Stochastic convex optimization from a continuous dynamical system perspective

Rodrigo MAULEN S.[1], Jalal FADILI[1], Hedy ATTOUCH[2],

[1]Normandie Université, ENSICAEN, UNICAEN, CNRS, GREYC
6 Boulevard Marechal Juin, 14000 Caen, France

[2]Institut Montpellierain Alexander Grothendieck, CNRS, Université Montpellier
641 Avenue du Doyen Gaston Giraud, 34000 Montpellier, France

rodrigo.maulen@ensicaen.fr, jalal.fadili@ensicaen.fr,hedy.attouch@umontpellier.fr

Résumé – Afin de résoudre des problèmes d'optimisation non contraints avec des objectives différentiables et convexes, et où le gradient est soumis à des erreurs, nous analysons le comportement des flots de type gradient sous des perturbations stochastiques. Plus précisément, nous étudions une équation différentielle stochastique où le terme de dérivé est l'opposé du gradient de la fonction objective, et le terme de diffusion est borné ou carré-intégrable. Dans ce contexte, sous des conditions de Lipschitz continuité du gradient, en plus d'assurer l'existence standard et l'unicité d'une solution, un premier résultat principal montre la convergence presque sûre de l'objectif et de la solution/processus vers un minimiseur. Nous menons ensuite une étude de complexité et établissons des taux de convergence ponctuels et ergodiques en espérance lorsque l'objective est convexe, fortement convexe ou vérifie (localement) l'inégalité de Polyak-Łojasiewicz. Cette dernière, qui implique une analyse locale, nécessite des arguments fins en théorie de la mesure.

Abstract – In order to solve differentiable and convex unconstrained optimization problems with a noisy (inexact) gradient input, we analyze the behavior of gradient-like flows under stochastic errors. More precisely, we study a stochastic differential equation where the drift term is minus the gradient of our objective function and the diffusion term is bounded or square-integrable. In this context, under Lipschitz continuity of the gradient, beside ensuring standard existence and uniqueness of a solution, our first main result shows almost sure convergence of the objective and the solution/process to a minimizer of the objective function. We also provide a comprehensive complexity analysis by establishing several new pointwise and ergodic convergence rates in expectation for the convex, strongly convex and local Polyak-Łojasiewicz case. The latter, which involves a local analysis, is very challenging and necessitates non-trivial arguments from measure theory.

1 Introduction

1.1 Problem Statement

Consider the unconstrained convex problem

$$\min_{x\in\mathbb{R}^d} f(x), \tag{P}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ (called the potential) is a continuously differentiable convex function with gradient Lipschitz. We will assume that

$$\operatorname{argmin}(f) \neq \varnothing. \tag{H_0}$$

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ be a filtered probability space. Our goal in this paper is to get deeper understanding into local and global convergence properties of stochastic gradient descent (SGD) through the lens of stochastic differential equations. Toward this goal, we will consider the following stochastic dynamic, defined for (deterministic) initial data $X_0 \in \mathbb{R}^d$ as

$$\begin{aligned} dX(t) &= -\nabla f(X(t))dt + \sigma(t, X(t))dB(t),\ t \geq 0 \quad \text{(SDE)}\\ X(0) &= X_0, \end{aligned}$$

where:

1. B is a $\mathcal{F}_t$-adapted $m-$dimensional Brownian motion.
2. The $d \times m$ volatility matrix $\sigma_{ik} : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ is measurable and

$$\sup_{t,x} |\sigma_{ik}(t, x)| < +\infty,\ |\sigma_{ik}(t, x') - \sigma_{ik}(t, x)| \leq l_0\|x' - x\|, \tag{H}$$

for some $l_0 > 0$ and for all $t \geq 0, x, x' \in \mathbb{R}^d$.

1.2 Contributions

For a function $f : \mathbb{R}^d \to \mathbb{R}$ that is differentiable, convex, and has Lipschitz continuous gradient, we study (SDE) under hypotheses (H_0) and (H). Assuming that the diffusion term is uniformly bounded, we present upper bounds of the quantity $\mathbb{E}[f(X(t)) - \min(f)]$ for the convex and strongly convex case. Moreover, we analyze the case when the diffusion term is square-integrable, showing the almost sure convergence of the process defined by the algorithm (SDE) to the set of minimizers of (P) (see Theorem 3.2), a result that is new to the best of our knowledge. Besides, we show new asymptotic conver-

gence rates of $\mathbb{E}[f(X(t))-\min(f)]$ for the convex and strongly convex case (see Theorem 3.3). Furthermore, we show rigorously local convergence properties of the objective under the Polyak-Łojasiewicz inequality for the first time (see Theorem 3.6). To show this precisely, let $\delta > 0$ be sufficiently small, and consider (SDE) under hypothesis (H), then there exists $\sigma_*^2 > 0$ such that: $\|\sigma(t,x)\|_F^2 \leq \sigma_*^2, \ \forall t \geq 0, \forall x \in \mathbb{R}^d$. Moreover, denoting $\sigma_\infty(t) := \sup_{x\in\mathbb{R}^d} \|\sigma(t,x)\|_F$ and assuming it is decreasing, then the asymptotic order $\mathcal{O}(\cdot)$ of the convergence rate of the objective in expectation, $\mathbb{E}[f(X(t)) - \min(f)]$, is summarized in the following table:

Property	**SDE** ($\sigma_\infty \leq \sigma_*$)	**SDE** ($\sigma_\infty \in \mathrm{L}^2$)
Conv.	$t^{-1} + \sigma_*^2$	t^{-1}
$\mu-$Str. Conv.	$e^{-2\mu t} + \sigma_*^2$	$\max\{e^{-\mu t}, \sigma_\infty^2(t)\}$
Conv.+ PL_{loc}	✘	$\max\{e^{-\mu t}, \sigma_\infty^2(t)\} + \sqrt{\delta}$

Although it is natural to think that we can take the limit when δ goes to 0^+, the time from which these convergence rates are valid depends on δ and increases (potentially to $+\infty$) as δ approaches 0^+.

Assuming just the boundedness of the diffusion and the local PL Inequality, we could not find better results (marked with ✘) than those shown in the convex case. In that case, we would like to localize the process in the long term with high probability. However, at this stage, it is not clear how to accomplish so.

1.3 Relation to prior work

Let $f : \mathbb{R}^d \to \mathbb{R}$ such that $\operatorname{argmin} f \neq \varnothing$, the Gradient Flow:

$$\dot{x} = -\nabla f(x) \tag{GF}$$

is a transcendental dissipative system in convex optimization since it turns the problem of minimizing f into one of analyzing the behavior of a process in the long term. Its Euler forward discretization (with stepsize $\gamma_k > 0$) is the celebrated gradient descent scheme

$$x_{k+1} = x_k - \gamma_k \nabla f(x_k). \tag{GD}$$

If f has L-Lipschitz continuous gradient, and for $(\gamma_k)_{k\in\mathbb{N}} \subset]0, 2/L[$, then one can ensure that $f(x_k) - \min(f) = \mathcal{O}(1/t)$ (in fact even $o(1/t)$). and the convergence of the iterates $(x_k)_{k\in\mathbb{N}}$ to a point in $\operatorname{argmin} f$. Moreover, if a Łojasiewicz Inequality (see [16]) is satisfied, then we can ensure faster convergence rates compared to just assuming convexity (see [1, 2])

Although (GD) is a classical algorithm, with the need to handle large-scale problems (such as in various areas of data science and machine learning), there has become necessary to find ways to get around the high computational cost per iteration that these problems entail. More precisely, considering the empirical risk minimization problem

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x),$$

when n is large, it is prohibitive to compute the full gradient of the objective, and SGD provides an alternative based on a noisy gradient evaluated from a mini-batch $M \subset \{1, \dots, n\}$ (see [11]):

$$\tilde{\nabla} f(x) \overset{\text{def}}{=} \frac{1}{|M|}\sum_{i\in M} \nabla f_i(x) = \nabla f(x) + \xi,$$

where M is sampled uniformly at random from $\{1, \dots, n\}$, thus ξ has mean-zero. Given an initial point $x_0 \in \mathbb{R}^d$, (SGD) updates the iterates according to

$$x_{k+1} = x_k - \gamma_k \tilde{\nabla} f(x_k) = x_k - \gamma_k(\nabla f(x_k) + \xi_k), \tag{SGD}$$

where ξ_k denotes the noise term at the $k-$th iteration.

The dynamic (SDE) is well-studied for MC sampling where the volatility in the diffusion term is not allowed to vanish. Here, we are interested in an optimization perspective. In this respect, recent work (see [4, 5, 6, 9, 10, 13, 14]) has linked algorithm (SGD) with dynamic (SDE), showing the context under which (SDE) can be seen as an approximation (under a specific error) of (SGD) and vice-versa. However, many questions are still open, including the global convergence behavior of the trajectory, as well as global and local complexity bounds. It is our aim here to settle these questions.

2 Preliminaries

Let $f : \mathbb{R}^d \to \mathbb{R}$, then $[f \leq r] := \{x \in \mathbb{R}^d : f(x) \leq r\}$. We denote $\Gamma_0(\mathbb{R}^d)$ as the class of convex and lower semi-continuous (l.s.c.) functions defined from $\mathbb{R}^d$ to $\mathbb{R}$, moreover, $\Gamma_\mu(\mathbb{R}^d)$ is the class of $\mu-$strongly convex functions. We also denote $C_L^{1,1}(\mathbb{R}^d)$ to the continuously differentiable functions defined from $\mathbb{R}^d$ to $\mathbb{R}$ whose gradient is $L-$Lipschitz.

An event $A \in \mathcal{F}$ happens almost surely if $\mathbb{P}(A) = 1$, and it will be denoted as "$A, \ \mathbb{P}-a.s.$" or "$A, \ -a.s.$". Besides

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & if \ \ \omega \in A, \\ 0 & else, \end{cases}$$

and $\mathbb{E}(\mathbb{1}_A) = \mathbb{P}(A)$.

We define the quotient space $S_d^0[0,T]$ as the space of progressively measurable continuous processes from $\Omega \times [0,T]$ to $\mathbb{R}^d$ under the equivalence relation $\mathcal{R}$. Where $X\mathcal{R}Y$ if

$$X(t) = Y(t), \ \ \forall t \in [0,T], \ \ \mathbb{P}-a.s.$$

And $S_d^0 := \bigcap_{T\geq 0} S_d^0[0,T]$. Furthermore, for $\nu > 0$, we define $S_d^\nu[0,T]$ as the subset of the processes $X(t)$ in $S_d^0[0,T]$ such that

$$S_d^\nu[0,T] := \left\{ X \in S_d^0[0,T] \middle| \mathbb{E}\left(\sup_{t\in[0,T]} \|X_t\|^\nu \right) < +\infty \right\}$$

And

$$S_d^\nu := \bigcap_{T\geq 0} S_d^\nu[0,T].$$

A useful theorem to show the convergence of the trajectory is the following:

Theorem 2.1. *[7, Ch. 1, Theorem 3.9] Let $\{A_t\}_{t\geq 0}$ and $\{U_t\}_{t\geq 0}$ be two continuous adapted increasing processes with $A_0 = U_0 = 0$ a.s. Let $\{M_t\}_{t\geq 0}$ be a real valued continuous local martingale with $M_0 = 0$ a.s. Let ξ be a nonnegative $\mathcal{F}_0$-measurable random variable. Define*

$$X_t = \xi + A_t - U_t + M_t \quad \text{for } t \geq 0.$$

If X_t is nonnegative and $\lim_{t\to\infty} A_t < +\infty$ a.s., then $\lim_{t\to\infty} X_t$ exists and is finite a.s., and $\lim_{t\to\infty} U_t < +\infty$ a.s..

Now we present the main key for the convergence rate results, a subtle variation of the Itô's formula, valid for $C_L^{1,1}$ functions and which can be proved using a mollification argument.

Proposition 2.2. *Consider X a solution of (SDE), $\phi \in C_L^{1,1}(\mathbb{R}^d)$. Then the process*

$$Y(t) = \phi(X(t)),$$

is an Itô Process, such that

$$\begin{aligned} Y(t) \leq Y(0) &- \int_0^t \langle \nabla\phi(X(s)), \nabla f(X(s))\rangle ds \\ &+ \int_0^t \langle \sigma^t(s, X(s))\nabla\phi(X(s)), dB(s)\rangle \\ &+ \frac{L}{2}\int_0^t \operatorname{tr}[\sigma(s, X(s))\sigma^t(s, X(s))]ds. \end{aligned} \tag{2.1}$$

3 Main results

Subsection 3.1 shows almost sure convergence of the trajectory generated by (SDE) to a minimizer of f, as well as global convergence rates under convexity and strong convexity. In subsection 3.2, we will provide convergence rates under the local Polyak-Łojasiewicz (PL) inequality.

3.1 Global convergence guarantees

Consider $f \in C_L^{1,1}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d)$ (called the potential) and the dynamic (SDE) under hypotheses (H_0) and (H).

Remark 3.1. (H) implies the existence of $\sigma_* > 0$ such that:

$$\|\sigma(t,x)\|_F^2 = \operatorname{tr}[\sigma(t,x)\sigma^t(t,x)] \leq \sigma_*^2, \quad \forall t \geq 0, x \in \mathbb{R}^d$$

Throughout the rest of the paper, we denote

$$\sigma_\infty(t) \overset{\text{def}}{=} \sup_{x\in\mathbb{R}^d} \|\sigma(t,x)\|_F \quad S \overset{\text{def}}{=} \operatorname{argmin}(f).$$

Theorem 3.2. *Consider $f \in C_L^{1,1}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d)$ and the dynamic (SDE) under the hypotheses (H_0) and (H). Then, there exists a unique solution $X \in S_d^\nu$, for every $\nu \geq 2$. Moreover, if $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$, then:*

(i) $\sup_{t\geq 0} \mathbb{E}[\|X(t)\|^2] < +\infty$.

(ii) $\forall x^\star \in S$, $\lim_{t\to\infty} \|X(t) - x^\star\|$ *exists a.s. and* $\sup_{t\geq 0} \|X(t)\| < +\infty$ *a.s.*

(iii) $\lim_{t\to\infty} \|\nabla f(X(t))\| = 0$ *a.s., in consequence,* $\lim_{t\to\infty} f(X(t)) = \min f$ *a.s.*

(iv) In addition to (iii), there exists an $S-$valued random variable $x^\star$ such that $X \lim_{t\to\infty} X(t) = x^\star$ a.s.

Sketch of proof. The existence and uniqueness of solution comes from [7, Theorem 2.4.1] and [15, Theorem 5.2.1]. The rest of the proof will consist of three steps. The first one is to use Itô's formula to conclude the first point, then Theorem 2.1 with $X_t = \frac{\|X(t)-x^\star\|}{2}$ ($x^\star \in S$) and Itô's formula to conclude that for every $x^\star \in S$, $\lim_{t\to\infty} \|X(t) - x^\star\|$ exists a.s., then a separability argument to conclude that almost surely, for every $x^\star \in S$, $\lim_{t\to\infty} \|X(t) - x^\star\|$ exists. The second step consists in using another conclusion of Theorem 2.1 to conclude that $\|\nabla f(X(\cdot))\|^2 \in \mathrm{L}^1(\mathbb{R}_+)$ a.s., then proving that this function is eventually uniformly continuous, we proceed as Barbalat's Lemma says (see [3]) to conclude that $\lim_{t\to\infty} \|\nabla f(X(t))\| = 0$ a.s. and by consequence of the convexity of f that $\lim_{t\to\infty} f(X(t)) = \min f$ a.s. Finally, the third step is to use Opial's Lemma to conclude that there exists an $S-$valued random variable $x^\star$ such that $\lim_{t\to\infty} X(t) = x^\star$ a.s. □

Theorem 3.3. *Let $f \in C_L^{1,1}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d)$, if we consider X the solution of the dynamic (SDE) under the hypotheses (H_0) and (H), then the following statements holds:*

(i) Let $\overline{X}(t) := t^{-1}\int_0^t X(s)ds$. Since $f \in \Gamma_0(\mathbb{R}^d)$, then

$$\mathbb{E}\left[f(\overline{X}(t)) - \min(f)\right] \leq \frac{\operatorname{dist}(X_0, S)^2}{2t} + \frac{\sigma_*^2}{2}, \quad \forall t > 0. \tag{3.1}$$

Besides, if σ_∞ is $\mathrm{L}^2(\mathbb{R}_+)$, then

$$\mathbb{E}\left[f(\overline{X}(t)) - \min(f)\right] = \mathcal{O}\left(\frac{1}{t}\right), \quad \forall t > 0. \tag{3.2}$$

(ii) If $f \in \Gamma_\mu(\mathbb{R}^d)$, then $S = \{x^\star\}$ and

$$\mathbb{E}\left(\frac{\|X(t) - x^\star\|^2}{2}\right) \leq \frac{\|X_0 - x^\star\|^2}{2} e^{-2\mu t} + \frac{\sigma_*^2}{4\mu}, \quad \forall t \geq 0. \tag{3.3}$$

Besides, if σ_∞ is decreasing and vanishes at infinity, then:

$$\begin{aligned} \mathbb{E}\left(\frac{\|X(t) - x^\star\|^2}{2}\right) \leq \frac{\|X_0 - x^\star\|^2}{2} e^{-2\mu t} + \frac{\sigma_*^2}{2} e^{-\mu t} \\ + \frac{\sigma_\infty^2\left(\frac{t}{2}\right)}{2}, \quad \forall t \geq 0. \end{aligned} \tag{3.4}$$

3.2 Local convergence rates under PL Inequality

Let us start with the definition.

Definition 3.4. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex such that $\operatorname{argmin}(f) \neq \varnothing$. Then f satisfies locally the Polyak-Łojasiewicz (PL) Inequality if there exists $r > \min f, \mu > 0$ such that:

$$2\mu(f(x) - \min(f)) \leq \|\nabla f(x)\|^2, \quad \forall x \in [f \leq r], \tag{3.5}$$

and it will be denoted $f \in \mathrm{PL}_{loc}(\mathbb{R}^d)$.

We will need the following lemma:

Lemma 3.5. *Consider $f \in C_L^{1,1}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d)$ and X a solution of (SDE) under hypotheses (H_0), (H) and such that $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$. Let us also consider $\delta \in (0,1)$, $\Omega_\delta \in \mathcal{F}$ such that $\mathbb{P}(\Omega_\delta) \geq 1-\delta$. Then, there exists $C_d, C_f > 0$:*

$$\mathbb{E}\left[\frac{\mathrm{dist}(X(t),S)^2}{2}\mathbb{1}_{\Omega\setminus\Omega_\delta}\right] \leq C_d\sqrt{\delta},$$
$$\mathbb{E}\left[(f(X(t)) - \min f)\mathbb{1}_{\Omega\setminus\Omega_\delta}\right] \leq C_f\sqrt{\delta}.$$

Now we are ready to state the main result about the local convergence rates.

Theorem 3.6. *Let $f \in C_L^{1,1}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d) \cap \mathrm{PL}_{loc}(\mathbb{R}^d)$, consider X the solution of the dynamic (SDE) under the hypotheses (H_0), (H), and such that $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$ ($C_\infty := \|\sigma_\infty\|_{\mathrm{L}^2}$) and decreasing. Consider also the positive constants C, C_*, C_f, μ, γ. Then, for all $\delta > 0$, there exists $\hat{t}_\delta > 0$ such that for all $t > \hat{t}_\delta$:*

$$\begin{aligned}\mathbb{E}\left(f(X(t)) - \min f\right) &\leq e^{-2\mu(t-\hat{t}_\delta)}\mathbb{E}(f(X(\hat{t}_\delta)) - \min f)\\ &+ \frac{LC_\infty^2}{2}e^{-\mu(t-\hat{t}_\delta)} + \frac{L}{4\mu}\sigma_\infty^2\left(\frac{t+\hat{t}_\delta}{2}\right) \qquad (3.6)\\ &+ \sqrt{\delta}\left[\frac{C}{4\mu}\frac{\sigma_\infty^2(\frac{t+\hat{t}_\delta}{2})}{\sqrt{\int_{\hat{t}_\delta}^{\frac{t+\hat{t}_\delta}{2}}\sigma_\infty^2(u)du}} + C_\infty Ce^{-\mu(t-\hat{t}_\delta)} + C_f\right].\end{aligned}$$

Moreover,

$$\begin{aligned}\mathbb{E}\left(\frac{\mathrm{dist}(X(t),S)^2}{2}\right) &\leq e^{-2\gamma(t-\hat{t}_\delta)}\mathbb{E}\left(\frac{\mathrm{dist}(X(\hat{t}_\delta),S)^2}{2}\right)\\ &+ C_\infty^2 e^{-\gamma(t-\hat{t}_\delta)} + \frac{1}{2\gamma}\sigma_\infty^2\left(\frac{t+\hat{t}_\delta}{2}\right) \qquad (3.7)\\ &+ \sqrt{\delta}\left[\frac{C_*}{4\gamma}\frac{\sigma_\infty^2\left(\frac{t+\hat{t}_\delta}{2}\right)}{\sqrt{\int_{\hat{t}_\delta}^{\frac{t+\hat{t}_\delta}{2}}\sigma_\infty^2(u)du}} + C_*C_\infty e^{-\gamma(t-\hat{t}_\delta)} + C_d\right].\end{aligned}$$

Sketch of proof. Consider that $f \in C_L^{1,1}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d)$ and $r > \min f$ such that f satisfies the PL Inequality with constant μ on $[f \leq r]$, use that $\lim_{t\to\infty} f(X(t)) = \min f$ a.s. in order to apply Egorov's Theorem [12, Chapter 3, Exercise 16]. This guarantees uniform convergence on a set $\Omega_\delta \in \mathcal{F}$ with $\mathbb{P}(\Omega_\delta) \geq 1-\delta$. Thus, there exists $\hat{t}_\delta > 0$ such that, after that time, we can localize the process on $[f \leq r]$ with a probability of at least $1-\delta$. We can use Proposition 2.2 for the function $\phi(x) = f(x) - \min f$ and then multiply the obtained equation by $\mathbb{1}_{\Omega_\delta}$. After, we take expectation and we can apply a slight change of Comparison Lemma [8, Proposition 2.3], and the integrating factor method to obtain a convergence rate on $\mathbb{E}\left([f(X(t) - \min f]\mathbb{1}_{\Omega_\delta}\right)$. Combining this with Lemma 3.5 shows a convergence rate on $\mathbb{E}\left([f(X(t) - \min f]\right)$. Moreover, since the local PL implies an Error bound Inequality, i.e., there exists $\gamma > 0$ such that:

$$f(x) - \min(f) \geq \gamma \mathrm{dist}(x,S)^2, \quad \forall x \in [f \leq r],$$

we proceed as before with the function $\tilde{\phi}(x) = \mathrm{dist}(x,S)^2$. □

If we have the global PL inequality, the statements of Theorem 3.6 would hold if we replace $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$ by σ_∞ decreasing and vanishing at infinity, δ by 0 and $\hat{t}_\delta$ by 0. It is of paramount importance to observe that a.s. convergence of $f(X(t))$ is not sufficient to get the local bounds, since this only gives that the time beyond which X is a.s. localized in $[f \leq r]$ is a random variable that cannot be made uniform. Unfortunately, this is a flawed argument that usually appears in the literature.

References

[1] Jerome Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. In Jon Lee and Sven Leyffer, editors, *Mathematical Programming*, volume 165, page 471–507. Springer, 2016.

[2] T. Colding and W. Minicozzi H. Lojasiewicz inequalities and applications. *Surveys in Differential Geometry*, XIX:63–82, 2014.

[3] Bálint Farkas and Sven-Ake Wegner. Variations on barbalat's lemma. *The American Mathematical Monthly*, 123:8:825–830, 2016.

[4] Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Lui. On the diffusion approximation of nonconvex stochastic gradient descent. *arXiv:1705.07562v2*, 2018.

[5] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. *arXiv:1511.06251*, 2017.

[6] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling sgd with stochastic differential equations. *arXiv:2102.12470*, 2021.

[7] Xuerong Mao. Stochastic differential equations and applications. *Elsevier*, 2007.

[8] Radoslaw Matusik, Andrzej Nowakowski, Slawomir Plaskacz, and Andrzej Rogoswski. Finite-time stability for differential inclusions with applications to neural networks. *arXiv:1804.08440v2*, 2019.

[9] Panayotis Mertikopoulos and Mathias Staudigl. On the convergence of gradient-like flows with noisy gradient input. *SIAM Journal on Optimization*, 28(1):163–197, 2018.

[10] Antonio Orvieto and Aurelien Lucchi. Continuous-time models for stochastic optimization algorithms. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[11] Herbert Robins and Sutton Monro. A stochastic approximation method. *Ann. Math. Statist. 22*, pages 400–407, 1951.

[12] Walter Rudin. Real and complex analysis. *McGraw-Hill*, 1987.

[13] Bin Shi, Weijie J. Su, and Michael I. Jordan. On learning rates and schrödinger operators. *arXiv:2004.06977*, 2020.

[14] S. Soatto and P. Chaudhari. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10, 2018.

[15] Bernt Øksendal. Stochastic differential equations. *Springer*, 2003.

[16] S. Łojasiewicz. Sur les trajectoires du gradient d'une fonction analytique. *Semin. Geom., Univ. Studi Bologna*, 1982/1983:115–117, 1984.