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 $\nabla = \nabla_\theta$.