right), & \dfrac{1}{2} < q < 1. \end{cases} \tag{4.5}$$

Proof. We consider the same event as before

$$E_k(r) := \bigcap_{j=0}^{k} \{|\theta_j - \theta_0| \leq r\} \tag{4.6}$$

for some $0 < r \leq R$. Taking similar beginning steps as in Theorem 3.1, we arrive to the same inequality as (3.7)

$$F(\theta_{k+1})\mathbb{1}_{E_{k+1}(r)} \leq \left(F(\theta_k) - \eta_k \nabla F(\theta_k)\cdot \nabla f(\theta_k;\xi_k) + \frac{\eta_k^2}{2} C_L |\nabla f(\theta_k;\xi_k)|^2\right)\mathbb{1}_{E_k(r)}. \tag{4.7}$$

By inserting $\nabla f(\theta;\xi) = \nabla F(\theta) + Z(\theta;\xi)$, we expand the right side to be

$$\begin{aligned} & F(\theta_{k+1})\mathbb{1}_{E_{k+1}(r)} \\ \leq & \Big(F(\theta_k) - \eta_k|\nabla F(\theta_k)|^2 + \big(\eta_k^2 C_L - \eta_k\big)\nabla F(\theta_k)\cdot Z(\theta_k;\xi_k) \\ & \quad + \frac{\eta_k^2}{2} C_L\big(|\nabla F(\theta_k;\xi_k)|^2 + |Z(\theta_k;\xi_k)|^2\big)\Big)\mathbb{1}_{E_k(r)}. \end{aligned} \tag{4.8}$$

Because $Z(\theta_k,\xi_k)$ has zero mean and its second moment is bounded by c^2 (2.9), in addition with (2.4), by taking the expectation we have that

$$\mathbb{E}[F(\theta_{k+1})\mathbb{1}_{E_{k+1}(r)}|\theta_0] \leq \big(1 - \alpha\eta_k + \eta_k^2 C_L^2\big)\mathbb{E}[F(\theta_k)\mathbb{1}_{E_k(r)}|\theta_0] + \frac{\eta_k^2 c^2 C_L}{2}. \tag{4.9}$$

We may view $\mathbb{E}[F(\theta_k)\mathbb{1}_{E_k(r)}|\theta_0]$ as b_k in (4.2). We may set n_0 to satisfy

$$n_0 \geq \left(\frac{2C_L^2\gamma}{\alpha}\right)^{\frac{1}{q}}, \tag{4.10}$$

so that

$$1 - \alpha\eta_k + \eta_k^2 C_L^2 \leq 1 - \frac{\alpha\eta_k}{2} = 1 - \frac{\alpha\gamma}{2(k+n_0)^q}. \tag{4.11}$$

Note that the higher order term in (4.9) has the expression

$$\frac{\eta_k^2 c^2 C_L}{2} = \frac{\gamma^2 c^2 C_L}{2(k+n_0)^{2q}}. \tag{4.12}$$

Thus, we apply Lemma 4.1 and obtain the convergence

1. If $q = 1$, we set $\gamma > 2/\alpha$ so that

$$\mathbb{E}\big[F(\theta_k)\mathbb{1}_{E_k(r)}|\theta_0\big] \leq \frac{\gamma^2 c^2 C_L}{\alpha\gamma - 2}\frac{1}{k} + o\left(\frac{1}{k}\right). \tag{4.13}$$

2. If $1/2 < q < 1$, we have

$$\mathbb{E}\big[F(\theta_k)\mathbb{1}_{E_k(r)}|\theta_0\big] \leq \frac{\gamma^2 c^2 C_L}{\alpha\gamma}\frac{1}{k^q} + o\left(\frac{1}{k^q}\right). \tag{4.14}$$

The proof is complete. □

Remark 4.1. Compared with the proof of Theorem 3.1, the decay rate (4.5) of $F(\theta_k)$ is not enough to ensure that $E_\infty(r)$ happens with a positive probability. In fact, we need both probabilities in (3.14) and (3.15) to be summable.

In [28], the authors assume a local convexity in a convex neighborhood $\mathcal{K}$ of a local minimizer x^*, so that there exists $0 < \alpha < \beta < \infty$ and

$$\alpha|x - x^*|^2 \leq \nabla F(x)^\top (x - x^*) \leq \beta|x - x^*|^2, \quad \forall\, x \in \mathcal{K}.$$

This ensures a strong contraction of $|x_k - x^*|^2$ through iterations. Just with bounded noises, they can show $\{x_k\}$ stays inside the neighborhood $\mathcal{K}$ with a positive probability under Robbins-Monro step sizes. Unfortunately, the local Łojasiewicz condition (2.2) we assume here is not sufficient to guarantee such a contraction.

In fact, we can show that the boundedness of noises is not enough to guarantee that iterates stay inside a ball all the time under the local Łojasiewicz condition (2.2). Let us present a negative result in the following. By constructing additive noise $\mathcal{O}_k Z(\theta_k; \xi) \equiv \mathcal{O}_k Z_k$ such that $\mathbb{E}[Z_k] = 0$ and $\mathbb{E}[|Z_k|] = \bar{m} > 0$ for $k \in \mathbb{N}$ with orthogonal matrices $\mathcal{O}_k$ so that all $\mathcal{O}_k Z_k$ have the same direction, we find that under step sizes (4.1), iterates will escape the ball $B(\theta_0, r)$ almost surely.

Theorem 4.1. *Consider the following stochastic gradient iteration:*

$$\theta_{k+1} = \theta_k - \eta_k\big(\nabla F(\theta_k) + \mathcal{O}_k Z_k\big), \tag{4.15}$$

where $\eta_k = \gamma/(k + n_0)^q, 1/2 < q \leq 1, n_0 \geq (2C_L^2\gamma/\alpha)^{1/q}$ and $\gamma > 2/\alpha$ as in Lemma 4.2. We assume that $Z_k \in \mathbb{R}^d$ are i.i.d. random vectors with $\mathbb{E}[Z_k] = 0, \mathbb{E}[|Z_k|] = \bar{m}$ for some $\bar{m} > 0$, and $\mathbb{E}[|Z_k|^2] \leq c^2$ for some $c > 0$. Furthermore, $\mathcal{O}_k \in \mathbb{R}^{d\times d}$ are orthogonal matrices which rotate Z_k to the direction of Z_1, and we set $O_0 = I$. Then, given a radius $r > 0$ such that (2.2) holds, $\{\theta_k\}_{k\geq 0}$ will exit the ball $B(\theta_0, r)$ almost surely.

Proof. We argue by contradiction. Suppose all iterates from (4.15) stay inside the ball $B(\theta_0, r)$, we aim to prove that

$$\left|\lim_{n\to\infty}\sum_{k=0}^{n}(\theta_{k+1} - \theta_k)\right| = \left|\lim_{n\to\infty}\theta_{n+1} - \theta_0\right| = \infty$$

almost surely. For any $n \geq 1$, we have

$$\begin{aligned} &\left| \sum_{k=0}^{n} \left(\theta_{k+1} - \theta_k - \eta_k \nabla F(\theta_k) \right) \right| \\ &= \left| \sum_{k=0}^{n} \eta_k \mathcal{O}_k Z_k \right| = \sum_{k=0}^{n} \eta_k |Z_k| \\ &\geq \sum_{k=0}^{n} \frac{\gamma}{k+n_0} |Z_k| \geq \tilde{c} \sum_{k=1}^{n} \frac{1}{k} |Z_k|, \end{aligned} \tag{4.16}$$

where the second inequality can be satisfied if we choose $\tilde{c} \in (0, \gamma/(n_0+1))$. We claim that

$$\sum_{k=1}^{n} \frac{1}{k} |Z_k| \;\; \to \;\; \infty \quad \text{as} \quad n \;\; \to \;\; \infty \quad \text{almost surely.} \tag{4.17}$$

If not, say there exists a finite number $\mu > 0$ such that

$$\lim_{n \to \infty} \sum_{k=1}^{n} \frac{1}{k} |Z_k| = \mu,$$

the Cesàro mean theorem implies that the Cesàro average also converges to the same limit,

$$\frac{1}{n} \sum_{j=1}^{n} \left(\sum_{k=1}^{j} \frac{1}{k} |Z_k| \right) \;\; \to \;\; \mu \quad \text{as} \quad n \;\; \to \;\; \infty \quad \text{almost surely.} \tag{4.18}$$

On the other hand,

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^{n} \left(\sum_{k=1}^{j} \frac{1}{k} |Z_k| \right) &= \frac{1}{n} \sum_{k=1}^{n} \left(\sum_{j=k}^{n} \frac{1}{k} |Z_k| \right) = \frac{1}{n} \sum_{k=1}^{n} \frac{n+1-k}{k} |Z_k| \\ &= \sum_{k=1}^{n} \frac{1}{k} |Z_k| + \frac{1}{n} \sum_{k=1}^{n} \frac{1}{k} |Z_k| - \frac{1}{n} \sum_{k=1}^{n} |Z_k|, \end{aligned} \tag{4.19}$$

which implies that as $n \to \infty$,

$$\frac{1}{n} \sum_{k=1}^{n} |Z_k| \;\; \to \;\; 0$$

almost surely. But by the law of large number and the construction of Z_k,

$$\frac{1}{n} \sum_{k=1}^{n} |Z_k| \;\; \to \;\; \bar{m} \neq 0,$$

we get the contradiction. Thus (4.17) is proved.

Back to (4.16), we note that

$$\left| \sum_{k=0}^{n} \left(\theta_{k+1} - \theta_k - \eta_k \nabla F(\theta_k)\right) \right| \leq \left| \sum_{k=0}^{n} (\theta_{k+1} - \theta_k) \right| + \sum_{k=0}^{n} \eta_k |\nabla F(\theta_k)|. \tag{4.20}$$

Due to the convergence result (4.5) in Lemma 4.2, we have

$$\lim_{n\to\infty} \sum_{k=1}^{n} \eta_k |\nabla F(\theta_k)| \leq \lim_{n\to\infty} \sum_{k=1}^{n} \frac{\gamma\sqrt{2C_L}}{(k+n_0)^q} \sqrt{F(\theta_k)} < \infty \tag{4.21}$$

almost surely. Combining (4.17), (4.20) and (4.21) together, we can deduce from (4.16) that

$$\lim_{n\to\infty} \left| \sum_{k=0}^{n} (\theta_{k+1} - \theta_k) \right| = \left| \lim_{n\to\infty} \theta_{n+1} - \theta_0 \right| = \infty$$

almost surely, which means that the iterates of the SGD algorithm (4.15) will exit the ball almost surely. □

Remark 4.2. This negative argument can be extended to other adaptive learning rate schedules η_k, provided the proposed schedule satisfies both (4.16) and (4.21). It is crucial to note that the convergence described in (4.21) depends on Lemma 4.2 and the auxiliary Lemma 4.1. Demonstrating non-convergence for a specific learning rate schedule η_k would require significant modifications to Lemmas 4.1 and 4.2 to ensure that η_k aligns with these conditions.

Appendix A Assumption 3 verification

Chatterjee [8] provides a feedforward neural network that satisfies the local Łojasiewicz condition (2.2). The purpose of this appendix is to show that, by restricting the training parameter θ in some subspace, this feedforward neural network construction satisfies Assumption 3 (2.5) as well.

The loss function that [8] considers is the squared error loss. For the dateset $\{(x_i, y_i)\}_{i=1}^n$, we consider

$$F(\theta) := \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \phi(x_i, \theta)\right)^2. \tag{A.1}$$

Here, $\phi(x, \theta)$ is the neural network written as

$$\phi(x, \theta) := \tilde{\sigma}_L\Big(W_L \tilde{\sigma}_{L-1}\big(\cdots\big(W_2\tilde{\sigma}_1(W_1 x + b_1) + b_2\big)\cdots\big) + b_L\Big) \tag{A.2}$$

with $\tilde{\sigma}_1 \cdots, \tilde{\sigma}_L$ being activation functions acting componentwisely, and $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$, $d_0 = d, d_L = 1, b_l \in \mathbb{R}^{d_l}$ for $1 \leq l \leq L$. The parameter to be trained is

$$\theta = (W_1, b_1, W_2, b_2, \cdots, W_L, b_L), \tag{A.3}$$

and it can be viewed as a vector in $\mathbb{R}^p$ with

$$p = \sum_{l=1}^{L} d_l(d_{l-1}+1).$$

We assume that the input data $x_1, \cdots, x_n \in \mathbb{R}^d$ are linearly independent, and consider $X^\top X/n$, where the matrix $X = (x_1, \cdots x_n) \in \mathbb{R}^{d\times n}$ is formed by column vectors x_i's. When $d \geq n$, the matrix $X^\top X/n$ has a positive minimum eigenvalue denoted by $\lambda_0 > 0$.

[8, Theorem 2.1] shows that, with an appropriate initialization and neural network setups, the landscape of $F(\theta)$ satisfies the local Łojasiewicz condition (2.2), and the gradient descent will converge to $F(\theta^*) = 0$ by the convergence analysis. However, the setup in [8] does not exclude the case of (globally) flat minima, where Assumption 3 (2.5) may fail. In order to allow assumptions (2.2) and (2.5) to coexist, we restrict the domain of training θ to some subspace containing θ_0 and θ^* where Assumption 3 (2.5) holds.

One choice for such subspace is the following: fix some $\tilde{\alpha} > 0$,

$$\Theta = \big\{\theta' \in \mathbb{R}^p : (W_1' x_i + b_1')_j \geq \tilde{\alpha}|\theta' - \theta_0|,\ 1 \leq i \leq n,\ 1 \leq j \leq d_1, \\ \text{and}\ \ b_l' \geq 0,\ 2 \leq l \leq L\big\}. \tag{A.4}$$

Such a $\tilde{\alpha} > 0$ can exist if we have a suitable data matrix X. The setup of Θ is to avoid training θ' in the nullspace of $\phi(x, \cdot)$. Θ looks somewhat artificial in order to serve the computational convenience (see Theorem A.1). In practice during the training for deep neural networks, all gradient descent iterates can be considered to stay in Θ, as otherwise the gradient becomes small and training would terminate. We leave the following remark as one example of such suitable neural networks.

Remark A.1. The feedforward neural networks in [8] include deep linear neural networks. Taking biases to be zeroes and activation functions to be identity, it writes as

$$\tilde{\phi}(\theta, x) := W_L W_{L-1} \cdots W_1 x. \tag{A.5}$$

The partial derivative of $F(\theta)$ with respect to W_l, for each $2 \leq l \leq L$ is given as

$$\frac{\partial F}{dW_l} = \frac{2}{n} \sum_{i=1}^{n} W_{l+1}^\top \cdots W_L^\top (W_L W_{L-1} \cdots W_1 x_i - y_i) x_i^\top W_1^\top \cdots W_{l-1}^\top, \tag{A.6}$$

and when $l = L$, we take $W_{L+1}^\top W_L^\top$ as an identity matrix by convention (so the above formula applies to every l). Then with the initialization in Theorem A.1, we see that $\partial F/dW_l$ will be small if $W_1 x_i$ is small of all $1 \leq i \leq n$. When it happens, we may stop training due to the vanishing ∇F.

Let us recall the setups in [8, Theorem 2.1] and show that the loss function $F(\theta)$ satisfies Assumption 3 (2.5) restricted in Θ.

Theorem A.1. *We consider the squared error loss* (A.1) *with a feedforward neural network* (A.2), *with depth* $L \geq 2$, $d_L = 1$ *and* $\tilde{\sigma}_L =$ *identity. Suppose for* $0 \leq l \leq L-1$, *the activation functions*

$\tilde{\sigma}_l \in C^2, \tilde{\sigma}_l(0) = 0$, and $c_l := \min_x \tilde{\sigma}_l'(x) > 0$. Suppose the input data $x_1, \cdots, x_n \in \mathbb{R}^d$ are linearly independent, and let λ_0 be the minimum eigenvalue of $X^\top X/n$, with the matrix $X = (x_1, \cdots x_n) \in \mathbb{R}^{d\times n}$ formed by column vectors x_i's. We can initialize

$$\theta_0 = (W_1, b_1, W_2, b_2, \cdots W_L, b_L)$$

such that $b_l = 0$ for all $1 \le l \le L$, $W_1 = 0$, and the entries of $W_2, \cdots, W_{L-1}$ are all strictly positive. In addition, let $R > 0$ be the minimum value of the entries of $W_2, \cdots, W_{L-1}$, and $A > R/2 > 0$ be the minimum value of the entries of W_L. Then for any $\theta' \in B(\theta_0, R/2) \cap \Theta$, we have the following lower bound:

$$F(\theta') \ge \frac{\tilde{\alpha}^2}{2}\left(A - \frac{R}{2}\right)^2 \left(\frac{R}{2}\right)^{2L-4} (c_{L-1}\cdots c_2 c_1 d_{L-1}\cdots d_1)^2 |\theta' - \theta_0|^2 - \frac{1}{n}\sum_{i=1}^n y_i^2. \quad \text{(A.7)}$$

Proof. We define

$$\phi_1(x, \theta) = \tilde{\sigma}_1(W_1 x + b_1), \quad \text{(A.8)}$$

and for $2 \le l \le L$,

$$\phi_l(x, \theta) = \tilde{\sigma}_l\Big(W_l \tilde{\sigma}_{l-1}\big(\cdots\big(W_2\tilde{\sigma}_1(W_1 x + b_1) + b_2\big)\cdots\big) + b_l\Big), \quad \text{(A.9)}$$

so that $\phi_L = \phi$.

Let θ_0 be the starting vector where the entries of $W_2, \cdots W_{L-1}$ are all strictly positive, and $b_1, \cdots, b_L$ and W_1 be zero, then one can find $\phi(x, \theta_0) = 0$ for each x. Thus

$$F(\theta_0) = \frac{1}{n}\sum_{i=1}^n y_i^2.$$

Using the Young's inequality, the loss function can be rewritten as

$$F(\theta') \ge \frac{1}{n}\sum_{i=1}^n \big(y_i^2 + \phi(x_i, \theta')^2 - 2|\phi(x_i, \theta')|\,|y_i|\big) \ge \frac{1}{2n}\sum_{i=1}^n \phi(x_i, \theta')^2 - \frac{1}{n}\sum_{i=1}^n y_i^2. \quad \text{(A.10)}$$

For each x_i and $1 \le j \le d_1$, since $\theta' \in \Theta$, we have that

$$\phi_1(x_i, \theta')_j \ge c_1 \tilde{\alpha} |\theta' - \theta_0|. \quad \text{(A.11)}$$

Given the choice of θ_0, we note that all the entries of $W_2, \cdots, W_{L-1}$ are bounded below by a constant $R > 0$, and $b_1, \cdots, b_L$ and W_1 are zero. If we take any $\theta' \in \Theta$ such that $|\theta' - \theta_0| \le R/2$, then for such

$$\theta' = \big(W_1', b_1', W_2', b_2', \cdots W_L', b_L'\big) \in \Theta,$$

all the entries of $W_2', \cdots, W_{L-1}'$ are bounded below by $R/2$, and $b_2', \cdots, b_L' \ge 0$. In the second layer, for $1 \le j \le d_2$, we have that

$$\phi_2(x_i, \theta')_j \ge c_2 c_1 d_1 \frac{R}{2}\tilde{\alpha}|\theta' - \theta_0|. \quad \text{(A.12)}$$

Iterate it up to $(L-1)$-th layer, we get that for $1 \le j \le d_{L-1}$,

$$\phi_{L-1}(x_i, \theta')_j \ge c_{L-1} \cdots c_2 c_1 d_{L-2} \cdots d_1 \left(\frac{R}{2}\right)^{L-2} \tilde{\alpha} |\theta' - \theta_0|. \tag{A.13}$$

In the last layer, as $A > R/2$ is a lower bound on the entries of W_L, the entries of W_L' are bounded below by $A - R/2$. Therefore,

$$\phi(x_i, \theta') = \phi_L(x_i, \theta') \ge \left(A - \frac{R}{2}\right) \left(\frac{R}{2}\right)^{L-2} c_{L-1} \cdots c_2 c_1 d_{L-1} \cdots d_1 \tilde{\alpha} |\theta' - \theta_0|, \tag{A.14}$$

and we can bound $F(\theta')$ by

$$F(\theta') \ge \frac{\tilde{\alpha}^2}{2} \left(A - \frac{R}{2}\right)^2 \left(\frac{R}{2}\right)^{2L-4} (c_{L-1} \cdots c_2 c_1 d_{L-1} \cdots d_1)^2 |\theta' - \theta_0|^2 - \frac{1}{n} \sum_{i=1}^{n} y_i^2. \tag{A.15}$$

The proof is complete. □

From the lower bound estimate above, if R is large enough compared with $\sum_{i=1}^n y_i^2/n$, we can find $M_0 > 0$ such that $F(\theta') \ge M_0$ for

$$\frac{R}{2} - 1 \le |\theta' - \theta_0| \le \frac{R}{2}.$$

Appendix B Proof of Lemma 2.1

Proof. The statement is trivially true of $\nabla F(\theta) = 0$, so we only consider the case where $\nabla F(\theta) \neq 0$. Let us define a function

$$g(t) = F(\theta - t\nu), \quad \nu = \frac{\nabla F(\theta)}{|\nabla F(\theta)|}, \tag{B.1}$$

then $g'(0) = -\nu \cdot \nabla F(\theta) = -|\nabla F(\theta)|$ and

$$|g'(t) - g'(0)| \le |\nabla F(\theta - t\nu) - \nabla F(\theta)| \le C_L t, \tag{B.2}$$

if $\theta, \theta - t\nu \in \mathcal{K}$ based on (2.3). With that,

$$g(t) = g(0) + \int_0^t g'(s) ds \le F(\theta) - |\nabla F(\theta)| t + \frac{C_L}{2} t^2. \tag{B.3}$$

By taking $t = -|\nabla F(\theta)|/C_L$, the right side bound achieves the minimum and it implies that

$$\text{RHS} = F(\theta) - \frac{|\nabla F(\theta)|^2}{2C_L} \ge 0. \tag{B.4}$$

Because we require both $\theta, \theta - t\nu \in \mathcal{K}$, due to the choice of t, we may choose a compact set $\tilde{\mathcal{K}} \subset \mathcal{K}$, with $\text{dist}(\partial\mathcal{K}, \partial\tilde{\mathcal{K}}) \ge \bar{C}/C_L$, so that $\theta, \theta - t\nu \in \mathcal{K}$ for all $\theta \in \tilde{\mathcal{K}}$. □

Acknowledgments

The work of J. Lu is supported in part by the US National Science Foundation (Grant No. DMS-2012286).

References

[1] Z. Allen-Zhu, How to make the gradients small stochastically: Even faster convex and nonconvex SGD, in: *Advances in Neural Information Processing Systems*, Vol. 31, Curran Associates, Inc., 1157–1167, 2018.

[2] Z. Allen-Zhu, Natasha 2: Faster non-convex optimization than SGD, in: *Advances in Neural Information Processing Systems*, Vol. 31, Curran Associates, Inc., 2675–2686, 2018.

[3] Z. Allen-Zhu, Y. Li, and Z. Song, A convergence theory for deep learning via over-parameterization, in: *International Conference on Machine Learning*, PMLR, 242–252, 2019.

[4] F. Bach and E. Moulines, Non-asymptotic analysis of stochastic approximation algorithms for machine learning, in: *Advances in Neural Information Processing Systems*, Vol. 24, Curran Associates, Inc., 451–459, 2011.

[5] M. Benaïm, Dynamics of stochastic approximation algorithms, in: *Seminaire de Probabilites XXXIII. Lecture Notes in Mathematics*, Vol. 1709, Springer, 1–68, 2006.

[6] D. P. Bertsekas, Incremental gradient, subgradient, and proximal methods for convex optimization: A survey, in: *Optimization for Machine Learning*, 85–115, 2012.

[7] L. Bottou, F. E. Curtis, and J. Nocedal, Optimization methods for large-scale machine learning, *SIAM Rev.*, 60(2):223–311, 2018.

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

[9] L. Chizat and F. Bach, On the global convergence of gradient descent for over-parameterized models using optimal transport, in: *Advances in Neural Information Processing Systems*, Vol. 31, Curran Associates, Inc., 3036–3046, 2018.

[10] L. Chizat, E. Oyallon, and F. Bach, On lazy training in differentiable programming, in: *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, Inc., 2914–2924, 2019.

[11] K. L. Chung, On a stochastic approximation method, *Ann. Math. Statist.*, 25(3):463–483, 1954.

[12] A. Cutkosky and F. Orabona, Momentum-based variance reduction in non-convex SGD, in: *Advances in Neural Information Processing Systems*, Vol. 32, Curran Associates, Inc., 15157–15166, 2019.

[13] C. De Sa, S. Kale, J. D. Lee, A. Sekhari, and K. Sridharan, From gradient flow on population loss to learning with stochastic gradient descent, in: *Advances in Neural Information Processing Systems*, Vol. 35, Curran Associates, Inc., 30963–30976, 2022.

[14] J. C. Duchi and F. Ruan, Stochastic methods for composite and weakly convex optimization problems, *SIAM J. Optim.*, 28(4):3229–3259, 2018.

[15] B. Fehrman, B. Gess, and A. Jentzen, Convergence rates for the stochastic gradient descent method for non-convex objective functions, *J. Mach. Learn. Res.*, 21(1):5354–5401, 2020.

[16] S. Ghadimi and G. Lan, Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.*, 156(1-2):59–99, 2016.

[17] R. Gower, O. Sebbouh, and N. Loizou, SGD for structured nonconvex functions: Learning rates, minibatching and interpolation, in: *International Conference on Artificial Intelligence and Statistics*, PMLR, 1315–1323, 2021.

[18] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik, SGD: General analysis and improved rates, in: *International Conference on Machine Learning*, PMLR, 5200–5209, 2019.

[19] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, Accurate, large minibatch SGD: Training imagenet in 1 hour, *arXiv:1706.02677*, 2017.

[20] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in: *Advances in Neural Information Processing Systems*, Vol. 31, Curran Associates, Inc., 8571–8580, 2018.

[21] A. Jentzen and A. Riekert, On the existence of global minima and convergence analyses for gradient descent methods in the training of deep neural networks, *arXiv:2112.09684*, 2021.
[22] H. Karimi, J. Nutini, and M. Schmidt, Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition, in: *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2016*, Vol. 9851, Springer, 795–811, 2016.
[23] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, in: *Stochastic Modelling and Applied Probability*, Vol. 35, Springer, 2003.
[24] L. Lei, C. Ju, J. Chen, and M. I. Jordan, Non-convex finite-sum optimization via SCSG methods, in: *Advances in Neural Information Processing Systems*, Vol. 30, Curran Associates, Inc., 2349–2359, 2017.
[25] C. Liu, L. Zhu, and M. Belkin, Loss landscapes and optimization in over-parameterized non-linear systems and neural networks, *Appl. Comput. Harmon. Anal.*, 59:85–116, 2022.
[26] F. Liu, H. Yang, S. Hayou, and Q. Li, From optimization dynamics to generalization bounds via Łojasiewicz gradient inequality, *Transactions on Machine Learning Research*, 2022. `https://openreview.net/forum?id=mW6nD3567x`
[27] S. Mei, A. Montanari, and P.-M. Nguyen, A mean field view of the landscape of two-layer neural networks, *Proc. Natl. Acad. Sci. USA*, 115(33):E7665–E7671, 2018.
[28] P. Mertikopoulos, N. Hallak, A. Kavis, and V. Cevher, On the almost sure convergence of stochastic gradient descent in non-convex problems, in *Advances in Neural Information Processing Systems*, Vol. 33, Curran Associates, Inc., 1117–1128, 2020.
[29] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, Robust stochastic approximation approach to stochastic programming, *SIAM J. Optim.*, 19(4):1574–1609, 2009.
[30] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, in: *Applied Optimization*, Vol. 87, Springer, 2003.
[31] B. Neyshabur, R. R. Salakhutdinov, and N. Srebro, Path-SGD: Path-normalized optimization in deep neural networks, in: *Advances in Neural Information Processing Systems*, Vol. 28, Curran Associates, Inc., 2422–2430, 2015.
[32] H. T. Pham and P.-M. Nguyen, Global convergence of three-layer neural networks in the mean field regime, in: *Proceedings of the 9th International Conference on Learning Representations* , ICLR, 2021.
[33] B. T. Polyak, *Introduction to Optimization*, Optimization Software, Publications Division, 1987.
[34] S. J. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola, Stochastic variance reduction for nonconvex optimization, in: *International Conference on Machine Learning*, PMLR, 314–323, 2016.
[35] H. Robbins and S. Monro, A stochastic approximation method, *Ann. Math. Stat.*, 22(3):400–407, 1951.
[36] M. Soltanolkotabi, A. Javanmard, and J. D. Lee, Theoretical insights into the optimization landscape of over-parameterized shallow neural networks, *IEEE Trans. Inf. Theory*, 65(2):742–769, 2018.
[37] S. U. Stich, Unified optimal analysis of the (stochastic) gradient method, *arXiv:1907.04232*, 2019.
[38] R. Ward, X. Wu, and L. Bottou, Adagrad stepsizes: Sharp convergence over nonconvex landscapes, *J. Mach. Learn. Res.*, 21(1):9047–9076, 2020.
[39] S. Wojtowytsch, Stochastic gradient descent with noise of machine learning type. Part I: Discrete time analysis, *J. Nonlinear Sci.*, 33(3):45, 2023.
[40] D. Zou, Y. Cao, D. Zhou, and Q. Gu, Gradient descent optimizes over-parameterized deep ReLU networks, *Mach. Learn.*, 109:467–492, 2020.