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

Convergence of Stochastic Gradient Descent Schemes for Łojasiewicz-Landscapes

Steffen Dereich * [1] and Sebastian Kassing † [2]

[1]Institute for Mathematical Stochastics, Faculty of Mathematics and Computer Science, University of Münster, Germany
[2]Faculty of Mathematics, University of Bielefeld, Germany

Abstract. In this article, we consider convergence of stochastic gradient descent schemes (SGD), including momentum stochastic gradient descent (MSGD), under weak assumptions on the underlying landscape. More explicitly, we show that on the event that the SGD stays bounded we have convergence of the SGD if there is only a countable number of critical points or if the objective function satisfies Łojasiewicz-inequalities around all critical levels as all analytic functions do. In particular, we show that for neural networks with analytic activation function such as softplus, sigmoid and the hyperbolic tangent, SGD converges on the event of staying bounded, if the random variables modelling the signal and response in the training are compactly supported.

Keywords:
Stochastic gradient descent,
Stochastic approximation,
Robbins-Monro,
Almost sure convergence,
Łojasiewicz-inequality.

Article Info.:
Volume: 3
Number: 3
Pages: 245 - 281
Date: September/2024
doi.org/10.4208/jml.240109

Article History:
Received: 09/01/2024
Accepted: 14/06/2024

Communicated by:
Arnulf Jentzen

1 Introduction

In this article, we analyse stochastic gradient descent schemes for C^1-objective functions $f : \mathbb{R}^d \to \mathbb{R}$ with $d \in \mathbb{N} := \{1, 2, \ldots\}$ being an arbitrary dimension. We denote by $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\in\mathbb{N}_0}, \mathbb{P})$ a filtered probability space and consider an $\mathbb{R}^d$-valued stochastic process $(X_n)_{n\in\mathbb{N}_0}$ that admits a representation

$$X_n = X_{n-1} + \gamma_n(\Gamma_n + D_n) \tag{1.1}$$

for $n \in \mathbb{N}$, where

- $(\gamma_n)_{n\in\mathbb{N}}$ is a sequence of strictly positive reals, the step-sizes or learning rates.
- $(\Gamma_n)_{n\in\mathbb{N}}$ is an $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$-predictable sequence of random variables, the drift.
- $(D_n)_{n\in\mathbb{N}}$ is an $(\mathcal{F}_n)_{n\in\mathbb{N}}$-adapted sequence of random variables, the perturbation.
- X_0 is an $\mathcal{F}_0$-measurable random variable, the initial value.

*steffen.dereich@uni-muenster.de
†Corresponding author. skassing@math.uni-bielefeld.de

The choice $(\Gamma_n)_{n\in\mathbb{N}} = (-\nabla f(X_{n-1}))_{n\in\mathbb{N}}$ leads to a standard representation of stochastic gradient descent. However, our results hold for a more general class of dynamical systems, including momentum stochastic gradient descent (see Section 2), assuming the drift is comparable in size and direction to the gradient vector field (the precise condition is given in Definition 1.1). Additional assumptions will be imposed in the theorems below.

Stochastic gradient descent schemes form a subclass of Robbins-Monro schemes which were introduced in 1951 [63] and have been highly influential since then. Their relevance stems from their applicability of finding zeros of functions $F : \mathbb{R}^d \to \mathbb{R}^d$ in the case where one has only simulations at hand which give approximations to the value of F in mean. Following the original papers a variety of results were derived, and we refer the reader to the mathematical accounts [9, 23, 41] on stochastic approximation methods.

In this article, we analyse convergence of $(X_n)_{n\in\mathbb{N}_0}$. This problem is intimately related to understanding the asymptotic behaviour of $(\nabla f(X_n))_{n\in\mathbb{N}_0}$. The classical analysis of Polyak and Tsypkin [60] yields existence of the limit $\lim_{n\to\infty} f(X_n)$ and proves $\liminf_{n\to\infty} |\nabla f(X_n)| = 0$ under appropriate assumptions. Later, Walk [70] showed that in an appropriate setting one has almost sure convergence $\lim_{n\to\infty} \nabla f(X_n) = 0$. Similar results were established in various settings, see [10, 26, 30, 43, 47, 49].

We will provide a short proof for $\lim_{n\to\infty} \nabla f(X_n) = 0$ under weak assumptions. Here, we include the case where ∇f is only Hölder continuous and where $(D_n)_{n\in\mathbb{N}}$ is an L^p-martingale difference sequence for a $p \in (1, 2]$. Then, we will conclude that we have almost sure convergence of $(X_n)_{n\in\mathbb{N}_0}$ on the event that $(X_n)_{n\in\mathbb{N}_0}$ stays bounded in the case where the set of critical points of f does not contain a continuum, see Theorem 1.1 below.

In the case where the set of critical points of f contains a continuum of points the situation is more subtle. In that case, Tadic [68] showed that under a Łojasiewicz-inequality stochastic gradient descent schemes converge under appropriate additional assumptions. In this article, our considerations are also based on the validity of certain Łojasiewicz-inequalities. However, here we allow the drift to be more general so that the new convergence theorem is applicable for a bigger class of optimisation methods such as momentum stochastic gradient descent (MSGD). We stress that the proofs developed in this article are significantly different from the ones in [68]. We mention that the asymptotic behaviour of a stochastic gradient descent scheme is tightly related to that of the first order differential equation

$$\dot{x}_t = -\nabla f(x_t) \tag{1.2}$$

(though many approximation results only hold for a finite time-horizon), see e.g. [50, Proposition 1]. Convergence for the latter differential equation is a non-trivial issue even in the case where the solution stays on a compact set: One can find C^∞-functions f together with solutions $(x_t)_{t\geq 0}$ that stay on compact sets but do not converge, see [54, Example 3, page 14]. Counterexamples of this structure have been known for a long time (see e.g. [16]) and include the famous Mexican hat function [1]. To guarantee convergence (at least in the case where the solution stays on a compact set), one needs to impose additional assumptions. An appropriate assumption is the validity of a Łojasiewicz-inequality, see Definition 1.2 below. This assumption has the appeal that it is satisfied by analytic functions, see [45, 46].