An SDE Perspective on Stochastic Convex Optimization

Rodrigo Maulen S.[*] Jalal Fadili[†] Hedy Attouch[‡]

Abstract. In this paper, we analyze the global and local behavior of gradient-like flows under stochastic errors towards the aim of solving convex optimization problems with noisy gradient input. We first study the unconstrained differentiable convex case, using a stochastic differential equation where the drift term is minus the gradient of the objective function and the diffusion term is either bounded or square-integrable. In this context, under Lipschitz continuity of the gradient, our first main result shows almost sure convergence of the objective and the trajectory process towards 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) Łojasiewicz case. The latter involves a challenging local analysis which requires non-trivial arguments from measure theory. Then, we extend our study to the constrained case and more generally to nonsmooth problems. We show that several of our results have natural extensions obtained by replacing the gradient of the objective function by a cocoercive monotone operator. This makes it possible to obtain similar convergence results for optimization problems with an additively "smooth + non-smooth" convex structure. Finally, we consider another extension of our results to non-smooth optimization which is based on the Moreau envelope.

Key words. Convex optimization, Stochastic Differential Equation, Stochastic gradient descent, Łojasiewicz inequality, KL inequality, Convergence rate, Asymptotic behavior.

AMS subject classifications. 37N40, 46N10, 49M99, 65B99, 65K05, 65K10, 90B50, 90C25, 60H10, 49J52, 90C53

1 Introduction

1.1 Problem statement

We aim to solve convex minimization problems by means of stochastic differential equations whose drift term is driven by the gradient of the objective function. This allows for noisy (inaccurate) gradient input to be taken into account. Consider the minimization problem

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

where the objective f satisfies the following standing assumptions:

$$\begin{cases} f \text{ is convex and continuously differentiable with } L\text{-Lipschitz continuous gradient;} \\ \mathcal{S} \overset{\text{def}}{=} \operatorname{argmin}(f) \neq \emptyset. \end{cases} \tag{H_0}$$

[*] Normandie Université, ENSICAEN, UNICAEN, CNRS, GREYC, France. E-mail: rodrigo.maulen@ensicaen.fr

[†] Normandie Université, ENSICAEN, UNICAEN, CNRS, GREYC, France. E-mail: Jalal.Fadili@ensicaen.fr

[‡] IMAG, CNRS, Université Montpellier, France. E-mail: hedy.attouch@umontpellier.fr

We will also later deal with the constrained case, and more generally with additively structured "smooth + nonsmooth" convex optimization.

Let us first recall some basic facts about the deterministic case. To solve (P), a fundamental dynamic to consider is the gradient flow of f, *i.e.* the gradient descent dynamic with initial condition $x_0 \in \mathbb{R}^d$:

$$\begin{cases} \dot{x}(t) = -\nabla f(x(t)), \quad t > 0 \\ x(0) = x_0. \end{cases} \tag{GF}$$

It is well known since the founding papers of Brezis, Baillon, Bruck in the 1970s that, if the solution set $\operatorname{argmin} f$ of (P) is non-empty, then each solution trajectory of (GF) converges, and its limit belongs to $\operatorname{argmin} f$. In fact, this result is true in a more general setting, simply assuming that the objective function f is convex, lower semicontinuous (lsc) and proper (in which case we must consider the differential inclusion obtained by replacing in (GF) the gradient of f by the sub-differential ∂f).

In many cases, the gradient input is subject to noise, for example, if the gradient cannot be evaluated directly, or due to some other exogenous factor. In such scenario, one can model the associated errors using a stochastic integral with respect to the measure defined by a continuous Itô martingale. This entails the following stochastic differential equation as a stochastic counterpart of (GF):

$$\begin{cases} dX(t) = -\nabla f(X(t))dt + \sigma(t, X(t))dW(t), \quad t > 0 \\ X(0) = X_0, \end{cases} \tag{SDE}$$

defined over a complete filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$, where the diffusion (volatility) term $\sigma : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^{d\times m}$ is matrix-valued measurable function, W is a $\mathcal{F}_t$-adapted m-dimensional Brownian motion, and the initial data X_0 is an $\mathcal{F}_0$-measurable $\mathbb{R}^d$-valued random variable.

Our goal is to study the dynamic of (SDE) and its long time behavior in order to solve (P). To identify the assumptions necessary to hope for such a behavior to occur, remember that when the diffusion term σ is a positive real constant, it is well-known that $X(t)$ in this case is a continuous-time diffusion process known as Langevin diffusion, and has a unique invariant probability measure π_σ with density $\propto e^{-2f(x)/\sigma^2}$ [10]. In fact, (SDE) can be interpreted as the pathwise solution to the Fokker-Planck equation (see [28]). It is also very well known that the measure π_σ gets concentrated around $\operatorname{argmin} f$ as σ tends to 0^+ with $\lim_{\sigma\to 0^+} \pi_\sigma(\operatorname{argmin} f) = 1$; see *e.g.* [14].

Motivated by this last observation, our paper will then mostly focus on the case where $\sigma(\cdot, x)$ vanishes sufficiently fast as $t \to +\infty$ uniformly in x, and some guarantees will also be provided for uniformly bounded σ. Therefore, throughout the paper, the entries σ_{ik} are assumed to satisfy:

$$\begin{cases} \sup_{t\geq 0, x\in\mathbb{R}^d} |\sigma_{ik}(t, x)| < +\infty, \\ |\sigma_{ik}(t, x') - \sigma_{ik}(t, x)| \leq l_0\|x' - x\|, \end{cases} \tag{H}$$

for some $l_0 > 0$ and for all $t \geq 0, x, x' \in \mathbb{R}^d$. The Lipschitz continuity assumption is mild and required to ensure the well-posedness of (SDE).

1.2 Contributions

We study the properties of the process $X(t)$ and $f(X(t))$ for the stochastic differential equation (SDE) from an optimization perspective, under the assumptions (H_0) and (H). When the diffusion term is uniformly bounded, we show convergence of $\mathbb{E}[f(X(t))-\min f]$ to a noise-dominated region both for the convex and strongly convex case. When the diffusion term is square-integrable, we show in Theorem 3.1 that $X(t)$ converges almost surely to a solution of (P), which is a new result to the best of our knowledge. Moreover, in Theorem 3.3 and Proposition 3.4, we provide new ergodic and pointwise convergence rates of the objective in expectation, again for both the convex and strongly convex case.

Then we turn to a local analysis relying on the Łojasiewicz inequality and its strong ties with error bounds. Since this property is most often satisfied only locally, we deepen the discussion on the long time localization of the process. This is fundamental, because in the recent literature on local convergence properties of stochastic gradient descent, strong assumptions are imposed, such as $X(t)$ or $f(X(t))$ is locally bounded almost surely. Such assumptions are unfortunately unrealistic due to the presence of the Brownian Motion. We manage to circumvent this problem by using arguments from measure theory, in particular Egorov's theorem. In turn, under the Łojasiewicz inequality assumption with exponent $q \geq 1/2$, this allows us to show local convergence rates of the objective and the trajectory itself in expectation over a set of events whose probability is arbitrarily close to 1 (see Theorem 4.5).

Table 1 summarizes the local and global convergence rates obtained for $\mathbb{E}[f(X(t))-\min f]$. In this table, $\delta > 0$ is a parameter which is intended to be taken arbitrarily close to 0 but different from it, $\sigma_* > 0$ and $\sigma_\infty(\cdot)$ are defined as

$$\|\sigma(t,x)\|_F^2 \leq \sigma_*^2, \quad \forall t \geq 0, \forall x \in \mathbb{R}^d, \qquad \text{and} \qquad \sigma_\infty(t) \overset{\text{def}}{=} \sup_{x\in\mathbb{R}^d} \|\sigma(t,x)\|_F, \tag{1}$$

and $\sigma_\infty(\cdot)$ is a non-increasing function. $\text{Ł}^q(\mathcal{S})$ is the class of functions satisfying the Łojasiewicz inequality with exponent $q \in [0,1]$ at each point of $\mathcal{S}$ (see Definition 4.1)[1].

Property of f	**Gradient Flow**	**SDE** $(\sup_{t\geq 0}\sigma_\infty(t) \leq \sigma_*)$	**SDE** $(\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+))$
Convex	t^{-1}	$t^{-1}+\sigma_*^2$	t^{-1}
μ-Strongly Convex	$e^{-2\mu t}$	$e^{-2\mu t}+\sigma_*^2$	$\max\{e^{-2\mu t},\sigma_\infty^2(t)\}$
Convex $\cap$ $\text{Ł}^{1/2}(\mathcal{S})$ (coef. μ)	$e^{-\mu^2 t}$	✘	$\max\{e^{-\mu^2 t},\sigma_\infty^2(t)\}+\sqrt{\delta}$
Convex $\cap$ $\text{Ł}^{q}(\mathcal{S})$, $q \in (\frac{1}{2},1)$	$t^{-\frac{1}{2q-1}}$	✘	$t^{-\frac{1}{2q-1}}$ [2]$+\sqrt{\delta}$

Table 1: Summary of local and global convergence rates obtained for $\mathbb{E}[f(X(t))-\min f]$.

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 only the boundedness of the diffusion and the Łojasiewicz inequality, we could not find better results (cells marked with ✘) than those presented in the convex case. Since the Łojasiewicz inequality is local, a natural approach would be to localize the process in the long term with high probability. However, it is not clear how to achieve this.

[1] Semialgebraic functions, and more generally, functions based on the class of analytic functions is a typical family of functions that verify the Łojasiewicz inequality at each point [38, 39].

[2] This is not yet proven, our conjecture is that it is true when $\sigma_\infty = \mathcal{O}((t+1)^{-\frac{q}{2q-1}})$ (see the detailed discussion in Conjecture 4.11).

In Section 5, we turn to extending some of the preceding results to the structured convex minimization problem

$$\min_{x\in\mathbb{R}^d} f(x) + g(x), \tag{P$_\mathrm{c}$}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ satisfies (H$_0$), $g : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is proper, lsc and convex and $\operatorname{argmin}(f+g) \neq \emptyset$. This obviously covers the case of constrained minimization of f over a non-empty closed convex set. We take two different routes leading to different SDEs.

The first approach consists in reformulating (P$_\mathrm{c}$) as finding for zeros of the operator $M_\mu : \mathbb{R}^d \to \mathbb{R}^d$

$$M_\mu(x) = \frac{1}{\mu}\left(x - \operatorname{prox}_{\mu g}(x - \mu\nabla f(x))\right),$$

where $\mu > 0$ and $\operatorname{prox}_{\mu g}$ is the proximal mapping of μg. It is well-known that the operator M_μ is cocoercive [4], hence monotone and Lipschitz continuous, and $M_\mu = \nabla f$ when g vanishes. The idea is then to replace the operator ∇f in (SDE) by M_μ leading to an SDE which will have many of the convergence properties obtained in the smooth convex case. This approach is in accordance with the deterministic theory for monotone cocoercive operators (see [11, 1, 4]).

The second approach regularizes the nonsmooth component g of the objective function using its Moreau envelope

$$g_\theta(x) = \min_{z\in\mathbb{R}^d} g(z) + \frac{1}{2\theta}\|x - z\|^2.$$

This leads to studying the dynamic (SDE) with the function $f + g_\theta$, which has a continuous Lipschitz gradient. This approximation method leads to an SDE with non-autonomous drift term. Note, however, that the noise in this case can be considered on the evaluation of $\nabla f(x)$, while it is on $M_\mu(x)$ in the first approach.

1.3 Relation to prior work

The gradient system (GF), which is valid on a general real Hilbert space $\mathcal{H}$, is a dissipative dynamical system, whose study dates back to Cauchy [15]. It plays a fundamental role in optimization: it transforms the problem of minimizing f into the study of the asymptotic behavior of the trajectories of (GF). This example was the precursor to the rich connection between continuous dissipative dynamical systems and optimization. Its Euler forward discretization (with stepsize $\gamma_k > 0$) is the celebrated gradient descent scheme

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

Under (H$_0$), and for $(\gamma_k)_{k\in\mathbb{N}} \subset]0, 2/L[$, then we have both the convergence of the values $f(x_k) - \min f = \mathcal{O}(1/k)$ (in fact even $o(1/k)$), and the weak convergence of iterates $(x_k)_{k\in\mathbb{N}}$ to a point in $\operatorname{argmin} f$. Moreover, if the Łojasiewicz inequality (24) (see [40]) is satisfied, then we can ensure the strong convergence of $(x_k)_{k\in\mathbb{N}}$ to a point in $\operatorname{argmin} f$ and faster convergence rates than those ensured by the simple convexity hypothesis (see [12, 18]).

Now, let us focus on the finite-dimensional case ($\mathcal{H} = \mathbb{R}^d$). Although the Gradient Descent is a classical algorithm to solve the convex minimization problem, 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. The Robbins-Monro stochastic approximation algorithm [52] is at the heart of Stochastic Gradient

Descent (SGD), which, roughly speaking, consists in cheaply and randomly approximating the gradient at the price of obtaining a random noise in the solutions. Given an initial point $x_0 \in \mathbb{R}^d$, $\gamma > 0$, (SGD) updates the iterates according to

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

where ξ_k denotes the (random) noise term on the gradient at the k-th iteration.

The SDE continuous-time approach is motivated by its relations to (SGD), where the latter can be viewed as a Euler forward time discretization, and the noise $\xi_k \sim \mathcal{N}(0, \sigma_k I_d)$ (hence not necessarily bounded). In fact, several recent works (see *e.g.* [36, 46, 29, 55, 37, 56, 21]) have 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. The continuous-time perspective offers a deep insight and unveils the key properties of the dynamic, without being tied to a specific discretization. This in turn enlightens the behavior of the sequence generated by some specific algorithm such as (SGD). One may also wonder whether (SDE) is a better continuous-time model for (SGD) than (GF). The answer is affirmative as has been shown recently in [21, Proposition 2.1]. There, the trajectory of the sequence $(x_k)_{k\in\mathbb{N}}$ of (SGD), with $\xi_k \sim \mathcal{N}(0, \sigma_k I_d)$, was proved to be accurately approximated by (SDE) with $\sigma(t, X(t)) = \sqrt{\gamma}\sigma(t)$. The approximation error is of order $\mathcal{O}(\gamma)$ which is much better than that of (GF) which is only $\mathcal{O}(\sqrt{\gamma})$.

The Euler forward discretization (with stepsize $\gamma > 0$) of (SDE) when $d = m$ and $\sigma = \sqrt{2}I_d$ is the following algorithm

$$X_{k+1} = X_k - \gamma\nabla f(X_k) + \sqrt{2\gamma}\xi_k, \tag{LMC}$$

where $\xi_k \sim \mathcal{N}(0, I_d)$ (multivariate standard normal distribution). This algorithm, which is known as Langevin Monte Carlo (see [49]), is a standard sampling scheme, whose purpose is to generate samples from an approximation of a target distribution, in our case, proportional to $e^{-f(x)}$. Under appropriate assumptions on f, when γ is small and k is large such that $k\gamma$ is large, the distribution of X_k converges in different topologies or is close in various metrics to the target distribution with density $\propto e^{-f(x)}$. Asymptotic and non-asymptotic (with convergence rates) results of this kind have been studied in a number of papers under various conditions; see [20, 19, 24, 25, 16, 30] and references therein. By rescaling the problem, relation between sampling (*i.e.* (LMC)) and optimization (*i.e.* (SGD)) has been also investigated for the strongly convex case in *e.g.* [20].

Concerning (SDE), one can easily infer from [9, Proposition 7.4] that assuming $\sup_{x\in\mathbb{R}^d} \|\sigma(t,x)\|_F = o(1/\sqrt{\log(t)})$, and conditioning on the event that $X(t)$, we have almost surely that the set of limits of convergent sequences $X(t_k)$, $t_k \to +\infty$ is contained in $\operatorname{argmin} f$. Using results on asymptotic pseudo-trajectories from [9], the work of [43, 51, 5] analyzed the behavior of the Stochastic Mirror Descent dynamics:

$$\begin{aligned} dY(t) &= -\nabla f(X(t))dt + \sigma(t, X(t))dW(t), \\ X(t) &= Q(\eta Y(t)), \end{aligned} \tag{SMD}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex feasible region, f is convex with Lipschitz continuous gradient on $\mathcal{X}$, $Q : \mathbb{R}^d \to \mathcal{X}$ is the mirror map induced by some strongly convex entropy, and $\eta > 0$ is a sensitivity parameter. In [43, Theorem 4.1], it is shown that if $\mathcal{X}$ is also assumed bounded, that $\sup_{x\in\mathbb{R}^d} \|\sigma(t,x)\|_F = o(1/\sqrt{\log(t)})$, and Q satisfies some continuity assumptions[3], then the process

[3] Compactness of $\mathcal{X}$ and the condition on $\sigma(\cdot,\cdot)$ are clearly reminiscent of [9, Proposition 7.4], though the latter is not discussed in [43].

$X(t)$ (SMD) converges to a point in $\operatorname{argmin} f$ almost surely. Similar assumptions can be found in [5] to obtain almost sure convergence on the objective. Let us observe that all these results do not apply to our setting. Indeed, if $\mathcal{X} = \mathbb{R}^d$ (unconstrained problem), $Q(x) = x$ and $\eta = 1$, we recover (SDE). Our work does not assume any boundedness whatsoever to establish our results. This comes however at somewhat stronger assumptions on $\sigma(\cdot, \cdot)$.

While finalizing this work, we became aware of the recent work of [22], which analyzes the behavior of (SDE) for $f \in C^2(\mathbb{R}^d)$ not necessarily convex and which satisfies $\sup_{x\in\mathbb{R}^d} \|\sigma(\cdot, x)\|_F \in \mathrm{L}^2(\mathbb{R}_+)$. Conditioning on the event that $\limsup_{t\to+\infty} \|X(t)\| < +\infty$, they showed that $\nabla f(X(t)) \to 0$ almost surely, almost sure convergence of $f(X(t))$, and if the objective f is semialgebraic (and more generally tame), they also showed almost sure convergence of $X(t)$ towards a critical point of f. They also made attempt to get local convergence rates under the Łojasiewicz inequality that are less transparent than ours. Our analysis on the other hand leverages convexity of f to establish stronger results.

1.4 Organization of the paper

Section 2 introduces notations and reviews some necessary material from convex and stochastic analysis. Section 3 states our main convergence results in the case of a convex differentiable objective function whose gradient is Lipschitz continuous. We first show the almost sure convergence of the process towards the set of minimizers, then we establish convergence rates for the values. Section 4 introduces further geometric properties of the objective functions, namely Łojasiewicz property and related error bound, which allows to obtain improved (local) convergence rates. This covers in particular the (locally) strongly convex case. In section 5, we extend some results to the nonsmooth case by considering the additively structured ”smooth + nonsmooth” convex minimization. We develop new stochastic differential equations that naturally lend themselves to splitting techniques. Technical lemmas and theorems that are needed throughout the paper are collected in the appendix.

2 Notation and preliminaries

We will use the following shorthand notations: given $d, n \in \mathbb{N}$, $[n] \overset{\text{def}}{=} \{1, \ldots, n\}$, $\mathbb{R}^{d\times n}$ is the set of real matrices of size $d\times n$, and I_d is the identity matrix of dimension d. For $M \in \mathbb{R}^{d\times n}$, $M^\top \in \mathbb{R}^{n\times d}$ is its transpose matrix and $\|M\|_F$ is its Frobenius norm. For $M, M' \in \mathbb{R}^{d\times d}$, $M \preccurlyeq M'$ if and only if $u^\top(M' - M)u \geq 0$ for every $u \in \mathbb{R}^d$. The notation $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ means that A is a set-valued operator from $\mathbb{R}^d$ to $\mathbb{R}^d$. Consider $f : \mathbb{R}^d \to \mathbb{R}$, the sublevel of f at height $r \in \mathbb{R}$ is denoted $[f \leq r] \overset{\text{def}}{=} \{x \in \mathbb{R}^d : f(x) \leq r\}$. For $1 \leq p \leq +\infty$, $\mathrm{L}^p([a, b])$ is the space of measurable functions $g : \mathbb{R} \to \mathbb{R}$ such that $\int_a^b |g(t)|^p dt < +\infty$, with the usual adaptation when $p = +\infty$. Functions obeying $\int_0^{+\infty} |g(t)|^p dt < +\infty$ belong to $\mathrm{L}^p(\mathbb{R}_+)$. On the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathrm{L}^p(\Omega; \mathbb{R}^d)$ denotes the (Bochner) space of $\mathbb{R}^d$-valued random variables whose p-th moment (with respect to the measure $\mathbb{P}$) is finite. Other notations will be explained when they first appear.

2.1 On convex analysis

Let us recall some important definitions and results from convex analysis in the finite-dimensional case; for a comprehensive coverage, we refer the reader to [53].

We denote by $\Gamma_0(\mathbb{R}^d)$ the class of proper lsc and convex functions on $\mathbb{R}^d$ taking values in $\mathbb{R}\cup\{+\infty\}$. For $\mu > 0$, $\Gamma_\mu(\mathbb{R}^d) \subset \Gamma_0(\mathbb{R}^d)$ is the class of μ-strongly convex functions, roughly speaking, this means that there exists a quadratic lower bound on the growth of these functions. We denote by $C^s(\mathbb{R}^d)$ the class of s-times continuously differentiable functions on $\mathbb{R}^d$. For $L \geq 0$, $C_L^{1,1}(\mathbb{R}^d) \subset C^1(\mathbb{R}^d)$ is the set of functions on $\mathbb{R}^d$ whose gradient is L-Lipschitz continuous.

The following *Descent Lemma* which is satisfied by this class of functions plays a central role in optimization.

Lemma 2.1. *Let $f \in C_L^{1,1}(\mathbb{R}^d)$, then*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y-x\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

Corollary 2.2. *Let $f \in C_L^{1,1}(\mathbb{R}^d)$ such that* $\operatorname{argmin} f \neq \emptyset$*, then*

$$\|\nabla f(x)\|^2 \leq 2L(f(x) - \min f), \quad \forall x \in \mathbb{R}^d.$$

Proof. Proof. Use Lemma 2.1 for an arbitrary $x \in \mathbb{R}^d$ and $y = x - \frac{1}{L}\nabla f(x)$. Then bound

$$f\left(x - \frac{1}{L}\nabla f(x)\right) \geq \min f.$$

□

The *subdifferential* of a function $f \in \Gamma_0(\mathbb{R}^d)$ is the set-valued operator $\partial f : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ such that, for every x in $\mathbb{R}^d$,

$$\partial f(x) = \{u \in \mathbb{R}^d : f(y) \geq f(x) + \langle u, y - x\rangle \quad \forall y \in \mathbb{R}^d\}.$$

When f is continuous, $\partial f(x)$ is non-empty convex and compact set for every $x \in \mathbb{R}^d$. If f is differentiable, then $\partial f(x) = \{\nabla f(x)\}$. For every $x \in \mathbb{R}^d$ such that $\partial f(x) \neq \emptyset$, the minimum norm selection of $\partial f(x)$ is the unique element $\partial^0 f(x) \overset{\text{def}}{=} \operatorname{argmin}_{u\in\partial f(x)} \|u\|$.

2.2 On stochastic differential equations

For the necessary notation and preliminaries on stochastic processes, see Section A.2 in the appendix.

We emphasize that Theorem A.7 in the appendix provides us with sufficient conditions to ensure the existence and uniqueness of the solution to (SDE). These conditions are met in our case under assumptions (H_0) and (H).

Let us now present Itô's formula which plays a central role in the theory of stochastic differential equations.

Proposition 2.3. *[44, Chapter 4] Consider X a solution of* (SDE)*, $\phi : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ be such that $\phi(\cdot, x) \in C^1(\mathbb{R}_+)$ for every $x \in \mathbb{R}^d$ and $\phi(t, \cdot) \in C^2(\mathbb{R}^d)$ for every $t \geq 0$. Then the process*

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

is an Itô Process such that for all $t \geq 0$

$$\begin{aligned}Y(t) = Y(0) + \int_0^t \frac{\partial \phi}{\partial t}(s, X(s))ds - \int_0^t \langle \nabla \phi(s, X(s)), \nabla f(X(s)) \rangle \, ds \\ + \int_0^t \left\langle \sigma^\top(s, X(s)) \nabla \phi(s, X(s)), dW(s) \right\rangle + \frac{1}{2} \int_0^t \operatorname{tr}\left(\sigma(s, X(s)) \sigma^\top(s, X(s)) \nabla^2 \phi(s, X(s)) \right) ds. \quad (2)\end{aligned}$$

Moreover, if $\mathbb{E}[Y(0)] < +\infty$*, and if for all* $T > 0$

$$\mathbb{E}\left(\int_0^T \|\sigma^\top(s, X(s)) \nabla \phi(s, X(s))\|^2 ds \right) < +\infty,$$

then $\int_0^t \left\langle \sigma^\top(s, X(s)) \nabla \phi(s, X(s)), dW(s) \right\rangle$ *is a square-integrable continuous martingale and*

$$\begin{aligned}\mathbb{E}[Y(t)] = \mathbb{E}[Y(0)] + \mathbb{E}\left(\int_0^t \frac{\partial \phi}{\partial t}(s, X(s))ds \right) - \mathbb{E}\left(\int_0^t \langle \nabla \phi(s, X(s)), \nabla f(X(s)) \rangle \, ds \right) \\ + \frac{1}{2} \mathbb{E}\left(\int_0^t \operatorname{tr}\left(\sigma(s, X(s)) \sigma^\top(s, X(s)) \nabla^2 \phi(s, X(s)) \right) ds \right). \quad (3)\end{aligned}$$

The C^2 assumption on $\phi(t, \cdot)$ in Itô's formula is crucial. This can be weakened in certain cases leading to the following inequality that will be useful in our context.

Proposition 2.4. *Consider* X *a solution of* (SDE), $\phi_1 \in C^1(\mathbb{R}_+)$, $\phi_2 \in C_L^{1,1}(\mathbb{R}^d)$ *and* $\phi(t, x) = \phi_1(t)\phi_2(x)$*. Then the process*

$$Y(t) = \phi(t, X(t)) = \phi_1(t)\phi_2(X(t)),$$

is an Itô Process such that

$$\begin{aligned}Y(t) \leq Y(0) + \int_0^t \phi_1'(s)\phi_2(X(s))ds - \int_0^t \phi_1(s) \langle \nabla \phi_2(X(s)), \nabla f(X(s)) \rangle \, ds \\ + \int_0^t \left\langle \sigma^\top(s, X(s)) \phi_1(s) \nabla \phi_2(X(s)), dW(s) \right\rangle + \frac{L}{2} \int_0^t \phi_1(s) \operatorname{tr}\left(\sigma(s, X(s)) \sigma^\top(s, X(s)) \right) ds. \quad (4)\end{aligned}$$

Moreover, if $\mathbb{E}[Y(0)] < +\infty$*, then*

$$\begin{aligned}\mathbb{E}[Y(t)] \leq \mathbb{E}[Y(0)] + \mathbb{E}\left(\int_0^t \phi_1'(s)\phi_2(X(s))ds \right) - \mathbb{E}\left(\int_0^t \phi_1(s) \langle \nabla \phi_2(X(s)), \nabla f(X(s)) \rangle \, ds \right) \\ + \frac{L}{2} \mathbb{E}\left(\int_0^t \phi_1(s) \operatorname{tr}\left(\sigma(s, X(s)) \sigma^\top(s, X(s)) \right) ds \right). \quad (5)\end{aligned}$$

Proof. Proof. Analogous to the proof of [43, Proposition C.2] using Rademacher's theorem instead of Alexandrov's. □

3 Convergence properties for convex differentiable functions

We consider f (called the potential) and study the dynamic (SDE) under hypotheses (H_0) (*i.e.* $f \in C_L^{1,1}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d)$) and (H). Recall the definitions of σ_* and $\sigma_\infty(t)$ from (1). Observe that from (H) one can take $\sigma_*^2 = md\ \sup_{i\in[d],k\in[m]} \sup_{t\geq 0, x\in\mathbb{R}^d} |\sigma_{ik}(t,x)|^2$. Throughout the rest of the paper, we will use the shorthand notation

$$\Sigma(t,x) \overset{\text{def}}{=} \sigma(t,x)\sigma(t,x)^\top.$$

3.1 Almost sure convergence of trajectory

Our first main result establishes almost sure convergence of $X(t)$ to an $\mathcal{S}$-valued random variable as $t \to +\infty$.

Theorem 3.1. *Consider the dynamic* (SDE)*, where f and σ satisfy the assumptions* (H_0) *and* (H)*, respectively. Additionally, let $\nu \geq 2$, $X_0 \in \mathrm{L}^\nu(\Omega;\mathbb{R}^d)$ and is $\mathcal{F}_0$-measurable. Then, there exists a unique solution $X \in S_d^\nu$ of* (SDE)*. Moreover, if $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$, then the following holds:*

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

(ii) $\forall x^\star \in \mathcal{S}$, $\lim_{t\to+\infty} \|X(t) - x^\star\|$ *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.. As a result,* $\lim_{t\to\infty} f(X(t)) = \min f$ *a.s..*

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

Remark 3.2. (i) The assumptions on the noise variance are compatible with the theory of asymptotic pseudotrajectories (APT) [9] and their weak version (WAPT) [8]. As we have already discussed in Section 1.3, the theory of APT can indeed be used to study convergence properties of $X(t)$. For instance, using [9, Proposition 7.4] one can show easily that assuming $\sup_{x\in\mathbb{R}^d} \|\sigma(t,x)\|_F = o(1/\sqrt{\log(t)})$ and that $X(t)$ is bounded almost surely, one has almost sure subsequential convergence of $X(t)$ to points in $\mathcal{S}$. Our work leverages convexity, does not need any boundedness assumption and shows almost sure global convergence of the process (not just subsequentially).

(ii) Ergodic properties of $X(t)$ can be derived from the theory of WAPT as developed in [8] under the weaker assumptions that $\|\sigma(x,t)\| \leq \alpha(t)$, for some decreasing function α such that $\alpha(t) \to 0$ as $t \to +\infty$. For instance, an immediate consequence of [8, Proposition 1 and Corollary 1] and Theorem 3.1(i) is that the fraction of time spent by X in an arbitrary neighborhood of $\mathcal{S}$ goes to one with probability one. We also have by combining [8, Corollary 2] and Theorem 3.1(i), and since $\mathcal{S}$ is convex in our case, that the average process $\frac{1}{t}\int_0^t X(s)ds$ converges almost surely to a point in $\mathcal{S}$.

Proof. Proof. The existence and uniqueness of a solution follows directly from the fact that the conditions of Theorem A.7 are satisfied under (H_0) and (H). The architecture of the proof of Theorem 3.1 consists of three steps that we briefly describe:

- The first step is based on Itô's formula (Proposition 2.3) and Burkholder-Davis-Gundy inequality (Proposition A.10) that let us prove a uniform bound (on time) for the $\nu-$moment of $X(t)$.

- The second step is also based on Itô's formula. Instead of the previous step, we use Theorem A.9 that allows us to conclude that for every $x^\star \in \mathcal{S}$, $\lim_{t\to+\infty}\|X(t)-x^\star\|$ exists a.s.. Then, a separability argument is used to conclude that almost surely, for all $x^\star \in \mathcal{S}$, $\lim_{t\to+\infty}\|X(t)-x^\star\|$ exists.

- The third step consists in using another conclusion of Theorem A.9 to conclude that $\|\nabla f(X(\cdot))\|^2 \in \mathrm{L}^1(\mathbb{R}_+)$ a.s.. After proving that this function is eventually uniformly continuous, we proceed according to Barbalat's Lemma (see [26]) to conclude that $\lim_{t\to+\infty}\|\nabla f(X(t))\| = 0$ a.s.. As a consequence of the convexity of f we deduce that $\lim_{t\to+\infty} f(X(t)) = \min f$ a.s..

- Finally, the fourth step consists in using Opial's Lemma to conclude that there exists an $\mathcal{S}$-valued random variable $X^\star$ such that $\lim_{t\to+\infty} X(t) = X^\star$ a.s..

(i) Let $x^\star$ be taken arbitrarily in $\mathcal{S}$. Let us define the corresponding anchor function $\phi(x) = \frac{\|x-x^\star\|^2}{2}$. Using Itô's formula we obtain

$$\phi(X(t)) = \underbrace{\frac{\|X_0-x^\star\|^2}{2}}_{\xi=\phi(X_0)} + \underbrace{\frac{1}{2}\int_0^t \mathrm{tr}\left(\Sigma(s,X(s))\right)ds}_{A_t} - \underbrace{\int_0^t \langle \nabla f(X(s)), X(s)-x^\star\rangle\, ds}_{U_t}$$
$$+ \underbrace{\int_0^t \left\langle \sigma^\top(s,X(s))\left(X(s)-x^\star\right), dW(s)\right\rangle}_{M_t}. \tag{6}$$

Let us now embark from (6) and use that

$$0 \le \mathrm{tr}\left(\Sigma(s,X(s))\right) \le \sigma_\infty^2(s) \;\text{ and }\; \langle \nabla f(X(s)), X(s)-x^\star\rangle \ge 0,$$

where the second inequality is due to the convexity of f, to get

$$\phi(X(t)) \le \phi(X_0) + \frac{1}{2}\int_0^{+\infty} \sigma_\infty^2(s)ds + M_t.$$

Taking power $\frac{\nu}{2}$ at both sides and using that $(a+b+c)^{\frac{\nu}{2}} \le 3^{\frac{\nu-2}{2}}(a^{\frac{\nu}{2}}+b^{\frac{\nu}{2}}+c^{\frac{\nu}{2}})$, we have that

$$\|X(t)-x^\star\|^\nu \le 3^{\frac{\nu-2}{2}}\left[\|X_0-x^\star\|^\nu + \left(\int_0^{+\infty}\sigma_\infty^2(s)ds\right)^{\frac{\nu}{2}} + 2^{\frac{\nu}{2}}|M_t|^{\frac{\nu}{2}}\right].$$

Let $T>0$, applying the supremum over $t\in[0,T]$ and then taking expectation, we obtain

$$\mathbb{E}\left(\sup_{t\in[0,T]}\|X(t)-x^\star\|^\nu\right) \le 3^{\frac{\nu-2}{2}}\left[\mathbb{E}(\|X_0-x^\star\|^\nu) + \left(\int_0^{+\infty}\sigma_\infty^2(s)ds\right)^{\frac{\nu}{2}} + 2^{\frac{\nu}{2}}\mathbb{E}\left(\sup_{t\in[0,T]}|M_t|^{\frac{\nu}{2}}\right)\right].$$

Letting $g(s) = \sigma^\top(s, X(s))(X(s) - x^\star)$ and $p = \frac{\nu}{2}$ in Proposition A.10, we get

$$\mathbb{E}\left(\sup_{t\in[0,T]} \|X(t) - x^\star\|^\nu\right) \leq 3^{\frac{\nu-2}{2}} \left[\mathbb{E}(\|X_0 - x^\star\|^\nu) + \left(\int_0^{+\infty} \sigma_\infty^2(s)ds\right)^{\frac{\nu}{2}}\right] + 3^{\frac{\nu-2}{2}} 2^{\frac{\nu}{2}} C_{\frac{\nu}{2}} \mathbb{E}\left(\sup_{t\in[0,T]} \|X(t) - x^\star\|^{\frac{\nu}{2}} \left(\int_0^{+\infty} \sigma_\infty^2(s)ds\right)^{\frac{\nu}{4}}\right).$$

Using that $ab \leq \frac{a^2}{2K} + \frac{Kb^2}{2}$ for every $K > 0$,

$$\mathbb{E}\left(\sup_{t\in[0,T]} \|X(t) - x^\star\|^\nu\right) \leq 3^{\frac{\nu-2}{2}} \left[\mathbb{E}(\|X_0 - x^\star\|^\nu) + \left(\int_0^{+\infty} \sigma_\infty^2(s)ds\right)^{\frac{\nu}{2}}\right] + \frac{1}{2}\mathbb{E}\left(\sup_{t\in[0,T]} \|X(t) - x^\star\|^\nu\right) + 6^{\frac{\nu-2}{2}} C_{\frac{\nu}{2}} \left(\int_0^{+\infty} \sigma_\infty^2(s)ds\right)^{\frac{\nu}{2}}.$$

And we end up with

$$\mathbb{E}\left(\sup_{t\in[0,T]} \|X(t) - x^\star\|^\nu\right) \leq 3^{\frac{\nu-2}{2}} 2 \left[\mathbb{E}(\|X_0 - x^\star\|^\nu) + (1 + 2^{\frac{\nu-2}{2}} C_{\frac{\nu}{2}}) \left(\int_0^{+\infty} \sigma_\infty^2(s)ds\right)^{\frac{\nu}{2}}\right].$$

Since the right-hand side is independent of T, we take $\liminf_{T\to+\infty}$ on the previous expression and apply Fatou's Lemma to show the first claim.

(ii) Observe that, since $\nu \geq 2$, we have that $\mathbb{E}(\sup_{t\geq 0} \|X(t)\|^2) < +\infty$. Moreover $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$, and therefore

$$\mathbb{E}\left(\int_0^{+\infty} \|\sigma^\top(s, X(s)) (X(s) - x^\star) \|^2 ds\right) \leq \mathbb{E}\left(\sup_{t\geq 0} \|X(t) - x^\star\|^2\right) \int_0^{+\infty} \sigma_\infty^2(s)ds < +\infty.$$

Therefore M_t is a square-integrable continuous martingale. It is also a continuous local martingale (see [41, Theorem 1.3.3]), which implies that $\mathbb{E}(M_t) = 0$.

On the other hand, A_t and U_t defined as in (6) are two continuous adapted increasing processes with $A_0 = U_0 = 0$ a.s.. Since $\phi(X(t))$ is nonnegative and $\sup_{x\in\mathbb{R}^d} \|\sigma(\cdot, x)\|_F \in \mathrm{L}^2(\mathbb{R}_+)$, we deduce that $\lim_{t\to+\infty} A_t < +\infty$. Then, we can use Theorem A.9 to conclude that

$$\int_0^{+\infty} \langle \nabla f(X(s)), X(s) - x^\star\rangle ds < +\infty \quad a.s. \tag{7}$$

and

$$\forall x^\star \in \mathcal{S}, \exists \Omega_{x^\star} \in \mathcal{F}, \text{such that } \mathbb{P}(\Omega_{x^\star}) = 1 \text{ and } \lim_{t\to+\infty} \|X(\omega, t) - x^\star\| \text{ exists } \forall \omega \in \Omega_{x^\star}. \tag{8}$$

Since $\mathbb{R}^d$ is separable, there exists a countable set $\mathcal{Z} \subseteq \mathcal{S}$, such that $\mathrm{cl}(\mathcal{Z}) = \mathcal{S}$. Let $\widetilde{\Omega} = \bigcap_{z\in\mathcal{Z}} \Omega_z$. Since $\mathcal{Z}$ is countable, a union bound shows

$$\mathbb{P}(\widetilde{\Omega}) = 1 - \mathbb{P}\left(\bigcup_{z\in\mathcal{Z}} \Omega_z^c\right) \geq 1 - \sum_{z\in\mathcal{Z}} \mathbb{P}(\Omega_z^c) = 1.$$

For arbitrary $x^\star \in \mathcal{S}$, there exists a sequence $(z_k)_{k\in\mathbb{N}} \subseteq \mathcal{Z}$ such that $z_k \to x^\star$. In view of (8), for every $k \in \mathbb{N}$ there exists $\tau_k : \Omega_{z_k} \to \mathbb{R}_+$ such that

$$\lim_{t\to+\infty} \|X(\omega,t) - z_k\| = \tau_k(\omega), \quad \forall \omega \in \Omega_{z_k}. \tag{9}$$

Now, let $\omega \in \widetilde{\Omega}$. Since $\widetilde{\Omega} \subset \Omega_{z_k}$ for any $k \in \mathbb{N}$, and using the triangle inequality and (9), we obtain that

$$\tau_k(\omega) - \|z_k - x^\star\| \leq \liminf_{t\to+\infty} \|X(\omega,t) - x^\star\| \leq \limsup_{t\to+\infty} \|X(\omega,t) - x^\star\| \leq \tau_k(\omega) + \|z_k - x^\star\|.$$

Now, passing to $k \to +\infty$, we deduce

$$\limsup_{k\to+\infty} \tau_k(\omega) \leq \liminf_{t\to+\infty} \|X(\omega,t) - x^\star\| \leq \limsup_{t\to+\infty} \|X(\omega,t) - x^\star\| \leq \liminf_{k\to+\infty} \tau_k(\omega),$$

whence we deduce that $\lim_{k\to+\infty} \tau_k(\omega)$ exists on the set $\widetilde{\Omega}$ of probability 1, and in turn $\lim_{t\to+\infty} \|X(\omega,t) - x^\star\| = \lim_{k\to+\infty} \tau_k(\omega)$.

Let us recall that there exists $\Omega_{\mathrm{cont}} \in \mathcal{F}$ such that $\mathbb{P}(\Omega_{\mathrm{cont}}) = 1$ and $X(\omega,\cdot)$ is continuous for every $\omega \in \Omega_{\mathrm{cont}}$. Now let $x^\star \in \mathcal{S}$ arbitrary, since the limit exists, for every $\omega \in \widetilde{\Omega} \cap \Omega_{\mathrm{cont}}$ there exists $T(\omega)$ such that $\|X(\omega,t) - x^\star\| \leq 1 + \lim_{k\to+\infty} \tau_k(\omega)$ for every $t \geq T(\omega)$. Besides, since $X(\omega,\cdot)$ is continuous, by Bolzano's theorem $\sup_{t\in[0,T(\omega)]} \|X(\omega,t)\| = \max_{t\in[0,T(\omega)]} \|X(\omega,t)\| \overset{\text{def}}{=} h(\omega) < +\infty$. Therefore, $\sup_{t\geq 0} \|X(\omega,t)\| \leq \max\{h(\omega), 1+\lim_{k\to+\infty} \tau_k(\omega) + \|x^\star\|\} < +\infty$.

(iii) Let $M_t = \displaystyle\int_0^t \sigma(s,X(s))dW(s)$. This is a continuous martingale (w.r.t. the filtration $\mathcal{F}_t$), which verifies

$$\mathbb{E}(|M_t|^2) = \mathbb{E}\left(\int_0^t \|\sigma(s,X(s))\|_F^2 ds\right) \leq \mathbb{E}\left(\int_0^{+\infty} \sigma_\infty^2(s) ds\right) < +\infty, \forall t \geq 0.$$

According to Theorem A.8, we deduce that there exists a random variable M_∞ w.r.t. $\mathcal{F}_\infty$, and which verifies: $\mathbb{E}(|M_\infty|^2) < +\infty$, and there exists $\Omega_M \in \mathcal{F}$ such that $\mathbb{P}(\Omega_M) = 1$ and

$$\lim_{t\to+\infty} M_t(\omega) = M_\infty(\omega) \text{ for every } \omega \in \Omega_M.$$

Besides, by convexity of f and (7), we have that there exists $\Omega_f \in \mathcal{F}$ such that $\mathbb{P}(\Omega_f) = 1$ and $(f(X(\omega,\cdot)) - \min f) \in \mathrm{L}^1(\mathbb{R}_+)$ for every $\omega \in \Omega_f$. By Corollary 2.2, we obtain that $\|\nabla f(X(\omega,\cdot))\| \in \mathrm{L}^2(\mathbb{R}_+)$ for every $\omega \in \Omega_f$.

Let $\Omega_{\text{conv}} \overset{\text{def}}{=} \widetilde{\Omega} \cap \Omega_{\text{cont}} \cap \Omega_f \cap \Omega_M$, hence $\mathbb{P}(\Omega_{\text{conv}}) = 1$. Let $\omega \in \Omega_{\text{conv}} \subseteq \Omega_f$ arbitrary, then $\liminf_{t\to+\infty} \|\nabla f(X(\omega,t))\| = 0$. If $\limsup_{t\to+\infty} \|\nabla f(X(\omega,t))\| = 0$ then we conclude. Suppose by contradiction that there exists $\omega_0 \in \Omega_{\text{conv}}$ such that $\limsup_{t\to+\infty} \|\nabla f(X(\omega_0,t))\| > 0$. Then, by Lemma A.3, there exists $\delta(\omega_0) > 0$ satisfying

$$0 = \liminf_{t\to+\infty} \|\nabla f(X(\omega_0,t))\| < \delta(\omega_0) < \limsup_{t\to+\infty} \|\nabla f(X(\omega_0,t))\|,$$

and there exists $(t_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ such that $\lim_{k\to+\infty} t_k = +\infty$,

$$\|\nabla f(X(\omega_0,t_k))\| > \delta \text{ and } t_{k+1} - t_k > 1, \quad \forall k \in \mathbb{N}.$$

We allow ourselves the abuse of notation $X(t) \overset{\text{def}}{=} X(\omega_0,t)$ and $\delta \overset{\text{def}}{=} \delta(\omega_0)$ during the rest of the proof from this point.

Let $\varepsilon \in \left]0, \min\left(\frac{\delta^2}{4L^2}, 1\right)\right[$. Note that this choice entails that the intervals $\left([t_k, t_k + \frac{\varepsilon}{2}]\right)_{k\in\mathbb{N}}$ are disjoint. On the other hand, according to the convergence property of M_t and the fact that $\|\nabla f(X(\cdot))\| \in \mathrm{L}^2(\mathbb{R}_+)$, there exists $k' > 0$ such that for every $k \geq k'$

$$\sup_{t\geq t_k} |M_t - M_{t_k}|^2 < \frac{\varepsilon}{4} \text{ and } \int_{t_k}^{+\infty} \|\nabla f(X(s))\|^2 ds \leq \frac{1}{2}.$$

Besides, for every $k \geq k'$, $t \in [t_k, t_k + \frac{\varepsilon}{2}]$

$$\|X(t) - X(t_k)\|^2 \leq 2(t - t_k) \int_{t_k}^{t} \|\nabla f(X(s))\|^2 ds + 2|M_t - M_{t_k}|^2 \leq (t - t_k) + \frac{\varepsilon}{2} \leq \varepsilon.$$

Since $f \in C_L^{1,1}(\mathbb{R}^d)$ and $L^2\varepsilon \leq \left(\frac{\delta}{2}\right)^2$ by assumption on ε, we have that for every $k \geq k'$ and $t \in [t_k, t_k + \frac{\varepsilon}{2}]$

$$\|\nabla f(X(t)) - \nabla f(X(t_k))\|^2 \leq L^2\|X(t) - X(t_k)\|^2 \leq \left(\frac{\delta}{2}\right)^2.$$

Therefore, for every $k \geq k'$, $t \in [t_k, t_k + \frac{\varepsilon}{2}]$

$$\|\nabla f(X(t))\| \geq \|\nabla f(X(t_k))\| - \underbrace{\|\nabla f(X(t)) - \nabla f(X(t_k))\|}_{\leq \frac{\delta}{2}} \geq \frac{\delta}{2}.$$

Finally,

$$\int_0^{+\infty} \|\nabla f(X(s))\|^2 ds \geq \sum_{k\geq k'} \int_{t_k}^{t_k+\frac{\varepsilon}{2}} \|\nabla f(X(s))\|^2 ds \geq \sum_{k\geq k'} \frac{\delta^2\varepsilon}{8} = +\infty,$$

which contradicts $\|\nabla f(X(\cdot))\| \in \mathrm{L}^2(\mathbb{R}_+)$. So,

$$\limsup_{t\to+\infty} \|\nabla f(X(\omega,t))\| = \liminf_{t\to+\infty} \|\nabla f(X(\omega,t))\| = \lim_{t\to+\infty} \|\nabla f(X(\omega,t))\| = 0, \quad \forall \omega \in \Omega_{\text{conv}}.$$

Let $x^\star \in \mathcal{S}$ and $\omega \in \Omega_{\text{conv}}$ taken arbitrary. By convexity and Cauchy-Schwarz inequality:

$$0 \leq f(X(\omega,t)) - \min f \leq \|\nabla f(X(\omega,t))\|\|X(\omega,t) - x^\star\|.$$

The claim then follows because we have already obtained that $\lim_{t\to+\infty}\|X(\omega,t) - x^\star\|$ exists, and $\lim_{t\to\infty} \|\nabla f(X(\omega,t))\| = 0$.

(iv) Let $\omega \in \Omega_{\text{conv}}$ and $\widetilde{X}(\omega)$ be a sequential limit point of $X(\omega, t)$. Equivalently, there exists an increasing sequence $(t_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ such that $\lim_{k\to+\infty} t_k = +\infty$ and

$$\lim_{k\to+\infty} X(\omega, t_k) = \widetilde{X}(\omega).$$

Since $\lim_{t\to+\infty} f(X(\omega, t)) = \min f$ and by continuity of f, we obtain directly that $\widetilde{X}(\omega) \in \mathcal{S}$. Finally, by Opial's Lemma (see [45]) we conclude that there exists $X^\star(\omega) \in \mathcal{S}$ such that $\lim_{t\to+\infty} X(\omega, t) = X^\star(\omega)$. In other words, since $\omega \in \Omega_{\text{conv}}$ was arbitrary, there exists an $\mathcal{S}$-valued random variable $X^\star$ such that $\lim_{t\to+\infty} X(t) = X^\star$ a.s..

□

3.2 Convergence rates of the objective

Our first result, stated below, summarizes the global convergence rates in expectation satisfied by the trajectories of (SDE).

Theorem 3.3. *Consider the dynamic (SDE) where f and σ satisfy the assumptions (H_0) and (H). Additionally, $X_0 \in \mathrm{L}^2(\Omega; \mathbb{R}^d)$ and is $\mathcal{F}_0$-measurable. The following statements are satisfied by the solution trajectory $X \in S_d^2$ of (SDE):*

(i) Let $\overline{f\circ X}(t) \overset{\text{def}}{=} t^{-1}\int_0^t f(X(s))ds$ and $\overline{X}(t) = t^{-1}\int_0^t X(s)ds$. Then

$$\mathbb{E}\left(f(\overline{X}(t)) - \min f\right) \le \mathbb{E}\left(\overline{f\circ X}(t) - \min f\right) \le \frac{\mathbb{E}\left(\operatorname{dist}(X_0, \mathcal{S})^2\right)}{2t} + \frac{\sigma_*^2}{2}, \quad \forall t > 0. \tag{10}$$

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

$$\mathbb{E}\left(f(\overline{X}(t)) - \min f\right) \le \mathbb{E}\left(\overline{f\circ X}(t) - \min f\right) = \mathcal{O}\left(\frac{1}{t}\right). \tag{11}$$

(ii) Moreover, if $f \in \Gamma_\mu(\mathbb{R}^d)$ with $\mu > 0$, then $\mathcal{S} = \{x^\star\}$ and

(a)

$$\mathbb{E}\left(\|X(t) - x^\star\|^2\right) \le \mathbb{E}\left(\|X_0 - x^\star\|^2\right)e^{-\mu t} + \frac{\sigma_*^2}{\mu}, \quad \forall t \ge 0. \tag{12}$$

Besides, if σ_∞ is non-increasing and vanishes at infinity, then for every $\lambda \in]0, 1[$:

$$\mathbb{E}\left(\|X(t) - x^\star\|^2\right) \le \mathbb{E}\left(\|X_0 - x^\star\|^2\right)e^{-\mu t} + \frac{\sigma_*^2}{\mu}e^{-\mu(1-\lambda)t} + \sigma_\infty^2(\lambda t), \quad \forall t \ge 0. \tag{13}$$

(b) Furthermore,

$$\mathbb{E}\left(f(X(t)) - \min f\right) \le \mathbb{E}\left(f(X_0) - \min f\right)e^{-2\mu t} + \frac{L\sigma_*^2}{4\mu}, \quad \forall t \ge 0. \tag{14}$$

Besides, if σ_∞ is non-increasing and vanishes at infinity, then for every $\lambda \in]0, 1[$:

$$\mathbb{E}\left(f(X(t)) - \min f\right) \le \mathbb{E}\left(f(X_0) - \min f\right)e^{-2\mu t} + \frac{L\sigma_*^2}{4\mu}e^{-2\mu(1-\lambda)t} + \frac{L}{2}\sigma_\infty^2(\lambda t), \quad \forall t \ge 0. \tag{15}$$

Proof. Proof.

(i) Let $x^\star \in \mathcal{S}$. Let $g(t) = \phi(X(t)) = \frac{\|X(t)-x^\star\|^2}{2}$ and $G(t) = \mathbb{E}(g(t))$. By applying Proposition 2.3 with ϕ, and using the convexity of f, we obtain

$$\begin{aligned} G(t) - G(0) &= \mathbb{E}\left(\int_0^t \langle \nabla f(X(s)), x^\star - X(s)\rangle ds\right) + \frac{1}{2}\mathbb{E}\left(\int_0^t \mathrm{tr}[\Sigma(s, X(s))]ds\right) \\ &\leq -\mathbb{E}\left(\int_0^t (f(X(s)) - \min f)ds\right) + \frac{1}{2}\mathbb{E}\left(\int_0^t \mathrm{tr}[\Sigma(s, X(s))]ds\right) \qquad (16) \\ &\leq -\mathbb{E}\left(\int_0^t (f(X(s)) - \min f)ds\right) + \frac{\sigma_*^2}{2}t. \end{aligned}$$

Then rearranging the terms in (16), using $G(t) \geq 0$, and dividing by $t > 0$, we obtain

$$\frac{1}{t}\mathbb{E}\left(\int_0^t (f(X(s)) - \min f)ds\right) \leq \frac{\mathbb{E}\left(\|X_0 - x^\star\|^2\right)}{2t} + \frac{\sigma_*^2}{2}, \quad \forall t > 0. \tag{17}$$

Since $x^\star$ is arbitrary, by taking the infimum with respect to $x^\star \in \mathcal{S}$ in (17), we obtain

$$\frac{1}{t}\mathbb{E}\left(\int_0^t (f(X(s)) - \min f)ds\right) \leq \frac{\mathbb{E}\left(\mathrm{dist}(X_0, \mathcal{S})^2\right)}{2t} + \frac{\sigma_*^2}{2}, \quad \forall t > 0. \tag{18}$$

Moreover, if $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$, then using inequality (16), we have

$$G(t) - G(0) \leq -\mathbb{E}\left(\int_0^t (f(X(s)) - \min f)ds\right) + \frac{1}{2}\left(\int_0^{+\infty} \sigma_\infty^2(s)ds\right).$$

Rearranging as before, we conclude that

$$\frac{1}{t}\mathbb{E}\left(\int_0^t (f(X(s)) - \min f)ds\right) \leq \frac{\mathbb{E}\left(\mathrm{dist}(X_0, \mathcal{S})^2\right)}{2t} + \frac{1}{2t}\int_0^{+\infty} \sigma_\infty^2(s)ds, \quad \forall t > 0. \tag{19}$$

Then complete the result with the inequality

$$\mathbb{E}\left(f(\overline{X}(t)) - \min f\right) \leq \mathbb{E}\left(\overline{f \circ X}(t) - \min f\right)$$

which follows from convexity of f and Jensen's inequality.

(ii) (a) Let $g(t) = \phi(X(t)) = \frac{\|X(t)-x^\star\|^2}{2}$, $G(t) = \mathbb{E}(g(t))$. By Proposition 2.3 with ϕ, we obtain

$$G(t) - G(0) = \mathbb{E}\left(\int_0^t \langle -\nabla f(X(s)), X(s) - x^\star\rangle ds\right) + \frac{1}{2}\mathbb{E}\left(\int_0^t \mathrm{tr}[\Sigma(s, X(s))]ds\right). \tag{20}$$

Using that $f \in \Gamma_\mu(\mathbb{R}^d)$, we deduce that

$$G(t) \leq G(0) - \mu\int_0^t G(s)ds + \int_0^t \frac{\sigma_*^2}{2}ds, \quad \forall t \geq 0. \tag{21}$$

In order to invoke Lemma A.2, we solve the ODE

$$\begin{cases} y'(t) & = -\mu y(t) + \frac{\sigma_*^2}{2}, \quad \forall t > 0, \\ y(0) & = \mathbb{E}\left(\frac{\|X_0 - x^\star\|^2}{2}\right). \end{cases}$$

Solving it by the integrating factor method, we conclude that

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

Suppose now that σ_∞ is non-increasing and vanishes at infinity. We can bound the trace term by σ_∞^2 in (20). To use Lemma A.2, we need to solve

$$\begin{cases} y'(t) & = -\mu y(t) + \frac{\sigma_\infty^2(t)}{2}, \quad \forall t > 0, \\ y(0) & = \mathbb{E}\left(\frac{\|X_0 - x^\star\|^2}{2}\right). \end{cases}$$

Let $\lambda \in]0,1[$, using the integrating factor method, we get

$$\begin{aligned} y(t) &\leq y(0)e^{-\mu t} + e^{-\mu t}\int_0^t \frac{\sigma_\infty^2(s)}{2} e^{\mu s} ds \\ &\leq y(0)e^{-\mu t} + e^{-\mu t}\left(\int_0^{\lambda t} \frac{\sigma_\infty^2(s)}{2} e^{\mu s} ds + \int_{\lambda t}^t \frac{\sigma_\infty^2(s)}{2} e^{\mu s} ds\right) \\ &\leq y(0)e^{-\mu t} + e^{-\mu t}\left(\frac{\sigma_*^2}{2}\int_0^{\lambda t} e^{\mu s} ds + \frac{\sigma_\infty^2(\lambda t)}{2}\int_{\lambda t}^t e^{\mu s} ds\right) \\ &\leq y(0)e^{-\mu t} + e^{-\mu t}\left(\frac{\sigma_*^2}{2\mu} e^{\mu\lambda t} + \frac{\sigma_\infty^2(\lambda t)}{2} e^{\mu t}\right), \quad \forall t \geq 0. \end{aligned}$$

According to Lemma A.2, we deduce that

$$G(t) \leq \mathbb{E}\left(\frac{\|X_0 - x^\star\|^2}{2}\right) e^{-\mu t} + \frac{\sigma_*^2}{2\mu} e^{-\mu(1-\lambda)t} + \frac{\sigma_\infty^2(\lambda t)}{2}, \quad \forall t \geq 0,$$

which is our claim (13).

(b) Since $f \in \Gamma_\mu(\mathbb{R}^d)$, it is well known that f satisfies the Polyak-Łojasiewicz inequality, i.e.

$$2\mu(f(x) - \min f) \leq \|\nabla f(x)\|^2, \quad \forall x \in \mathbb{R}^d,$$

(see Section 4 for an explanation of this inequality). Besides, since $f \in \Gamma_0(\mathbb{R}^d) \cap C_L^{1,1}(\mathbb{R}^d)$ and $X_0 \in \mathrm{L}^2(\Omega; \mathbb{R}^d)$, we have that $\mathbb{E}(f(X_0) - \min f) < +\infty$.

We take the function $\widehat{\phi}(x) = f(x) - \min f$ and apply Proposition 2.3. Then, defining $\widehat{g}(t) = f(X(t)) - \min f$ and $\widehat{G}(t) = \mathbb{E}(g(t))$, we obtain

$$\widehat{G}(t) - \widehat{G}(0) \leq -\mathbb{E}\left(\int_0^t \|\nabla f(X(s))\|^2 ds\right) + \frac{L}{2}\int_0^t \sigma_\infty^2(s) ds.$$

Using the Polyak-Łojasiewicz inequality, we end up having

$$\widehat{G}(t) - \widehat{G}(0) \leq -2\mu\left(\int_0^t \widehat{G}(s) ds\right) + \frac{L}{2}\int_0^t \sigma_\infty^2(s) ds. \quad (22)$$

And we conclude by continuing the analysis as in the previous item after arriving to (21).

□

Under a stronger assumption on σ_∞, we also have the following pointwise sublinear convergence rate in expectation.

Proposition 3.4. *Consider the dynamic (SDE) where f and σ satisfy the assumptions (H_0) and (H), respectively. Additionally, $X_0 \in \mathrm{L}^2(\Omega;\mathbb{R}^d)$ and is $\mathcal{F}_0$-measurable. Assume that there exists $K \geq 0, \beta \in [0,1[$ such that*

$$\int_0^t (s+1)\sigma_\infty^2(s)ds \leq Kt^\beta, \quad \forall t \geq 0. \tag{23}$$

Then the solution trajectory $X \in S_d^2$ of (SDE) satisfies

$$\mathbb{E}\left(f(X(t)) - \min f\right) = \mathcal{O}(t^{\beta-1}).$$

Proof. Proof. Given $x^\star \in \mathcal{S}$, let us apply Proposition 2.4 successively with $V_1(t,x) = t(f(x) - \min f)$, then with $V_2(x) = \frac{1}{2}\|x - x^\star\|^2$. Taking the expectation and adding the two results, we get

$$\begin{aligned}\mathbb{E}\left(V_1(t,X(t)) + V_2(X(t))\right) &\leq \mathbb{E}\left(\frac{\|X_0 - x^\star\|^2}{2}\right) + \frac{L}{2}\int_0^t s\sigma_\infty^2(s)ds + \frac{1}{2}\int_0^t \sigma_\infty^2(s)ds \\ &\leq \mathbb{E}\left(\frac{\|X_0 - x^\star\|^2}{2}\right) + \frac{\max\{1,L\}}{2}\left(\int_0^t (s+1)\sigma_\infty^2(s)ds\right),\end{aligned}$$

where we have used the convexity of f in the first inequality. Then we conclude that

$$\mathbb{E}(f(X(t)) - \min f) \leq \frac{\mathbb{E}\left(\|X_0 - x^\star\|^2\right)}{2t} + \frac{K\max\{1,L\}}{2}t^{\beta-1} = \mathcal{O}(t^{\beta-1}).$$

□

When f is also C^2 and the first order moment of σ_∞^2 is bounded, we get an improved $o(t^{-1})$ global convergence rate on the objective in almost sure sense.

Theorem 3.5. *Consider the dynamic (SDE). Assume that $f \in C^2(\mathbb{R}^d)$ and σ satisfy the assumptions (H_0) and (H), respectively. Additionally, $X_0 \in \mathrm{L}^2(\Omega;\mathbb{R}^d)$ and is $\mathcal{F}_0$-measurable, and that $t \mapsto t\sigma_\infty^2(t) \in \mathrm{L}^1(\mathbb{R}_+)$. Then, the solution trajectory $X \in S_d^2$ of (SDE) obeys:*

(i) $t \mapsto t\|\nabla f(X(t))\|^2 \in \mathrm{L}^1(\mathbb{R}_+)$ *a.s..*

(ii) $f(X(t)) - \min f = o(t^{-1})$ *a.s..*

Proof. Proof. By applying Itô's formula in Proposition 2.3 with $\phi(t,x) = t(f(x) - \min f)$ we get

$$\begin{aligned}t(f(X(t)) - \min f) = &\int_0^t (f(X(s)) - \min f)ds + \frac{1}{2}\int_0^t \mathrm{tr}[\Sigma(s,X(s))\,\nabla^2 f(X(s))]sds \\ &- \int_0^t s\|\nabla f(X(s))\|^2 ds + \int_0^t \langle s\sigma^\top(s,X(s))\nabla f(X(s)), dW(s)\rangle.\end{aligned}$$

By (7) and convexity of f, we deduce that $f(X(\cdot)) - \min f \in \mathrm{L}^1(\mathbb{R}_+)$ a.s.. Moreover,

$$\int_0^{+\infty} s\mathrm{tr}[\Sigma(s,X(s))\,\nabla^2 f(X(s))]ds \leq L\int_0^{+\infty} s\sigma_\infty^2(s)ds < +\infty.$$

Then by Theorem A.9, we have that $\lim_{t\to+\infty} t(f(X(t)) - \min f)$ exists a.s. and $\int_0^{+\infty} t\|\nabla f(X(t))\|^2 dt < +\infty$ a.s.. Finally, by Lemma A.1, we conclude that $\lim_{t\to+\infty} t(f(X(t)) - \min f) = 0$ a.s.. □

4 Convergence rates under Łojasiewicz inequality

The local convergence rate of the first-order descent methods can be understood using the Łojasiewicz property and the associated Łojasiewicz exponent, see [3, 27]. The Łojasiewicz property has its roots in algebraic geometry, and it essentially describes a relationship between the objective value and its gradient (or subgradient).

Definition 4.1 (Łojasiewicz inequality). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function with $\mathcal{S} = \mathrm{argmin}(f) \neq \emptyset$ and $q \in [0, 1[$. f satisfies the Łojasiewicz inequality with exponent q at $\bar{x} \in \mathcal{S}$ if there exists a neighborhood $\mathcal{V}_{\bar{x}}$ of $\bar{x}$, $r > \min f$ and $\mu > 0$ such that

$$\mu(f(x) - \min f)^q \leq \|\nabla f(x)\|, \quad \forall x \in \mathcal{V}_{\bar{x}} \cap [\min f < f < r]. \tag{24}$$

The function f has the Łojasiewicz property on $\mathcal{S}$ if it obeys (24) at each point of $\mathcal{S}$ with the same constant μ and exponent q, and we will write $f \in \text{Ł}^q(\mathcal{S})$.

Error bounds have also been successfully applied to various branches of optimization, and in particular to complexity analysis, see [47]. Of particular interest in our setting is the Hölderian error bound.

Definition 4.2 (Hölderian error bound). Let $f : \mathbb{R}^d \to \mathbb{R}$ be a proper function such that $\mathcal{S} = \mathrm{argmin}(f) \neq \emptyset$. f satisfies a Hölderian (or power-type) error bound inequality with exponent $p \geq 1$, and we write $f \in \mathrm{EB}^p$, if there exists $\gamma > 0$ and $r > \min f$ such that

$$f(x) - \min f \geq \gamma \mathrm{dist}(x, \mathcal{S})^p, \quad \forall x \in [\min f \leq f \leq r]. \tag{25}$$

For a given $r > \min f$ such that (25) holds, we will use the shorthand notation $f \in \mathrm{EB}^p([f \leq r])$.

A deep result due to Łojasiewicz states that for arbitrary continuous semi-algebraic functions, the Hölderian error bound inequality holds on any compact set, and the Łojasiewicz inequality holds at each point; see [38, 39]. In fact, for convex functions, the Łojasiewicz property and Hölderian error bound are actually equivalent.

Proposition 4.3. *Assume that $f \in \Gamma_0(\mathbb{R}^d) \cap C^1(\mathbb{R}^d)$ with $\mathcal{S} = \mathrm{argmin}(f) \neq \emptyset$. Let $q \in [0, 1[$, $p \overset{\text{def}}{=} \frac{1}{1-q} \geq 1$ and $r > \min f$. Then f verifies the Łojasiewicz inequality (24) at $\bar{x} \in \mathcal{S}$ if and only if the Hölderian error bound (25) holds on $\mathcal{V}_{\bar{x}} \cap [\min f < f < r]$.*

Proof. Proof. Combine [12, Lemma 4 and Theorem 5]. □

We are now ready to state the following ergodic local convergence rate.

Proposition 4.4. *Consider the hypotheses of Theorem 3.3 and let $\varepsilon > 0$. If $f \in \mathrm{EB}^p([f \leq r_\varepsilon])$ for $r_\varepsilon > \min f + \frac{\sigma_*^2}{2} + \varepsilon$, then $\exists t_\varepsilon > 0$ such that*

$$\mathrm{dist}\left(\mathbb{E}(\overline{X}(t)), S\right) = \mathcal{O}(t^{-\frac{1}{p}}) + \mathcal{O}\left(\sigma_*^{\frac{2}{p}}\right), \quad \forall t \geq t_\varepsilon.$$

Proof. Proof. There exists $t_\varepsilon > 0$ such that for all $t \geq t_\varepsilon$, $\frac{\mathbb{E}(\mathrm{dist}(X_0,\mathcal{S})^{\in})}{2t} < \varepsilon$. Thus, from (10) and Jensen's inequality, we have

$$f\left(\mathbb{E}[\overline{X}(t)]\right) \leq \mathbb{E}[f(\overline{X}(t))] \leq \min f + \frac{\sigma_*^2}{2} + \varepsilon \leq r_\varepsilon, \quad \forall t \geq t_\varepsilon.$$

This reflects the fact that, $\mathbb{E}[\overline{X}(t)] \in [f \leq r_\varepsilon]$ for $t \geq t_\varepsilon$. Using Theorem 3.3 and that $f \in \mathrm{EB}^p([f \leq r_\varepsilon])$, letting $\gamma > 0$ the coefficient of the error bound, we have

$$\gamma \mathrm{dist}(\mathbb{E}(\overline{X}(t)), \mathcal{S})^p \leq f(\mathbb{E}[\overline{X}(t)]) - \min f \leq \frac{\mathbb{E}\left(\mathrm{dist}(X_0, \mathcal{S})^2\right)}{2t} + \frac{\sigma_*^2}{2}, \quad \forall t \geq t_\varepsilon.$$

Dividing by $\gamma > 0$, then taking the power $\frac{1}{p}$ on both sides of the previous inequality, and finally using the subadditivity of the power function $(\cdot)^{1/p}$ on $[0, +\infty[$ (recall $p \geq 1$), we obtain

$$\mathrm{dist}(\mathbb{E}(\overline{X}(t)), \mathcal{S}) \leq \left(\frac{\mathbb{E}\left(\mathrm{dist}(X_0, \mathcal{S})^2\right)}{2\gamma}\right)^{\frac{1}{p}} t^{-\frac{1}{p}} + \left(\frac{\sigma_*^2}{2\gamma}\right)^{\frac{1}{p}}, \quad \forall t \geq t_\varepsilon.$$

□

4.1 Discussion on the localization of the process

Let us take a moment to elaborate on the localization of the process $X(t)$ generated by (SDE) when $f \in C_L^{1,1}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d)$ and $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$. This discussion is essential to understand the challenges underlying the analysis of the local convergence properties and rates in a stochastic setting under (local) error bounds. First, observe that the hypothesis of Lipschitz continuity of the gradient is incompatible with a global hypothesis of error bound or Łojasiewicz inequality unless the exponent is $p = 2$ or $q = \frac{1}{2}$, respectively. Therefore, we can only ask for these inequalities to be locally satisfied. Even though, thanks to convexity, we could introduce a global desingularizing function (see [12, Theorem 3]), this function would not be concave nor convex, a fundamental property usually at the heart of the local analysis. In recent literature on stochastic processes and local properties, it is usual to find hypotheses about the almost sure localization of the process or that it is essentially bounded. Nevertheless, these assumptions are unrealistic or outright false due to the behavior of the Brownian Motion. Hence, we will avoid making these kinds of assumptions.

What we will do is to consider that by Theorem 3.1 we have that $\lim_{t \to +\infty} f(X(t)) = \min f$ a.s., which means that there exists $\Omega_{\mathrm{conv}} \in \mathcal{F}$ such that $\mathbb{P}(\Omega_{\mathrm{conv}}) = 1$, and $(\forall r > \min f, \forall \omega \in \Omega_{\mathrm{conv}})$, $(\exists t_r(\omega) > 0)$ such that $(\forall t > t_r(\omega))$, $X(\omega, t) \in [f \leq r]$. However, one should not infer from this that $X(t) \in [f \leq r]$ a.s. for t large enough. Indeed, t_r is a random variable which cannot be in general bounded uniformly on Ω_{conv}. Rather, in this paper, we will invoke measure theoretic arguments to pass from a.s. convergence to almost uniform convergence thanks to Egorov's theorem (see Theorem A.4). More precisely, we will show that

$$\boxed{\begin{array}{c}(\forall \delta > 0, \forall r > \min f), (\exists \Omega_\delta \in \mathcal{F} \text{ s.t. } \mathbb{P}(\Omega_\delta) \geq 1 - \delta \text{ and } \exists \hat{t}_{r,\delta} > 0), (\forall \omega \in \Omega_\delta, \forall t > \hat{t}_{r,\delta}), \\ X(\omega, t) \in [f \leq r].\end{array}}$$

Hence, this property will allow us to localize $X(t)$ in the sublevel set of f at r for t large enough with probability at least $1 - \delta$. In turn, we will be able to invoke the error bound (or Łojasiewicz) inequality.

4.2 Convergence rates under Łojasiewicz Inequality

Let $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+), L > 0, \delta > 0, \beta \in [0,1[$ and some positive constants C_*, C_{**}, C_K. Consider the functions $h_\delta, l_\delta, k_\delta : \mathbb{R}_+ \to \mathbb{R}$ defined by:

$$h_\delta(t) = \sigma_\infty^2(t) + C_*\sqrt{\delta}\frac{\sigma_\infty^2(t)}{2\sqrt{\int_{\hat{t}_\delta}^t \sigma_\infty^2(u)du}}, \tag{26}$$

$$l_\delta(t) = \frac{L}{2}\sigma_\infty^2(t) + C_{**}\sqrt{\delta}\frac{\sigma_\infty^2(t)}{2\sqrt{\int_{\hat{t}_\delta}^t \sigma_\infty^2(u)du}}, \tag{27}$$

$$k_\delta(t) = \frac{L}{2}\sigma_\infty^2(t) + C_K\sqrt{\delta}\frac{\sigma_\infty^2(t)t^{\beta-1}}{2\sqrt{\int_{\hat{t}_\delta}^t \sigma_\infty^2(u)u^{\beta-1}du}}. \tag{28}$$

We are now ready to state our main local convergence result.

Theorem 4.5. *Consider (SDE) where f and σ satisfy the assumptions (H_0) and (H), respectively. Additionally, $X_0 \in \mathrm{L}^4(\Omega;\mathbb{R}^d)$ and is $\mathcal{F}_0$-measurable. Let $X \in S_d^4$ the unique solution trajectory of (SDE). Suppose also that $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$ ($C_\infty \overset{\text{def}}{=} \|\sigma_\infty\|_{\mathrm{L}^2(\mathbb{R}_+)}$). Let $p \geq 2$ and $q \overset{\text{def}}{=} 1 - \frac{1}{p} \in [\frac{1}{2}, 1[$, and assume that $f \in \text{Ł}^q(\mathcal{S})$. Consider also the positive constants $C_*, C_{**}, C_K, C_d, C_f$ (detailed in the proof). Then, for all $\delta > 0$, there exists a measurable set Ω_δ such that $\mathbb{P}(\Omega_\delta) \geq 1-\delta$ and $\hat{t}_\delta > 0$ such that the following statements hold.*

(i) If $p = 2$ and σ_∞ is non-increasing, then σ_∞ vanishes at infinity and

(a) there exists $\gamma > 0$ such that for every $\lambda \in]0,1[$,

$$\begin{aligned}\mathbb{E}\left(\frac{\mathrm{dist}(X(t),\mathcal{S})^2}{2}\right) \leq{}& e^{-2\gamma(t-\hat{t}_\delta)}\mathbb{E}\left(\frac{\mathrm{dist}(X(\hat{t}_\delta),\mathcal{S})^2}{2}\right) \\ &+ e^{-2\gamma(1-\lambda)(t-\hat{t}_\delta)}(C_\infty^2 + C_*C_\infty\sqrt{\delta}) \\ &+ \frac{h_\delta(\hat{t}_\delta + \lambda(t-\hat{t}_\delta))}{2\gamma} + C_d\sqrt{\delta}, \qquad \forall t > \hat{t}_\delta;\end{aligned} \tag{29}$$

(b) there exists $\mu > 0$ such that for every $\lambda \in]0,1[$,

$$\begin{aligned}\mathbb{E}\left(f(X(t)) - \min f\right) \leq{}& e^{-\mu^2(t-\hat{t}_\delta)}\mathbb{E}([f(X(\hat{t}_\delta)) - \min f]\mathbb{1}_{\Omega_\delta}) \\ &+ e^{-\mu^2(1-\lambda)(t-\hat{t}_\delta)}\left(\frac{LC_\infty^2}{2} + C_{**}C_\infty\sqrt{\delta}\right) \\ &+ \frac{l_\delta(\hat{t}_\delta + \lambda(t-\hat{t}_\delta))}{\mu^2} + C_f\sqrt{\delta}, \qquad \forall t > \hat{t}_\delta.\end{aligned} \tag{30}$$

Moreover, if (23) holds, then

$$\begin{aligned}\mathbb{E}\left(f(X(t)) - \min f\right) \leq{}& e^{-\mu^2(t-\hat{t}_\delta)}\mathbb{E}([f(X(\hat{t}_\delta)) - \min f]\mathbb{1}_{\Omega_\delta}) \\ &+ e^{-\mu^2(1-\lambda)(t-\hat{t}_\delta)}\left(\frac{LC_\infty^2}{2} + C_KC_\infty\sqrt{\hat{t}_\delta^{\beta-1}}\sqrt{\delta}\right) \\ &+ \frac{k_\delta(\hat{t}_\delta + \lambda(t-\hat{t}_\delta))}{\mu^2} + C_f\sqrt{\delta}, \qquad \forall t > \hat{t}_\delta.\end{aligned} \tag{31}$$

(ii) If $p > 2$:

(a) There exists $\gamma > 0$ such that

$$\mathbb{E}\left(\frac{\operatorname{dist}(X(t),\mathcal{S})^2}{2}\right) \leq y_\delta^\star(t) + C_d\sqrt{\delta}, \qquad \forall t > \hat{t}_\delta, \tag{32}$$

where $y_\delta^\star$ is the solution of the Cauchy problem

$$\text{(C.1)} \begin{cases} y'(t) & = -2^{\frac{p}{2}}\gamma y^{\frac{p}{2}} + h_\delta(t), \quad \forall t > \hat{t}_\delta \\ y(\hat{t}_\delta) & = \mathbb{E}\left(\frac{\operatorname{dist}(X(\hat{t}_\delta,\mathcal{S}))^2}{2}\mathbb{1}_{\Omega_\delta}\right). \end{cases}$$

(b) There exists $\mu > 0$ such that

$$\mathbb{E}\left[f(X(t)) - \min f\right] \leq w_\delta^\star(t) + C_f\sqrt{\delta}, \qquad \forall t > \hat{t}_\delta, \tag{33}$$

where $w_\delta^\star$ is the solution of the Cauchy problem

$$\text{(C.2)} \begin{cases} y'(t) & = -\mu^2 y(t)^{2q} + l_\delta(t), \quad t > \hat{t}_\delta \\ y(\hat{t}_\delta) & = \mathbb{E}([f(X(\hat{t}_\delta) - \min f]\mathbb{1}_{\Omega_\delta}). \end{cases}$$

Moreover, if (23) holds, then

$$\mathbb{E}\left[f(X(t)) - \min f\right] \leq z_\delta^\star(t) + C_f\sqrt{\delta}, \qquad \forall t > \hat{t}_\delta, \tag{34}$$

where $z_\delta^\star$ is the solution of the Cauchy problem

$$\text{(C.3)} \begin{cases} y'(t) & = -\mu^2 y(t)^{2q} + k_\delta(t), \quad \forall t > \hat{t}_\delta \\ y(\hat{t}_\delta) & = \mathbb{E}([f(X(\hat{t}_\delta) - \min f]\mathbb{1}_{\Omega_\delta}). \end{cases}$$

Before proceeding with the proof, a few remarks are in order.

Remark 4.6. The hypothesis that f has a Lipschitz continuous gradient restricts the Łojasiewicz exponent q to be in $[\frac{1}{2}, 1[$.

Remark 4.7. If we have a *global* error bound (or Łojasiewicz inequality), then as noted in the discussion of Section 4.1, one necessarily has $p = 2$ (or $q = \frac{1}{2}$). In this case, the statements (i) of Theorem 4.5 will hold if we replace $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$ by σ_∞ non-increasing and vanishing at infinity, δ by 0 and $\hat{t}_\delta$ by 0. Clearly, one recovers (13).

Remark 4.8. It is important to highlight the trade-off in the selection of δ. Although δ can be arbitrarily small, the time from which the inequalities are satisfied, $\hat{t}_\delta$, surely increases when δ approaches 0^+. Besides, let $q_{\delta,\hat{t}_\delta} : \mathbb{R}_+ \to \mathbb{R}$ be a decreasing function. Our convergence rates in Theorem 4.5 are of the form $\mathbb{E}[m(X(t))] \leq q_{\delta,\hat{t}_\delta}(t) + C\sqrt{\delta}, \quad \forall t > t_\delta$, where $m(x) = f(x) - \min f$ or $m(x) = \operatorname{dist}(x,\mathcal{S})^2/2$. Let $\varepsilon \in]0, 2C[$ and $\delta^\star = \frac{\varepsilon^2}{4C^2}$. Then one gets an ε-optimal solution for $t > \max\{q^\star(\varepsilon), \hat{t}_{\delta^\star}\}$.

Remark 4.9. Referring again to the discussion of Section 4.1, we have that there exists $\delta > 0$ and $\Omega_\delta \in \mathcal{F}$ with $\mathbb{P}(\Omega_\delta) \geq 1 - \delta$ over which we have uniform convergence of the objective. If δ could be 0 (a.s. uniform convergence), there would be a $\hat{t} > 0$ such that $X(t) \in [f \leq r], \forall t > \hat{t}$ a.s.. Thus, the statements in Theorem 4.5 would hold if we replace δ by 0 and $\hat{t}_\delta$ by $\hat{t}$. The proof is far easier in this case. It is however not easy to ensure the existence of such $\hat{t}$ in general.

Remark 4.10. In order to find explicit convergence rates in Theorem 4.5 we have to solve or bound the solution of the Cauchy problems (C.1), (C.2) and (C.3). We can generalize these problems as follows: Let $a > 0, b > 1, \hat{t}_\delta > 0, \delta > 0, y_0(\hat{t}_\delta, \delta) > 0$ and p_δ a nonnegative integrable function. Consider

$$\text{(C.0)} \begin{cases} y'(t) & = -ay^b(t) + p_\delta(t), \quad \forall t > \hat{t}_\delta \\ y(\hat{t}_\delta) & = y_0(\hat{t}_\delta, \delta). \end{cases}$$

Although one could give an explicit ad-hoc p_δ in order to find a particular solution of (C.0), the dependence of this function on $\hat{t}_\delta$ is unavoidable, which is a problem, since p_δ is explicitly related to σ_∞, and this in turn is the one that defines $\hat{t}_\delta$ in the first place.

To the best of our knowledge, there is no way to arithmetically solve this non linear ODE, not even a sharp bound of the solution.

Nevertheless, if $y(t) = \mathcal{O}\left((t+1)^{-\frac{1}{b-1}}\right)$, then $p_\delta(t) = \mathcal{O}\left((t+1)^{-\frac{b}{b-1}}\right)$. Which leads us to make the following conjecture:

Conjecture 4.11. If $p_\delta = \mathcal{O}(\sigma_\infty^2)$ and $\sigma_\infty^2(t) = \mathcal{O}\left((t+1)^{-\frac{b}{b-1}}\right)$ (for constants independent of δ and $\hat{t}_\delta$), then $y(t) = \mathcal{O}\left((t+1)^{-\frac{1}{b-1}}\right)$.

Proof. Proof of Theorem 4.5. Given that $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$, if it is non-increasing, we have immediately that it vanishes at infinity. Let $x^\star \in \mathcal{S}$. Let us recall that by claim (i) of Theorem 3.1, there exists $C^* > 0$ such that

$$\sup_{t\geq 0} \mathbb{E}\left(\operatorname{dist}(X(t), \mathcal{S})^2\right) \leq \sup_{t\geq 0} \mathbb{E}\left(\|X(t) - x^\star\|^2\right) \leq C^*.$$

$$\sup_{t\geq 0} \mathbb{E}\left(f(X(t)) - \min f\right) \leq \frac{1}{2}\mathbb{E}(\|\nabla f(X(t)) - \nabla f(x^\star)\|^2) + \frac{1}{2}\mathbb{E}(\|X(t) - x^\star\|^2) \leq \frac{L^2+1}{2}C^* < +\infty.$$

On the other hand, by Theorem 3.1(iii), there exists a set $\Omega_{\text{conv}} \in \mathcal{F}$ such that $\mathbb{P}(\Omega_{\text{conv}}) = 1$ where, for all $\omega \in \Omega_{\text{conv}}$: $\lim_{t\to+\infty} f(X(\omega,t)) = \min f$, $t \mapsto f(X(\omega,t))$ is continuous, and $\lim_{t\to+\infty} \operatorname{dist}(X(\omega,t), \mathcal{S}) = 0$. Then, by Theorem A.4 for every $\delta > 0$ there exists $\Omega_\delta \in \mathcal{F}$ such that $\Omega_\delta \subset \Omega_{\text{conv}}$, $\mathbb{P}(\Omega_\delta) > 1-\delta$ and $f(X(\cdot,t))$ (resp. $\operatorname{dist}(X(\cdot,t), \mathcal{S})$) converges uniformly to $\min f$ (resp. to 0) on Ω_δ. This means that given $r \geq \min f$, and for every $\delta > 0$, there exist $\hat{t}_\delta > 0$ and $\Omega_\delta \in \mathcal{F}$ with $\mathbb{P}(\Omega_\delta) > 1-\delta$ such that $X(\omega,t) \in [f \leq r] \cap \mathcal{V}_\mathcal{S}$ for all $t \geq \hat{t}_\delta$ and $\omega \in \Omega_\delta$, where $\mathcal{V}_\mathcal{S}$ is a neighbourhood of $\mathcal{S}$. On the other hand, since $f \in \text{Ł}^q(\mathcal{S})$, by Proposition 4.3, there exists $r > \min f$ and a neighbourhood $\mathcal{V}_\mathcal{S}$ of $\mathcal{S}$ such that f verifies the p-Hölderian error bound inequality (25) on $[\min f < f < r] \cap \mathcal{V}_\mathcal{S}$. Consequently, for any $\delta > 0$, there exists $t \geq \hat{t}_\delta$ large enough such that the p-Hölderian error bound inequality holds at $X(\omega,t)$ for all $t \geq \hat{t}_\delta$ and $\omega \in \Omega_\delta$.

We are now ready to start. Let $x^\star \in \mathcal{S}$, $\delta > 0$, and $t \geq \hat{t}_\delta$.

(i) $p = 2$:

(a) Let $\widehat{g}(t) = \widehat{\phi}(X(t)) = \frac{\mathrm{dist}(X(t),\mathcal{S})^2}{2}$, $\widehat{G}(t) = \mathbb{E}(\widehat{g}(t)\mathbb{1}_{\Omega_\delta})$, and $\mu > 0$ be the coefficient of the error bound inequality. We have

$$\nabla\widehat{\phi}(X(t)) = X(t) - P_{\mathcal{S}}(X(t)),$$

where $P_{\mathcal{S}}(x)$ is the projection of x on $\mathcal{S}$, so $\widehat{\phi} \in C_1^{1,1}(\mathbb{R}^d)$. We use Proposition 2.4 to obtain

$$\begin{aligned}\widehat{g}(t) - \widehat{g}(\hat{t}_\delta) \leq &- \int_{\hat{t}_\delta}^t \langle \nabla f(X(s), X(s) - P_{\mathcal{S}}(X(s))\rangle \, ds \\ &+ \int_{\hat{t}_\delta}^t \mathrm{tr}[\Sigma(s, X(s))]ds + \int_{\hat{t}_\delta}^t \left\langle \sigma^\top(s, X(s))(X(s) - P_{\mathcal{S}}(X(s))), dW(s)\right\rangle. \quad (35)\end{aligned}$$

We have that $\mathrm{tr}[\Sigma(s, X(s))] \leq \sigma_\infty^2(s)$ and by convexity

$$- \langle \nabla f(X(s), X(s) - P_{\mathcal{S}}(X(s))\rangle \leq - (f(X(s)) - \min f).$$

Therefore,

$$\begin{aligned}\widehat{g}(t) - \widehat{g}(\hat{t}_\delta) \leq &- \int_{\hat{t}_\delta}^t (f(X(s)) - \min f)ds \\ &+ \int_{\hat{t}_\delta}^t \sigma_\infty^2(s)ds + \int_{\hat{t}_\delta}^t \langle \sigma^\top(s, X(s))(X(s) - P_{\mathcal{S}}(X(s))), dW(s)\rangle.\end{aligned}$$

Then, multiplying this inequality by $\mathbb{1}_{\Omega_\delta}$, and taking expectation we obtain

$$\begin{aligned}\widehat{G}(t) - \widehat{G}(\hat{t}_\delta) \leq &-\mathbb{E}\left[\int_{\hat{t}_\delta}^t (f(X(s)) - \min f)\, \mathbb{1}_{\Omega_\delta} ds\right] + \int_{\hat{t}_\delta}^t \sigma_\infty^2(s)ds \\ &+ \mathbb{E}\left[\mathbb{1}_{\Omega_\delta} \int_{\hat{t}_\delta}^t \left\langle \sigma^\top(s, X(s))(X(s) - P_{\mathcal{S}}(X(s))), dW(s)\right\rangle\right].\end{aligned}$$

On the other hand, since $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$, we have for all $T > 0$

$$\begin{aligned}\mathbb{E}\left(\int_0^T \|\sigma^\top(s, X(s))(X(s) - P_{\mathcal{S}}(X(s)))\|^2 ds\right) &\leq \mathbb{E}\left(\int_0^T \sigma_\infty^2(s)\|X(s) - P_{\mathcal{S}}(X(s))\|^2 ds\right) \\ &= \int_0^T \sigma_\infty^2(s)\mathbb{E}(\mathrm{dist}(X(t), \mathcal{S})^2)ds \\ &\leq C^* \int_0^{+\infty} \sigma_\infty^2(s)ds < +\infty.\end{aligned}$$

Letting $Y(s) = \sigma^\top(s, X(s))(X(s) - P_{\mathcal{S}}(X(s)))$, then

$$\mathbb{E}\left[\int_{\hat{t}_\delta}^t \langle Y(s), dW(s)\rangle\right] = 0.$$

This immediately implies

$$\mathbb{E}\left[\mathbb{1}_{\Omega_\delta}\int_{\hat{t}_\delta}^t \langle Y(s), dW(s)\rangle\right] = -\mathbb{E}\left[\mathbb{1}_{\Omega_{\text{conv}}\setminus\Omega_\delta}\int_{\hat{t}_\delta}^t \langle Y(s), dW(s)\rangle\right].$$

The right-hand side can be bounded using Cauchy-Schwarz inequality as follows

$$\begin{aligned}
&\left|\mathbb{E}\left[\mathbb{1}_{\Omega_{\text{conv}}\setminus\Omega_\delta}\int_{\hat{t}_\delta}^t \langle Y(s), dW(s)\rangle\right]\right| \\
&= \left|\mathbb{E}\left[\mathbb{1}_{\Omega_{\text{conv}}\setminus\Omega_\delta}\int_{\hat{t}_\delta}^t \left\langle \sigma^\top(s, X(s))(X(s) - P_{\mathcal{S}}(X(s))), dW(s)\right\rangle\right]\right| \\
&\leq \sqrt{\mathbb{E}(\mathbb{1}_{\Omega_{\text{conv}}\setminus\Omega_\delta})}\sqrt{\mathbb{E}\left[\left(\int_{\hat{t}_\delta}^t \langle \sigma^\top(s, X(s))(X(s) - P_{\mathcal{S}}(X(s))), dW(s)\rangle\right)^2\right]} \\
&\leq \sqrt{\delta}\sqrt{\mathbb{E}\left[\int_{\hat{t}_\delta}^t \|\sigma^\top(s, X(s))(X(s) - P_{\mathcal{S}}(X(s)))\|^2 ds\right]} \\
&\leq \sqrt{C^*\delta}\sqrt{\int_{\hat{t}_\delta}^t \sigma_\infty^2(s)ds} = \sqrt{C^*\delta}\int_{\hat{t}_\delta}^t \frac{\sigma_\infty^2(s)}{2\sqrt{\int_{\hat{t}_\delta}^s \sigma_\infty^2(u)du}}ds,
\end{aligned}$$

where we have used the fundamental theorem of calculus to arrive at the last equality. Set $C_* = \sqrt{C^*}$, and recall that $C_\infty = \sqrt{\int_0^{+\infty} \sigma_\infty^2(s)ds}$. Thus, for every $t > \hat{t}_\delta$

$$\widehat{G}(t) \leq \widehat{G}(\hat{t}_\delta) - \int_{\hat{t}_\delta}^t \mathbb{E}\left[(f(X(s)) - \min f)\mathbb{1}_{\Omega_\delta}\right] ds + \int_{\hat{t}_\delta}^t \sigma_\infty^2(s)ds + C_*\sqrt{\delta}\int_{\hat{t}_\delta}^t \frac{\sigma_\infty^2(s)}{2\sqrt{\int_{\hat{t}_\delta}^s \sigma_\infty^2(u)du}}ds. \tag{36}$$

Recall $h_\delta(t)$ from (26). Then, we can rewrite (36) as

$$\widehat{G}(t) \leq \widehat{G}(\hat{t}_\delta) - \int_{\hat{t}_\delta}^t \mathbb{E}\left[(f(X(s)) - \min f)\mathbb{1}_{\Omega_\delta}\right] ds + \int_{\hat{t}_\delta}^t h_\delta(s)ds, \quad \forall t > \hat{t}_\delta. \tag{37}$$

Using that $f \in \mathrm{EB}^2([f \leq r])$, we obtain

$$\widehat{G}(t) \leq \widehat{G}(\hat{t}_\delta) - 2\gamma\int_{\hat{t}_\delta}^t \widehat{G}(s)ds + \int_{\hat{t}_\delta}^t h_\delta(s)ds, \quad \forall t > \hat{t}_\delta.$$

Observe that $h_\delta \in \mathrm{L}^1([\hat{t}_\delta, \infty[)$ since

$$\int_{\hat{t}_\delta}^{+\infty} h_\delta(s)ds \leq C_\infty^2 + C_*C_\infty\sqrt{\delta}.$$

The goal now is to apply the comparison lemma to $\widehat{G}(t)$ (see Lemma A.2) which necessitates to solve the following ODE

$$\begin{cases} y'(t) = -2\gamma y(t) + h_\delta(t) & \forall t > \hat{t}_\delta, \\ y(\hat{t}_\delta) = \widehat{G}(\hat{t}_\delta). \end{cases}$$

Let $\lambda \in]0,1[$. Using the integrating factor method, we obtain

$$\begin{aligned}
y(t) &= e^{-2\gamma(t-\hat{t}_\delta)}y(\hat{t}_\delta) + e^{-2\gamma t}\int_{\hat{t}_\delta}^{\hat{t}_\delta+\lambda(t-\hat{t}_\delta)} h_\delta(s)e^{2\gamma s}ds + e^{-2\gamma t}\int_{\hat{t}_\delta+\lambda(t-\hat{t}_\delta)}^{t} h_\delta(s)e^{2\gamma s}ds \\
&\leq e^{-2\gamma(t-\hat{t}_\delta)}\mathbb{E}(\widehat{g}(\hat{t}_\delta)) + e^{-2\gamma(1-\lambda)(t-\hat{t}_\delta)}\int_{\hat{t}_\delta}^{\hat{t}_\delta+\lambda(t-\hat{t}_\delta)} h_\delta(s)ds \\
&+ h_\delta(\hat{t}_\delta+\lambda(t-\hat{t}_\delta))e^{-2\gamma t}\int_{\hat{t}_\delta+\lambda(t-\hat{t}_\delta)}^{t} e^{2\gamma s}ds \\
&\leq e^{-2\gamma(t-\hat{t}_\delta)}\mathbb{E}(\widehat{g}(\hat{t}_\delta)) + e^{-2\gamma(1-\lambda)(t-\hat{t}_\delta)}(C_\infty^2 + C_* C_\infty\sqrt{\delta}) + \frac{h_\delta(\hat{t}_\delta+\lambda(t-\hat{t}_\delta))}{2\gamma}.
\end{aligned}$$

where in the first inequality, we used that σ^2 is non-increasing and so is h_δ. Lemma A.2 then gives

$$\mathbb{E}\left(\frac{\operatorname{dist}(X(t),\mathcal{S})^2}{2}\mathbb{1}_{\Omega_\delta}\right) \leq e^{-2\gamma(t-\hat{t}_\delta)}\mathbb{E}\left(\frac{\operatorname{dist}(X(\hat{t}_\delta),\mathcal{S})^2}{2}\right) + e^{-2\gamma(1-\lambda)(t-\hat{t}_\delta)}(C_\infty^2 + C_* C_\infty\sqrt{\delta}) + \frac{h_\delta(\hat{t}_\delta+\lambda(t-\hat{t}_\delta))}{2\gamma}.$$

According to Corollary A.6 we obtain that for all $t > \hat{t}_\delta$

$$\mathbb{E}\left(\frac{\operatorname{dist}(X(t),\mathcal{S})^2}{2}\right) \leq e^{-2\gamma(t-\hat{t}_\delta)}\mathbb{E}\left(\frac{\operatorname{dist}(X(\hat{t}_\delta),\mathcal{S})^2}{2}\right) + e^{-2\gamma(1-\lambda)(t-\hat{t}_\delta)}(C_\infty^2 + C_* C_\infty\sqrt{\delta}) + \frac{h_\delta(\hat{t}_\delta+\lambda(t-\hat{t}_\delta))}{2\gamma} + C_d\sqrt{\delta}.$$

(b) Denote $\widetilde{g}(t) = \widetilde{\phi}(X(t)) = f(X(t)) - \min f$ and $\widetilde{G}(t) = \mathbb{E}(\mathbb{1}_{\Omega_\delta}\widetilde{g}(t))$. By Proposition 2.4

$$\widetilde{g}(t) \leq \widetilde{g}(\hat{t}_\delta) - \int_{\hat{t}_\delta}^{t}\left\langle \nabla f(X(s)), \nabla\widetilde{\phi}(X(s))\right\rangle ds + \frac{L}{2}\int_{\hat{t}_\delta}^{t}\operatorname{tr}[\Sigma(s,X(s))]ds + \int_{\hat{t}_\delta}^{t}\left\langle \sigma^\top(s,X(s))\nabla f(X(s)), dW(s)\right\rangle. \quad (38)$$

Multiplying both sides by $\mathbb{1}_{\Omega_\delta}$ and taking expectation we obtain

$$\widetilde{G}(t) - \widetilde{G}(\hat{t}_\delta) \leq -\mathbb{E}\left[\int_{\hat{t}_\delta}^{t}\|\nabla f(X(s))\|^2\mathbb{1}_{\Omega_\delta}ds\right] + \frac{L}{2}\mathbb{E}\left[\int_{\hat{t}_\delta}^{t}\operatorname{tr}[\Sigma(s,X(s))]ds\right] + \mathbb{E}\left[\mathbb{1}_{\Omega_\delta}\int_{\hat{t}_\delta}^{t}\left\langle \sigma^\top(s,X(s))\nabla f(X(s)), dW(s)\right\rangle\right]. \quad (39)$$

On the other hand, we have

$$\begin{aligned}
\mathbb{E}\left(\int_0^T\|\sigma^\top(s,X(s))\nabla f(X(s))\|^2 ds\right) &\leq L^2\mathbb{E}\left(\int_0^T\sigma_\infty^2(s)\|X(s)) - x^\star\|^2 ds\right) \\
&\leq L^2 C^* \int_0^{+\infty}\sigma_\infty^2(s)ds < +\infty, \quad \forall T > 0.
\end{aligned}$$

Since $\mathbb{E}\left[\int_{\hat{t}_\delta}^t \langle \sigma^\top(s, X(s))\nabla f(X(s)), dW(s)\rangle\right] = 0$, we have

$$\mathbb{E}\left[\mathbb{1}_{\Omega_\delta} \int_{\hat{t}_\delta}^t \left\langle \sigma^\top(s, X(s))(\nabla f(X(s))), dW(s)\right\rangle\right] = -\mathbb{E}\left[\mathbb{1}_{\Omega_{\mathrm{conv}}\setminus\Omega_\delta} \int_{\hat{t}_\delta}^t \left\langle \sigma^\top(s, X(s))(\nabla f(X(s)), dW(s)\right\rangle\right].$$

The last term can be bounded as

$$\begin{aligned}
&\left|\mathbb{E}\left[\mathbb{1}_{\Omega_{\mathrm{conv}}\setminus\Omega_\delta} \int_{\hat{t}_\delta}^t \left\langle \sigma^\top(s, X(s))(\nabla f(X(s))), dW(s)\right\rangle\right]\right| \\
&\leq \sqrt{\mathbb{E}(\mathbb{1}_{\Omega_{\mathrm{conv}}\setminus\Omega_\delta})}\sqrt{\mathbb{E}\left[\left(\int_{\hat{t}_\delta}^t \langle \sigma^\top(s, X(s))(\nabla f(X(s))), dW(s)\rangle\right)^2\right]} \\
&\leq L\sqrt{\delta}\sqrt{\mathbb{E}\left[\int_{\hat{t}_\delta}^t \sigma_\infty^2(s)\|X(s) - x^\star\|^2 ds\right]} \\
&\leq L\sqrt{C^*}\sqrt{\delta}\sqrt{\int_{\hat{t}_\delta}^t \sigma_\infty^2(s)ds} = L\sqrt{C^*}\sqrt{\delta}\int_{\hat{t}_\delta}^t \frac{\sigma_\infty^2(s)}{2\sqrt{\int_{\hat{t}_\delta}^s \sigma_\infty^2(u)du}}ds,
\end{aligned}$$

where we have used the fundamental theorem of calculus to arrive at the last equality. Let us notice that if (23) holds, then Proposition 3.4 tells us that $\mathbb{E}(f(X(t)) - \min f) \leq K't^{\beta-1}$ with $\beta \in [0, 1[$, and for some $K' > 0$. In this case, Cauchy-Schwarz inequality and Corollary 2.2 yield

$$\left|\mathbb{E}\left[\mathbb{1}_{\Omega_{\mathrm{conv}}\setminus\Omega_\delta} \int_{\hat{t}_\delta}^t \left\langle \sigma^\top(s, X(s))(\nabla f(X(s))), dW(s)\right\rangle\right]\right| \leq \sqrt{2LK'}\sqrt{\delta}\int_{\hat{t}_\delta}^t \frac{\sigma_\infty^2(s)s^{\beta-1}}{2\sqrt{\int_{\hat{t}_\delta}^s \sigma_\infty^2(u)u^{\beta-1}du}}ds.$$

Injecting this into (39), we have for all $t > \hat{t}_\delta$

$$\begin{aligned}
\widetilde{G}(t) \leq \widetilde{G}(\hat{t}_\delta) &- \mathbb{E}\left[\int_{\hat{t}_\delta}^t \|\nabla f(X(s))\|^2 \mathbb{1}_{\Omega_\delta} ds\right] + \frac{L}{2}\int_{\hat{t}_\delta}^t \sigma_\infty^2(s)ds \\
&+ \begin{cases} C_K\sqrt{\delta}\int_{\hat{t}_\delta}^t \frac{\sigma_\infty^2(s)s^{\beta-1}}{2\sqrt{\int_{\hat{t}_\delta}^s \sigma_\infty^2(u)u^{\beta-1}du}}ds, \quad \forall t > \hat{t}_\delta & \text{if (23) holds,} \\ C_{**}\sqrt{\delta}\int_{\hat{t}_\delta}^t \frac{\sigma_\infty^2(s)}{2\sqrt{\int_{\hat{t}_\delta}^s \sigma_\infty^2(u)du}}ds & \text{otherwise,} \end{cases}
\end{aligned} \tag{40}$$

where $C_{**} = L\sqrt{C^*}$, $C_K = \sqrt{2LK'}$ and recall that $C_\infty = \sqrt{\int_0^{+\infty} \sigma_\infty^2(s)ds}$. Recalling $l_\delta(t)$ and $k_\delta(t)$ from (27)-(28), and by Fubini's theorem, (40) becomes

$$\widetilde{G}(t) \leq \widetilde{G}(\hat{t}_\delta) - \int_{\hat{t}_\delta}^t \mathbb{E}\left[\|\nabla f(X(s))\|^2 \mathbb{1}_{\Omega_\delta}\right] ds + \begin{cases} \int_{\hat{t}_\delta}^t k_\delta(s)ds & \text{if (23) holds,} \\ \int_{\hat{t}_\delta}^t l_\delta(s)ds & \text{otherwise.} \end{cases} \tag{41}$$

Since $f \in Ł^{1/2}(\mathcal{S})$, there exists $\mu > 0$ such that

$$\widetilde{G}(t) \leq \widetilde{G}(\hat{t}_\delta) - \mu^2\int_{\hat{t}_\delta}^t \widetilde{G}(s)ds + \begin{cases} \int_{\hat{t}_\delta}^t k_\delta(s)ds & \text{if (23) holds,} \\ \int_{\hat{t}_\delta}^t l_\delta(s)ds & \text{otherwise.} \end{cases} \tag{42}$$

To get an explicit bound in (42), we use Lemma A.2, which involves solving

$$\text{(E.2)}\ \begin{cases} y'(t) &= -\mu^2 y(t) + l_\delta(t), \quad \forall t > \hat{t}_\delta \\ y(\hat{t}_\delta) &= \widetilde{G}(\hat{t}_\delta) \end{cases}$$

$$\text{(E.3)}\ \begin{cases} y'(t) &= -\mu^2 y(t) + k_\delta(t), \quad \forall t > \hat{t}_\delta \\ y(\hat{t}_\delta) &= \widetilde{G}(\hat{t}_\delta) \end{cases}$$

Let $\lambda \in]0,1[$. Using the integrating factor method as in (i), we get for (E.2)

$$y(t) \leq e^{-\mu^2(t-\hat{t}_\delta)}\mathbb{E}(\widetilde{g}(\hat{t}_\delta)) + \begin{cases} e^{-\mu^2(1-\lambda)(t-\hat{t}_\delta)}\left(\frac{LC_\infty^2}{2} + C_{**}C_\infty\sqrt{\delta}\right) + \frac{l_\delta(\hat{t}_\delta+\lambda(t-\hat{t}_\delta))}{\mu^2} & \text{for (E.2)} \\ e^{-\mu^2(1-\lambda)(t-\hat{t}_\delta)}\left(\frac{LC_\infty^2}{2} + C_K C_\infty\sqrt{\hat{t}_\delta^{\beta-1}}\sqrt{\delta}\right) + \frac{k_\delta(\hat{t}_\delta+\lambda(t-\hat{t}_\delta))}{\mu^2} & \text{for (E.3).} \end{cases}$$

Using Lemma A.2 and then Corollary A.6, we obtain

$$\begin{aligned} \mathbb{E}\left[f(X(t)) - \min f\right] &\leq y(t) + C_f\sqrt{\delta} \\ &\leq e^{-\mu^2(t-\hat{t}_\delta)}\mathbb{E}\left[f(X(\hat{t}_\delta)) - \min f\right] + C_f\sqrt{\delta} \\ &+ \begin{cases} e^{-\mu^2(1-\lambda)(t-\hat{t}_\delta)}\left(\frac{LC_\infty^2}{2} + C_{**}C_\infty\sqrt{\delta}\right) + \frac{l_\delta(\hat{t}_\delta+\lambda(t-\hat{t}_\delta))}{\mu^2} & \text{for (E.2)} \\ e^{-\mu^2(1-\lambda)(t-\hat{t}_\delta)}\left(\frac{LC_\infty^2}{2} + C_K C_\infty\sqrt{\hat{t}_\delta^{\beta-1}}\sqrt{\delta}\right) + \frac{k_\delta(\hat{t}_\delta+\lambda(t-\hat{t}_\delta))}{\mu^2} & \text{for (E.3).} \end{cases} \end{aligned}$$

(ii) $p > 2$:

(a) We embark from inequality (37) and we now use that $f \in \mathrm{EB}^p([f \leq r])$ with $p > 2$, to get

$$\begin{aligned} \widehat{G}(t) &\leq \widehat{G}(\hat{t}_\delta) - \int_{\hat{t}_\delta}^t \mathbb{E}\left[(f(X(s)) - \min f)\mathbb{1}_{\Omega_\delta}\right] ds + \int_{\hat{t}_\delta}^t h_\delta(s) ds \\ &\leq \widehat{G}(\hat{t}_\delta) - 2^{p/2}\gamma \int_{\hat{t}_\delta}^t \widehat{G}(s)^{p/2} + \int_{\hat{t}_\delta}^t h_\delta(s) ds. \end{aligned} \tag{43}$$

In the last inequality, we used that $p > 2$ and Jensen's inequality.

The idea is again to use the comparison lemma (Lemma A.2), which will now involve solving the Cauchy problem (C.1), and finally invoke Corollary A.6.

(b) The reasoning is similar to the previous point using now that $f \in Ł^q(\mathcal{S})$ and the computations of (i)(b). We omit the details for the sake of brevity.

□

5 SDE for nonsmooth structured convex optimization

In this section, we turn to the composite convex minimization problem with additive structure

$$\min_{x \in \mathbb{R}^d} f(x) + g(x), \tag{44}$$

where

$$\begin{cases} f \in C_L^{1,1}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d) \text{ and } g \in \Gamma_0(\mathbb{R}^d); \\ \mathcal{S} = \operatorname{argmin}(f+g) \neq \emptyset. \end{cases} \tag{H$_0'$}$$

The importance of this class of problems comes from its wide spectrum of applications ranging from data processing, to machine learning and statistics to name a few.

We consider two different approaches leading to different SDE's. The first is based on a fixed point argument and the use of the notion of cocoercive monotone operator. The second approach is based on a regularization/smoothing argument, for instance the Moreau envelope.

5.1 Fixed point approach via cocoercive monotone operators

Let us start with some classical definitions concerning monotone operators.

Definition 5.1. An operator $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is monotone if

$$\langle u - v, x - y\rangle \geq 0, \qquad \forall (x, u) \in \operatorname{graph}(A), (y, v) \in \operatorname{graph}(A).$$

It is maximally monotone if there exists no monotone operator whose graph properly contains graph(A). Moreover, A is γ-strongly monotone with modulus $\gamma > 0$ if

$$\langle u - v, x - y\rangle \geq \gamma\|x - y\|^2, \quad \forall (x, u) \in \operatorname{graph}(A), (y, v) \in \operatorname{graph}(A).$$

Remark 5.2. If A is maximally monotone and strongly monotone, then $A^{-1}(0) \overset{\text{def}}{=} \{x \in \mathbb{R}^d : A(x) = 0\}$ is non-empty and reduced to a singleton.

Remark 5.3. The subdifferential operator ∂g of $g \in \Gamma_0(\mathbb{R}^d)$ is maximally monotone.

Definition 5.4. A single-valued operator $M : \mathbb{R}^d \to \mathbb{R}^d$ is cocoercive with constant $\rho > 0$ if

$$\langle M(x) - M(y), x - y\rangle \geq \rho\|M(x) - M(y)\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

Remark 5.5. It is clear that a ρ-cocoercive operator is ρ^{-1}-Lipschitz continuous. In turn, a cocoercive operator is maximally monotone.

Remark 5.6. If $f \in C_L^{1,1}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d)$, then the operator ∇f is L^{-1}-cocoercive.

Our interest now is to solve the structured monotone inclusion problem

$$0 \in A(x) + B(x),$$

where A is maximally monotone, and B is cocoercive with $(A + B)^{-1}(0) \neq \emptyset$. This is of course a generalization of (44) by taking $A = \partial g$ and $B = \nabla f$.

A favorable situation occurs when one can compute the resolvent operator of A

$$J_{\mu A} = (I + \mu A)^{-1}, \quad \mu > 0.$$

In this case, we can develop a strategy parallel to the one which consists in replacing a maximally monotone operator by its Yosida approximation. Indeed, given $\mu > 0$, we have

$$(A + B)(x) \ni 0 \iff x - J_{\mu A}(x - \mu B(x)) = 0 \iff M_{A,B,\mu}(x) = 0, \tag{45}$$

where $M_{A,B,\mu} : \mathbb{R}^d \to \mathbb{R}^d$ is the single-valued operator defined by

$$M_{A,B,\mu}(x) = \frac{1}{\mu}\left(x - J_{\mu A}(x - \mu B(x))\right). \tag{46}$$

$M_{A,B,\mu}$ is closely tied to the well-known forward-backward fixed point operator. Moreover, when $B = 0$, $M_{A,B,\mu} = \frac{1}{\mu}(I - J_{\mu A})$ which is nothing but the Yosida regularization of A with index μ. As a remarkable property, for the μ parameter properly set, the operator $M_{A,B,\mu}$ is cocoercive. This is made precise in the following result.

Proposition 5.7. *[4, Lemma B.1] Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a general maximally monotone operator, and let $B : \mathbb{R}^d \to \mathbb{R}^d$ be a monotone operator which is λ-cocoercive. Assume that $\mu \in]0, 2\lambda[$. Then, $M_{A,B,\mu}$ is ρ-cocoercive with*

$$\rho = \mu \left(1 - \frac{\mu}{4\lambda}\right).$$

We first focus on finding the zeros of M, where

$$M : \mathbb{R}^d \to \mathbb{R}^d \text{ is cocoercive and } M^{-1}(0) \neq \emptyset. \tag{H_0^M}$$

We will then specialize our results to the case of a structured operator of the form $M_{A,B,\mu}$.

Our goal is to handle the situation where M can be evaluated up to a stochastic error. We therefore consider the following SDE with an $\mathcal{F}_0$-measurable initial data X_0:

$$\begin{cases} dX(t) = -M(X(t))dt + \sigma(t, X(t))dW(t), \; t \geq 0 \\ X(0) = X_0. \end{cases} \tag{SDE^M}$$

As in Section 1.1, we will assume that the volatility matrix $\sigma : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^{d \times m}$ satisfies (H), W is a $\mathcal{F}_t$-adapted m-dimensional Brownian motion.

Remark 5.8. The motivation of (SDE^M) comes again from the Robbins-Monro stochastic approximation algorithm where the martingale difference noise/error is induced by randomly approximating the action of the whole fixed point operator $M_{A,B,\mu}$. This allows for instance for inexact computation of the resolvent of A with random noise. However, the situation is more intricate when the noise is solely on B (*i.e.* inside the resolvent), as it is standard in many applications (think of $B = \nabla f$ and the latter is accessible only some unbiased stochastic estimator). In this case, to justify moving the noise outside of the resolvent, one has to modify (SDE^M) to a limiting continuous-time process of a forward-backward scheme, which would take us to the land of *stochastic differential inclusions* (SDI). SDI's were only introduced in the early 80's by [32, 33], where the notion of solutions with pathwise uniqueness of a solution for a certain class of maximal monotone operators A was introduced. Theory of SDI's has subsequently received much attention with general applications including beyond the maximal monotone case; see *e.g.* [31]. We point out in particular the results of [50] who was the first to show existence and uniqueness of a solution to SDI's with maximal monotone A and Lipschitz continuous B using the Yosida approximation of A, hence extending known results of Brézis [13] in the deterministic case. This is yet another justification behind our second approach using Moreau-Yosida regularization, though restricted to functions. However, handling SDI's properly would necessitate much more care and many new techniques and notions that will deserve a whole dedicated paper which is the subject of ongoing work.

Let us now state the natural extensions of our main results to this situation.

Theorem 5.9. *Let $M : \mathbb{R}^d \to \mathbb{R}^d$ be a cocoercive operator. Consider the stochastic differential equation* (SDE^M)*, with M and σ under the hypotheses* (H$_0^M$) *and* (H)*, respectively. Additionally, let $\nu \geq 2$, $X_0 \in \mathrm{L}^\nu(\Omega; \mathbb{R}^d)$ and is $\mathcal{F}_0$-measurable. Then, there exists a unique solution $X \in S_d^\nu$. Moreover, if $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$, then:*

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

(ii) $\forall x^\star \in M^{-1}(0)$, $\lim_{t\to+\infty} \|X(t) - x^\star\|$ *exists a.s. and* $\sup_{t\geq 0} \|X(t)\| < +\infty$ *a.s..*

(iii) $\lim_{t\to\infty} \|M(X(t))\| = 0$ *a.s..*

(iv) There exists an $M^{-1}(0)$-valued random variable $X^\star$ such that $\lim_{t\to+\infty} X(t) = X^\star$ a.s..

Proof. Proof. Existence and uniqueness follow from Theorem A.7 since M is Lipschitz continuous and σ verifies (H). The proof of the first three items remains the same as for Theorem 3.1, where we use the cocoercivity of M instead of the convexity of f in the third item to prove that $\lim_{t\to\infty} \|M(X(t))\| = 0$ a.s.. For the last item, it suffices to use that the operator M is continuous (since it is Lipschitz continuous) to conclude with Opial's Lemma. □

Theorem 5.10. *Consider the dynamic* (SDE^M) *where M and σ satisfy the assumptions* (H$_0^M$) *and* (H)*. Moreover, let M be a ρ-cocoercive operator. Additionally, $X_0 \in \mathrm{L}^2(\Omega; \mathbb{R}^d)$ and is $\mathcal{F}_0$-measurable. Let $X \in S_d^2$ be the unique solution of* (SDE^M)*, then the following properties are satisfied:*

(i) Let $\overline{M \circ X}(t) \overset{\text{def}}{=} t^{-1} \int_0^t M(X(s))ds$ *and* $\overline{\|M(X(t))\|^2} \overset{\text{def}}{=} t^{-1} \int_0^t \|M(X(s))\|^2 ds$*. We have*

$$\mathbb{E}\left[\|\overline{M \circ X}(t)\|^2\right] \leq \mathbb{E}\left[\overline{\|M(X(t))\|^2}\right] \leq \frac{\mathbb{E}\left(\mathrm{dist}(X_0, M^{-1}(0))^2\right)}{2\rho t} + \frac{\sigma_*^2}{2\rho}, \quad \forall t > 0. \tag{47}$$

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

$$\mathbb{E}\left[\|\overline{M \circ X}(t)\|^2\right] \leq \mathbb{E}\left[\overline{\|M(X(t))\|^2}\right] = \mathcal{O}\left(\frac{1}{t}\right), \quad \forall t > 0. \tag{48}$$

(ii) If M is γ-strongly monotone, then $M^{-1}(0) = \{x^\star\}$ and

$$\mathbb{E}\left(\|X(t) - x^\star\|^2\right) \leq \mathbb{E}\left(\|X_0 - x^\star\|^2\right) e^{-\gamma t} + \frac{\sigma_*^2}{\gamma}, \quad \forall t \geq 0. \tag{49}$$

If, moreover, σ_∞ is non-increasing and vanishes at infinity, then for every $\lambda \in]0,1[$

$$\mathbb{E}\left(\|X(t) - x^\star\|^2\right) \leq \mathbb{E}\left(\|X_0 - x^\star\|^2\right) e^{-\gamma t} + \frac{\sigma_*^2}{\gamma} e^{-\gamma t(1-\lambda)} + \sigma_\infty^2(\lambda t), \quad \forall t > 0. \tag{50}$$

Proof. Proof. Analogous to Theorem 3.3. □

We now turn to the local convergence properties. To this end, we need an extension of the Hölderian error bound inequality (or Łojasiewicz inequality) to the operator setting. For convex functions, it is known that error bound inequalities are closely related to metric subregularity of the subdifferential [2, 34, 35]. This leads to the following definition.

Definition 5.11. Let $M : \mathbb{R}^d \to \mathbb{R}^d$ be a single-valued operator. We say that M satisfies the Hölder metric subregularity property with exponent $p \geq 2$ at $x^\star \in M^{-1}(0)$ if there exists $\gamma > 0$ and a neighbourhood $\mathcal{V}_{x^\star}$ such that

$$\|M(x)\|^2 \geq \gamma \mathrm{dist}(x, M^{-1}(0))^p, \quad \forall x \in \mathcal{V}_{x^\star}. \tag{51}$$

If this inequality holds for any $x^\star \in M^{-1}(0)$ with the same γ, we will write $M \in \mathrm{HMS}^p(\mathbb{R}^d)$.

Theorem 5.12. *Consider the dynamic* (SDE^M) *where M and σ satisfy the assumptions* (H_0^M) *and* (H)*. Moreover, let $M \in \mathrm{HMS}^2(\mathbb{R}^d)$ be a ρ-cocoercive operator. Additionally, $X_0 \in \mathrm{L}^4(\Omega; \mathbb{R}^d)$ and is $\mathcal{F}_0$-measurable. Suppose that $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$ ($C_\infty \overset{\mathrm{def}}{=} \|\sigma_\infty\|_{\mathrm{L}^2(\mathbb{R}_+)}$) and σ_∞ is non-increasing. Let $X \in S_d^2$ be the unique solution of* (SDE^M)*. Consider also the positive constants C, C_d, γ. Then, for all $\delta > 0$, there exists $\hat{t}_\delta > 0$ such that for every $\lambda \in (0,1)$:*

$$\begin{aligned}\mathbb{E}\left(\frac{\mathrm{dist}(X(t), M^{-1}(0))^2}{2}\right) &\leq e^{-2\gamma\rho(t-\hat{t}_\delta)}\mathbb{E}\left(\frac{\mathrm{dist}(X(\hat{t}_\delta), M^{-1}(0))^2}{2}\right) \\ &\quad + e^{-2\gamma\rho(1-\lambda)(t-\hat{t}_\delta)}(C_\infty^2 + C_\infty C\sqrt{\delta}) \\ &\quad + \frac{h_\delta(\hat{t}_\delta + \lambda(t-\hat{t}_\delta))}{2\gamma\rho} + C_d\sqrt{\delta}, \quad \forall t > \hat{t}_\delta,\end{aligned} \tag{52}$$

where $h_\delta(t) = \sigma_\infty^2(t) + C\sqrt{\delta}\frac{\sigma_\infty^2(t)}{2\sqrt{\int_{\hat{t}_\delta}^t \sigma_\infty^2(u)du}}$.

Proof. Proof. The proof is essentially the same as that of Theorem 4.5(i)(a), where instead of convexity in (35), we use cocoercivity of M, and in (37) we invoke Theorem 5.9 and Hölder metric subregularity. □

Remark 5.13. We can naturally extend the previous result for $p > 2$ as in Theorem 4.5(ii). Nevertheless, since that bound is not explicit, we will skip this extension.

As an immediate consequence of the above result, by considering the cocoercive operator $M_{A,B,\mu}$ defined in (46), we obtain the following result.

Corollary 5.14. *Let $A : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be a maximally monotone operator and $B : \mathbb{R}^d \to \mathbb{R}^d$ be a λ-cocoercive operator, $\lambda > 0$. Let $M_{A,B,\mu}$ be the operator defined in* (46)*. Assume that $\mu \in]0, 2\lambda[$ and $(A+B)^{-1}(0) \neq \emptyset$. Then, the operator $M_{A,B,\mu}$ is ρ-cocoercive with $\rho = \mu\left(1 - \frac{\mu}{4\lambda}\right)$, letting $\nu \geq 2$ and considering the SDE with initial data $X_0 \in \mathrm{L}^\nu(\Omega; \mathbb{R}^d)$ which is $\mathcal{F}_0$-measurable:*

$$\begin{cases} dX(t) = -M_{A,B,\mu}(X(t))dt + \sigma(t, X(t))dW(t), \ t \geq 0 \\ X(0) = X_0, \end{cases} \tag{$\mathrm{SDE}^{M_{A,B,\mu}}$}$$

we can conclude the same results as in Theorem 5.9 and Theorem 5.10. In particular, if $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$, there exists an $(A+B)^{-1}(0)$-valued random variable $X^\star$ such that $\lim_{t\to+\infty} X(t) = X^\star$ a.s..

This result naturally applies to problem (44) when $\mathcal{S} = \operatorname{argmin}(f+g) \neq \emptyset$ by taking $A = \partial g$ and $B = \nabla f$. In this case, one has that $X(t)$ converges a.s. to an $\mathcal{S}$-valued random variable. Moreover, using standard inequalities, see *e.g.* [6], one can show that

$$\mathbb{E}\left[(f+g)\left(t^{-1}\int_0^t \left(\operatorname{prox}_{\mu g}(X(s) - \mu\nabla f(X(s)))\right) ds\right) - \min(f+g)\right] = \mathcal{O}\left(\sqrt{\mathbb{E}\left[\overline{\|M_{\partial g,\nabla f,\mu}(X(t))\|^2}\right]}\right),$$

where $\operatorname{prox}_{\mu g} = (I + \mu\partial g)^{-1}$ is the proximal mapping of g. From this, one can deduce an $\mathcal{O}(t^{-1/2})$ rate thanks to (47) and (48).

5.2 Approach via Moreau-Yosida regularization

The previous approach, though it is able to deal with more general setting (that of monotone inclusions), took us out of the framework of convex optimization by considering instead a dynamic governed by a cocoercive operator. In particular, the perturbation/noise is considered on the whole operator evaluation and not on a part of it (*i.e.* B) as it is standard in many applications. Moreover this approach led to a pessimistic convergence rate estimate when specialized to convex function minimization. By contrast, the following approach will operate directly on problem (44) and is based on a standard smoothing approach, replacing the non-smooth part g by its Moreau envelope [7].

5.2.1 Moreau envelope

Let us start by recalling some basic facts concerning the Moreau envelope.

Definition 5.15. Let $g \in \Gamma_0(\mathbb{R}^d)$. Given $\theta > 0$, the Moreau envelope of g of parameter θ is the function

$$g_\theta(x) \overset{\text{def}}{=} \inf_{y\in\mathbb{R}^d}\left(g(y) + \frac{1}{2\theta}\|x-y\|^2\right) = \left(g \,\square\, \frac{1}{\theta}q\right)(x)$$

where $\square$ is the infimal convolution operator and $q(x) = \frac{1}{2}\|x\|^2$.

The Moreau envelope has remarkable approximation and regularization properties, as summarized in the following statement.

Proposition 5.16. *Let $g \in \Gamma_0(\mathbb{R}^d)$.*

(i) $g_\theta(x) \downarrow \inf g(\mathbb{R}^d)$ *as* $\theta \uparrow +\infty$.

(ii) $g_\theta(x) \uparrow g(x)$ *as* $\theta \downarrow 0$.

(iii) $g_\theta(x) \leq g(x)$ *for any* $\theta > 0$ *and* $x \in \mathbb{R}^d$,

(iv) $\operatorname{argmin}(g_\theta) = \operatorname{argmin}(g)$ *for any* $\theta > 0$,

(v) $g(x) - g_\theta(x) \leq \frac{\theta}{2}\|\partial^0 g(x)\|^2$ *for any* $\theta > 0$ *and* $x \in \operatorname{dom}(\partial g)$,

(vi) $g_\theta \in C^{1,1}_{\frac{1}{\theta}}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d)$ *for any* $\theta > 0$.

We use the following notation in the rest of the section: $F \overset{\text{def}}{=} f + g, \mathcal{S} \overset{\text{def}}{=} \operatorname{argmin} F$, $F_\theta \overset{\text{def}}{=} f + g_\theta$ and $\mathcal{S}_\theta \overset{\text{def}}{=} \operatorname{argmin} F_\theta$.

Note that $F_\theta \in C^{1,1}_{L+\frac{1}{\theta}}(\mathbb{R}^d) \cap \Gamma_0(\mathbb{R}^d)$. Thus we will use F_θ as the potential driving in (SDE), that is

$$\begin{cases} dX(t) = -\nabla F_\theta(X(t))dt + \sigma(t, X(t))dW(t), \ t \geq 0 \\ X(0) = X_0. \end{cases} \tag{SDE$_\theta$}$$

Throughout this section, we assume that $X_0 \in \mathrm{L}^2(\Omega; \mathbb{R}^d)$ and is $\mathcal{F}_0$-measurable. Under (H_0') and (H), we will show almost sure convergence of the trajectory and corresponding convergence rates.

Remark 5.17. Though we focus here on the Moreau envelope, our convergence results, in particular, Proposition 5.20, still hold with infimal-convolution based smoothing using more general smooth kernels beyond the norm squared; see [7, Section 4.4].

5.2.2 Convergence of the trajectory

Applying Theorem 3.1 to F_θ, we have the following result.

Proposition 5.18. *For any $\theta > 0$, let $X_0 \in \mathrm{L}^2(\Omega; \mathbb{R}^d)$ and $X_\theta \in S_d^2$ be the unique solution of the dynamic (SDE$_\theta$) governed by the potential F_θ, and make assumptions (H_0'), $\mathcal{S}_\theta \neq \emptyset$, (H) and $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$. Then there exists an $\mathcal{S}_\theta$-valued random variable $X_\theta^\star$ such that*

$$\lim_{t \to +\infty} X_\theta(t) = X_\theta^\star, \quad a.s..$$

If $f = 0$, then $\mathcal{S}_\theta = \mathcal{S}$ (see Proposition 5.16(iv)), and Proposition 5.18 provides almost sure convergence to a solution of (44). On the other hand for $f \neq 0$, $\mathcal{S} \neq \mathcal{S}_\theta$ in general and we only obtain an "approximate" solution of (44); see Proposition 5.19(ii) for a quantitative estimate of this approximation when f is strongly convex. To obtain a true solution of the initial problem, a common device consists in using a diagonalization process which combines the dynamic with the approximation. Specifically, one considers

$$\begin{cases} dX(t) = -\nabla F_{\theta(t)}(X(t))dt + \sigma(t, X(t))dW(t), \ t \geq 0 \\ X(0) = X_0, \end{cases} \tag{SDE$_{\theta(t)}$}$$

where $\theta(t) \downarrow 0$ as $t \to +\infty$. In the deterministic case, an abundant literature has been devoted to the convergence of this type of systems. Note that unlike the cocoercive approach, we are now faced with a non-autonomous stochastic differential equation, making this a difficult problem, a subject for further research (see also Remark 5.8).

5.2.3 Convergence rates

We start with the following uniform bound on $\mathcal{S}_\theta$ which holds under slightly reinforced, but reasonable assumptions on f and g.

Proposition 5.19. *Consider f, g where f and g and are proper lsc and convex, and g is also L_0-Lipschitz continuous.*

(i) Assume that $F = f + g$ is coercive. Then, there exists $C > 0$, such that for any $\theta \in [0,1]$

$$\sup_{z \in \mathcal{S}_\theta} \|z\| \leq C. \tag{53}$$

(ii) Assume that $f \in \Gamma_\mu(\mathbb{R}^d)$ for $\mu > 0$. Then (53) holds for every $\theta \in [0,1]$. Moreover, $\mathcal{S} = \{x^\star\}$, $\mathcal{S}_\theta = \{x^\star_\theta\}$ and

$$\|x^\star_\theta - x^\star\|^2 \leq \frac{L_0^2}{\mu}\theta. \tag{54}$$

Proof. Proof.

(i) Since F is coercive, so is F_θ. Thus both $\mathcal{S}$ and $\mathcal{S}_\theta$ are non-empty compact sets. Let $x^\star_\theta \in \mathcal{S}_\theta$ and $x^\star \in \mathcal{S}$. By Proposition 5.16(v) and Lipschitz continuity of g, we obtain

$$F(x^\star_\theta) - F_\theta(x^\star_\theta) = g(x^\star_\theta) - g_\theta(x^\star_\theta) \leq \frac{L_0^2}{2}\theta.$$

Moreover,

$$F_\theta(x^\star_\theta) + \frac{L_0^2}{2}\theta \leq F_\theta(x^\star) + \frac{L_0^2}{2}\theta \leq F(x^\star) + \frac{L_0^2}{2}\theta \leq \min(F) + \frac{L_0^2}{2} \stackrel{\text{def}}{=} \widetilde{C},$$

where the second inequality is given by Proposition 5.16(iv). On the other hand, the coercivity of F implies that there exists $a > 0, b \in \mathbb{R}$ such that for any $x \in \mathbb{R}^d$

$$a\|x\| + b \leq F(x).$$

Therefore, collecting the above inequalities yields

$$a\|x^\star_\theta\| + b \leq F(x^\star_\theta) \leq \widetilde{C}.$$

Using that $x^\star_\theta$ is arbitrary in $\mathcal{S}_\theta$, and defining $C \stackrel{\text{def}}{=} \frac{\widetilde{C}-b}{a} \geq 0$, we obtain (53), or equivalently that the set of approximate minimizers is bounded independently of θ.

(ii) Since f is μ-strongly convex, so are F and F_θ. In turn, F is coercive and thus (53) holds by claim (i). Strong convexity implies uniqueness of minimizers of F and F_θ. Moreover,

$$\frac{\mu}{2}\|x^\star_\theta - x^\star\|^2 \leq F_\theta(x^\star) - F_\theta(x^\star_\theta). \tag{55}$$

From Proposition 5.16(iii)-(v) and Lipschitz continuity of g, we infer that

$$F_\theta(x^\star) - F_\theta(x^\star_\theta) \leq F(x^\star) - F_\theta(x^\star_\theta) \leq F(x^\star_\theta) - F_\theta(x^\star_\theta) \leq \frac{L_0^2}{2}\theta. \tag{56}$$

Combining (55) and (55), we get the claimed bound.

□

We are now ready to establish complexity results.

Proposition 5.20. *Suppose that in addition to* (H_0') *and* (H), $F = f + g$ *is coercive and* g *is* L_0-*Lipschitz continuous. Let* $X_0 \in \mathrm{L}^2(\Omega; \mathbb{R}^d)$ *and* $X_\theta \in S_d^2$ *be the unique solution of* (SDE_θ) *governed by* F_θ *with* $\theta \in]0,1]$. *Let* $C_0 = \mathbb{E}[(\|X_0\| + C)^2]$, *where* C *is the constant defined in* (53). *Then the following statements hold for any* $t > 0$.

(i) *Let* $\overline{X_\theta}(t) = t^{-1} \int_0^t X_\theta(s)ds$, *then*

$$\mathbb{E}\left(F\left(\overline{X_\theta}(t)\right) - \min F\right) \leq \frac{C_0}{2t} + \frac{\sigma_*^2}{2} + \theta \frac{L_0^2}{2}.$$

Besides, if $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$, *then*

$$\mathbb{E}\left(F\left(\overline{X_\theta}(t)\right) - \min F\right) = \frac{C_0 + \int_0^{+\infty} \sigma_\infty^2(s)ds}{2t} + \theta \frac{L_0^2}{2}.$$

(ii) *If* σ_∞ *verifies* (23) *with* $\beta \in [0,1[$, *and* $\theta \in]0,1]$, *then*

$$\mathbb{E}\left(F(X(t)) - \min F\right) = \frac{C_0}{2t} + \frac{K(1+L)}{2t^{1-\beta}\theta} + \theta \frac{L_0^2}{2}.$$

(iii) *If, in addition,* $f \in \Gamma_\mu(\mathbb{R}^d)$ *for some* $\mu > 0$, *then* $\mathcal{S} = \{x^\star\}$ *and*

$$\mathbb{E}\left(\|X_\theta(t) - x^\star\|^2\right) \leq 2C_0 e^{-\mu t} + \frac{2\sigma_*^2}{\mu} + 2\frac{L_0^2}{\mu}\theta.$$

Besides, if σ_∞ *is non-increasing and vanishes at infinity, then* $\forall \lambda \in]0,1[$:

$$\mathbb{E}\left(\|X_\theta(t) - x^\star\|^2\right) \leq 2C_0 e^{-\mu t} + \frac{2\sigma_*^2}{\mu} e^{-\mu(1-\lambda)t} + 2\sigma_\infty^2(\lambda t) + 2\frac{L_0^2}{\mu}\theta.$$

Remark 5.21. Observe that when $f = 0$, then $\mathcal{S}_\theta = \mathcal{S} = \{x^\star\}$. Therefore in Proposition 5.20, the last term in θ can be dropped.

Proof. Proof.

(i) Combine Theorem 3.3(i) applied to F_θ, Proposition 5.16(iii) and (v), and Proposition 5.19(i).

(ii) Argue as in claim (i) using Proposition 3.4 instead of Theorem 3.3(i), and use the fact that ∇F_θ is Lipschitz continuous with constant

$$L + \frac{1}{\theta} \leq \frac{L+1}{\theta} \quad \text{for} \quad \theta \in]0,1].$$

(iii) Combine Theorem 3.3(ii) applied to F_θ, Proposition 5.19(ii) and Jensen's inequality.

□

6 Conclusion, Perspectives

This work was intended to uncover the global and local convergence properties of trajectories of gradient-like flows under stochastic errors. The aim is to solve convex optimization problems with noisy gradient input with vanishing variance. We have shed light on these properties and provided a comprehensive local and global complexity analysis both in the smooth and non-smooth case. We believe that this work paves the way to many important extensions and research avenues. Among them, we mention the following ones:

- Let $\mathbb{H}$ and $\mathbb{K}$ be two real separable Hilbert spaces, we can extend naturally every result of this paper to the case where the data belongs to $\mathbb{H}$, W is a $\mathbb{K}$-valued Brownian motion, and the volatility term is a linear Hilbert-Schmidt operator from $\mathbb{K}$ to $\mathbb{H}$.
- Extension beyond the convex case, and for instance to the quasi-convex case, we refer to the recent work of [17] which offers us perspective concerning the extension of our work to the non-convex KL setting.
- Analyzing the non-smooth case, and more generally, the situations involving the sum of two maximal monotone operators one of which is merely Lipschitz continuous. This covers many important practical cases (*e.g.* primal-dual splitting, ADMM), and will take us to the land of stochastic differential inclusions. This will in turn allow us to understand the behavior of the solution X_θ of (SDE_θ) as θ vanishes. These are very interesting but challenging problems.
- Studying second-order dynamics with inertia in view of understanding the behavior of accelerated dynamics in presence of stochastic errors.

A Auxiliary results

A.1 Deterministic results

The following lemma is straightforward to prove. We omit the details.

Lemma A.1. *Let $t_0 > 0$ and $g : [t_0, +\infty[\to \mathbb{R}_+$. Suppose that $\lim_{t\to\infty} g(t)$ exists and $\int_{t_0}^{\infty} \frac{g(s)}{s} ds < +\infty$. Then $\lim_{t\to\infty} g(t) = 0$.*

The next result is an adaptation of [42, Proposition 2.3] to our specific context but under slightly less stringent assumptions.

Lemma A.2 (Comparison Lemma). *Let $t_0 \geq 0$ and $T > t_0$. Assume that $h : [t_0, +\infty[\to \mathbb{R}_+$ is measurable with $h \in \mathrm{L}^1([t_0, T])$, that $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous and non-decreasing, $\varphi_0 > 0$ and the Cauchy problem*

$$\begin{cases} \varphi'(t) = -\psi(\varphi(t)) + h(t) & \textit{for almost all } t \in [t_0, T] \\ \varphi(t_0) = \varphi_0 \end{cases}$$

has an absolutely continuous solution $\varphi : [t_0, T] \to \mathbb{R}_+$. If a lower semicontinuous function $\omega : [t_0, T] \to \mathbb{R}_+$ is bounded from below and satisfies

$$\omega(t) \leq \omega(s) - \int_s^t \psi(\omega(\tau))d\tau + \int_s^t h(\tau)d\tau$$

for $t_0 \leq s < t \leq T$ *and* $\omega(t_0) = \varphi_0$*, then*

$$\omega(t) \leq \varphi(t) \quad \textit{for } t \in [t_0, T].$$

Lemma A.3. *Let* $f : \mathbb{R}_+ \to \mathbb{R}$ *and* $\liminf_{t\to+\infty} f(t) \neq \limsup_{t\to+\infty} f(t)$*. Then, for every* α *satisfying* $\liminf_{t\to+\infty} f(t) < \alpha < \limsup_{t\to+\infty} f(t)$*, and for every* $\beta > 0$*, we can define a sequence* $(t_k)_{k\in\mathbb{N}} \subset \mathbb{R}_+$ *such that*

$$f(t_k) > \alpha, \quad t_{k+1} > t_k + \beta, \quad \forall k \in \mathbb{N}.$$

Proof. Proof. Since $\liminf_{t\to+\infty} f(t)$ and $\limsup_{t\to+\infty} f(t)$ are different real numbers, α in the lemma obviously exists. Moreover, by definition of lim sup, there exists a sequence $(t_k)_{k\in\mathbb{N}}$ such that $\lim_{k\to+\infty} t_k = +\infty$ and $f(t_k) > \alpha$. Let $\beta > 0$ and $n_0 = 0$, let us define recursively for $j \geq 1$, $n_j = \min\{n > n_{j-1} : t_n - t_{n_{j-1}} > \beta\}$. Let $j' \in \mathbb{N}$ be the first natural such that $n_{j'} = +\infty$. This implies that for every $n > n_{j'-1}$, $t_n \leq \beta + t_{n_{j'-1}} < +\infty$, a contradiction since $\lim_{n\to+\infty} t_n = +\infty$, then for every $j \in \mathbb{N}$, $n_j < +\infty$. Thus, we can define $(t_{n_j})_{j\in\mathbb{N}}$ a subsequence of $(t_k)_{k\in\mathbb{N}}$ such that $\lim_{j\to+\infty} t_{n_j} = +\infty$ and for every $j \in \mathbb{N}$, $t_{n_{j+1}} - t_{n_j} > \beta$. □

A.2 Stochastic and measure-theoretic results

Let us recall some elements of stochastic analysis; for a more complete account, we refer to [44, 48, 41]. Throughout the paper, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\{\mathcal{F}_t | t \geq 0\}$ is a filtration of the σ-algebra $\mathcal{F}$. Given $\mathcal{C} \subseteq \otimes$, we will denote $\sigma(\mathcal{C})$ the σ-algebra generated by $\mathcal{C}$. We denote $\mathcal{F}_\infty \overset{\text{def}}{=} \sigma\left(\bigcup_{t\geq 0} \mathcal{F}_t\right) \in \mathcal{F}$.

The expectation of an $\mathbb{R}^d$-valued random variable $\xi : \Omega \to \mathbb{R}^d$ is denoted by

$$\mathbb{E}(\xi) \overset{\text{def}}{=} \int_\Omega \xi(\omega) d\mathbb{P}(\omega).$$

As said in Section 2, for $1 \leq p \leq +\infty$, $\mathrm{L}^p(\Omega; \mathbb{R}^d)$ is the space of $\mathbb{R}^d$-valued random variables ξ such that $\mathbb{E}(\|\xi\|^p) < +\infty$, with the usual adaptation when $p = +\infty$.

An event $E \in \mathcal{F}$ happens almost surely if $\mathbb{P}(E) = 1$, and it will be denoted as "E, $\mathbb{P}$-a.s." or simply "E, a.s.". The characteristic function of an event $E \in \mathcal{F}$ is denoted by

$$\mathbb{1}_E(\omega) \overset{\text{def}}{=} \begin{cases} 1 & \text{if } \omega \in E, \\ 0 & \text{otherwise.} \end{cases}$$

An $\mathbb{R}^d$-valued stochastic process is a function $X : \Omega \times \mathbb{R}_+ \to \mathbb{R}^d$. It is said to be continuous if $X(\omega, \cdot) \in C(\mathbb{R}_+; \mathbb{R}^d)$ for almost all $\omega \in \Omega$. We will denote $X(t) \overset{\text{def}}{=} X(\cdot, t)$. We are going to study (SDE), and in order to ensure the uniqueness of a solution, we introduce a relation over stochastic processes. Two stochastic processes $X, Y : \Omega \times [0, T] \to \mathbb{R}^d$ are said to be equivalent if $X(t) = Y(t)$, $\forall t \in [0, T]$, $\mathbb{P}$-a.s.. This leads us to define the equivalence relation $\mathcal{R}$, which associates the equivalent stochastic processes in the same class.

Furthermore, we will need some properties about the measurability of these processes. A stochastic process $X : \Omega \times \mathbb{R}_+ \to \mathbb{R}^d$ is progressively measurable if for every $t \geq 0$, the map $\Omega \times [0, t] \to \mathbb{R}^d$ defined by $(\omega, s) \to X(\omega, s)$ is $\mathcal{F}_t \otimes \mathcal{B}([0, t])$-measurable, where $\otimes$ is the product σ-algebra and $\mathcal{B}$ is the Borel σ-algebra. On the other hand, X is $\mathcal{F}_t$-adapted if $X(t)$ is $\mathcal{F}_t$-measurable for every

$t \geq 0$. It is a direct consequence of the definition that if X is progressively measurable, then X is $\mathcal{F}_t$-adapted.

Let us define the quotient space:

$$S_d^0[0,T] \overset{\text{def}}{=} \left\{ X : \Omega \times [0,T] \to \mathbb{R}^d : \; X \text{ is a prog. measurable cont. stochastic process} \right\} \Big/ \mathcal{R}.$$

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

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

We define $S_d^\nu \overset{\text{def}}{=} \bigcap_{T \geq 0} S_d^\nu[0,T]$.

Theorem A.4 (Egorov's Theorem). *[54, Chapter 3, Exercise 16] Let (X, Σ, μ) be a measure space with $\mu(X) < +\infty$, and $(f_t)_{t \in \mathbb{R}_+}$ is a family of real measurable functions such that for μ-almost all $x \in X$:*

1. *$\lim_{t \to +\infty} f_t(x) = f(x)$, and*
2. *$t \mapsto f_t(x)$ is continuous.*

Then, for every $\delta > 0$, there exists a measurable set $E_\delta \subset X$, with $\mu(X \setminus E_\delta) < \delta$, such that $(f_t)_{t \in \mathbb{R}_+}$ converges uniformly on E_δ.

Lemma A.5. *Let $\delta > 0, \Omega_\delta \in \mathcal{F}$ such that $\mathbb{P}(\Omega_\delta) \geq 1 - \delta$ and $h : \Omega \times \mathbb{R}_+ \to \mathbb{R}$ a stochastic process such that $\sup_{t \geq 0} \mathbb{E}[h(t)^2] < +\infty$. Then, there exists a constant $C_h > 0$ (independent of δ) such that*

$$\mathbb{E}[h(t) \mathbb{1}_{\Omega \setminus \Omega_\delta}] \leq C_h \sqrt{\delta}.$$

Proof. Proof. Note that $\mathbb{P}(\Omega \setminus \Omega_\delta) \leq \delta$ and thus Cauchy-Schwarz inequality gives

$$\mathbb{E}[h(t)\mathbb{1}_{\Omega \setminus \Omega_\delta}] = \int_\Omega h(\omega, t) \mathbb{1}_{\Omega \setminus \Omega_\delta}(\omega) d\mathbb{P}(\omega) \leq \sqrt{\delta} \sqrt{\mathbb{E}[h(t)^2]} \leq \sqrt{\delta} \underbrace{\sqrt{\sup_{t \geq 0} \mathbb{E}[h(t)^2]}}_{\overset{\text{def}}{=} C_h}.$$

□

Corollary A.6. *Let $\delta > 0, \Omega_\delta \in \mathcal{F}$ such that $\mathbb{P}(\Omega_\delta) \geq 1 - \delta$. Consider (SDE) where f and σ satisfy the assumptions (H_0) and (H), respectively. Assume that $X_0 \in \mathrm{L}^4(\Omega; \mathbb{R}^d)$ and is $\mathcal{F}_0$-measurable. Moreover, suppose that $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$. Let $X \in S_d^4$ be the unique solution of (SDE), then there exists $C_d, C_f > 0$ (independent of δ) such that*

$$\mathbb{E}\left[\frac{\operatorname{dist}(X(t), \mathcal{S})^2}{2} \right] - \mathbb{E}\left[\frac{\operatorname{dist}(X(t), \mathcal{S})^2}{2} \mathbb{1}_{\Omega_\delta} \right] \leq C_d \sqrt{\delta},$$

and

$$\mathbb{E}\left[f(X(t)) - \min f \right] - \mathbb{E}\left[(f(X(t)) - \min f) \mathbb{1}_{\Omega_\delta} \right] \leq C_f \sqrt{\delta}.$$

Proof. Proof. Let $x^\star \in \mathcal{S}$. Using Proposition 2.4 with $\widehat{\phi}(x) = \frac{\mathrm{dist}(x,\mathcal{S})^2}{2}$, squaring the obtained inequality and taking expectation, we obtain

$$\begin{aligned}\mathbb{E}\left[\frac{\mathrm{dist}(X(t),\mathcal{S})^4}{4}\right] &\leq \frac{3}{4}\mathbb{E}(\mathrm{dist}(X_0,\mathcal{S})^4) + \frac{3}{4}\left(\int_0^t \sigma_\infty^2(s)ds\right)^2 \\ &+ 3\mathbb{E}\left[\left(\int_0^t \langle \sigma^\top(s,X(s))(X(s)-P_{\mathcal{S}}(X(s))), dW(s)\rangle\right)^2\right] \\ &\leq \frac{3}{4}\mathbb{E}(\mathrm{dist}(X_0,\mathcal{S})^4) + \frac{3}{4}\left(\int_0^t \sigma_\infty^2(s)ds\right)^2 + 3\sup_{t\geq 0}\mathbb{E}[\|X(t)-x^\star\|^2]\left[\int_0^t \sigma_\infty^2(s)ds\right].\end{aligned}$$

Taking the supremum over $t \geq 0$, we obtain

$$\begin{aligned}\sup_{t\geq 0}\mathbb{E}\left[\left(\frac{\mathrm{dist}(X(t),\mathcal{S})^2}{2}\right)^2\right] &\leq \frac{3}{4}\mathbb{E}(\mathrm{dist}(X_0,\mathcal{S})^4) + \frac{3}{4}\left(\int_0^{+\infty} \sigma_\infty^2(s)ds\right)^2 \\ &+ 3\sup_{t\geq 0}\mathbb{E}[\|X(t)-x^\star\|^2]\left[\int_0^{+\infty} \sigma_\infty^2(s)ds\right] \overset{\mathrm{def}}{=} C_d < +\infty.\end{aligned}$$

In the above estimation we used that $\sigma_\infty \in \mathrm{L}^2(\mathbb{R}_+)$ and $\sup_{t\geq 0}\mathbb{E}[\|X(t)-x^\star\|^2] < +\infty$ by Theorem 3.1(i).

On the other hand, since $f \in \Gamma_0(\mathbb{R}^d) \cap C_L^{1,1}(\mathbb{R}^d)$ and $X_0 \in \mathrm{L}^4(\Omega;\mathbb{R}^d)$, we have that

$$\mathbb{E}([f(X_0)-\min f]^2) \leq \frac{1}{2}\mathbb{E}(\|\nabla f(X_0)-\nabla f(x^\star)\|^4)+\frac{1}{2}\mathbb{E}(\|X(t)-x^\star\|^4) < \frac{L^4+1}{2}\mathbb{E}(\|X_0-x^\star\|^4) < +\infty.$$

Then using Proposition 2.4 with $\widetilde{\phi}(x) = f(x) - \min f$, squaring it, and taking expectation, we obtain

$$\begin{aligned}\mathbb{E}\left[[f(X(t)-\min f]^2\right] &\leq 3\mathbb{E}([f(X_0)-\min f]^2) + \frac{3L^2}{4}\left(\int_0^t \sigma_\infty^2(s)ds\right)^2 \\ &+ 3\mathbb{E}\left[\left(\int_0^t \langle \sigma^\top(s,X(s))(\nabla f(X(s))), dW(s)\rangle\right)^2\right] \\ &\leq 3\mathbb{E}([f(X_0)-\min f]^2) + \frac{3L^2}{4}\left(\int_0^t \sigma_\infty^2(s)ds\right)^2 \\ &+ 3L^2\sup_{t\geq 0}\mathbb{E}[\|X(t)-x^\star\|^2]\left[\int_0^t \sigma_\infty^2(s)ds\right].\end{aligned}$$

Taking the supremum over $t \geq 0$, we obtain

$$\begin{aligned}\sup_{t\geq 0}\mathbb{E}\left[[f(X(t)-\min f]^2\right] &\leq 3\mathbb{E}([f(X_0)-\min f]^2) + \frac{3L^2}{4}\left(\int_0^{+\infty} \sigma_\infty^2(s)ds\right)^2 \\ &+ 3L^2\sup_{t\geq 0}\mathbb{E}[\|X(t)-x^\star\|^2]\left[\int_0^{+\infty} \sigma_\infty^2(s)ds\right] \overset{\mathrm{def}}{=} C_f < +\infty.\end{aligned}$$

And we have proved the hypothesis of Lemma A.5 in both cases, applying this lemma, we conclude the proof. □

Let us consider $\nu \geq 2$ and the SDE with initial data $X_0 \in \mathrm{L}^\nu(\Omega;\mathbb{R}^d)$ which is $\mathcal{F}_0$-measurable:

$$\begin{cases} dX(t) = F(t,X(t))dt + G(t,X(t))dW(t), \quad t \geq 0, \\ X(0) = X_0, \end{cases} \tag{57}$$

where $F : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$, $G : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^{d\times m}$ are measurable functions and W is an $\mathcal{F}_t$-adapted m-dimensional Brownian Motion.

Theorem A.7. *(See [44, Theorem 5.2.1], [41, Theorem 2.3.1 and Theorem 2.4.4]) Let $F : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ and $G : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^{d\times m}$ be measurable functions satisfying*

$$\sup_{t\geq 0} \left(\|F(t,0)\| + \|G(t,0)\|_F\right) < +\infty, \tag{58}$$

and for every $T > 0$ and some constant $C_1 \geq 0$,

$$\|F(t,x) - F(t,y)\| + \|G(t,x) - G(t,y)\|_F \leq C_1\|x-y\|, \quad \forall x,y \in \mathbb{R}^d, \forall t \in [0,T]. \tag{59}$$

Then (57) has a unique solution $X \in S^\nu_d$.

Proof. Proof. Conditions (58) and (59) implies that there exists $C_2 \geq 0$ such that

$$\|F(t,x)\|^2 + \|G(t,x)\|_F^2 \leq C_2(1+\|x\|^2), \quad \forall x \in \mathbb{R}^d, \forall t \in [0,T],$$

which is the necessary inequality to use [41, Theorem 2.4.4] and deduce that $X \in S^\nu_d$.

□

A.3 On martingales

Theorem A.8. *[23] Let $(M_t)_{t\geq 0} : \Omega \to \mathbb{R}$ be a continuous martingale such that $\sup_{t\geq 0} \mathbb{E}\left(|M_t|^p\right) < +\infty$ for some $p > 1$. Then there exists a random variable $M_\infty \in \mathrm{L}^p(\Omega;\mathbb{R})$ such that $\lim_{t\to+\infty} M_t = M_\infty$ a.s..*

Theorem A.9. *[41, Theorem 1.3.9] 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*

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

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

Proposition A.10. *(see [41, Theorem 1.7.3]) Let $p > 0$, W be a m-dimensional Brownian motion defined over a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ and $g : \Omega \times \mathbb{R}_+ \to \mathbb{R}^m$ (with our usual notation $g(t) \overset{\text{def}}{=} g(\cdot,t)$) be such that*

$$\mathbb{E}\left[\int_0^T \|g(s)\|^2 ds\right] < +\infty, \quad \forall T > 0.$$

Then, there exists $C_p > 0$ (only depending on p) for every $T > 0$ such that:

$$\mathbb{E}\left[\sup_{t\in[0,T]} \left|\int_0^t \langle g(s), dW(s)\rangle\right|^p\right] \leq C_p \mathbb{E}\left[\left(\int_0^T \|g(s)\|^2 ds\right)^{\frac{p}{2}}\right].$$

In particular, we have that $C_2 = 4$.

References

[1] A.S. Antipin. Minimization of convex functions on convex sets by means of differential equations. *Differ. Uravn.*, 30(9):1475–1486, 1994.

[2] Francisco Javier Aragón Artacho and Michel Geoffroy. Characterization of metric regularity of subdifferentials. *Journal of Convex Analysis*, 15:365–380, 01 2008.

[3] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, Forward–Backward splitting, and regularized Gauss–Seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[4] Hedy Attouch and Alexandre Cabot. Convergence of a relaxed inertial forward-backward algorithm for structured monotone inclusions. *Applied Mathematics and Optimization, special issue on Games, Dynamics and Optimization*, 80 (3):547–598, 2019.

[5] Peter Bartlett and Walid Krichene. Acceleration and averaging in stochastic mirror descent dynamics. *arXiv: 1707.06219*, 2017.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[7] Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

[8] M. Benaïm and S.J. Schreiber. Ergodic properties of weak asymptotic pseudotrajectories for semiflows. *Journal of Dynamics and Differential Equations*, 12:579–598, 2000.

[9] Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pages 1–69. Springer, 1999.

[10] R. N. Bhattacharya. Criteria for recurrence and existence of invariant measures for multidimensional diffusions. *Ann. Prob.*, 6(4):541–553, 1978.

[11] Jérôme Bolte. Continuous gradient projection method in Hilbert spaces. *Journal of Optimization Theory and its Applications*, 119(2):235–259, 2003.

[12] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165:471–507, 2016.

[13] H. Brézis. *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, volume 5 of *Mathematics studies*. North-Holland, New York, 1973.

[14] Olivier Catoni. Simulated annealing algorithms and Markov chains with rare transitions. In *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pages 70–119. Springer, 1999.

[15] A. Cauchy. Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptes Rendus de l'Académie des Sciences de Paris*, 25:536–538, 1847.

[16] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin mcmc: A non-asymptotic analysis. *arXiv:1707.03663*, 2017.

[17] Emilie Chouzenoux, Jean-Baptiste Fest, and Audrey Repetti. A kurdyka-lojasiewicz property for stochastic optimization algorithms in a non-convex setting. *arXiv:2302.06447*, 2023.

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

[19] Arnak S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J.R. Stat. Soc. Series B. Stat. Methodol.*, 79(3):651–676, 2017.

[20] Arnak S. Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin monte carlo with inaccurate gradient. *arXiv:1710.00095v3*, 2018.

[21] Marc Dambrine, Ch Dossal, Bénédicte Puig, and Aude Rondepierre. Stochastic Differential Equations for modeling first order optimization methods. hal-03630785, April 2022.

[22] Steffen Dereich and Sebastian Kassing. Cooling down stochastic differential equations: Almost sure convergence. *arXiv:2106.03510*, 2021.

[23] J.L. Doob. Stochastic processes. *Wiley*, 1991.

[24] A. Durmus and E Moulines. High-dimensional bayesian inference via the unadjusted Langevin algorithm. *arXiv:1605.01559*, 2016.

[25] A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.

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

[27] Pierre Frankel, Guillaume Garrigos, and Juan Peypouquet. Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2014.

[28] C.W. Gardiner. Handbook of stochastic methods. *Springer*, 3, 1985.

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

[30] J. Huggins and J. Zou. Quantifying the accuracy of approximate diffusions and Markov chains. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54 of Proceedings of Machine Learning Research:382–391, 2017.

[31] M. Kisielewicz. *Stochastic Differential Inclusions and Applications*, volume 80 of *Springer Optimization and Its Applications*. Springer, 2013.

[32] P. Krée. Diffusion equation for multivalued stochastic differential equations. *J. Func. Anal.*, 49:73–90, 1982.

[33] P. Krée and C. Soize. *Mathematics of random phenomena*. Reidel Publishing Company, 1986.

[34] Alexander Y. Kruger. Error bounds and hölder metric subregularity. *Set-Valued and Variational Analysis*, 23(4):705–736, 2015.

[35] Alexander Y. Kruger. Error bounds and metric subregularity. *Optimization*, 64(1):49–79, 2015.

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

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

[38] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles*, pages 87–89. Editions du Centre National de la Recherche Scientifique, 1963.

[39] S. Łojasiewicz. Ensembles semi-analytiques. *Lectures Notes IHES (Bures-sur-Yvette)*, 1965.

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

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

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

[43] 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.

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

[45] Z. Opial. Weak convergence of the sequence of successive approximations for nonexpansive mappings. *Bull. Amer. Math. Soc.*, 73:591–597, 1967.

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

[47] Jong-Shi Pang. Error bounds in mathematical programming. *Mathematical Programming*, 79(1):299–332, 1997.

[48] Etienne Pardoux and Aurel Rascanu. *Stochastic differential equations, backward SDEs, partial differential equations*. Springer, 2014.

[49] G. Parisi. Correlation functions and computer simulations. *Nucl. Phys. B*, 180(3):378–384, 1981.

[50] R. Pettersson. Yosida approximations for multivalued stochastic differential equations. *Stochastics and Stochastics Reports*, 52:107–120, 1995.

[51] Maxim Raginsky and Jake Bouvrie. Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence. *2012 IEEE 51st IEEE Conference on Decision and Control*, 2012.

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

[53] R.T. Rockafellar. Convex analysis. *Princeton university press*, 28, 1997.

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

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

[56] 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.