Stochastic Inertial Dynamics Via Time Scaling and Averaging

Rodrigo Maulen-Soto* Jalal Fadili† Hedy Attouch‡ Peter Ochs§

Abstract. Our work is part of the close link between continuous-time dissipative dynamical systems and optimization algorithms, and more precisely here, in the stochastic setting. We aim to study stochastic convex minimization problems through the lens of stochastic inertial differential inclusions that are driven by the subgradient of a convex objective function. This will provide a general mathematical framework for analyzing the convergence properties of stochastic second-order inertial continuous-time dynamics involving vanishing viscous damping and measurable stochastic subgradient selections. Our chief goal in this paper is to develop a systematic and unified way that transfers the properties recently studied for first-order stochastic differential equations to second-order ones involving even subgradients in lieu of gradients. This program will rely on two tenets: time scaling and averaging, following an approach recently developed in the literature by one of the co-authors in the deterministic case. Under a mild integrability assumption involving the diffusion term and the viscous damping, our first main result shows that almost surely, there is weak convergence of the trajectory towards a minimizer of the objective function and fast convergence of the values and gradients. We also provide a comprehensive complexity analysis by establishing several new pointwise and ergodic convergence rates in expectation for the convex, strongly convex, and (local) Polyak-Łojasiewicz case. Finally, using Tikhonov regularization with a properly tuned vanishing parameter, we can obtain almost sure strong convergence of the trajectory towards the minimum norm solution.

Key words. Stochastic convex optimization; Inertial gradient systems; Stochastic Differential Equation; Tikhonov regularization; Time-dependent viscosity.

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

1 Introduction

1.1 Problem Statement

We consider the minimization problem

$$\min_{x\in\mathbb{H}} F(x) \stackrel{\text{def}}{=} f(x) + g(x), \tag{P}$$

where $\mathbb{H}$ is a separable real Hilbert space, and the objective F satisfies the following standing assumptions:

$$\begin{cases} f:\mathbb{H}\to\mathbb{R} \text{ is continuously differentiable and convex with } L\text{-Lipschitz continuous gradient;} \\ g:\mathbb{H}\to\mathbb{R}\cup\{\pm\infty\} \text{ is proper, lower semi-continuous (lsc) and convex;} \\ \mathcal{S}_F \stackrel{\text{def}}{=} \operatorname{argmin}(F) \neq \emptyset. \end{cases} \tag{H$_0$}$$

*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
§Department of Mathematics and Computer Science, Saarland University, Germany, E-mail: ochs@math.uni-sb.de

First-order in time systems. To solve (P) when $g \equiv 0$, a fundamental dynamic is the gradient flow system:

$$\begin{cases} \dot{x}(t) + \nabla f(x(t)) = 0, & t > t_0; \\ x(t_0) = x_0. \end{cases} \tag{GF}$$

The gradient system (GF) is a dissipative dynamical system, whose study dates back to Cauchy [25] in finite dimension. 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. 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 weakly, and its (weak) limit belongs to $\operatorname{argmin}(f)$. Moreover, this dynamic is known to yield a convergence rate of $\mathcal{O}(t^{-1})$ (in fact even $o(t^{-1})$) on the values.

The Euler forward (a.k.a. Euler-Maruyama) discretization (GF), with stepsize sequence $h_k > 0$, is the celebrated gradient descent scheme

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

Under (H_0), and for $(h_k)_{k\in\mathbb{N}} \subset]0, 2/L[$, 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 $\operatorname{argmin}(f)$. This convergence rate can be refined under various additional geometrical properties on the objective f such as error bounds (and the closely related Kurdyka-Łojasiewicz property in the convex case; see [19]).

Second-order in time systems: Key role of inertia. Second-order in time inertial dynamical systems have been introduced to provably accelerate the convergence behavior compared to (GF). An abundant literature has been devoted to the study of inertial dynamics

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t)) = 0, \quad t > t_0. \tag{IGS$_\gamma$}$$

The importance of working with an asymptotically vanishing viscosity coefficient $\gamma(t)$ to obtain acceleration was stressed by several authors; see for instance [7, 22]. Most of the literature focuses on the case $\gamma(t) = \frac{\alpha}{t}$, originating from the seminal work of [80] who showed the rate of convergence $\mathcal{O}(1/t^2)$ of the values for $\alpha = 3$, thus making the link with the Nesterov accelerated gradient method [67]. Since then, an important body of literature has been devoted to this important case and the subtle tuning of the parameter α. Taking $\alpha \geq 3$ is mandatory for provable accelerated rate $\mathcal{O}\left(1/t^2\right)$ of the values [11], and $\alpha > 3$ provides an even better rate $o\left(1/t^2\right)$ and weak convergence of the trajectory [16, 64]. On the other hand, $\alpha < 3$ necessarily leads to a slower rate $\mathcal{O}\left(1/t^{2\alpha/3}\right)$ [3, 12]. Another remarkable instance of (IGS$_\gamma$) corresponds to the well-known Heavy Ball with Friction (HBF) method, where $\gamma(t)$ is a constant, first introduced by Polyak in [73]. When the objective is strongly convex, it was shown that the trajectory converges exponentially with an optimal convergence rate for a properly chosen constant γ [15].

Geometric Hessian-driven damping. Because of the inertial aspects, and the asymptotic vanishing viscous damping coefficient, (IGS$_\gamma$) may exhibit many small oscillations which are not desirable from an optimization point of view. To remedy this, a powerful tool consists in introducing a geometric damping driven by the Hessian of f into the dynamic. This yields the Inertial System with Explicit Hessian-driven Damping [17, 10]

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \beta(t)\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0, \tag{ISEHD}$$

and the Inertial System with Implicit Hessian-driven Damping [1, 66]

$$\left\{ \ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t) + \beta(t)\dot{x}(t)) = 0. \right. \tag{ISIHD}$$

The rationale behind the use of the term "implicit" in (ISIHD) comes from a Taylor expansion of the gradient term (as $t \to +\infty$ we expect $\dot{x}(t) \to 0$). Following the physical interpretation of these ODEs, we call the non-negative parameters γ and β as the viscous and geometric damping parameters, respectively.

At first glance, the presence of the Hessian in (ISEHD) may seem to entail numerical difficulties. However, this is not the case as the term $\nabla^2 f(x(t))\dot{x}(t)$ is nothing but the time derivative of $t \mapsto \nabla f(x(t))$. This explains why the time discretization of this dynamic provides efficient first-order algorithms [10]. On the other hand, (ISEHD) can be argued to be truly of second-order nature, i.e., close to Newton's and Levenberg-Marquardt's dynamics [23]. This understanding suggests that (ISIHD) may reflect the nature of first-order algorithms more faithfully than (ISEHD). Moreover, when it will come to our stochastic setting, which is the focus of this paper, the approach we propose will only make sense for the implicit form of the Hessian-driven damping. Indeed, in our stochastic setting, we do not have direct access to the evaluation of the gradient of f. Instead, we model the associated errors with a continuous Itô martingale (denoted as $M(t)$). But for the explicit form of the Hessian-driven damping, this would entail a term involving the time derivative of $\nabla f(X(t)) + M(t)$. This is meaningless as (non-constant) martingales are not differentiable a.s.. This is why from now on, we will solely consider (ISIHD).

1.2 Motivations

In the following, $\mathbb{H}, \mathbb{K}$ are real separable Hilbert spaces. In many practical situations, the (sub-)gradient evaluation is subject to stochastic errors. This is for example the case if the cost per iteration is very high and thus cheap and random approximations of the (sub-)gradient are necessary. These errors can also be due to some other exogenous factor. The continuous-time approach through stochastic differential equations (SDE) is a powerful way to model these errors in a unified way, and stochastic algorithms can then be viewed as time- discretizations. In fact, several recent works have used the SDE perspective to model SGD-type algorithms motivated by various reasons; (see *e.g.* [51, 70, 42, 78, 53, 79, 65, 62, 30, 63]). The continuous-time perspective offers a deep insight and unveils the key properties of the dynamic without being tied to a specific discretization.

In this context, the first SDE that comes to mind is

$$\begin{cases} dX(t) &= -\nabla f(X(t)) + \sigma(t, X(t))dW(t), \quad t > t_0; \\ X(t_0) &= X_0, \end{cases} \tag{SGF}$$

defined over a complete filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t \geq t_0}, \mathbb{P})$, where the diffusion (volatility) term $\sigma : \mathbb{R}^+ \times \mathbb{H} \to \mathcal{L}_2(\mathbb{K}; \mathbb{H})$ is a measurable function, W is a $\mathcal{F}_t$-adapted $\mathbb{K}$-valued cylindrical Brownian motion, and the initial data X_0 is an $\mathcal{F}_{t_0}-$measurable $\mathbb{H}$-valued random variable. All these notations and concepts are introduced in Section 2. (SGF) is a stochastic counterpart of (GF) where the error is modeled as a stochastic integral with respect to the measure defined by a continuous Itô martingale.

SDE modeling of SGD. To simplify the discussion, let us focus in this section on the finite-dimensional case ($\mathbb{H} = \mathbb{R}^d$, $\mathbb{K} = \mathbb{R}^m$). In various areas of science and engineering, in particular in machine learning, the Stochastic Gradient Descent (SGD) is a powerful alternative to gradient descent, and consists in replacing the

full gradient computation by a cheap random version, serving as an unbiased estimator. The SGD updates the iterates according to

$$x_{k+1} = x_k - h(\nabla f(x_k) + e_k) \tag{SGD}$$

where $h \in \mathbb{R}_+$ is the stepsize and e_k is the random noise term on the gradient at the k-th iteration. As such, (SGD) can be viewed as instance of the Robbins-Monro stochastic approximation algorithm [74]. When the objective takes the form $f(x) = \mathbb{E}[\hat{f}(x, \xi)]$, where the expectation is w.r.t. to the random variable ξ the single-batch version of SGD reads

$$x_{k+1} = x_k - h\nabla \hat{f}(x_k, \xi_k), \tag{SGD$_{\text{SB}}$}$$

where $(\xi_k)_{k\in\mathbb{N}}$ are i.i.d. random variables with the same distribution as ξ. Of course, (SGD$_{\text{SB}}$) is an instance of (SGD) with $e_k = \nabla \hat{f}(x_k, \xi_k) - \nabla f(x_k)$.

The SDE continuous-time approach is motivated by its close relation to (SGD) or (SGD$_{\text{SB}}$). We first note that when the noise e_k in (SGD) is $\mathcal{N}(0, \sigma_k I_d)$, (SGF) is a better continuous-time model for (SGD) than (GF), as has been shown recently in [30, Proposition 2.1]. There, the sequence $(x_k)_{k\in\mathbb{N}}$ provided by (SGD), with $e_k \sim \mathcal{N}(0, \sigma_k I_d)$, was proved to be accurately approximated by (SGF) with $\sigma(t, X(t)) = \sqrt{h}\sigma(t)$ and $\sigma(kh) = \sigma_k I_d$.

For the standard single-batch SGD (SGD$_{\text{SB}}$), the argument is more involved. Actually, many recent works (see *e.g.* [60, 51, 70, 42, 78, 53, 79, 82, 49, 52]) have linked algorithm (SGD$_{\text{SB}}$) with continuous-time first-order stochastic diffusion dynamics such as (SGF). These works show either empirically or theoretically under which conditions (appropriate drift and diffusion terms, regularity of f, etc.) (SGF) can be seen as a good approximation model of (SGD$_{\text{SB}}$) for fixed stepsize. By a good model we mean that (SGF) is a continuum limit of (SGD$_{\text{SB}}$) as the stepsize h goes to zero, or equivalently that the approximation of (SGD$_{\text{SB}}$) via the diffusion process (SGF) is precise in some weak sense; see [51, 42, 53].

As a consequence, using (SGF) as a proxy of (SGD) or (SGD$_{\text{SB}}$) allows to capitalize on the wealth of results in the field of SDEs, Itô calculus and measure theory, and this in turn opens the door to new insights in the behavior of (SGD) or (SGD$_{\text{SB}}$) and to transfer all the convergence results that one can prove for (SGF) to (SGD). Actually, this is one of the main messages we want to convey in this work. Our motivation and results are also complementary to those in the literature. Indeed, most, if not all, of works cited in the previous paragraph are primarily motivated by the fact that continuous-time SDE approximation of (SGD) is a crucial tool to study its escape behavior of bad saddle points (a.k.a. traps) in the non-convex smooth case. Our standpoint, which is line with [62, 65, 30, 63], is complementary and we argue that the continuous-time perspective offers a deep insight and unveils the key properties of the dynamic, without being tied to a specific discretization. This enlightens the behavior of the sequence generated by some specific algorithm. In turn, studying the continuous-time SDE will allow to predict the convergence behavior of stochastic algorithms seen as a discretization of the corresponding continuous-time dynamics.

Second-order SDE modeling of inertial SGD. Using a lifting argument to get an equivalent first-order reformulation, a natural generalization of (ISIHD) to the non-smooth case yields the differential inclusion

$$\begin{cases} \dot{x}(t) = v(t), \quad t > t_0; \\ \dot{v}(t) \in -[\gamma(t)v(t) + \partial F(x(t) + \beta(t)v(t))], \quad t > t_0; \\ x(t_0) = x_0, \quad \dot{x}(t_0) = v_0, \end{cases} \tag{ISIHD$_{\text{NS}}$}$$

where ∂F is the convex subdifferential of F. In this setting, keeping in mind that we want to give a rigorous meaning to (ISIHD$_{\text{NS}}$), we 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 inclusion (SDI for short), which is the stochastic counterpart of (ISIHD$_{\rm NS}$):

$$\begin{cases} dX(t) &= V(t)dt, \\ dV(t) &\in -\gamma(t)V(t)dt - \partial F(X(t)+\beta(t)V(t))dt + \sigma(t, X(t)+\beta(t)V(t))dW(t), \\ X(t_0) &= X_0, \quad V(t_0) = V_0. \end{cases} \tag{S – ISIHD$_{\rm NS}$}$$

This SDI is defined on a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$, where $\sigma : [t_0, +\infty[\times\mathbb{H} \to \mathcal{L}_2(\mathbb{K};\mathbb{H})$ is a measurable function; and W is a $\mathbb{K}$-valued cylindrical Brownian motion. When $g \equiv 0$, this reduces to the stochastic counterpart of (ISIHD), given by the following SDE

$$\begin{cases} dX(t) &= V(t)dt, \\ dV(t) &= -\gamma(t)V(t)dt - \nabla f(X(t)+\beta(t)V(t))dt + \sigma(t, X(t)+\beta(t)V(t))dW(t), \\ X(t_0) &= X_0, \quad V(t_0) = V_0. \end{cases} \tag{S – ISIHD}$$

In the smooth but non-convex case, and motivated again by strict saddle points avoidance, the authors in [43, 44] study the convergence behavior of a stochastic discrete heavy ball method from its approximating SDE. The latter is a randomly perturbed nonlinear oscillator in [43] and a coupled system of nonlinear oscillators in [44]. These SDEs are very different from the ones considered here. By contrast, other works do not employ SDE techniques and instead analyze the stochastic iteration directly; see, for instance, [18, 37, 50], with the last reference also addressing the non-smooth case. There are several advantages of adopting the SDE perspective in our setting, among which a more accurate approximation of the discrete stochastic dynamics than the ODE method and the ability to quantify stochastic fluctuations and describe the long-run distribution of the iterates; the approximation aspect is detailed in Appendix A.

Extending what we have discussed above for (SGF) as a good model of (SGD), we will show in Proposition A.2 (see Appendix A for details) that (S – ISIHD) is a good model of the natural stochastic inertial algorithm (A.1) obtained by simple Euler forward (a.k.a. Euler-Maruyama) discretization of (S – ISIHD). The corresponding convergence rate as a function of the time step-size h is $\mathcal{O}(h)$ and show that this is much better that the approximation rate of (ISIHD) which is only $\mathcal{O}(\sqrt{h})$. As a consequence, this justifies and motivates the continuous-time dynamics (S – ISIHD) (and (S – ISIHD$_{\rm NS}$)) as a good proxy of stochastic inertial algorithms and opens the door to new insights in the behavior of such algorithms, and that their convergence properties can be easily derived from those of (S – ISIHD) with minimal effort. For instance, when f is smooth with Lipschitz-continuous gradient, it is easy to see from the descent lemma that

$$\mathbb{E}[f(X_k) - \min f] = \mathbb{E}[f(X(kh)) - \min f] + \mathcal{O}(h),$$

where X_k are the iterates of (A.1) and $X(t)$ the solution to (S – ISIHD). This means that any rate proved on $\mathbb{E}[f(X(t)) - \min f]$ can be directly transferred to $\mathbb{E}[f(X_k) - \min f]$.

1.3 Objectives and Contributions

In this work, our goal is to provide a general mathematical framework for analyzing the convergence properties of (S – ISIHD) and (S – ISIHD$_{\rm NS}$). In this context, considering inertial dynamics with a time-dependent vanishing viscosity coefficient γ is a key ingredient to obtain fast convergent methods. We will develop a systematic and unified way that transfers the properties of stochastic first-order dynamics recently studied by [62, 63] to second-order ones. Our program will then rely on two pillars: *time scaling* and *averaging*, following the methodology recently developed in [6] for the deterministic case.

More precisely, we study the stochastic dynamics ($\mathrm{S-ISIHD_{NS}}$) and its long-time behavior in order to solve (P). We conduct a new analysis using specific and careful arguments that are much more involved than in the deterministic case. To get some intuition, keeping the discussion informal at this stage, let us first identify the assumptions needed to expect that the position state of ($\mathrm{S-ISIHD}$) "approaches" $\mathrm{argmin}(f)$ in the long run. In the case where $\mathbb{H} = \mathbb{K}$, $\gamma(\cdot) \equiv \gamma > 0, \beta \equiv 0$, and $\sigma = \tilde{\sigma} I_{\mathbb{H}}$, where $\tilde{\sigma}$ is a positive real constant; under mild assumptions one can show that ($\mathrm{S-ISIHD}$) has a unique invariant distribution $\pi_{\tilde{\sigma}}$ in (x, v) with density proportional to $\exp\left(-\frac{2\gamma}{\tilde{\sigma}^2}\left(f(x) + \frac{\|v\|^2}{2}\right)\right)$, see e.g., [71, Proposition 6.1]. Clearly, as $\tilde{\sigma} \to 0^+$, $\pi_{\tilde{\sigma}}$ gets concentrated around $\mathrm{argmin}(f) \times \{0_{\mathbb{H}}\}$, with $\lim_{\tilde{\sigma}\to 0^+} \pi_{\tilde{\sigma}}(\mathrm{argmin}(f) \times \{0_{\mathbb{H}}\}) = 1$; see also Section 1.4 for further discussion. Motivated by these observations and the fact that we aim to exactly solve (P), our paper will then mainly focus on the case where $\sigma(\cdot, x)$ vanishes fast enough as $t \to +\infty$ uniformly in x.

Our main contributions are summarized as follows:

- We show almost sure weak convergence of the trajectory (see Theorem 3.1) and convergence rates (see Theorem 3.2) in expectation in the case of time-dependent coefficients $\gamma(t)$ and a proper choice of $\beta(t)$ (depending on γ). For this analysis, we transfer the results from the Lyapunov analysis of the first-order in-time stochastic (sub-)gradient system studied in [62, 63] from which our inertial system is built through time scaling and averaging.
- We obtain almost sure and ergodic convergence results which correspond precisely to the best-known results in the deterministic case. In particular, if we let $\alpha > 3, \gamma(t) = \frac{\alpha}{t}, \beta(t) = \frac{t}{\alpha-1}$, then under appropriate assumptions on the diffusion (volatility) term σ, we obtain the rate of convergence $o(1/t^2)$ of the values as well as fast convergence of the gradient in almost sure sense (see Corollary 3.5). This corresponds to the best known results of second-order in-time Hessian driven damping dynamics in the deterministic case.
- As far as the interplay between viscous damping, geometric damping, and noise level, our results reveal that viscous damping, which allows for acceleration, enters in the control of the noise and may impose a more stringent summability condition on the diffusion coefficient (see e.g., Theorem 3.1, Theorem 3.2 and Corollary 3.5). The geometric damping, which is designed to reduce oscillations, does not enter directly in this control and can be chosen in a more flexible way.
- We then turn to providing a local analysis with a local linear convergence rate under the Polyak-Łojasiewicz inequality (See Theorem 3.7). This is much more challenging in the stochastic case, and even more for second-order systems, as localizing the process in this case is very delicate. Leveraging time scaling and averaging offers an elegant framework to achieve such a local convergence analysis while it is barely possible otherwise. To the best of our knowledge, this is the first result of this kind for these systems in the literature.
- We also show almost sure strong convergence of the trajectory to the minimal norm solution when adding a Tikhonov regularization to our systems (see Theorem 4.1). Moreover, we show convergence rates in expectation for the objective and the trajectory for a particular Tikhonov regularizer (see Theorem 4.5).

It is worth observing that since our approach is based on an averaging technique, it will involve Jensen's inequality at some point to get fast convergence rates. In this respect, the convexity assumption on the objective function appears unavoidable, at least in this proof strategy. It is also worth stressing the fact that the approach only makes sense for the implicit form of the Hessian-driven damping. Indeed, as explained above, the explicit form of the Hessian-driven damping has a term involving the time derivative of the (sub)gradient at the trajectory. As the noise, modeled here as an Itô martingale, in practice stems from the (sub)gradient evaluation, this time derivative is meaningless with explicit Hessian-driven damping, as (non-constant) mar-

tingales are a.s. non-differentiable.

1.4 Relation to prior work

Kinetic diffusion dynamics for sampling. Let us consider (S − ISIHD) in the case where $\mathbb{H} = \mathbb{K} = \mathbb{R}^n$, $\gamma(t) = \gamma > 0, \beta \equiv 0$, and $\sigma = \sqrt{2\gamma}I$. Then one recovers the kinetic Langevin diffusion (or second-order Langevin process). In this case, the continuous-time Markov process $(X(t), V(t))$ is positive recurrent and has a unique invariant distribution which has the density $\propto \exp\left(-f(x) - \frac{\|v\|^2}{2}\right)$ with respect to the Lebesgue measure on $\mathbb{R}^{2n}$. Time-discretized versions of this Langevin diffusion process have been studied in the literature to (approximately) sample from $\propto \exp(-f(x))$ with asymptotic and non-asymptotic convergence guarantees in various topologies and under various conditions; see [26, 59, 29] and references therein. By rescaling the problem, relation between sampling and optimization with quantitive estimates has been investigated in *e.g.* [28] for the strongly convex case.

First-order stochastic (sub)gradient systems. In [62] (resp. [63]), the authors studied (SGF) (resp. its non-smooth version as an SDI) as a proxy for (SGD). One of their main results is the almost sure weak convergence of the trajectory to the set of minimizers under integrability conditions on σ, as well as convergence rates in expectation and in almost sure sense of the dynamic under different geometries of f. Our goal here is to take these results to second-order inertial dynamics featuring both viscous and geometric dampings. This turns out to be a challenging task. We will show that the convergence rate on the objective can be achieved provided that the noise vanishes sufficiently fast.

Inexact inertial gradient systems. There is an abundant literature regarding the dynamics (ISIHD) and (ISEHD), either in the exact case or with errors but only deterministic ones; see [1, 13, 41, 77, 8, 11, 32, 76, 81, 5, 10, 9, 24]. Only a few papers have been devoted to studying stochastic versions of (IGS_γ), either with vanishing damping $\gamma(t) = \alpha/t$ or constant damping $\gamma(t)$ (stochastic HBF); see e.g. [36, 37, 30, 69]. One of the advantages of considering the stochastic version of (ISIHD) is the possible reduction of oscillations thanks to the introduction of the geometric damping. We show that this effect could be preserved in the stochastic setting, hence obtaining more stable trajectory solutions than other stochastic second-order dynamics. Additionally, having the flexibility to choose the viscous damping would eventually allow us to achieve faster convergence results. For a stochastic version of (IGS_γ), [30] provide asymptotic $\mathcal{O}(1/t^2)$ convergence rate on the objective values in expectation under integrability conditions on the diffusion term as well as other rates under additional geometrical properties of the objective. These geometrical assumptions come in the form of *global* growth and flatness of the objective which is very restrictive. Rather, here, our geometrical assumptions on f will be only local. The corresponding stochastic algorithms for these two choices of γ, whose mathematical formulation and analysis are simpler, have been the subject of active research work; see *e.g.* [54, 35, 45, 4, 2, 83, 37, 55, 47, 48, 31, 33, 40, 14].

Time scaling and averaging. The authors in [6] proposed time scaling and averaging to link (GF) and (ISIHD) with a general viscous damping function γ and a properly adjusted geometric damping function β (related to γ). Our aim is to extend the results of [6] to the stochastic case. Leveraging these techniques with a general function γ and an appropriate β, we will be able to transfer all the results we obtained in [63] for a first-order SDI to the second-order one ($\mathrm{S-ISIHD_{NS}}$). This avoids in particular to go through an intricate and a dedicated Lyapunov analysis for ($\mathrm{S-ISIHD_{NS}}$). A local convergence analysis becomes also

easily accessible through this lens while it is barely possible otherwise. We also specialize our results to the standard case where $\gamma(t) = \frac{\alpha}{t}$ and $\beta(t) = \frac{t}{\alpha-1}$.

The idea of passing from a first-order system to a second-order one via time scaling is not new. In the smooth case ($g \equiv 0$), the author of [21] propose time scaling and a tricky change of variables to show that (IGS$_\gamma$) is equivalent to an averaged gradient system, *i.e.* the steepest gradient system (GF) where the instantaneous value of $\nabla f(x(t))$ is replaced by some average of the gradients $\nabla f(x(s))$ over all past positions $s \leq t$. See also [39] for more general gradient systems with memory terms involving kernels. This gives rise to an integro-differential equation whose asymptotic behavior and the equivalent second-order dynamic have been investigated in [21]. A stochastic version of this integro-differential equation has been studied in [36] where the long time behavior of the resulting process, in particular its invariant distribution and occupation measure, was investigated under ellipticity assumptions on f and σ and proper behavior of the averaging gradient function. Clearly, the motivation of that work is not on the minimizing properties of the process while it is our focus here. We remark that after simple integration on (S − ISIHD), we can get that

$$\begin{cases} dX(t) &= V(t)dt \\ V(t) &= \frac{V_0}{p(t)} - \frac{1}{p(t)} \int_{t_0}^t p(s)\nabla f(X(s) + \beta(s)V(s))ds + \frac{1}{p(t)} \int_{t_0}^t p(s)\sigma(s, X(s) + \beta(s)V(s))dW(s), \end{cases}$$

where $p(t) = \exp\left(\int_{t_0}^t \gamma(s)ds\right)$. This representation highlights the link with integro-differential equations.

1.5 Organization of the paper

Section 2 introduces notations, definitions and preliminaries that are essential to our exposition. Section 3 is the main part of our study. We develop the passage from the first-order system to the second-order inertial system by using the time scaling and averaging in a stochastic framework. Almost sure and ergodic convergence rates are provided under different geometric properties of the objective function, such as convexity and Polyak-Łojasiewicz geometry. Finally, we show a strong convergence result when adding a Tikhonov regularization. Additional technical results that are needed throughout the paper are gathered in Appendix B.

2 Notation and Preliminaries

2.1 Notation

We will use the following shorthand notations: Given $n \in \mathbb{N}$, $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$. Consider $\mathbb{H}, \mathbb{K}$ real separable Hilbert spaces endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and $\langle \cdot, \cdot \rangle_{\mathbb{K}}$, respectively, and norm $\|\cdot\|_{\mathbb{H}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathbb{H}}}$ and $\|\cdot\|_{\mathbb{K}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathbb{K}}}$, respectively (we will omit the subscripts $\mathbb{H}$ and $\mathbb{K}$ whenever it is clear from the context). $I_{\mathbb{H}}$ is the identity operator from $\mathbb{H}$ to $\mathbb{H}$. $\mathcal{L}(\mathbb{K}; \mathbb{H})$ is the space of bounded linear operators from $\mathbb{K}$ to $\mathbb{H}$, $\mathcal{L}_1(\mathbb{K})$ is the space of trace-class operators, and $\mathcal{L}_2(\mathbb{K}; \mathbb{H})$ is the space of bounded linear Hilbert-Schmidt operators from $\mathbb{K}$ to $\mathbb{H}$. For $M \in \mathcal{L}_1(\mathbb{K})$, the trace is defined by

$$\operatorname{tr}(M) \stackrel{\text{def}}{=} \sum_{i \in I} \langle M e_i, e_i \rangle < +\infty,$$

where $I \subseteq \mathbb{N}$ and $(e_i)_{i \in I}$ is an orthonormal basis of $\mathbb{K}$. Besides, for $M \in \mathcal{L}(\mathbb{K}; \mathbb{H})$, $M^\star \in \mathcal{L}(\mathbb{H}; \mathbb{K})$ is the adjoint operator of M, and for $M \in \mathcal{L}_2(\mathbb{K}; \mathbb{H})$,

$$\|M\|_{\text{HS}} \stackrel{\text{def}}{=} \sqrt{\operatorname{tr}(MM^\star)} < +\infty$$

is its Hilbert-Schmidt norm (in the finite-dimensional case is equivalent to the Frobenius norm). We denote by w-lim (resp. s-lim) the limit for the weak (resp. strong) topology of $\mathbb{H}$. The notation $A : \mathbb{H} \rightrightarrows \mathbb{H}$ means that A is a set-valued operator from $\mathbb{H}$ to $\mathbb{H}$. For $f : \mathbb{H} \to \mathbb{R}$, the sublevel of f at height $r \in \mathbb{R}$ is denoted $[f \leq r] \stackrel{\text{def}}{=} \{x \in \mathbb{H} : f(x) \leq r\}$. 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 $p = +\infty$. On the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathrm{L}^p(\Omega; \mathbb{H})$ denotes the (Bochner) space of $\mathbb{H}$-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.

Let us recall some important definitions and results from convex analysis; for a comprehensive coverage, we refer the reader to [75].

We denote by $\Gamma_0(\mathbb{H})$ the class of proper lsc and convex functions on $\mathbb{H}$ taking values in $\mathbb{R} \cup \{+\infty\}$. For $\mu > 0$, $\Gamma_\mu(\mathbb{H}) \subset \Gamma_0(\mathbb{H})$ is the class of μ-strongly convex functions, *i.e.*, functions f such that $f - \frac{\mu}{2}\|\cdot\|^2$ is convex. We denote by $C^s(\mathbb{H})$ the class of s-times continuously differentiable functions on $\mathbb{H}$. For $L \geq 0$, $C_L^{1,1}(\mathbb{H}) \subset C^1(\mathbb{H})$ is the set of functions on $\mathbb{H}$ whose gradient is L-Lipschitz continuous, and $C_L^2(\mathbb{H})$ is the subset of $C_L^{1,1}(\mathbb{H})$ whose functions are twice differentiable.

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

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

which is non-empty for every point in the relative interior of the domain of f. When f is finite-valued, then f is continuous, and $\partial f(x)$ is a non-empty convex and compact set for every $x \in \mathbb{H}$. If f is differentiable, then $\partial f(x) = \{\nabla f(x)\}$. For every $x \in \mathbb{H}$ such that $\partial f(x) \neq \emptyset$, the minimum norm selection of $\partial f(x)$ is the unique element $\{\partial^0 f(x)\} \stackrel{\text{def}}{=} \operatorname{argmin}_{u \in \partial f(x)} \|u\|$. The projection of a point $x \in \mathbb{H}$ onto a non-empty closed convex set $C \subseteq \mathbb{H}$ is denoted by $\mathrm{P}_C(x)$.

2.2 Assumptions on volatility and damping parameters

Recall that our focus in this paper is on an optimization perspective, and as we argued in the introduction, we will study the long time behavior of our SDEs and SDIs (in particular ($\mathrm{S-ISIHD}$) and ($\mathrm{S-ISIHD_{NS}}$)) as the diffusion term vanishes when $t \to +\infty$. Therefore, throughout the paper, we assume that the diffusion (volatility) term σ satisfies:

$$\begin{cases} \sup_{t \geq t_0, x \in \mathbb{H}} \|\sigma(t,x)\|_{\mathrm{HS}} < +\infty, \\ \|\sigma(t,x') - \sigma(t,x)\|_{\mathrm{HS}} \leq l_0 \|x' - x\|, \end{cases} \tag{H_σ}$$

for some $l_0 > 0$ and for all $t \geq t_0, x, x' \in \mathbb{H}$. The Lipschitz continuity assumption is mild and classical and will be required to ensure the well-posedness of ($\mathrm{S-ISIHD}$) and ($\mathrm{S-ISIHD_{NS}}$). Let us also define $\sigma_\infty : [t_0, +\infty[\to \mathbb{R}_+$ as

$$\sigma_\infty(t) \stackrel{\text{def}}{=} \sup_{x \in \mathbb{H}} \|\sigma(t,x)\|_{\mathrm{HS}}.$$

Remark 2.1. (H_σ) implies the existence of $\sigma_* > 0$ such that:

$$\|\sigma(t,x)\|_{\mathrm{HS}}^2 = \mathrm{tr}[\Sigma(t,x)] \leq \sigma_*^2,$$

for all $t \geq t_0, x \in \mathbb{H}$, where $\Sigma \stackrel{\text{def}}{=} \sigma\sigma^\star$.

For $t_0 > 0$, let $\gamma : [t_0, +\infty[\to \mathbb{R}_+$ be a viscous damping and define:

$$p(t) \stackrel{\text{def}}{=} \exp\left(\int_{t_0}^{t} \gamma(s)ds\right), \quad \Gamma(t) \stackrel{\text{def}}{=} p(t)\int_t^{\infty} \frac{ds}{p(s)} \text{ and } I[h](t) \stackrel{\text{def}}{=} \exp\left(-\int_{t_0}^{t} \frac{du}{\Gamma(u)}\right)\int_{t_0}^{t} h(u)\frac{\exp\left(\int_{t_0}^{u} \frac{ds}{\Gamma(s)}\right)}{\Gamma(u)}du.$$

We assume that

$$\begin{cases} \gamma \text{ is upper bounded by a non-increasing function for every } t \geq t_0; \\ \int_{t_0}^{\infty} \frac{ds}{p(s)} < +\infty. \end{cases} \tag{H_γ}$$

Remark 2.2. Let us notice that Γ satisfies the relation $\Gamma' = \gamma\Gamma - 1$.

2.3 Reminder of main previous results

Before we delve into our core contributions, it is important to note that we will require some specific results gleaned from [62, 63]. These are the subject of the following paragraphs.

Results on (SGF). Since we are going to show results in the smooth case, we rewrite (H_0) when $g \equiv 0$,

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

Theorem 2.3 ([62, Theorem 3.1]). *Consider f and σ that satisfy Assumptions (H_0') and (H_σ). Let $\nu \geq 2$ and consider the SDE:*

$$\begin{cases} dX(t) = -\nabla f(X(t))dt + \sigma(t, X(t))dW(t), \\ X(t_0) = X_0, \end{cases} \tag{2.1}$$

where $X_0 \in \mathrm{L}^\nu(\Omega; \mathbb{H})$. Then, there exists a unique solution $X \in S^\nu_{\mathbb{H}}[t_0]$ (see Section B.2 for the notation) of (2.1). Additionally, if $\sigma_\infty \in \mathrm{L}^2([t_0, +\infty[)$, then:

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

(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) There exists an $\mathcal{S}$-valued random variable $X^\star$ such that $\text{w-}\lim_{t\to+\infty} X(t) = X^\star$ *a.s..*

Theorem 2.4 ([62, Theorem 3.4]). *Let $\nu \geq 2$ and consider the SDE (2.1) with initial data $X_0 \in \mathrm{L}^\nu(\Omega; \mathbb{H})$, where f and σ satisfy Assumptions (H_0) and (H_σ). Moreover, we assume that σ satisfies $t \mapsto t\sigma_\infty^2(t) \in \mathrm{L}^1([t_0, +\infty[)$ and that either $\mathbb{H}$ is finite-dimensional or $f \in C^2(\mathbb{H})$. Then, the solution trajectory $X \in S^\nu_{\mathbb{H}}[t_0]$ is unique and we have that:*

(i) $\mathbb{E}[f(X(t)) - \min f] = \mathcal{O}(t^{-1})$.

Moreover, if $f \in C^2(\mathbb{H})$, then the following hold:

(ii) $t \mapsto t\|\nabla f(X(t))\|^2 \in \mathrm{L}^1([t_0, +\infty[)$ *a.s..*

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

Results on (SGI) The far more intricate non-smooth version of (SGF) has been recently studied in [63]. Let F, σ satisfy (H_0) and (H_σ). We consider the stochastic differential inclusion

$$\begin{cases} dX(t) \in -\partial F(X(t))dt + \sigma(t, X(t))dW(t), & t > t_0, \\ X(t_0) = X_0. \end{cases} \tag{SGI}$$

The following definition makes precise the notion of solution that we are interested in.

Definition 2.5. A solution of (SGI) is a couple (X, η) of $\mathcal{F}_t$-adapted processes such that almost surely:

(i) X is continuous with sample paths in the domain of ∂g.

(ii) $\eta : [t_0, +\infty[\to \mathbb{H}$ is absolutely continuous, such that $\eta(t_0) = 0$, and $\forall T > t_0$, $\eta' \in \mathrm{L}^2([t_0, T]; \mathbb{H})$, $\eta'(t) \in \partial g(X(t))$ for almost all $t \geq t_0$;

(iii) For $t > t_0$,

$$\begin{cases} X(t) & = X_0 - \int_{t_0}^t \nabla f(X(s))ds - \eta(t) + \int_{t_0}^t \sigma(s, X(s))dW(s), \\ X(t_0) & = X_0. \end{cases} \tag{2.2}$$

For brevity, we sometimes omit the process η and say that X is a solution of (SGI), meaning that, there exists a process η such that (X, η) satisfies the previous definition. The definition of uniqueness for the process X is given in Section B.2.

In order to show the main results for (SGI), we consider the sequence of solutions $(X_\lambda)_{\lambda>0}$ of the SDEs

$$\begin{cases} dX_\lambda(t) = -\nabla(f + g_\lambda)(X_\lambda(t))dt + \sigma(t, X_\lambda(t))dW(t), & t > t_0, \\ X_\lambda(t_0) = X_0, \end{cases} \tag{SDE_λ}$$

where g_λ is the Moreau envelope of g with parameter $\lambda > 0$. Observe that system (SDE_λ) was also directly studied in [62] to solve (P), where a complete discussion is provided. It is worth emphasizing that the Moreau envelope of g, and the corresponding system (SDE_λ) is only needed as a means to establish existence and uniqueness of solution of (SGI). This follows the same reasoning as in the deterministic case where the Moreau-Yosida regularization and nonlinear semigroup theory have proven very useful to show existence and uniqueness for differential inclusions [20]. However, in the convergence analysis, we never replace g by its Moreau envelope.

We define the integrability condition that for every $T > t_0$,

$$\limsup_{\lambda \downarrow 0} \int_{t_0}^T \mathbb{E}(\|\nabla g_\lambda(X_\lambda(t))\|^2)dt < +\infty, \tag{H_λ}$$

Remark 2.6. Some conditions on g for (H_λ) to be satisfied are discussed in [72]. Of particular interest, which covers a broad class of functions, if when ∂g has full domain and there exists $C_0 > 0$ such that:

$$\left\|\partial^0 g(x)\right\| \leq C_0(1 + \|x\|), \quad \forall x \in \mathbb{H}.$$

This is for instance the case when g is Lipschitz continuous.

The following results on (SGI) were proved in [63].

Theorem 2.7 ([63]). *Consider $F = f + g$ and σ satisfying (H_0) and (H_σ). Suppose further that g verifies (H_λ). Let $\nu \geq 2$, $t_0 \geq 0$, and consider the dynamic (SGI) with initial data $X_0 \in \mathrm{L}^\nu(\Omega; \mathbb{H})$. Then, (SGI) has a unique solution in the sense of Definition 2.5 $(X, \eta) \in S^\nu_{\mathbb{H}}[t_0] \times C^1([t_0, +\infty[; \mathbb{H})$.*

Moreover, if $\sigma_\infty \in \mathrm{L}^2([t_0, +\infty[)$, then the following holds:

(i) $\mathbb{E}[\sup_{t\geq t_0}\|X(t)\|^\nu] < +\infty$.

(ii) $\forall x^\star \in \mathcal{S}_F$, $\lim_{t\to+\infty}\|X(t)-x^\star\|$ *exists a.s. and* $\sup_{t\geq t_0}\|X(t)\| < +\infty$ *a.s.*.

(iii) If g is continuous , then ∇f is constant on $\mathcal{S}_F$, there exists $x^\star \in \mathcal{S}_F$ such that $\text{s-lim}_{t\to\infty}\nabla f(X(t)) = \nabla f(x^\star)$ *a.s., and*

$$\int_{t_0}^{+\infty} F(X(t)) - \min F\, dt < +\infty \quad a.s..$$

(iv) There exists an $\mathcal{S}_F$-valued random variable $X^\star$ such that $\text{w-lim}_{t\to+\infty} X(t) = X^\star$.

Tikhonov regularization. Let us now turn to a Tikhonov regularization of (SGI), *i.e.*,

$$\begin{cases} dX(t) \in -\partial F(X(t)) - \varepsilon(t)X(t) + \sigma(t, X(t))dW(t), \quad t \geq t_0, \\ X(t_0) = X_0. \end{cases} \tag{SGI – TA}$$

Solution existence and uniqueness for (SGI – TA) is established in [63, Theorem 3.3]. We also have the following.

Theorem 2.8 ([63, Theorem 4.1]). *Let $\nu \geq 2$ and consider the dynamic (SGI – TA) with initial data $X_0 \in \mathrm{L}^\nu(\Omega;\mathbb{H})$, where $F = f + g$ and σ satisfy Assumptions (H_0) and (H_σ). Furthermore, assume that g satisfies (H_λ). Then, there exists a unique solution $X \in S^\nu_{\mathbb{H}}[t_0]$ of (SGI – TA). Let $x^\star = \mathrm{P}_{\mathcal{S}_F}(0)$ be the minimum norm solution, and for $\varepsilon > 0$ let x_ε be the unique minimizer of $F_\varepsilon(x) \stackrel{\text{def}}{=} F(x) + \frac{\varepsilon}{2}\|x\|^2$. Suppose that $\sigma_\infty \in \mathrm{L}^2([t_0,+\infty[)$, and that $\varepsilon : [t_0,+\infty[\to \mathbb{R}_+$ satisfies the conditions:*

(T_1) $\varepsilon(t) \to 0$ *as* $t \to +\infty$;

(T_2) $\displaystyle\int_{t_0}^{+\infty} \varepsilon(t)dt = +\infty$;

(T_3) $\displaystyle\int_{t_0}^{+\infty} \varepsilon(t)\left(\|x^\star\|^2 - \|x_{\varepsilon(t)}\|^2\right) dt < +\infty$.

Then, we have

(i) $\sup_{t\geq t_0}\|X(t)\| < +\infty$ *a.s., and*

(ii) $\text{s-lim}_{t\to+\infty} X(t) = x^\star$ *a.s.*.

This means that we can obtain almost sure strong convergence of the trajectory to a particular (non-random) solution: the one of minimal norm.

3 From first-order to second-order systems

3.1 Time scaling and averaging

We apply a time scaling and then an averaging technique to the system (SGI) to derive an insightful reparametrization of a particular case of our second-order system ($\mathrm{S-ISIHD_{NS}}$), specifically, the case when $\beta \equiv \Gamma$. The main advantage of this method is that the results of (SGI) directly carry over to obtain results on the convergence behavior of ($\mathrm{S-ISIHD_{NS}}$) without passing through a dedicated Lyapunov analysis. However, as discussed in the introduction, the averaging technique is restricted to convex objectives, as it heavily relies on Jensen's inequality.

Let $\nu \geq 2$, $s_0 > 0$. We consider the potential $F = f + g$ where g satisfies (H_λ). Let σ_1 be a diffusion term in the time parametrization by s. We will study the dynamic (SGI) in s, starting at s_0, with diffusion term σ_1 under hypotheses (H_0) and (H_σ). Let $\sigma_{1_*} > 0$ be such that

$$\|\sigma_1(s,x)\|_{\mathrm{HS}} \leq \sigma_{1_*}^2, \quad \forall s \geq s_0, \forall x \in \mathbb{H},$$

and $\sigma_{1_\infty}(s) \stackrel{\text{def}}{=} \sup_{x\in\mathbb{H}} \|\sigma_1(s,x)\|_{\mathrm{HS}}$. We rewrite (SGI) in a format adapted to our case,

$$\begin{cases} dZ(s) \in -\partial F(Z(s))ds + \sigma_1(s, Z(s))dW(s), & s > s_0, \\ Z(s_0) = Z_0, \end{cases} \tag{3.1}$$

where $Z_0 \in \mathrm{L}^\nu([s_0, +\infty[; \mathbb{H})$.

Let us make the change of time $s = \tau(t)$ in the dynamic (3.1), where τ is an increasing function from $[t_0, +\infty[$ to $[s_0, +\infty[$, which is twice differentiable, and which satisfies $\lim_{t\to+\infty} \tau(t) = +\infty$. Denote $Y(t) \stackrel{\text{def}}{=} Z(s)$ and t_0 be such that $s_0 = \tau(t_0)$. By the chain rule and [68, Theorem 8.5.7], we have

$$\begin{cases} dY(t) \in -\tau'(t)\partial F(Y(t))dt + \sqrt{\tau'(t)}\sigma_1(\tau(t), Y(t))dW(t), & t > t_0, \\ Y(t_0) = Z_0. \end{cases} \tag{3.2}$$

Consider the smooth case, i.e. when $g \equiv 0$ and the hypotheses of Theorem 2.4 ($f \in C_L^2(\mathbb{H})$ and $\sigma_1 \in \mathrm{L}^2([s_0, +\infty[)$), then we can conclude that the convergence rate of (3.2) (when $g \equiv 0$) is the following

$$f(Y(t)) - \min f = o\left(\frac{1}{\tau(t)}\right) \text{ a.s..} \tag{3.3}$$

By introducing a function τ that grows faster than the identity ($\tau(t) \geq t$), we have accelerated the dynamic, passing from the asymptotic convergence rate $1/s$ for (3.1) to $1/\tau(t)$ for (3.2). The price to pay is that the drift term in (3.2) is non-autonomous, furthermore, when the coefficient in front of the gradient tends to infinity as $t \to +\infty$, it will preclude the use of an explicit discretization in time. To overcome this, we adapt from [6] the following approach, which is called averaging.

Consider (3.2) and define the two stochastic processes $X, V : \Omega \times [t_0, +\infty[\to \mathbb{H}$ as

$$\begin{cases} dX(t) & = V(t)dt, \quad t > t_0, \\ Y(t) & = X(t) + \tau'(t)V(t), \quad t > t_0, \\ X(t_0) & = X_0, \quad V(t_0) = V_0, \end{cases} \tag{3.4}$$

where $Y(t)$ is the process in (3.2), and $X_0, V_0 \in \mathrm{L}^\nu(\Omega; \mathbb{H})$ are initial data. This leads us to set $Z_0 \stackrel{\text{def}}{=} X_0 + \tau'(t_0)V_0$ in order for the equations to fit. According to the averaging, the differential form of $Y(t)$ is

$$dY(t) = dX(t) + \tau''(t)V(t)dt + \tau'(t)dV(t).$$

Combining the previous equation with (3.2), we have that

$$-\tau'(t)\partial F(Y(t))dt + \sqrt{\tau'(t)}\sigma_1(\tau(t), Y(t))dW(t) \ni dX(t) + \tau''(t)V(t)dt + \tau'(t)dV(t).$$

Using that $dX(t) = V(t)dt$ and dividing by τ', we then have

$$-\partial F(X(t) + \tau'(t)V(t))dt + \frac{1}{\sqrt{\tau'(t)}}\sigma_1(\tau(t), X(t) + \tau'(t)V(t))dW(t) \ni \frac{1 + \tau''(t)}{\tau'(t)}V(t)dt + dV(t).$$

Therefore, after the time scaling and averaging, we obtain the following dynamic:

$$\begin{cases} dX(t) &= V(t)dt, \quad t > t_0, \\ dV(t) &\in -\frac{1+\tau''(t)}{\tau'(t)}V(t)dt - \partial F(X(t)+\tau'(t)V(t))dt \\ & \quad + \frac{1}{\sqrt{\tau'(t)}}\sigma_1(\tau(t), X(t)+\tau'(t)V(t))dW(t), \quad t > t_0, \\ X(t_0) &= X_0, \quad V(t_0) = V_0. \end{cases} \tag{ISIHD-S.1}$$

Let $\gamma : [t_0, +\infty[\to \mathbb{R}_+$ satisfying (H_γ). We are going to determine τ in order to obtain a viscous damping coefficient equal to γ, i.e.,

$$\frac{1+\tau''(t)}{\tau'(t)} = \gamma(t).$$

Clearly, τ' solves the following ODE in ζ

$$\zeta' = \gamma\zeta - 1.$$

As observed in Remark 2.2, the function Γ also satisfies the same ODE, and thus we can adjust the initial condition of τ' to obtain

$$\tau'(t) = \Gamma(t) = p(t)\int_t^\infty \frac{du}{p(u)} \quad \forall t \geq t_0.$$

We then integrate and take $\tau(t) = s_0 + \int_{t_0}^t \Gamma(u)du$ to get $\tau(t_0) = s_0$ as required. This is a valid selection of τ since $t \mapsto s_0 + \int_{t_0}^t \Gamma(u)du$ is an increasing function from $[t_0, +\infty[$ to $[s_0, +\infty[$, twice differentiable and $\Gamma \notin \mathrm{L}^1([t_0, +\infty[)$ because Γ is lower bounded by a non-decreasing function since γ is upper bounded by a non-increasing function (see [7, Proposition 2.2]) by (H_γ). For this particular selection of τ, and defining $\tilde{\sigma}_1(t,\cdot) \stackrel{\text{def}}{=} \frac{\sigma_1(\tau(t),\cdot)}{\sqrt{\Gamma(t)}}$, we have that (ISIHD-S.1) is equivalent to

$$\begin{cases} dX(t) &= V(t)dt, \quad t > t_0, \\ dV(t) &\in -\gamma(t)V(t)dt - \partial F(X(t)+\Gamma(t)V(t))dt + \tilde{\sigma}_1(t, X(t)+\Gamma(t)V(t))dW(t), \quad t > t_0, \\ X(t_0) &= X_0, \quad V(t_0) = V_0. \end{cases} \tag{ISIHD-S.2}$$

Clearly, (ISIHD-S.2) is nothing but ($\mathrm{S-ISIHD_{NS}}$) when $\beta \equiv \Gamma$ and $\sigma \equiv \tilde{\sigma}_1$.

In order to be able to transfer the convergence results on Z in (3.1) (via (3.2)) to X in (ISIHD-S.2), it remains to express X in terms of Y only. For this, let

$$a(t) \stackrel{\text{def}}{=} \frac{1}{\tau'(t)}, \quad A(t) \stackrel{\text{def}}{=} \int_{t_0}^t a(u)du.$$

Recalling the averaging in (3.4), we need to integrate the following equation

$$V(t) + a(t)X(t) = a(t)Y(t). \tag{3.5}$$

Multiplying both sides by $e^{A(t)}$ and using (3.4), we get equivalently

$$d\left(e^{A(t)}X(t)\right) = a(t)e^{A(t)}Y(t)dt. \tag{3.6}$$

Integrating and using again (3.4), we obtain

$$\begin{aligned} X(t) &= e^{-A(t)}X(t_0) + e^{-A(t)}\int_{t_0}^{t} a(u)e^{A(u)}Y(u)du \\ &= e^{-A(t)}Y(t_0) + e^{-A(t)}\int_{t_0}^{t} a(u)e^{A(u)}Y(u)du - e^{-A(t)}\tau'(t_0)V(t_0). \end{aligned}$$

Then we can write

$$X(t) = \int_{t_0}^{t} Y(u)d\mu_t(u) + \xi(t), \tag{3.7}$$

where μ_t is the probability measure on $[t_0, t]$ defined by

$$\mu_t \stackrel{\text{def}}{=} e^{-A(t)}\delta_{t_0} + a(u)e^{A(u)-A(t)}du, \tag{3.8}$$

where δ_{t_0} is the Dirac measure at t_0, $a(u)e^{A(u)-A(t)}du$ is the measure with density $a(\cdot)e^{A(\cdot)-A(t)}$ with respect to the Lebesgue measure on $[t_0, t]$, and $\xi(t)$ is a random process since V_0 is a random variable, i.e.,

$$\xi(t) \stackrel{\text{def}}{=} \xi(\omega, t) = -e^{-A(t)}\tau'(t_0)V_0(\omega) \quad \forall \omega \in \Omega. \tag{3.9}$$

3.2 Convergence of the trajectory and convergence rates under general γ, and $\beta \equiv \Gamma$

We here state the main results of this section. We first show almost sure convergence of the trajectory of ($\mathrm{S-ISIHD_{NS}}$) to a random variable taking values in the set of minimizers of F. When $g \equiv 0$, we also provide convergence rates.

The following result imposes an integrability condition on the diffusion term, which is essential for ensuring the almost sure weak convergence of the trajectory. More precisely, this integrability condition allows to show that the trajectory asymptotically converges almost surely, in the weak topology, to a random variable that takes values in the set of minimizers of F.

Theorem 3.1. *Let $\nu \geq 2$ and consider the dynamic ($\mathrm{S-ISIHD_{NS}}$) with initial data $X_0, V_0 \in \mathrm{L}^\nu(\Omega; \mathbb{H})$, where $\gamma : [t_0, +\infty[\to \mathbb{R}_+$ satisfies (H_γ), and $\beta \equiv \Gamma$. Besides, $F = f + g$ and σ satisfy Assumptions (H_0) and (H_σ). Moreover, suppose that g satisfies (H_λ). Then, there exists a unique solution $(X, V) \in S^\nu_{\mathbb{H}\times\mathbb{H}}[t_0]$ of ($\mathrm{S-ISIHD_{NS}}$). Additionally, if $\Gamma\sigma_\infty \in \mathrm{L}^2([t_0, +\infty[)$, then there exists an $\mathcal{S}_F$-valued random variable $X^\star$ such that $\text{w-lim}_{t\to\infty} X(t) = X^\star$ a.s. and $\text{w-lim}_{t\to\infty} \Gamma(t)V(t) = 0$. a.s..*

Proof. Let $\theta(t) \stackrel{\text{def}}{=} \int_{t_0}^{t} \Gamma(u)du$, $\tilde{\sigma}(s, \cdot) \stackrel{\text{def}}{=} \sigma(\theta^{-1}(s), \cdot)\sqrt{\Gamma(\theta^{-1}(s))}$, and $\tilde{\sigma}_\infty(s) \stackrel{\text{def}}{=} \sup_{x\in\mathbb{H}} \|\tilde{\sigma}(s, x)\|_{\mathrm{HS}}$. Then $\Gamma\sigma_\infty \in \mathrm{L}^2([t_0, +\infty[)$ is equivalent to $\tilde{\sigma}_\infty \in \mathrm{L}^2(\mathbb{R}_+)$. Consider the dynamic:

$$\begin{cases} dZ(s) & \in -\partial F(Z(s)) + \tilde{\sigma}(s, Z(s))dW(s), \quad s > 0, \\ Z(0) & = X_0 + \Gamma(t_0)V_0. \end{cases} \tag{3.10}$$

By Theorem 2.7, we have that there exists a unique solution $(Z, \eta) \in S^\nu_{\mathbb{H}} \times C^1(\mathbb{R}_+; \mathbb{H})$ of (3.10), and an $\mathcal{S}_F$-valued random variable $X^\star$ such that $\text{w-lim}_{s\to\infty} Z(s) = X^\star$ a.s.. Moreover, using the time scaling $\tau \equiv \theta$ and the averaging described in this section, we end up with the dynamic ($\mathrm{S-ISIHD_{NS}}$) in the case where $\beta \equiv \Gamma$.

It is direct to check that the time scaling and averaging preserves the uniqueness of a solution $(X,V) \in S^0_{\mathbb{H}\times\mathbb{H}}[t_0]$. Now let us validate $(X,V) \in S^\nu_{\mathbb{H}\times\mathbb{H}}[t_0]$. Since

$$\mathbb{E}\left(\sup_{s\in[0,T]} \|Z(s)\|^\nu\right) < +\infty, \quad \forall T > 0,$$

we directly obtain

$$\mathbb{E}\left(\sup_{t\in[t_0,T]} \|Y(t)\|^\nu\right) < +\infty, \quad \forall T > t_0.$$

Thanks to the relation (3.7), the following holds

$$\begin{aligned}\|X(t)\|^\nu &\le \nu\left(\left\|X(t) - \int_{t_0}^t Y(u)d\mu_t(u)\right\|^\nu + \left\|\int_{t_0}^t Y(u)d\mu_t(u)\right\|^\nu\right)\\ &\le \nu\left(\|\xi(t)\|^\nu + (t-t_0)^{\nu-1}\int_{t_0}^t \|Y(u)\|^\nu d\mu_t(u)\right).\end{aligned}$$

Let $T > t_0$ be arbitrary. Taking supremum over $[t_0,T]$ and then expectation at both sides, we obtain that

$$\mathbb{E}\left(\sup_{t\in[t_0,T]} \|X(t)\|^\nu\right) \le \nu\left(\mathbb{E}(\|V_0\|^\nu)\|\Gamma(t_0)\|^\nu + (T-t_0)^{\nu-1}\mathbb{E}\left(\sup_{t\in[t_0,T]} \|Y(t)\|^\nu\right)\right) < +\infty.$$

Since $V(t) = \frac{Y(t)-X(t)}{\Gamma(t)}$, we have

$$\|V(t)\|^\nu \le \frac{\nu}{\Gamma^\nu(t)}(\|Y(t)\|^\nu + \|X(t)\|^\nu).$$

Similarly as before, we let $T > t_0$ be arbitrary, and take the supremum over $[t_0,T]$ and then expectation at both sides to obtain

$$\mathbb{E}\left(\sup_{t\in[t_0,T]} \|V(t)\|^\nu\right) \le \nu \sup_{t\in[t_0,T]} \frac{1}{\Gamma^\nu(t)}\left(\mathbb{E}\left(\sup_{t\in[t_0,T]} (\|Y(t)\|^\nu + \|X(t)\|^\nu)\right)\right).$$

Since Γ is a continuous positive function, by the extreme value theorem, we have that there exists $t_T \in [t_0,T]$ such that $\sup_{t\in[t_0,T]} \frac{1}{\Gamma^\nu(t)} = \frac{1}{\Gamma^\nu(t_T)} < +\infty$, and we conclude that $(X,V) \in S^\nu_{\mathbb{H}\times\mathbb{H}}[t_0]$.

Now we prove that there exists an $\mathcal{S}_F-$valued random variable $X^\star$ such that w-$\lim_{t\to\infty} X(t) = X^\star$ a.s.. By virtue of Theorem 2.7, there exists an $\mathcal{S}_F-$valued random variable $X^\star$ such that w-$\lim_{s\to\infty} Z(s) = X^\star$ a.s.. We also notice that we have directly w-$\lim_{t\to\infty} Y(t) = X^\star$ a.s.. Let $h \in \mathbb{H}$ be arbitrary and use the relation (3.7) as follows:

$$\begin{aligned}|\langle X(t) - X^\star, h\rangle| &\le \left|\left\langle X(t) - \int_{t_0}^t Y(u)d\mu_t(u), h\right\rangle\right| + \left|\left\langle \int_{t_0}^t Y(u)d\mu_t(u) - X^\star, h\right\rangle\right|\\ &= |\langle \xi(t), h\rangle| + \left|\left\langle \int_{t_0}^t (Y(u) - X^\star)d\mu_t(u), h\right\rangle\right|\\ &= |\langle \xi(t), h\rangle| + \left|\int_{t_0}^t \langle Y(u) - X^\star, h\rangle d\mu_t(u)\right|\\ &\le \|\xi(t)\|\|h\| + \int_{t_0}^t |\langle Y(u) - X^\star, h\rangle| d\mu_t(u),\end{aligned}$$

where the second equality comes from the dominated convergence theorem, since $\sup_{s>t_0}\|Y(s)\| < +\infty$ a.s. (by (ii) of Theorem 2.7).
Now let $a(t) = \frac{1}{\Gamma(t)}$ and $A(t) = \int_{t_0}^t \frac{du}{\Gamma(u)}$. By Lemma B.3 (defined in Section B.1) and the fact that $V_0 \in \mathrm{L}^\nu(\Omega;\mathbb{H})$, we have that $\lim_{t\to\infty}\|\xi(t)\| = 0$ a.s.. On the other hand, it holds that

$$\int_{t_0}^t |\langle Y(u) - X^\star, h\rangle| d\mu_t(u) \leq e^{-A(t)}|\langle Y(t_0) - X^\star, h\rangle| + e^{-A(t)}\int_{t_0}^t a(u)e^{A(u)}|\langle Y(u) - X^\star, h\rangle| du.$$

Now let $b(u) = |\langle Y(u) - X^\star, h\rangle|$. Since we already proved that $\lim_{u\to\infty} b(u) = 0$ a.s., and we have that $a \notin \mathrm{L}^1([t_0, +\infty[)$ by Lemma B.3, we utilize Lemma B.2 (also defined in Section B.1) with our respective a, b functions. This let us conclude that

$$\lim_{t\to+\infty} |\langle X(t) - X^\star, h\rangle| = 0 \quad a.s..$$

Thus, $\text{w-}\lim_{t\to\infty} X(t) = X^\star$ a.s.. Finally, since

$$Y(t) = X(t) + \Gamma(t)V(t),$$

and the fact that X and Y have (a.s.) the same limit, we conclude that

$$\text{w-}\lim_{t\to\infty} \Gamma(t)V(t) = 0 \quad a.s..$$

□

In the smooth case, we also have convergence rates on the objective value and the gradient. In particular, the following two theorems will provide general abstract convergence rates under the same integrability condition on the diffusion term. These results will be specialized to specific choice of the parameters in Section 3.3.

Theorem 3.2. *Let $\nu \geq 2$ and consider the dynamic (S − ISIHD) with initial data $X_0, V_0 \in \mathrm{L}^\nu(\Omega;\mathbb{H})$, such that f and σ satisfy (H_0') and (H_σ), and in the case where γ satisfies (H_γ), $\beta \equiv \Gamma$. Moreover, suppose that either $\mathbb{H}$ is finite dimensional or $f \in C^2(\mathbb{H})$, and*

$$t \mapsto \sqrt{\theta(t)}\Gamma(t)\sigma_\infty(t) \in \mathrm{L}^2([t_0, +\infty[),$$

where $\theta(t) \stackrel{\text{def}}{=} \int_{t_0}^t \Gamma(u)du$. Then the solution trajectory $(X, V) \in S^\nu_{\mathbb{H}\times\mathbb{H}}[t_0]$ is unique and satisfies:

$$\mathbb{E}[f(X(t)) - \min f] = \mathcal{O}\left(\max\left\{e^{-A(t)}, I\left[\frac{1}{\theta}\right](t)\right\}\right), \quad \forall t > t_0,$$

where $A(t) \stackrel{\text{def}}{=} \int_{t_0}^t \frac{du}{\Gamma(u)}$ and we recall that $I[\frac{1}{\theta}](t) = e^{-A(t)}\int_{t_0}^t \frac{1}{\theta(u)}\frac{e^{A(u)}}{\Gamma(u)}du$.

From hypothesis (H_γ) we have that $\lim_{t\to+\infty} e^{-A(t)} = 0$, and since $\Gamma \notin \mathrm{L}^1([t_0, +\infty[)$, we can use Lemma B.2 to check that $\lim_{t\to\infty} I\left[\frac{1}{\theta}\right](t) = 0$.

Remark 3.3. If $\mathbb{H}$ is finite-dimensional and g is a L_0-Lipschitz function, we could obtain convergence rates for the nonsmooth setting (P) by considering (S − ISIHD) with potential $F_\rho \stackrel{\text{def}}{=} f + g_\rho$, where g_ρ is the Moreau envelope of g with parameter $\rho > 0$. With the same hypotheses on the initial conditions, σ, γ, β as in Theorem 3.2 and proceeding as in [62, Section 5], we could obtain

$$\mathbb{E}[F(X_\rho(t)) - \min F] = \mathcal{O}\left(\max\left\{e^{-A(t)}, I\left[\frac{1}{\theta}\right](t)\right\}\right) + \frac{L_0^2}{2}\rho, \quad \forall t > t_0,$$

where $(X_\rho, V_\rho) \in S^\nu_{\mathbb{H}\times\mathbb{H}}[t_0]$ is the solution of (S − ISIHD) with potential F_ρ.

Proof. We will utilize the averaging technique used in Theorem 3.1 and Jensen's inequality. First, we have

$$\mathbb{E}(f(X(t)) - \min f) = \mathbb{E}\left(f(X(t)) - f\left(\int_{t_0}^{t} Y(u)d\mu_t(u)\right)\right) + \mathbb{E}\left(f\left(\int_{t_0}^{t} Y(u)d\mu_t(u)\right) - \min f\right).$$

Let us recall that $\mathbb{E}(\sup_{s\geq 0}\|Z(s)\|) < +\infty$, which implies that $\mathbb{E}(\sup_{t\geq t_0}\|X(t)\|) < +\infty$. We bound the first term using the convexity of f and Cauchy-Schwarz inequality, *i.e.*, that for every $x, y \in \mathbb{H}$,

$$f(x) - f(y) \leq \langle \nabla f(x), x - y\rangle \leq \|\nabla f(x)\|\|y - x\|,$$

we get that

$$\begin{aligned} f(X(t)) - f\left(\int_{t_0}^{t} Y(u)d\mu_t(u)\right) &\leq \|\nabla f(X(t))\|\|\xi(t)\| \\ &\leq \|\xi(t)\|(L\|X(t)\| + \|\nabla f(0)\|) \\ &\leq \|\xi(t)\|\left(L\sup_{t\geq t_0}\|X(t)\| + \|\nabla f(0)\|\right), \end{aligned}$$

and we conclude that

$$\mathbb{E}\left(f(X(t)) - f\left(\int_{t_0}^{t} Y(u)d\mu_t(u)\right)\right) = \mathcal{O}(e^{-A(t)}).$$

For the second term, we use Jensen's inequality to obtain

$$\begin{aligned} \mathbb{E}\left(f\left(\int_{t_0}^{t} Y(u)d\mu_t(u)\right) - \min f\right) &\leq \int_{t_0}^{t} \mathbb{E}[f(Y(u)) - \min f]d\mu_t(u) \\ &\leq e^{-A(t)}\mathbb{E}[f(Y(t_0)) - \min f] + e^{-A(t)}\int_{t_0}^{t} \frac{e^{A(u)}}{\Gamma(u)}\mathbb{E}(f(Y(u)) - \min f)du. \end{aligned}$$

Since $\sqrt{\theta}\Gamma\sigma_\infty \in \mathrm{L}^2([t_0, +\infty[)$ is equivalent to $s \mapsto s\tilde{\sigma}_\infty^2(s) \in \mathrm{L}^1(\mathbb{R}_+)$, by Theorem 2.4, we have that there exists $C > 0$ such that $\mathbb{E}(f(Z(s)) - \min f) \leq \frac{C}{s}$. Then, we have $\mathbb{E}(f(Y(t)) - \min f) \leq \frac{C}{\theta(t)}$. Hence, there exists $C_0 > 0$ such that

$$\mathbb{E}\left(f(X(t)) - \min f\right) \leq C_0 e^{-A(t)} + CI\left[\frac{1}{\theta}\right](t).$$

□

Theorem 3.4. *Let $\nu \geq 2$ and consider the dynamic (S − ISIHD) with initial data $X_0, V_0 \in \mathrm{L}^\nu(\Omega; \mathbb{H})$, such that f and σ satisfy (H_0') and (H_σ), and in the case where γ satisfies (H_γ), $\beta \equiv \Gamma$. Moreover, suppose that $f \in C^2(\mathbb{H})$ and*

$$t \mapsto \sqrt{\theta(t)}\Gamma(t)\sigma_\infty(t) \in \mathrm{L}^2([t_0, +\infty[),$$

where $\theta(t) \stackrel{\text{def}}{=} \int_{t_0}^{t}\Gamma(u)du$. Then the solution trajectory $(X, V) \in S^\nu_{\mathbb{H}\times\mathbb{H}}[t_0]$ is unique and satisfies

$$\int_{t_0}^{\infty} \theta(u)\Gamma(u)\|\nabla f(X(u) + \Gamma(u)V(u))\|^2 du < +\infty \quad a.s.. \tag{3.11}$$

Proof. Consider (3.10) and the technique used in Theorem 3.1. We have that $t \mapsto \theta(t)\Gamma^2(t)\sigma_\infty^2(t) \in \mathrm{L}^1([t_0,+\infty[)$ is equivalent to $s \mapsto s\tilde{\sigma}_\infty^2(s) \in \mathrm{L}^1(\mathbb{R}_+)$. Therefore, we can use Theorem 2.4 to state that

$$\int_0^\infty s\|\nabla f(Z(s))\|^2 ds < +\infty \quad a.s..$$

Using the time scaling $\tau \equiv \theta$ and making the change of variable $\theta(t) = s$ in the previous integral, we obtain

$$\int_{t_0}^\infty \theta(t)\Gamma(t)\|\nabla f(Y(t))\|^2 dt < +\infty \quad a.s..$$

Recalling that in the averaging we impose that $Y = X + \Gamma V$, we conclude. □

3.3 Fast convergence under $\alpha > 3, \gamma(t) = \frac{\alpha}{t}$ and $\beta(t) = \frac{t}{\alpha-1}$

In the following, we show fast convergence results in expectation and in almost sure sense. This result imposes an integrability condition on the diffusion term to ensure a convergence rate of $\frac{1}{t^2}$ for the function values. This is highly desirable, as it represents the fastest convergence rate we can expect to achieve for gradient-based second-order dynamics featuring inertia when applied to a general convex function. Besides, these results match those in [6] when there is no noise.

Corollary 3.5 (Case $\frac{\alpha}{t}$). *Let $\nu \geq 2, \alpha > 3$ and consider the dynamic (S − ISIHD) with initial data $X_0, V_0 \in \mathrm{L}^\nu(\Omega;\mathbb{H})$, in the case where $\gamma(t) = \frac{\alpha}{t}$ and $\beta(t) = \frac{t}{\alpha-1}$. Besides, consider that f and σ satisfy (H_0') and (H_σ). Moreover, let $f \in C^2(\mathbb{H})$ and $t \mapsto t^2\sigma_\infty(t) \in \mathrm{L}^2([t_0,+\infty[)$. Then the solution trajectory $(X,V) \in S^\nu_{\mathbb{H}\times\mathbb{H}}[t_0]$ is unique and satisfies:*

(i) $f(X(t)) - \min f = o(t^{-2})$ *a.s..*

(ii) $\mathbb{E}[f(X(t)) - \min f] = \mathcal{O}(t^{-2})$.

(iii)

$$\int_{t_0}^\infty t^3\left\|\nabla f\left(X(t) + \frac{t}{\alpha-1}V(t)\right)\right\|^2 dt < +\infty \quad a.s..$$

Proof. Consider (3.10) with $\Gamma(t) = \frac{t}{\alpha-1}$ and $\theta(t) = \frac{t^2-t_0^2}{2(\alpha-1)}$. Let $\tilde{\sigma}(s,\cdot) = \sigma(\theta^{-1}(s),\cdot)\sqrt{\Gamma(\theta^{-1}(s))}$. Notice that $t \mapsto t^2\sigma_\infty(t) \in \mathrm{L}^2([t_0,+\infty[)$ is equivalent to $s \mapsto s\tilde{\sigma}_\infty^2(s) \in \mathrm{L}^1(\mathbb{R}_+)$. We apply Theorem 2.4 to deduce that

$$f(Z(s)) - \min f = o(s^{-1}) \text{ a.s..}$$

Using the time scaling $\tau \equiv \theta$ and then the averaging technique as in the proof of Theorem 3.1, we have that

$$f(Y(t)) - \min f = o(t^{-2}) \text{ a.s..}$$

Moreover, it holds that

$$X(t) = \int_{t_0}^t Y(u)d\mu_t(u) + \xi(t).$$

(i) Now we prove the first point in the following way:

$$t^2(f(X(t)) - \min f) = t^2\left(f(X(t)) - f\left(\int_{t_0}^t Y(u)d\mu_t(u)\right)\right) + t^2\left(f\left(\int_{t_0}^t Y(u)d\mu_t(u)\right) - \min f\right).$$

Let us bound the first term using the convexity of f:

$$\begin{aligned} t^2\left(f(X(t)) - f\left(\int_{t_0}^t Y(u)d\mu_t(u)\right)\right) &\leq t^2\|\nabla f(X(t))\|\|\xi(t)\| \\ &\leq t^2\|\xi(t)\|(L\|X(t)\| + \|\nabla f(0)\|) \\ &\leq t^2\|\xi(t)\|\left(L\sup_{t\geq t_0}\|X(t)\| + \|\nabla f(0)\|\right). \end{aligned}$$

Let us recall that $\sup_{s\geq 0}\|Z(s)\| < +\infty$ a.s.. Due to the time scaling and averaging, it is direct to check that $\sup_{t\geq t_0}\|X(t)\| < +\infty$ a.s.. On the other hand, $\|\xi(t)\| = \mathcal{O}(t^{1-\alpha})$ a.s.. Therefore, we have

$$t^2\left(f(X(t)) - f\left(\int_{t_0}^t Y(u)d\mu_t(u)\right)\right) = \mathcal{O}(t^{3-\alpha}) \quad a.s.. \tag{3.12}$$

Now let us bound the second term using Jensen's inequality,

$$\begin{aligned} t^2\left(f\left(\int_{t_0}^t Y(u)d\mu_t(u)\right) - \min f\right) &\leq t^2\left(\int_{t_0}^t [f(Y(u)) - \min f]d\mu_t(u)\right) \\ &= \frac{t_0^{\alpha-1}}{t^{\alpha-3}}[f(Y(t_0)) - \min f] \\ &+ \frac{\alpha-1}{t^{\alpha-3}}\int_{t_0}^t u^{\alpha-4}(u^2(f(Y(u)) - \min f))du. \end{aligned}$$

In order to calculate the limit of this second term, let $a(t) = \frac{\alpha-1}{t}, b(u) = u^2(f(Y(u)) - \min f)$, by Lemma B.2 we have that

$$\lim_{t\to\infty}\frac{\alpha-1}{t^{\alpha-1}}\int_{t_0}^t u^{\alpha-2}b(u)du = 0 \quad a.s..$$

Since $\alpha > 3$, we also have that

$$\lim_{t\to\infty}\frac{\alpha-3}{t^{\alpha-3}}\int_{t_0}^t u^{\alpha-4}b(u)du = 0 \quad a.s.. \tag{3.13}$$

Therefore, we conclude that

$$\lim_{t\to\infty} t^2(f(X(t)) - \min f) = 0 \quad a.s..$$

(ii) By Theorem 3.2 in the case $\gamma(t) = \frac{\alpha}{t}$, we have that $e^{-A(t)} = t_0^{\alpha-1}t^{1-\alpha}$ and $\theta(t) = \frac{t^2-t_0^2}{2(\alpha-1)}$. On the other hand

$$I\left[\frac{1}{\theta}\right](t) = 2(\alpha-1)^2t^{1-\alpha}\int_{t_0}^t \frac{u^{\alpha-2}}{u^2-t_0^2} = \mathcal{O}(t^{1-\alpha} + t^{-2}).$$

Since $\alpha > 3$, we have that $\mathcal{O}(t^{1-\alpha})$ is also $\mathcal{O}(t^{-2})$, and we conclude that

$$\mathbb{E}(f(X(t)) - \min f) = \mathcal{O}(t^{-2}).$$

(iii) This point follows directly from Theorem 3.4 in the case $\gamma(t) = \frac{\alpha}{t}$.

□

3.4 Convergence rate under Polyak-Łojasiewicz inequality

In this subsection, we show a local convergence rate under Polyak-Łojasiewicz inequality. The Polyak-Łojasiewicz property is a special case of the Łojasiewicz property (see [56, 57, 58]) and is commonly used to prove linear convergence of gradient descent algorithms.

Definition 3.6 (Polyak-Łojasiewicz inequality). Let $f : \mathbb{H} \to \mathbb{R}$ be a differentiable function with $\mathcal{S} \neq \emptyset$. Then, f satisfies the Polyak-Łojasiewicz (PŁ) inequality on $\mathcal{S}$, if there exists $r > \min f$ and $\mu > 0$ such that

$$2\mu(f(x) - \min f) \leq \|\nabla f(x)\|^2, \quad \forall x \in [\min f < f < r], \tag{3.14}$$

and we will write $f \in \text{PŁ}_\mu(\mathcal{S})$.

Theorem 3.7. *Let $\nu \geq 2$ and consider the dynamic ($\mathrm{S-ISIHD}$) with initial data $X_0, V_0 \in \mathrm{L}^\nu(\Omega; \mathbb{H})$, where f satisfies (H_0'), and σ satisfies (H_σ). Besides, $f \in \text{PŁ}_\mu(\mathcal{S})$ and suppose that either $\mathbb{H}$ is finite dimensional or $f \in C^2(\mathbb{H})$. Let also, $\gamma \equiv \sqrt{2\mu}$, $\beta \equiv \Gamma \equiv \frac{1}{\sqrt{2\mu}}$, and such that $\sigma_\infty \in \mathrm{L}^2([t_0, +\infty[)$.*

Then the solution trajectory $(X, V) \in S^\nu_{\mathbb{H}\times\mathbb{H}}[t_0]$ is unique. Moreover, letting $\delta > 0$, then there exists $\hat{t}_\delta > t_0, \hat{s}_\delta > 0, K_{\mu,\delta}, C_l, C_f > 0$ such that:

$$\mathbb{E}(f(X(t)) - \min f) \leq K_{\mu,\delta} e^{-\frac{\mu}{2}(t-\hat{t}_\delta)} + \frac{1}{\mu} l_\delta \left(\frac{t + 3\hat{t}_\delta - 4t_0}{4\mu} \right) + C_f\sqrt{\delta}, \quad \forall t > \hat{t}_\delta, \tag{3.15}$$

where

$$l_\delta(s) = \frac{L}{2}\sigma^2_\infty(s) + C_l\sqrt{\delta}\frac{\sigma^2_\infty(s)}{2\sqrt{\int_{\hat{s}_\delta}^s \sigma^2_\infty(u)du}}.$$

Besides, if $f \in \text{PŁ}_\mu(\mathcal{S})$ holds on the entire space (i.e. $r = +\infty$), then we have that there exists $K_\mu > 0$ such that:

$$\mathbb{E}(f(X(t)) - \min f) \leq K_\mu e^{-\frac{\mu}{2}(t-t_0)} + \frac{L}{2\mu}\sigma^2_\infty\left(\frac{t - t_0}{4\mu}\right), \quad \forall t > t_0, \tag{3.16}$$

Remark 3.8. If f is $\mu-$strongly convex, then $f \in \text{PŁ}_\mu(\mathcal{S})$ holds on the entire space (i.e. $r = +\infty$).

Remark 3.9. For a proper discussion on the arbitrarily small (but non-zero) term $C_f\sqrt{\delta}$, we refer to [62, Section 4].

Proof. Consider the dynamic ($\mathrm{S-ISIHD}$) with $\gamma \equiv c, \beta \equiv \Gamma \equiv \frac{1}{c}$, where $c > 0$ is a constant that will be fixed later.

Let us also define $\theta(t) \overset{\text{def}}{=} \int_{t_0}^t \Gamma(u)du = \frac{t-t_0}{c}$ and $\tilde{\sigma}(s, \cdot) \overset{\text{def}}{=} \sigma(\theta^{-1}(s), \cdot)\sqrt{\Gamma(\theta^{-1}(s))}$. Then $\sigma_\infty \in \mathrm{L}^2([t_0, +\infty[)$ is equivalent to $\tilde{\sigma}_\infty \in \mathrm{L}^2(\mathbb{R}_+)$. Now consider the dynamic:

$$\begin{cases} dZ(s) &= -\nabla f(Z(s)) + \tilde{\sigma}(s, Z(s))dW(s), \quad s > 0, \\ Z(0) &= X_0 + \Gamma(t_0)V_0. \end{cases} \tag{3.17}$$

Let $\delta > 0$ and apply the result of [62, Theorem 4.5(i-b)] (with coefficient $\sqrt{2\mu}$), that is, there exists $\hat{s}_\delta > 0$ such that for every $\lambda \in]0, 1[$,

$$\begin{aligned} \mathbb{E}\left(f(Z(s)) - \min f\right) \leq\ & e^{-2\mu(s-\hat{s}_\delta)}\mathbb{E}(f(Z(\hat{s}_\delta)) - \min f) \\ & + e^{-2\mu(1-\lambda)(s-\hat{s}_\delta)}\left(\frac{LC^2_\infty}{2} + C_l C_\infty\sqrt{\delta}\right) \\ & + \frac{l_\delta(\hat{s}_\delta + \lambda(s - \hat{s}_\delta))}{2\mu} + C_f\sqrt{\delta}, \qquad \forall s > \hat{s}_\delta, \end{aligned} \tag{3.18}$$

where $C_\infty, C_l, C_f > 0$ and the establishment of l_δ are detailed in [62, Section 4.2], in particular $C_\infty = \sqrt{\int_{t_0}^{+\infty} \sigma_\infty^2(s)ds}$.

Consider the time scaling $\tau \equiv \theta$, $Y(t) = Z(\theta(t))$ and $\hat{t}_\delta > t_0$ such that $\theta(\hat{t}_\delta) = \hat{s}_\delta$ (i.e. $\hat{t}_\delta = c\hat{s}_\delta + t_0$), we have that:

$$\begin{aligned}\mathbb{E}\left(f(Y(t)) - \min f\right) \leq\ & e^{-2\mu(\theta(t)-\hat{s}_\delta)}\mathbb{E}(f(Y(\hat{t}_\delta)) - \min f) \\ & + e^{-2\mu(1-\lambda)(\theta(t)-\hat{s}_\delta)}\left(\frac{LC_\infty^2}{2} + C_l C_\infty\sqrt{\delta}\right) \\ & + \frac{l_\delta(\hat{s}_\delta + \lambda(\theta(t) - \hat{s}_\delta))}{2\mu} + C_f\sqrt{\delta}, \qquad \forall t > \hat{t}_\delta.\end{aligned} \tag{3.19}$$

Let $a(t) = c$ and $A(t) = c(t - \hat{t}_\delta)$. Now, we consider the averaging as in (3.6) but change the initial condition to $\hat{t}_\delta$. Thus, we have

$$X(t) = \int_{\hat{t}_\delta}^t Y(u)d\tilde{\mu}_t(u) + \tilde{\xi}(t), \tag{3.20}$$

where $\tilde{\mu}_t$ is the probability measure on $[\hat{t}_\delta, t]$ defined by

$$\tilde{\mu}_t = e^{-c(t-\hat{t}_\delta)}\delta_{\hat{t}_\delta} + ce^{c(u-t)}du, \tag{3.21}$$

where $\delta_{\hat{t}_\delta}$ is the Dirac measure at $\hat{t}_\delta$ and

$$\tilde{\xi}(t) \stackrel{\text{def}}{=} -\frac{1}{c}e^{-c(t-\hat{t}_\delta)}V(\hat{t}_\delta). \tag{3.22}$$

Then

$$\mathbb{E}(f(X(t)) - \min f) = \mathbb{E}\left(f(X(t)) - f\left(\int_{\hat{t}_\delta}^t Y(u)d\mu_t(u)\right)\right) + \mathbb{E}\left(f\left(\int_{\hat{t}_\delta}^t Y(u)d\mu_t(u)\right) - \min f\right).$$

We can bound the first term using convexity and Cauchy-Schwarz inequality in the following way

$$\begin{aligned}\mathbb{E}\left(f(X(t)) - f\left(\int_{\hat{t}_\delta}^t Y(u)d\tilde{\mu}_t(u)\right)\right) &\leq \sqrt{\mathbb{E}(\|\nabla f(X(t))\|^2)}\sqrt{\mathbb{E}(\|\tilde{\xi}(t)\|^2)} \\ &\leq \frac{\sqrt{\mathbb{E}(\|V(\hat{t}_\delta)\|^2)}}{c}\sqrt{2\|\nabla f(0)\|^2 + 2L^2\mathbb{E}\left(\sup_{t\geq t_0}\|X(t)\|^2\right)}e^{-c(t-\hat{t}_\delta)},\end{aligned}$$

where $\mathbb{E}(\sup_{t\geq t_0}\|X(t)\|^2) < +\infty$ as mentioned in Corollary 3.5.

On the other hand, we can bound the second term using Jensen's inequality and then (3.19)

$$
\begin{aligned}
\mathbb{E}\left(f\left(\int_{\hat{t}_\delta}^t Y(u)d\tilde{\mu}_t(u)\right)-\min f\right) &\leq \int_{\hat{t}_\delta}^t \mathbb{E}(f(Y(u)-\min f))d\tilde{\mu}_t(u)\\
&\leq \int_{\hat{t}_\delta}^t e^{-2\mu(\theta(u)-\hat{s}_\delta)}\mathbb{E}(f(Y(\hat{t}_\delta))-\min f)d\tilde{\mu}_t(u)\\
&+\int_{\hat{t}_\delta}^t e^{-2\mu(1-\lambda)(\theta(u)-\hat{s}_\delta)}\left(\frac{LC_\infty^2}{2}+C_lC_\infty\sqrt{\delta}\right)d\tilde{\mu}_t(u)\\
&+\int_{\hat{t}_\delta}^t \frac{l_\delta(\hat{s}_\delta+\lambda(\theta(u)-\hat{s}_\delta))}{2\mu}d\tilde{\mu}_t(u)+C_f\sqrt{\delta}\\
&=\left(\mathbb{E}(f(Y(\hat{t}_\delta))-\min f)+\left(\frac{LC_\infty^2}{2}+C_lC_\infty\sqrt{\delta}\right)+\frac{l_\delta(\hat{s}_\delta)}{2\mu}\right)e^{-c(t-\hat{t}_\delta)}\\
&+c\mathbb{E}(f(Y(\hat{t}_\delta))-\min f)e^{-ct}\int_{\hat{t}_\delta}^t e^{-2\mu(\theta(u)-\hat{s}_\delta)}e^{cu}du\\
&+c\left(\frac{LC_\infty^2}{2}+C_lC_\infty\sqrt{\delta}\right)e^{-ct}\int_{\hat{t}_\delta}^t e^{-2\mu(1-\lambda)(\theta(u)-\hat{s}_\delta)}e^{cu}du\\
&+\frac{c}{2\mu}e^{-ct}\int_{\hat{t}_\delta}^t l_\delta(\hat{s}_\delta+\lambda(\theta(u)-\hat{s}_\delta))e^{cu}du+C_f\sqrt{\delta},\quad \forall t>\hat{t}_\delta.
\end{aligned}
$$

We bound the first integral as follows:

$$
e^{-ct}\int_{\hat{t}_\delta}^t e^{-2\mu(\theta(u)-\hat{s}_\delta)}e^{cu}du \leq e^{2\mu\left(\frac{t_0}{c}+\hat{s}_\delta\right)}e^{-\frac{2\mu}{c}t}.
$$

The second integral in the same way

$$
e^{-ct}\int_{\hat{t}_\delta}^t e^{-2\mu(1-\lambda)(\theta(u)-\hat{s}_\delta)}e^{cu}du \leq e^{2\mu(1-\lambda)\left(\frac{t_0}{c}+\hat{s}_\delta\right)}e^{-\frac{2\mu(1-\lambda)}{c}t}.
$$

To treat the third integral we are going to split the integral in two in order to find a useful convergence rate. Let us recall that $l_\delta \in \mathrm{L}^1([\hat{s}_\delta,+\infty[)$ and that l_δ is decreasing. Let us define $\varphi_{\lambda,c,\delta}(t) \stackrel{\text{def}}{=} \hat{s}_\delta+\lambda\left(\frac{t+\hat{t}_\delta-2t_0}{2c}-\hat{s}_\delta\right)$, then

$$
\begin{aligned}
e^{-ct}\int_{\hat{t}_\delta}^t l_\delta(\hat{s}_\delta+\lambda(\theta(u)-\hat{s}_\delta))e^{cu}du &= e^{-ct}\int_{\hat{t}_\delta}^{\frac{\hat{t}_\delta+t}{2}} l_\delta(\hat{s}_\delta+\lambda(\theta(u)-\hat{s}_\delta))e^{cu}du\\
&+e^{-ct}\int_{\frac{\hat{t}_\delta+t}{2}}^t l_\delta(\hat{s}_\delta+\lambda(\theta(u)-\hat{s}_\delta))e^{cu}du\\
&\leq \frac{c}{\lambda}e^{\frac{c\hat{t}_\delta}{2}}C_\infty e^{-\frac{ct}{2}}+l_\delta(\varphi_{\lambda,c,\delta}(t)).
\end{aligned}
$$

Now that we have bounded all the terms, we have the following bound

$$\begin{aligned}
\mathbb{E}(f(X(t)) - \min f) \leq & \frac{\sqrt{\mathbb{E}(\|V(\hat{t}_\delta)\|^2)}}{c}\sqrt{2\|\nabla f(0)\|^2 + 2L^2\mathbb{E}\left(\sup_{t\geq t_0}\|X(t)\|^2\right)}e^{-c(t-\hat{t}_\delta)} \\
& + \left(\mathbb{E}(f(Y(\hat{t}_\delta)) - \min f) + \left(\frac{LC_\infty^2}{2} + C_l C_\infty\sqrt{\delta}\right) + \frac{l_\delta(\hat{s}_\delta)}{2\mu}\right)e^{-c(t-\hat{t}_\delta)} \\
& + c\mathbb{E}(f(Y(\hat{t}_\delta)) - \min f)e^{2\mu\left(\frac{t_0+\hat{t}_\delta}{c}+\hat{s}_\delta\right)}e^{-\frac{2\mu}{c}(t-\hat{t}_\delta)} \\
& + c\left(\frac{LC_\infty^2}{2} + C_l C_\infty\sqrt{\delta}\right)e^{2\mu(1-\lambda)\left(\frac{t_0+\hat{t}_\delta}{c}+\hat{s}_\delta\right)}e^{-\frac{2\mu(1-\lambda)}{c}(t-\hat{t}_\delta)} \\
& + \frac{c}{2\mu}\left(\frac{c}{\lambda}C_\infty e^{-\frac{c(t-\hat{t}_\delta)}{2}} + l_\delta(\varphi_{\lambda,c,\delta}(t))\right) + C_f\sqrt{\delta}, \quad \forall t > \hat{t}_\delta.
\end{aligned}$$

Letting $\lambda = \frac{1}{2}$ and $c = \sqrt{2\mu}$ we obtain

$$\begin{aligned}
\mathbb{E}(f(X(t)) - \min f) \leq & \frac{\sqrt{\mathbb{E}(\|V(\hat{t}_\delta)\|^2)}}{\sqrt{2\mu}}\sqrt{2\|\nabla f(0)\|^2 + 2L^2\mathbb{E}\left(\sup_{t\geq t_0}\|X(t)\|^2\right)}e^{-\sqrt{2\mu}(t-\hat{t}_\delta)} \\
& + \left(\mathbb{E}(f(Y(\hat{t}_\delta)) - \min f) + \left(\frac{LC_\infty^2}{2} + C_l C_\infty\sqrt{\delta}\right) + \frac{l_\delta(\hat{s}_\delta)}{2\mu}\right)e^{-\sqrt{2\mu}(t-\hat{t}_\delta)} \\
& + \sqrt{2\mu}\mathbb{E}(f(Y(\hat{t}_\delta)) - \min f)e^{2\mu\left(\frac{t_0+\hat{t}_\delta}{\sqrt{2\mu}}+\hat{s}_\delta\right)}e^{-\sqrt{2\mu}(t-\hat{t}_\delta)} \\
& + \sqrt{2\mu}\left(\frac{LC_\infty^2}{2} + C_l C_\infty\sqrt{\delta}\right)e^{\mu\left(\frac{t_0+\hat{t}_\delta}{\sqrt{2\mu}}+\hat{s}_\delta\right)}e^{-\frac{\sqrt{2\mu}}{2}(t-\hat{t}_\delta)} \\
& + 2C_\infty e^{-\frac{\sqrt{2\mu}(t-\hat{t}_\delta)}{2}} + \frac{1}{\sqrt{2\mu}}l_\delta(\varphi_{\frac{1}{2},\sqrt{2\mu},\delta}(t)) + C_f\sqrt{\delta}, \quad \forall t > \hat{t}_\delta.
\end{aligned}$$

Letting $K_{\mu,\delta} \stackrel{\text{def}}{=} \sqrt{2\mu}\left(\frac{LC_\infty^2}{2} + C_l C_\infty\sqrt{\delta}\right)e^{\mu\left(\frac{t_0+\hat{t}_\delta}{\sqrt{2\mu}}+\hat{s}_\delta\right)} + 2C_\infty$, we conclude that

$$\mathbb{E}(f(X(t)) - \min f) \leq K_{\mu,\delta}e^{-\frac{\sqrt{2\mu}}{2}(t-\hat{t}_\delta)} + \frac{1}{\sqrt{2\mu}}l_\delta(\varphi_{\frac{1}{2},\sqrt{2\mu},\delta}(t)) + C_f\sqrt{\delta}, \quad \forall t > \hat{t}_\delta. \tag{3.23}$$

□

4 From weak to strong convergence under general γ and $\beta \equiv \Gamma$

4.1 General result

We consider the Tikhonov regularization of the dynamic ($\mathrm{S-ISIHD_{NS}}$), i.e., for $t > 0$,

$$\begin{cases}
dX(t) & = V(t)dt, \\
dV(t) & \in -\gamma(t)V(t)dt - \partial F(X(t) + \beta(t)V(t))dt - \varepsilon(t)(X(t) + \beta(t)V(t))dt \\
& \quad + \sigma(t, X(t) + \beta(t)V(t))dW(t), \\
X(t_0) & = X_0, \quad V(t_0) = V_0.
\end{cases} \tag{$\mathrm{S-ISIHD_{NS}-TA}$}$$

We show some conditions (on $\gamma, \beta, \varepsilon$) under which we can obtain strong convergence of the trajectory.

Theorem 4.1. *Consider that $\gamma : [t_0, +\infty[\to \mathbb{R}_+$ satisfies (H_γ). Besides, $F = f + g$ and σ satisfy assumptions (H_0) and (H_σ). Moreover, suppose that g satisfies (H_λ) and let $\nu \geq 2$. Consider ($\mathrm{S-ISIHD_{NS}-TA}$) with $\beta \equiv \Gamma$ and initial data $X_0, V_0 \in \mathrm{L}^\nu(\Omega; \mathbb{H})$.*

Then, there exists a unique solution $(X, V) \in S^\nu_{\mathbb{H}\times\mathbb{H}}[t_0]$ of ($\mathrm{S-ISIHD_{NS}-TA}$). Additionally, let $x^\star \overset{\text{def}}{=} \mathrm{P}_{\mathcal{S}_F}(0)$ be the minimum norm solution, and for $\varepsilon > 0$ let x_ε be the unique minimizer of $F_\varepsilon(x) \overset{\text{def}}{=} F(x) + \frac{\varepsilon}{2}\|x\|^2$. If we suppose that $\Gamma\sigma_\infty \in \mathrm{L}^2([t_0, +\infty[)$, and that $\varepsilon : [t_0, +\infty[\to \mathbb{R}_+$ satisfies the conditions:

(T_1') $\varepsilon(t) \to 0$ *as* $t \to +\infty$;

(T_2') $\displaystyle\int_{t_0}^{+\infty} \varepsilon(t)\Gamma(t)dt = +\infty$;

(T_3') $\displaystyle\int_{t_0}^{+\infty} \varepsilon(t)\Gamma(t)\left(\|x^\star\|^2 - \|x_{\varepsilon(t)}\|^2\right)dt < +\infty$.

Then $\text{s-lim}_{t\to+\infty} X(t) = x^\star$ *a.s., and* $V(t) = o\left(\frac{1}{\Gamma(t)}\right)$ *a.s..*

Proof. Let $s_0 > 0$, $\theta(t) \overset{\text{def}}{=} s_0 + \int_{t_0}^t \Gamma(u)du$; $\tilde{\varepsilon}(t) = \varepsilon(\theta^{-1}(t))$; and $\tilde{\sigma}(s, \cdot) \overset{\text{def}}{=} \sigma(\theta^{-1}(s), \cdot)\sqrt{\Gamma(\theta^{-1}(s))}$. Then ε satisfying (T_1'),(T_2'), and (T_3') is equivalent to $\tilde{\varepsilon}$ satisfying (T_1),(T_2), and (T_3) defined in Theorem 2.8. Besides, $\Gamma\sigma_\infty \in \mathrm{L}^2([t_0, +\infty[)$ is equivalent to $\tilde{\sigma}_\infty \in \mathrm{L}^2(\mathbb{R}_+)$. Consider the dynamic:

$$\begin{cases} dZ(s) & \in -\partial F(Z(s)) - \tilde{\varepsilon}(s)Z(s) + \tilde{\sigma}(s, Z(s))dW(s), \quad s > s_0, \\ Z(s_0) & = X_0 + \Gamma(t_0)V_0. \end{cases} \tag{4.1}$$

By Theorem 2.8, we have that there exists a unique solution $Z \in S^\nu_{\mathbb{H}}[s_0]$, and that $\lim_{s\to\infty} Z(s) = x^\star$ a.s. (Recall that $x^\star \overset{\text{def}}{=} \mathrm{P}_{S_F}(0)$). Using the time scaling $\tau \equiv \theta$ and the averaging described at the beginning of this section, we end up with the dynamic ($\mathrm{S-ISIHD_{NS}-TA}$) in the case where $\beta \equiv \Gamma$. The existence and uniqueness of solution, and the fact that $(X, V) \in S^\nu_{\mathbb{H}\times\mathbb{H}}[t_0]$ goes analogously as in the proof of Theorem 3.1.

Now we prove the claim, since $\lim_{s\to\infty} Z(s) = x^\star$ a.s., this implies directly that $\lim_{t\to\infty} Y(t) = x^\star$ a.s.. Besides, we have the relation (3.7), i.e.

$$X(t) = \int_{t_0}^t Y(u)d\mu_t(u) + \xi(t),$$

where μ_t and ξ are defined in (3.8) and (3.9), respectively. Consequently, we have

$$\begin{aligned} \|X(t) - x^\star\| &\leq \left\|X(t) - \int_{t_0}^t Y(u)d\mu_t(u)\right\| + \left\|\int_{t_0}^t Y(u)d\mu_t(u) - x^\star\right\| \\ &\leq \|\xi(t)\| + \left\|\int_{t_0}^t Y(u)d\mu_t(u) - x^\star\right\|. \end{aligned}$$

Let $a(t) = \frac{1}{\Gamma(t)}$ and $A(t) = \int_{t_0}^t \frac{du}{\Gamma(u)}$. By Lemma B.3, we have that $\lim_{t\to\infty} \|\xi(t)\| = 0$. On the other hand

$$\begin{aligned} \left\|\int_{t_0}^t Y(u)d\mu_t(u) - x^\star\right\| &= \left\|\int_{t_0}^t (Y(u) - x^\star)d\mu_t(u)\right\| \\ &\leq \int_{t_0}^t \|Y(u) - x^\star\|d\mu_t(u) \\ &= e^{-A(t)}\|Y(t_0) - x^\star\| + e^{-A(t)}\int_{t_0}^t a(u)e^{A(u)}\|Y(u) - x^\star\|du. \end{aligned}$$

Let $b(u) = \|Y(u) - x^\star\|$. Since we already proved that $\lim_{u\to\infty} b(u) = 0$ a.s., and we have that $a \notin \mathrm{L}^1([t_0, +\infty[)$ by Lemma B.3, we utilize Lemma B.2 with our respective a, b functions. This let us conclude that

$$\lim_{t\to\infty} \left\| \int_{t_0}^{t} Y(u) d\mu_t(u) - x^\star \right\| = 0 \quad a.s..$$

Thus, $\lim_{t\to\infty} X(t) = x^\star$ a.s.. Finally, since

$$Y(t) = X(t) + \Gamma(t)V(t),$$

and the fact that X and Y have (a.s.) the same limit, we conclude that

$$\lim_{t\to\infty} \Gamma(t)V(t) = 0 \quad a.s..$$

□

4.2 Concrete cases

In order to give some conditions when $(T_1^{'})$, $(T_2^{'})$, and $(T_3^{'})$ of Theorem 4.1 are satisfied we need to introduce the following definition:

Definition 4.2 (Hölderian error bound). Let $f : \mathbb{H} \to \mathbb{R}$ be a proper function such that $\mathcal{S} \neq \emptyset$. f satisfies a Hölderian (or power-type) error bound inequality on $\mathcal{S}$ with exponent $p \geq 1$, if there exists $\kappa > 0$ and $r > \min f$ such that

$$f(x) - \min f \geq \kappa \mathrm{dist}(x, \mathcal{S})^p, \quad \forall x \in [\min f \leq f \leq r], \tag{4.2}$$

and we will write $f \in \mathrm{EB}^p(\mathcal{S})$,

Remark 4.3. Let $f : \mathbb{H} \to \mathbb{R}$ be a differentiable function such that $\mathcal{S} \neq \emptyset$. If f satisfies the Polyak-Łojasiewicz inequality on $\mathcal{S}$, then f satisfies a Hölderian error bound inequality with exponent $p = 2$.

Theorem 4.4. *Consider the setting of Theorem 4.1 and suppose that $F = f + g \in \mathrm{EB}^p(\mathcal{S}_F)$. Let $s_0 > 0$ and denote $\theta(t) \stackrel{\text{def}}{=} s_0 + \int_{t_0}^{t} \Gamma(s)ds$, then taking the Tikhonov parameter $\varepsilon(t) = \frac{1}{\theta^r(t)}$ with*

$$1 \geq r > \frac{2p}{2p+1},$$

then the three conditions $(T_1^{'})$, $(T_2^{'})$, and $(T_3^{'})$ of Theorem 4.1 are satisfied simultaneously. In particular, for any solution $(X, V) \in S^{\nu}_{\mathbb{H}\times\mathbb{H}}[t_0]$ of (4.1), we get almost sure (strong) convergence of $X(t)$ to the minimal norm solution named $x^\star = P_{\mathcal{S}_F}(0)$ and that $V(t) = o\left(\frac{1}{\Gamma(t)}\right)$.

Proof. We proceed as in the proof of Theorem 4.1 and arrive to the dynamic (4.1), since $\tilde{\varepsilon}(t) = \varepsilon(\theta^{-1}(t)) = \frac{1}{t^r}$, the proof goes as in [63, Theorem 4.8]. □

Theorem 4.5. *Let $\nu \geq 2$, $f \in \Gamma_0(\mathbb{H}) \cap C^2_L(\mathbb{H})$ such that $\mathcal{S}$ is nonempty, and also $f \in \mathrm{EB}^p(\mathcal{S})$, σ satisfying (H_σ), and $\Gamma\sigma_\infty \in \mathrm{L}^2([t_0, +\infty[)$ and is non-increasing. Let us consider $\varepsilon(t) = \frac{1}{t^r}$ where $0 < r < 1$, then we evaluate $(\mathrm{S} - \mathrm{ISIHD}_{\mathrm{NS}} - \mathrm{TA})$ in the case where γ satisfies (H_γ), $g \equiv 0$, $\beta \equiv \Gamma$, and with initial data $X_0, V_0 \in \mathrm{L}^\nu(\Omega; \mathbb{H})$, i.e., for $t > t_0$,*

$$\begin{cases} dX(t) &= V(t)dt, \\ dV(t) &= -\gamma(t)V(t)dt - \nabla f(X(t)+\Gamma(t)V(t))dt - \varepsilon(t)(X(t)+\Gamma(t)V(t)) \\ & \quad + \sigma(t, X(t)+\Gamma(t)V(t))dW(t), \\ X(t_0) &= X_0, \quad V(t_0) = V_0. \end{cases} \tag{4.3}$$

For $\varepsilon > 0$, let us define $f_\varepsilon(x) \stackrel{\text{def}}{=} f(x) + \frac{\varepsilon}{2}\|x\|^2$, and let x_ε be the unique minimizer of f_ε. Moreover, let $s_0 > 0$ and for $s_1 > s_0$ consider,

$$R(s) \stackrel{\text{def}}{=} e^{-\frac{s^{1-r}}{1-r}} \int_{s_1}^{s} e^{\frac{u^{1-r}}{1-r}} \sigma_\infty^2(\theta^{-1}(u))\Gamma(\theta^{-1}(u))du, \tag{4.4}$$

where $\theta(t) \stackrel{\text{def}}{=} s_0 + \int_{t_0}^{t} \Gamma(u)du$. Let also $x^\star \stackrel{\text{def}}{=} \mathrm{P}_{\mathcal{S}}(0)$, $A(s) \stackrel{\text{def}}{=} \int_{s_1}^{s} \frac{du}{\Gamma(u)}$, and $t_1 \stackrel{\text{def}}{=} \theta^{-1}(s_1)$. Then, the solution trajectory $(X, V) \in S^{\nu}_{\mathbb{H}\times\mathbb{H}}[t_0]$ is unique, and we have that:

(i) $R(\theta(t)) \to 0$ as $t \to +\infty$.

(ii) Let $\bar{\sigma}(t) = \Gamma(t)\sigma_\infty^2(t)$, then

$$R(\theta(t)) = \mathcal{O}\left(\exp(-\theta^r(t)(1-2^{-r})) + \theta^r(t)\bar{\sigma}\left(\frac{s_1+\theta(t)}{2}\right)\right).$$

Moreover, if $\bar{\sigma}(t) = \mathcal{O}(\theta^{-\Delta}(t))$ for $\Delta > 1$, then $R(\theta(t)) = \mathcal{O}(\theta^{r-\Delta}(t))$.

Besides, we have the following convergence rate in expectation:

(iii) For the values, we have:

$$\mathbb{E}[f(X(t)) - \min(f)] = \mathcal{O}\left(\max\{e^{-A(t)}, I[h_1](t)\}\right),$$

where $h_1(t) = \frac{1}{\theta^r(t)} + R(\theta(t))$.

(iv) And for the trajectory, we obtain:

$$\mathbb{E}[\|X(t) - x^\star\|^2] = \mathcal{O}\left(\max\{e^{-A(t)}, I[h_2](t)\}\right),$$

where $h_2(t) = \theta^{r-1}(t) + \theta^{-\frac{r}{p}}(t) + \theta^r(t)R(\theta(t))$.

Proof. We proceed as in the proof of Theorem 3.1 and define analogously $\tilde{\sigma}$, $\tilde{\varepsilon}$, we also consider the dynamic (3.10). By [63, Theorem 4.11] we obtain that

$$R(s) = e^{-\frac{s^{1-r}}{1-r}} \int_{s_1}^{s} e^{\frac{u^{1-r}}{1-r}} \tilde{\sigma}_\infty^2(u)du,$$

where $\tilde{\sigma}_\infty^2 \in \mathrm{L}^2([s_0, +\infty[)$, satisfies the following:

- $R(s) \to +\infty$ as $t \to +\infty$.
- $R(s) = \mathcal{O}\left(\exp(-s^r(1-2^{-r})) + s^r\tilde{\sigma}_\infty^2\left(\frac{s_1+s}{2}\right)\right)$. Moreover if $\tilde{\sigma}_\infty^2(s) = \mathcal{O}(s^{-\Delta})$ for $\Delta > 1$, then $R(s) = \mathcal{O}(s^{r-\Delta})$.

And evaluating at $s = \theta(t)$ we obtain the first two items of the theorem. For the third and fourth items we used that

- $\mathbb{E}[f(Z(s)) - \min(f)] = \mathcal{O}\left(\frac{1}{s^r} + R(s)\right)$.
- $\mathbb{E}[\|Z(s) - x^\star\|^2] = \mathcal{O}\left(\frac{1}{s^{1-r}} + \frac{1}{s^{\frac{r}{p}}} + s^r R(s)\right)$.

Then, proceeding as in the proof of Theorem 3.2, we obtain the desired results. □

Corollary 4.6. *Consider Theorem 4.4 in the case where $\gamma(t) = \frac{\alpha}{t}$ for $\alpha > 1$, $\beta(t) = \frac{t}{\alpha-1}$ then we have that:*

1. *If $\sigma_\infty^2(t) = \mathcal{O}(t^{-2(\Delta+1)})$ for $\Delta > 1$, and $\alpha \neq \{1+2r, 1+2(\Delta-r)\}$, then*

$$\mathbb{E}[f(X(t)) - \min(f)] = \mathcal{O}\left(\max\{t^{-(\alpha-1)}, t^{-2r}, t^{-2(\Delta-r)}\}\right).$$

In particular, if $\alpha > 3$

2. *If $\sigma_\infty^2(t) = \mathcal{O}(t^{-2(\Delta+1)})$ for $\Delta > \max\{1, 2r\}$, and $\alpha \neq \{3-2r, 1+\frac{2r}{p}, 1+2(2r-\Delta)\}$, then*

$$\mathbb{E}[\|X(t) - x^\star\|^2] = \mathcal{O}\left(\max\{t^{-(\alpha-1)}, t^{-2(1-r)}, t^{-\frac{2r}{p}}, t^{-2(2r-\Delta)}\}\right).$$

5 Conclusion

This work uncovers the global and local convergence properties of trajectories of the Inertial System with Implicit Hessian-driven Damping under stochastic errors both in the smooth and non-smooth setting. 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 case where the Hessian damping parameter β was dependent on the geometric damping γ and when it was zero. We believe that this work, along with the technique of time scaling and averaging, paves the way for important extensions and research avenues. Among them, we mention extension to the situation where the drift term is a non-potential co-coercive operator.

A SDEs are better approximations to stochastic inertial algorithms

In this section, we will argue that (S − ISIHD) is an accurate continuous-time model of a natural stochastic inertial algorithm with an accuracy error that scales as $\mathcal{O}(h)$, and this is much better than (ISIHD) whose accuracy is only $\mathcal{O}(\sqrt{h})$. This generalizes [30, Proposition 2.1] and holds for a more general model of the diffusion coefficient and under weaker regularity assumptions. The proof is reminiscent of strong consistency bounds of Euler-type methods for nonlinear SDEs [46], that we will refine by exploiting the structure of our SDE (S − ISIHD). We also restrict our discussion to the finite-dimensional case where $\mathbb{H} = \mathbb{K} = \mathbb{R}^d$. Similar approximation results can be found, for instance, in [42, 53, 34], while for momentum-based algorithms, the tools developed in [51] may provide useful insights for deriving alternative approximation results, once adapted to our Hessian-driven damping setting.

Given a step-size $h > 0$, the Euler-Maruyama discretization applied to (S − ISIHD) computes approximations X_k and V_k of $X(t_k)$ and $V(t_k)$, where $t_k = t_0 + kh$, $k \in \mathbb{N}$, by initializing $(X_0, V_0) = (X(t_0), V(t_0))$ and forming the following iterative scheme:

$$\begin{cases} X_{k+1} &= X_k + hV_k, \\ V_{k+1} &= (1-\gamma_k h)V_k - h\nabla f(X_k + \beta_k V_k) + \sqrt{h}\sigma_k G_k, \end{cases} \tag{A.1}$$

where $G_k \sim \mathcal{N}(0, I_d)$, $\gamma_k = \gamma(t_k)$, $\sigma_k = \sigma(t_k, X_k + \beta_k V_k)$ and $\beta_k = \beta(t_k)$. This is a stochastic version of the algorithm proposed in [1]. Setting $Y_k = (X_k, V_k)$, it will be convenient to rewrite (A.1) in the product space $\mathbb{R}^{2d}$ as

$$\begin{cases} Y_{k+1} &= Y_k + h\Phi(t_k, Y_k) + \sqrt{h}\varsigma(t_k, Y_k)(W(t_{k+1}) - W(t_k)), \\ Y_0 &= (X_0, V_0), \end{cases} \tag{A.2}$$

where W is a $\mathbb{R}^{2d}-$valued Brownian motion and

$$\Phi(t, (x, v)) = (v, -\gamma(t)v - \nabla f(x + \beta(t)v)), \quad \text{and} \quad \varsigma(t, (x, v)) = \begin{pmatrix} 0_{d\times d} & 0_{d\times d} \\ 0_{d\times d} & \sigma(t, x + \beta(t)v) \end{pmatrix}.$$

This then motivates the SDE

$$\begin{cases} dY(t) &= \Phi(t, Y(t))dt + \sqrt{h}\varsigma(t, Y(t))dW(t), \\ Y(t_0) &= (X_0, V_0), \end{cases} \tag{A.3}$$

where we set $Y(t) = (X(t), V(t))$ which is ($S-$ ISIHD) with an extra $\sqrt{h}$ in front the diffusion.

As classical in numerical analysis of evolution equations, we define the continuous-time piece-wise linear extension of the sequence $(Y_k)_{k\in\mathbb{N}}$

$$\begin{aligned} \overline{Y}(t) &\stackrel{\text{def}}{=} Y_k + (t - t_k)\Phi(t_k, Y_k) + \sqrt{h}\varsigma(t_k, Y_k)(W(t) - W(t_k)) \quad \text{for} \quad t \in [t_k, t_{k+1}[\\ &= Y_0 + \int_{t_0}^{t} \widehat{\Phi}(s, \widehat{Y}(s))ds + \sqrt{h}\int_{t_0}^{t} \widehat{\varsigma}(s, \widehat{Y}(s))dW(s), \end{aligned}$$

where for $t \in [t_k, t_{k+1}[$, $\widehat{Y}(t) \stackrel{\text{def}}{=} Y_k$, $\widehat{\Phi}(t, \widehat{Y}(t)) \stackrel{\text{def}}{=} \Phi(t_k, Y_k)$, $\widehat{\varsigma}(t, \widehat{Y}(t)) \stackrel{\text{def}}{=} \varsigma(t_k, Y_k)$. We, also define $\widehat{\gamma}(t)$ and $\widehat{\beta}(t)$ similarly. We thus have $\widehat{\sigma}(t, \widehat{X}(t) + \widehat{\beta}(t)\widehat{V}(t)) = \sigma_k$ for $t \in [t_k, t_{k+1}[$.

Our result requires the following assumption.

Assumption A.1. *$f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable with L-Lipschitz continuous gradient. For $T > t_0$, γ and β are respectively L_γ- and L_β-Lipschitz continuous on $[t_0, T]$. σ verifies, $\forall t, s \in [t_0, T]$ and $\forall x, x' \in \mathbb{R}^d$,*

$$\left\|\sigma(t, x) - \sigma(t, x')\right\| \le L_\sigma \left\|x - x'\right\|, \|\sigma(t, x)\| \le K_\sigma(1+\|x\|) \ \text{ and } \ \|\sigma(t, x) - \sigma(s, x)\| \le L_\sigma(1+\|x\|)|t-s|^{1/2}.$$

where all constants $L_\gamma, L_\beta, L_\sigma, K_\sigma$ do not depend on h.

All these assumptions are verified in the instances studied in Section 3 when f is smooth. Observe that these assumptions also ensure existence and uniqueness of the solution $(X(t), V(t))$ to (A.3) (hence ($S-$ ISIHD)).

We have the strong consistency results which shows that (A.3) is a better continuous-time approximation to (A.1) than (ISIHD).

Proposition A.2. *Suppose that Assumption A.1 holds and that $X_0, V_0 \in \mathrm{L}^2(\Omega; \mathbb{R}^d)$. Let $(X(t), V(t))$ be the solution to (A.3) and $(x(t), v(t) = \dot{x}(t))$ the solution to (ISIHD). Then the iterates of algorithm (A.1) satisfy, as $h \to 0$,*

$$\mathbb{E}\left[\sup_{0\le k\le N-1} \|X_k - X(t_k)\| + \|V_k - V(t_k)\|\right] = \mathcal{O}(h), \tag{A.4}$$

$$\mathbb{E}\left[\sup_{0\le k\le N-1} \|X_k - x(t_k)\| + \|V_k - v(t_k)\|\right] = \mathcal{O}(\sqrt{h}). \tag{A.5}$$

Proof. By Assumption A.1, it is not difficult to see that Φ and ς verify the following the Lipschitz continuity and linear growth properties,

$$\begin{aligned}
\|\Phi(t,y)-\Phi(t,z)\|^2 &\le \max(4L^2, 1+2\gamma_{\max}^2+4L^2\beta_{\max}^2)\,\|y-z\|^2\,, && \forall t\in[t_0,T], y,z\in\mathbb{R}^d,\\
\|\Phi(t,y)-\Phi(s,y)\|^2 &\le (2L_\gamma^2+4L^2L_\beta^2)\,\|y\|^2\,|t-s|^2, && \forall s,t\in[t_0,T], y\in\mathbb{R}^d,\\
\|\Phi(t,y)\|^2 &\le \max(8L^2, 1+2\gamma_{\max}^2+8L^2\beta_{\max}^2)\,\|y\|^2+4\,\|\nabla f(0)\|^2\,, && \forall t\in[t_0,T], y\in\mathbb{R}^d.
\end{aligned} \tag{A.6}$$

In the rest of the proof, C is any positive constant that does not depend on h (and may depend on d, $T-t_0$, the different Lipschitz and growth constants in Assumption A.1), and which may change from one line to another. Using Jensen's inequality, Doob's martingale inequality and [61, Theorem 1.5.21], we have for any $\tau\in[t_0,T]$

$$\mathbb{E}\left[\sup_{t_0\le t\le\tau}\left\|\overline{Y}(t)-Y(t)\right\|^2\right] \tag{A.7}$$

$$=\mathbb{E}\left[\sup_{t_0\le t\le\tau}\left\|\int_{t_0}^t\left(\widehat{\Phi}(s,\widehat{Y}(s))-\Phi(s,Y(s))\right)ds+\sqrt{h}\int_{t_0}^t\left(\widehat{\varsigma}(s,\widehat{Y}(s))-\varsigma(s,Y(s))\right)dW(s)\right\|^2\right] \tag{A.8}$$

$$\le C\mathbb{E}\left[\int_{t_0}^\tau\left\|\widehat{\Phi}(s,\widehat{Y}(s))-\Phi(s,Y(s))\right\|^2ds\right]+Ch\mathbb{E}\left[\int_{t_0}^\tau\left\|\widehat{\sigma}(s,\widehat{X}(s)+\widehat{\beta}(s)\widehat{V}(s))-\sigma(s,X(s)+\beta(s)V(s))\right\|^2ds\right].$$

Using Jensen's inequality, (A.6) and Assumption A.1 on σ, we get

$$\begin{aligned}
&\mathbb{E}\left[\sup_{t_0\le t\le\tau}\left\|\overline{Y}(t)-Y(t)\right\|^2\right]\\
\le& C\mathbb{E}\left[\int_{t_0}^\tau\left\|\widehat{\Phi}(s,\widehat{Y}(s))-\Phi(s,\widehat{Y}(s))\right\|^2ds\right]+Ch\mathbb{E}\left[\int_{t_0}^\tau\left\|\widehat{\sigma}(s,\widehat{X}(s)+\widehat{\beta}(s)\widehat{V}(s))-\sigma(s,\widehat{X}(s)+\widehat{\beta}(s)\widehat{V}(s))\right\|^2ds\right]\\
&+C(1+h)\int_{t_0}^\tau\mathbb{E}\left[\sup_{t_0\le r\le s}\left\|\overline{Y}(r)-Y(r)\right\|^2\right]ds+C(1+h)\mathbb{E}\left[\int_{t_0}^\tau\left\|\overline{Y}(t)-\widehat{Y}(t)\right\|^2ds\right]\\
&+Ch\mathbb{E}\left[\sup_{t_0\le t\le T}\|Y(t)\|^2\right]\int_{t_0}^\tau|\beta(s)-\widehat{\beta}(s)|^2ds,
\end{aligned}$$

where the expectation in the last display is bounded by [46, Theorem 4.5.4]. Let $N=\lfloor(T-t_0)/h\rfloor$. Lipschitz continuity of β gives

$$\int_{t_0}^\tau|\beta(s)-\widehat{\beta}(s)|^2ds\le\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}|\beta(s)-\beta(t_k)|^2ds\le L_\beta^2\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}(s-t_k)^2ds=\frac{L_\beta^2(T-t_0)}{3}h^2.$$

Now, we have for any $s\in[t_k,t_{k+1}[$

$$\overline{Y}(s)-\widehat{Y}(s)=-(s-t_k)\Phi(t_k,Y_k)-\sqrt{h}\varsigma(t_k,Y_k)(W(s)-W(t_k)).$$

It then follows from (A.6), Assumption A.1 on σ and [46, Theorem 4.5.4] that

$$\mathbb{E}\left[\int_{t_0}^\tau\left\|\overline{Y}(t)-\widehat{Y}(t)\right\|^2ds\right]\le\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\mathbb{E}\left[\left\|(s-t_k)\Phi(t_k,Y_k)+\sqrt{h}\varsigma(t_k,Y_k)(W(s)-W(t_k))\right\|^2\right]ds\le Ch^2.$$

With similar arguments, we have

$$\mathbb{E}\left[\int_{t_0}^{\tau}\left\|\widehat{\Phi}(s,\widehat{Y}(s))-\Phi(s,\widehat{Y}(s))\right\|^2 ds\right]=\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\mathbb{E}\left[\|\Phi(t_k,Y_k)-\Phi(s,Y_k)\|^2\right]ds\leq Ch^2$$

and using Assumption A.1 on σ

$$\mathbb{E}\left[\int_{t_0}^{\tau}\left\|\widehat{\sigma}(s,\widehat{X}(s)+\widehat{\beta}(s)\widehat{V}(s))-\sigma(s,\widehat{X}(s)+\widehat{\beta}(s)\widehat{V}(s))\right\|^2 ds\right]$$
$$=\sum_{k=0}^{N-1}\int_{t_k}^{t_{k+1}}\mathbb{E}\left[\|\sigma(t_k,X_k+\beta_kV_k)-\sigma(s,X_k+\beta_kV_k)\|^2\right]ds\leq Ch.$$

Collecting all bounds, we get

$$\mathbb{E}\left[\sup_{t_0\leq t\leq\tau}\left\|\overline{Y}(t)-Y(t)\right\|^2\right]\leq C(h^2+h^3)+C(1+h)\int_{t_0}^{\tau}\mathbb{E}\left[\sup_{t_0\leq r\leq s}\left\|\overline{Y}(r)-Y(r)\right\|^2\right]ds.$$

Applying the Jensen and Gronwall inequalities then gives

$$\mathbb{E}\left[\sup_{t_0\leq t\leq T}\left\|\overline{Y}(t)-Y(t)\right\|\right]\leq\mathbb{E}\left[\sup_{t_0\leq t\leq T}\left\|\overline{Y}(t)-Y(t)\right\|^2\right]^{1/2}\leq C(h+h^{3/2})e^{C(1+h)T}=\mathcal{O}(h).$$

In turn

$$\mathbb{E}\left[\sup_{0\leq k\leq N-1}\|Y_k-Y(t_k)\|\right]=\mathbb{E}\left[\sup_{0\leq k\leq N-1}\left\|\overline{Y}(t_k)-Y(t_k)\right\|\right]\leq\mathbb{E}\left[\sup_{t_0\leq t\leq T}\left\|\overline{Y}(t)-Y(t)\right\|\right]=\mathcal{O}(h),$$

which gives (A.4). For (A.5), it is sufficient to see that with the system (A.3), the term $\varsigma(s,y(s))$ disappears in the main inequality (A.8). Starting from there, this becomes the leading term that scales as $\mathcal{O}(h)$ (instead of $\mathcal{O}(h^2)$ previously) hence entailing (A.5). □

B Auxiliary results

B.1 Deterministic results

Lemma B.1. *Let $t_0>0$ and $a,b:[t_0,+\infty[\to\mathbb{R}_+$. If $\lim_{t\to\infty}a(t)$ exists, $b\notin\mathrm{L}^1([t_0,+\infty[)$ and $\int_{t_0}^{\infty}a(s)b(s)ds<+\infty$, then $\lim_{t\to\infty}a(t)=0$.*

Lemma B.2. *Let $a,b:[t_0,+\infty[\to\mathbb{R}_+$ be two functions such that $a\notin\mathrm{L}^1([t_0,+\infty[)$, $\lim_{u\to+\infty}b(u)=0$, and define $A(t)\stackrel{\text{def}}{=}\int_{t_0}^{t}a(u)du$ and $B(t)\stackrel{\text{def}}{=}e^{-A(t)}\int_{t_0}^{t}a(u)e^{A(u)}b(u)du$. Then $\lim_{t\to+\infty}B(t)=0$.*

Proof. Let $\varepsilon>0$ arbitrary, let us take T_ε such that $t_0<T_\varepsilon$ and $b(u)\leq\varepsilon$ for $u\geq T_\varepsilon$. For $t>T_\varepsilon$, let us write

$$B(t)=e^{-A(t)}\int_{t_0}^{T_\varepsilon}a(u)e^{A(u)}b(u)du+e^{-A(t)}\int_{T_\varepsilon}^{t}a(u)e^{A(u)}b(u)du$$
$$\leq e^{-A(t)}\int_{t_0}^{T_\varepsilon}a(u)e^{A(u)}b(u)du+\varepsilon.$$

Since $a \notin \mathrm{L}^1([t_0, +\infty[)$, then $\lim_{t\to+\infty} e^{-A(t)} = 0$, we get

$$\limsup_{t\to+\infty} B(t) \leq \varepsilon.$$

This being true for any $\varepsilon > 0$, we infer that $\lim_{t\to+\infty} B(t) = 0$, which gives the claim. □

Lemma B.3. *Under hypothesis* (H_γ)*, then*

$$\int_{t_0}^{\infty} \frac{ds}{\Gamma(s)} = +\infty.$$

Proof. Let $q(t) \stackrel{\text{def}}{=} \int_t^\infty \frac{ds}{p(s)}$, since $\int_{t_0}^\infty \frac{ds}{p(s)} < +\infty$, then $\lim_{t\to\infty} q(t) = 0$ and $q'(t) = -\frac{1}{p(t)}$. On the other hand

$$\int_{t_0}^{\infty} \frac{ds}{\Gamma(s)} = -\int_{t_0}^{\infty} \frac{q'(t)}{q(t)} = \ln(q(t_0)) - \lim_{t\to\infty} \ln(q(t)) = +\infty.$$

□

B.2 On stochastic processes

Let us recall some elements of stochastic analysis. Throughout the paper, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\{\mathcal{F}_t | t \geq 0\}$ is a filtration of the $\sigma-$algebra $\mathcal{F}$. Given $\mathcal{C} \in 2^\Omega$, we will denote $\sigma(\mathcal{C})$ the $\sigma-$algebra generated by $\mathcal{C}$. We denote $\mathcal{F}_\infty \stackrel{\text{def}}{=} \sigma\left(\bigcup_{t\geq 0} \mathcal{F}_t\right) \in \mathcal{F}$.

The expectation of a random variable $\xi : \Omega \to \mathbb{H}$ is denoted by

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

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 indicator function of an event $E \in \mathcal{F}$ is denoted by

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

An $\mathbb{H}$-valued stochastic process starting at $t_0 \geq 0$ is a function $X : \Omega \times [t_0, +\infty[\to \mathbb{H}$. It is said to be continuous if $X(\omega, \cdot) \in C([t_0, +\infty[; \mathbb{H})$ for almost all $\omega \in \Omega$. We will denote $X(t) \stackrel{\text{def}}{=} X(\cdot, t)$. We are going to study SDEs and SDIs, and in order to ensure the uniqueness of a solution, we introduce a relation over stochastic processes. Two stochastic processes $X, Y : \Omega \times [t_0, T] \to \mathbb{H}$ are said to be equivalent if $X(t) = Y(t)$, $\forall t \in [t_0, T]$, $\mathbb{P}$-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 $X : \Omega \times [t_0, +\infty[\to \mathbb{H}$ is progressively measurable if for every $t \geq t_0$, the map $\Omega \times [t_0, t] \to \mathbb{H}$ defined by $(\omega, s) \to X(\omega, s)$ is $\mathcal{F}_t \otimes \mathcal{B}([t_0, t])$-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 $t \geq t_0$. 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:

$$S^0_{\mathbb{H}}[t_0,T] \stackrel{\text{def}}{=} \{X : \Omega \times [t_0,T] \to \mathbb{H},\ X \text{ is a prog. measurable cont. stochastic process}\} \Big/ \mathcal{R}.$$

Set $S^0_{\mathbb{H}}[t_0] \stackrel{\text{def}}{=} \bigcap_{T \geq t_0} S^0_{\mathbb{H}}[t_0,T]$. For $\nu > 0$, we define $S^\nu_{\mathbb{H}}[t_0,T]$ as the subset of processes $X(t)$ in $S^0_{\mathbb{H}}[t_0,T]$ such that

$$S^\nu_{\mathbb{H}}[t_0,T] \stackrel{\text{def}}{=} \left\{ X \in S^0_{\mathbb{H}}[t_0,T] :\ \mathbb{E}\left(\sup_{t \in [t_0,T]} \|X_t\|^\nu \right) < +\infty \right\}.$$

We define $S^\nu_{\mathbb{H}}[t_0] \stackrel{\text{def}}{=} \bigcap_{T \geq t_0} S^\nu_{\mathbb{H}}[t_0,T]$.

Following the notation of [38, Section 2.1.2], we say that W_t is a $\mathbb{K}$-valued cylindrical Brownian motion defined on on the filtered space $(\Omega, \mathcal{F}, \mathcal{F}_t, \mathbb{P})$ if:

1. For an arbitrary $t \geq 0$, the mapping $W_t : \mathbb{K} \to \mathrm{L}^2(\Omega;\mathbb{R})$ is linear;
2. For an arbitrary $k \in \mathbb{K}$, $W_t(k)$ is an $\mathcal{F}_t$ Brownian motion;
3. For arbitrary $k, k' \in \mathbb{K}$ and $t \geq 0$, $\mathbb{E}[W_t(k)W_t(k')] = t\langle k, k'\rangle_{\mathbb{K}}$.

Remark B.4. There is no $\mathbb{K}$-valued process $\tilde{W}_t$ such that:

$$W_t(k)(\omega) = \langle \tilde{W}_t(\omega), k\rangle_{\mathbb{K}}.$$

However, if Q is a non-negative definite symmetric trace-class operator on $\mathbb{K}$, then a $\mathbb{K}-$valued $Q-$Brownian motion can be defined (see e.g. [38, Definition 2.6], [27, Section 4.1]). Moreover, if $\mathbb{K} = \mathbb{R}^m$ then $W_t(k) = \langle \tilde{W}_t, k\rangle_{\mathbb{K}}$, where $\tilde{W}_t$ corresponds to the standard m-dimensional Brownian motion.

Besides, let $G : \Omega \times \mathbb{R}_+ \to \mathcal{L}_2(\mathbb{K};\mathbb{H})$ be a measurable and $\mathcal{F}_t$-adapted process, then we can define $\int_0^t G(s)dW(s)$ which is the stochastic integral of G, and we have that $G \to \int_0^\cdot G(s)dW(s)$ is an isometry between the measurable and $\mathcal{F}_t-$adapted $\mathcal{L}_2(\mathbb{K};\mathbb{H})-$valued processes and the space of $\mathbb{H}$-valued continuous square-integrable martingales (see [38, Theorem 2.4]).

References

[1] C. Alecsa, S.C. László, and T. Pinta. An extension of the second order dynamical system that models Nesterov's convex gradient method. *Appl. Math. Optim.*, 84:1687–1716, 2021.

[2] Z. Allen-Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. *J. Mach. Learn. Res.*, 18:Paper No. 221, 51, 2017.

[3] V. Apidopoulos, J-F. Aujol, and Ch. Dossal. The differential inclusion modeling the FISTA algorithm and optimality of convergence rate in the case $b \leq 3$. *SIAM J. Optim.*, 28(1):551—574, 2018.

[4] M. Assran and M. Rabbat. On the convergence of Nesterov's accelerated gradient method in stochastic settings. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 410–420, 2020.

[5] H. Attouch, A. Balhag, Z. Chbani, and H. Riahi. Fast convex optimization via inertial dynamics combining viscous and Hessian-driven damping with time rescaling. *Evol. Equ. Control Theory*, 11(2):487–514, 2022.

[6] H. Attouch, R.I. Bot, and D-K. Nguyen. Fast convex optimization via time scale and averaging of the steepest descent. *arXiv:2208.08260*, 2022.

[7] H. Attouch and A. Cabot. Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity. *Journal of Differential Equations*, 263-9:5412–5458, 2017.

[8] H. Attouch, A. Cabot, Chbani Z., and H. Riahi. Accelerated forward-backward algorithms with perturbations: Application to Tikhonov regularization. *J. Optim. Theory Appl.*, 179:1–36, 2018.

[9] H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. Convergence of iterates for first-order optimization algorithms with inertia and Hessian driven damping. *Optimization*, 2021.

[10] H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. First-order optimization algorithms via inertial systems with Hessian driven damping. *Math. Program.*, 193(1, Ser. A):113–155, 2022.

[11] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Math. Program. Ser. B*, 168:123–175, 2018.

[12] H. Attouch, Z. Chbani, and H. Riahi. Rate of convergence of the Nesterov accelerated gradient method in the subcritical case $\alpha \leq 3$. *ESAIM-COCV*, 25(2), 2019.

[13] H. Attouch, J. Fadili, and V. Kungurtsev. On the effect of perturbations in first-order optimization methods with inertia and Hessian driven damping. *Evolution equations and Control*, 12(1):71–117, 2023.

[14] H. Attouch, J. Fadili, and V. Kungurtsev. The stochastic ravine accelerated gradient method with general extrapolation coefficients. arXiv:2403.04860, 2024.

[15] H. Attouch, X. Goudou, and P. Redont. The heavy ball with friction method. i- the continuous dynamical system. *Communications in Contemporary Mathematics*, 2 (1), 2011.

[16] H. Attouch and J. Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $\frac{1}{k^2}$. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

[17] H. Attouch, J. Peypouquet, and P. Redont. Fast convex optimization via inertial dynamics with Hessian driven damping. *J. Differential Equations*, 261(10):5734–5783, 2016.

[18] A. Barakat, P. Bianchi, W. Hachem, and S. Schechtman. Stochastic optimization with momentum: Convergence, fluctuations, and traps avoidance. *Journal of statistics*, 15 (2):3892–3947, 2021.

[19] J. Bolte, T.P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165:471–507, 2016.

[20] 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.

[21] A. Cabot. Asymptotics for a gradient system with memory term. *Proceedings of the American Mathematical Society*, 137 (9):3013–3024, 2009.

[22] A. Cabot, H. Engler, and Gadat S. On the long time behavior of second order differential equations with asymptotically small dissipation. *Transactions of the American Mathematical Society*, 361 (11):5983–6017, 2009.

[23] C. Castera, H. Attouch, J. Fadili, and P. Ochs. Continuous Newton-like methods featuring inertia and variable mass. *SIAM Journal on Optimization*, 34(1):251–277, 2024.

[24] C. Castera, J. Bolte, C. Févotte, and E. Pauwels. An inertial Newton algorithm for deep learning. *J. Mach. Learn. Res.*, 22:1–31, 2021.

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

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

[27] Giuseppe Da Prato and Jerzy Zabszyk. Stochastic equations in infinite dimensions. *Cambridge University Press*, 1992.

[28] A. S. Dalalyan and A. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *arXiv:1710.00095v3*, 2018.

[29] A.S. Dalalyan, L. Riou-Durand, and A.G. Karagulyan. Bounding the error of discretized Langevin algorithms for non-strongly log-concave targets. *J. Mach. Learn. Res.*, 23:235:1–235:38, 2019.

[30] M. Dambrine, Ch. Dossal, B. Puig, and A. Rondepierre. Stochastic differential equations for modeling first order optimization methods. *SIAM J. Optim.*, 2022.

[31] A. Defazio and S. Jelassi. Adaptivity without compromise: a momentumized, adaptive, dual averaged gradient method for stochastic optimization. *J. Mach. Learn. Res.*, 23:Paper No. [144], 34, 2022.

[32] Ch. Dossal and J.F. Aujol. Stability of over-relaxations for the forward-backward algorithm, application to FISTA. *SIAM J. Optim.*, 25:2408–2433, 2015.

[33] D. Driggs, M. J. Ehrhardt, and C.B. Schönlieb. Accelerating variance-reduced stochastic gradient methods. *Math. Program.*, 191(2, Ser. A):671–715, 2022.

[34] X. Fontaine, V. De Bortoli, and A. Durmus. Convergence rates and approximation results for sgd and its continuous-time counterpart. *Proceedings of Machine Learning Research*, 134:1965–2058, 2021.

[35] R. Frostig, R. Ge, S. Kakade, and A. Sidford. Un-regularizing: approximate proximal point and faster stochastic algorithms for empirical risk minimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 2540–2548, 2015.

[36] S. Gadat and F. Panloup. Long time behaviour and stationary regime of memory gradient diffusions. *Annales de l'Institut Henri Poincaré - Probabilités et Statistiques*, 2014.

[37] S. Gadat, F. Panloup, and S. Saadane. Stochastic heavy ball. *Electron. J. Stat.*, 12(1):461–529, 2018.

[38] L. Gawarecki and V. Mandrekar. Stochastic differential equations in infinite dimensions. *Springer*, 2011.

[39] X. Goudou and J. Munier. Asymptotic behavior of solutions of a gradient-like integrodifferential Volterra inclusion. *Adv. Math. Sci. Appl.*, 15(2):509–525, 2005.

[40] A. Hamadouche, Y. Wu, A. M. Wallace, and J.F. C. Mota. Sharper bounds for proximal gradient algorithms with errors. *SIAM Journal on Optimization*, 34(1):278–305, 2024.

[41] A. Haraux and M.A. Jendoubi. On a second order dissipative ODE in Hilbert space with an integrable source term. *Acta Math. Sci.*, 32:155–163, 2012.

[42] W. Hu, C.J. Li, L. Li, and J-G. Lui. On the diffusion approximation of nonconvex stochastic gradient descent. *Ann. Math. Sci.Appl.*, 4(1):3–32, 2019.

[43] W. Hu, C.J. Li, and W. Su. On the global convergence of a randomly perturbed dissipative nonlinear oscillator. arXiv:1712.05733, 2017.

[44] W. Hu, C.J. Li, and X. Zhou. On the global convergence of continuous-time stochastic heavy-ball method for nonconvex optimization. In Chaitanya Baru, Jun Huan, and Latifur Khan, editors, *Proceedings - 2019 IEEE International Conference on Big Data*, pages 94–104, December 2019.

[45] P. Jain, P. Netrapalli, S. M. Kakade, R. Kidambi, and A. Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *J. Mach. Learn. Res.*, 18:Paper No. 223, 42, 2017.

[46] Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.

[47] M. Laborde and A. Oberman. A lyapunov analysis for accelerated gradient methods: from deterministic to stochastic case. In S. Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 602–612. PMLR, 26–28 Aug 2020.

[48] G. Lan. First-order and stochastic optimization methods for machine learning. *Springer Series in the Data Sciences*, pages xiii+582, 2020.

[49] J. Latz. Analysis of stochastic gradient descent in continuous time. *Statistics and Computing*, pages 31–39, 2021.

[50] T. Le. Nonsmooth nonconvex stochastic heavy ball. *Journal of Optimization Theory and Applications*, 201 (2):699–719, 2024.

[51] Q. Li, C. Tai, and E. Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, 2017.

[52] X. Li, Z. Shen, L. Zhang, and N. He. A Hessian-aware stochastic differential equation for modelling SGD. *arXiv:2405.18373*, 2024.

[53] Z. Li, S. Malladi, and S. Arora. On the validity of modeling SGD with stochastic differential equations. In *Advances in Neural Information Processing Systems*, 2021.

[54] H. Lin, J. Mairal, and Z. Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *J. Mach. Learn. Res.*, 18:Paper No. 212, 54, 2017.

[55] N. Loizou and P. Richtárik. Momentum and stochastic momentum for stochastic gradient, Newton, proximal point and subspace descent methods. *Comput. Optim. Appl.*, 77(3):653–710, 2020.

[56] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

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

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

[59] Y-A. Ma, N. Chatterji, X. Cheng, N. Flammarion, P. Bartlett, and M. I. Jordan. Is there an analog of Nesterov acceleration for MCMC? *Bernoulli*, 27 (3):1942–1992, 2021.

[60] S. Mandt, M. Hoffman, and D. Blei. A variational analysis of stochastic gradient algorithms. In *International Conference on Machine Learning*, 2016.

[61] X. Mao. Stochastic differential equations and applications. *Elsevier*, 2007.

[62] R. Maulen-Soto, J. Fadili, and H. Attouch. An SDE perspective on stochastic convex optimization. *Mathematics of Operations Research (revised)*, 2022. arXiv:2207.02750.

[63] R. Maulen-Soto, J. Fadili, and H. Attouch. Tikhonov regularization for stochastic non-smooth convex optimization in Hilbert spaces. *Open Journal of Mathematical Optimization (submitted)*, 2024. arXiv:2403.06708.

[64] R. May. Asymptotic for a second order evolution equation with convex potential and vanishing damping term. *Turkish Journal of Mathematics*, 41, 2017.

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

[66] M. Muehlebach and M. I. Jordan. Optimization with momentum: dynamical, control-theoretic, and symplectic perspectives. *J. Mach. Learn. Res.*, 22:Paper No. 73, 50, 2021.

[67] Y. Nesterov. A method of solving a convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

[68] B. Öksendal. Stochastic differential equations. *Springer*, 2003.

[69] A. Orvieto, J. Kohler, and A. Lucchi. The role of memory in stochastic optimization. *Proceeding of Machine Learning Research*, 115:356–366, 2020.

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

[71] G. A. Pavliotis. Stochastic processes and applications. *Springer*, 2014.

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

[73] B. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 1964.

[74] H. Robins and S. Monro. A stochastic approximation method. *Ann. Math. Statist. 22*, pages 400–407, 1951.

[75] R.T. Rockafellar. Convex analysis. *Princeton univerisity press*, 28, 1997.

[76] M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *NIPS'11*, 25th Annual Conference, 2011.

[77] B. Shi, S.S. Du, M.I. Jordan, and Su W.J. Understanding the acceleration phenomenon via high resolution differential equations. *Math. Program.*, 2021.

[78] B. Shi, W. J. Su, and M. I. Jordan. On learning rates and Schrödinger operators. *Journal of Machine Learning Research*, 24:1–53, 2023.

[79] 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.

[80] W. Su, S. Boyd, and E.J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17:1–43, 2016.

[81] S. Villa, S. Salzo, and L. Baldassarres. Accelerated and inexact forward-backward. *SIAM J. Optim.*, 23:1607–1633, 2013.

[82] Z. Xie, I. Sato, and M. Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2021.

[83] B. Yan. *Theoretical Analysis for Convex and Non-Convex Clustering Algorithms*. ProQuest LLC, Ann Arbor, MI, 2018. Thesis (Ph.D.)–The University of Texas at Austin.