Stochastic Inertial Dynamics Via Time Scaling and Averaging

Rodrigo Maulen-Soto, Jalal Fadili, Hedy Attouch, Peter Ochs

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 optimization, Inertial (sub)gradient systems, Convex optimization, Stochastic Differential Equation, Stochastic Differential Inclusion, Tikhonov regularization, Time-dependent viscosity, Łojasiewicz inequality, KL inequality, Convergence rate, Asymptotic behavior.

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

1 Introduction

1.1 Problem Statement

Let us consider the minimization problem

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

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

$$\begin{cases} f : \mathbb{H} \to \mathbb{R} \text{ is continuously differentiable and convex with } L\text{-Lipschitz continuous gradient;} \\ g : \mathbb{H} \to \mathbb{R} \cup \{\pm\infty\} \text{ is proper, lower semi-continuous (lsc) and convex;} \\ \mathcal{S}_F \overset{\text{def}}{=} \operatorname{argmin}(F) \neq \emptyset. \end{cases} \tag{H_0}$$

*Normandie Université, ENSICAEN, UNICAEN, CNRS, GREYC, France. E-mail: rodrigo.maulen@ensicaen.fr
†Normandie Université, ENSICAEN, UNICAEN, CNRS, GREYC, France. E-mail: Jalal.Fadili@ensicaen.fr
‡IMAG, CNRS, Université Montpellier, France. E-mail: hedy.attouch@umontpellier.fr
§Department of Mathematics and Computer Science, Saarland University, Germany, E-mail: ochs@math.uni-sb.de

To solve (P) when $g \equiv 0$, a fundamental dynamic is the gradient flow system:

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

This dynamic is known to yield a convergence rate of $\mathcal{O}(t^{-1})$ (in fact even $o(t^{-1})$) on the values. Second-order inertial dynamical systems have been introduced to provably accelerate the convergence behaviour. Among them, the Inertial System with Implicit Hessian Damping (ISIHD) is the following differential equation starting at $t_0 > 0$ with initial condition $x_0, v_0 \in \mathbb{H}$:

$$\begin{cases} \ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla f(x(t) + \beta(t)\dot{x}(t)) = 0, \quad t > t_0; \\ x(t_0) = x_0, \quad \dot{x}(t_0) = v_0. \end{cases} \tag{ISIHD}$$

where $\gamma, \beta : [t_0, +\infty[\to \mathbb{R}_+$. (ISIHD) is inspired by [1]; see also [2, 3]. Following the physical interpretation of this ODE, we call the non-negative parameters γ and β as the viscous and geometric damping parameters, respectively. The rationale behind the use of the term "implicit" comes from a a Taylor expansion of the gradient term (as $t \to +\infty$ we expect $\dot{x}(t) \to 0$), which make the Hessian damping appear indirectly in (ISIHD). This ODE was found to have a smoothing effect on the energy error and oscillations [1, 2, 3].

Let us now turn the case where g is non-smooth. We are naturally led to consider the generalization of (ISIHD) to the non-smooth case, which yields the differential inclusion

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

where ∂F is the convex subdifferential of F.

In 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 dynamic (3.1) to model SGD-type algorithms; (see *e.g.* [4, 5, 6, 7, 8, 9, 10, 11, 12, 13]). In fact, 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 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$_{\text{NS}}$):

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

(S $-$ ISIHD$_{\text{NS}}$)

This SDI is defined over a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$, where $X_0, V_0 \in \mathrm{L}^\nu(\Omega; \mathbb{H})$ (for some $\nu \geq 2$) are the initial data; the diffusion (volatility) term $\sigma : [t_0, +\infty[\times\mathbb{H} \to \mathcal{L}_2(\mathbb{K}; \mathbb{H})$ is a measurable function with $\mathbb{K}$ a separable real Hilbert space; and W is a $\mathbb{K}$-valued Brownian motion (see definition in Section A.2.1). When $g \equiv 0$, we recover the stochastic counterpart of (ISIHD) as the following SDE

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

In this work, our goal is to provide a general mathematical framework for analyzing the convergence properties of ($\mathrm{S-ISIHD_{NS}}$). 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 to develop a systematic and unified way that transfers the properties of stochastic first-order dynamics recently studied by Maulen-Soto, Fadili, and Attouch [11, 13] to second-order ones. Our program will then rely on two pillars: time scaling and averaging, following the methodology recently developed by Attouch, Bot, and Nguyen in [14] in the deterministic gradient 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 that 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" $\operatorname{argmin}(f)$ 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., [15, Proposition 6.1]. Clearly, as $\tilde{\sigma} \to 0^+$, $\pi_{\tilde{\sigma}}$ gets concentrated around $\operatorname{argmin}(f) \times \{0_{\mathbb{H}}\}$, with $\lim_{\tilde{\sigma}\to 0^+} \pi_{\tilde{\sigma}}(\operatorname{argmin}(f) \times \{0_{\mathbb{H}}\}) = 1$, see Section 1.3 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, and some guarantees to a "noise-dominated region" will also be provided when σ is uniformly bounded.

(ISIHD) is one of the most recent developments regarding the use of second-order gradient-based dynamical systems for optimization. Let us briefly recall the steps that led to its emergence. In this regard, let us stress the importance of working with a time-dependent viscosity coefficient $\gamma(t)$. An abundant literature has been devoted to the study of inertial dynamics with time-dependent viscosity coefficient with $\beta \equiv 0$,

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

See for instance [16, 17] for general parameter γ. Most of the literature focuses on the case $\gamma(t) = \frac{\alpha}{t}$, originating from the seminal work of Su, Boyd, and Candès [18] who showed the rate of convergence $\mathcal{O}(1/t^2)$ of the values for $\alpha = 3$, thus making the link with the accelerated gradient method of Nesterov [19]. Since then, an important body of literature has been devoted to this important case and the subtle tuning of the parameter α. Indeed, α must be taken greater than or equal to 3 for getting the rate of convergence $\mathcal{O}\left(1/t^2\right)$ of the values (see [20]), and $\alpha > 3$ provides an even better rate of convergence with little-o instead of big-$\mathcal{O}$ (see [21, 22]). On the other hand, $\alpha < 3$ necessarily leads to a slower rate $\mathcal{O}\left(1/t^{2\alpha/3}\right)$ [23, 24].

Another remarkable instance of (IGS_γ) arises when $\gamma(t)$ is a constant function. In this scenario, the resulting dynamic corresponds to the well-known Heavy Ball with friction (HBF) method, first introduced (in its discrete and continuous form) by Polyak in [25] where it was shown a linear rate of convergence for the trajectory when the objective function is strongly convex (we also refer to the insights given in [26]). There is also a stochastic version of this method, we refer to [27]) for further details.

However, 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

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

where γ and β are, the already presented, damping parameters, and b is a time scale parameter. This dynamic is the explicit version of (ISIHD). The time discretization of this system has been studied by Attouch, Chbani, Fadili, and Riahi [28]. It provides a rich family of first-order methods for minimizing f. At first glance, the presence of the Hessian may seem to entail numerical difficulties. However, this is not the case as the Hessian intervenes in the above ODE in the form $\nabla^2 f(x(t))\dot{x}(t)$, which is nothing but the time derivative of $t \mapsto \nabla f(x(t))$. This explains why the time discretization of this dynamic provides first-order algorithms. On the contrary, the time-continuous dynamics can be argued to be truly of second-order nature, i.e., close to Newton's and Levenberg-Marquardt's dynamics [29]. This understanding suggests that (ISIHD) may represent the nature of first-order algorithms better than (ISEHD). However, 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 $M(t)$). Therefore, it is meaningless to ask for the time derivative of $\nabla f(X(t)) + M(t)$ because (non-constant) martingales are not differentiable a.s.. This is why we focus on the implicit form of the Hessian-driven damping.

1.2 Contributions

Our main contributions pertain to the solution trajectories of the dynamics (S − ISIHD) and (S − ISIHD$_{\text{NS}}$) under integrability conditions on the noise. They 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 $\beta(t)$. For this analysis, we transfer the results from the Lyapunov analysis of the first-order in-time stochastic (sub-)gradient system studied in [11, 13] 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 in almost sure sense (see Corollary 3.4), which corresponds to the known result for the accelerated gradient method of Nesterov in the deterministic case.
- We then turn to providing a local analysis with a local linear convergence rate under the Polyak-Łojasiewicz inequality (See Theorem 3.6). 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.
- 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 condition on the objective function appears unavoidable, at least in this proof way. It is also worth mentioning 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) martingales are a.s. not differentiable.

1.3 Relation to prior work

Kinetic diffusion dynamics for sampling Let's 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 have been studied; see [30, 31, 32] and references therein.

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, 3, 33, 34, 35, 20, 36, 37, 38, 39, 40, 41]). We are not aware of any such work in the stochastic case. Only a few papers have been devoted to studying the second-order in-time inertial stochastic gradient systems with viscous damping, *i.e.* stochastic versions of (IGS_γ), either with vanishing damping $\gamma(t) = \alpha/t$ or constant damping $\gamma(t)$ (stochastic HBF); see e.g. [12, 42, 43]. For instance, [12] 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 of the objective[1]. The corresponding stochastic algorithms for these two choices of γ, whose mathematical formulation and analysis is simpler, have been the subject of an active research work; see *e.g.* [44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57].

Time scaling and averaging A SDI to solve (P) has been thoroughly studied in [13]; see also [11] for the SDE case with $g \equiv 0$ recalled in (2.2). This SDI has the form

$$\begin{cases} dX(t) \in -\partial F(X(t))dt + \sigma(t, X(t))dW(t), \quad t \geq t_0, \\ X(t_0) = X_0. \end{cases} \tag{1.1}$$

The authors in [14] 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 [14] 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 [13] for (1.1) to the SDI ($\text{S} - \text{ISIHD}_{\text{NS}}$). This avoids in particular to go through an intricate and a dedicated Lyapunov analysis for ($\text{S} - \text{ISIHD}_{\text{NS}}$). A local convergence analysis becomes also easily accessible through this lens while it is barely possible otherwise. We also specialized our results to a 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 authors of [58, 59] propose time scaling and tricky change of variables to show

[1] 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.

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 [60] for more general gradient systems with memory terms involving kernels. This gives rise to an integro-differential equation. The asymptotic behaviour of the dynamic associated to this equation and the equivalent second-order dynamic have been investigated in [58, 59]. A stochastic version of this integro-differential equation has been studied in [43] where the long time behaviour of the resulting process, in particular its invariant distribution and occupation measure, was investigated under ellipticity assumptions on f and σ and proper behaviour 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.

1.4 Organization of the paper

Section 2 introduces notations, and it reviews some definitions and results of convex and stochastic analysis that will be used in the paper. 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. Technical lemmas and theorems that are needed throughout the paper will be collected in Appendix A.

2 Notation and Preliminaries

2.1 Notation

We will use the following shorthand notations: Given $n \in \mathbb{N}$, $[n] \overset{\text{def}}{=} \{1, \dots, n\}$. 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

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

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 $M \in \mathcal{L}_2(\mathbb{K}; \mathbb{H})$,

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

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 $[f \leq r] \overset{\text{def}}{=} \{x \in \mathbb{H} : f(x) \leq r\}$. 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 [61].

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 Other assumptions

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 behaviour of our SDE's and SDI's (in particular ($\mathrm{S}-\mathrm{ISIHD}$) and ($\mathrm{S}-\mathrm{ISIHD}_{\mathrm{NS}}$)) as the diffusion term vanishes when $t \to +\infty$. Therefore, throughout the paper, we assume that the diffusion (volatility) term σ satisfies:

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

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 ($\mathrm{S}-\mathrm{ISIHD}$) and ($\mathrm{S}-\mathrm{ISIHD}_{\mathrm{NS}}$). 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) 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 denote:

$$p(t) \stackrel{\text{def}}{=} \exp\left(\int_{t_0}^t \gamma(s)ds\right).$$

If

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

We define $\Gamma : [t_0, +\infty[\to \mathbb{R}_+$ by

$$\Gamma(t) \stackrel{\text{def}}{=} p(t) \int_t^\infty \frac{ds}{p(s)}. \tag{2.1}$$

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

We denote

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

Before we delve into our core contributions, it is important to note that we will require some specific results gleaned from [11, 13]. These are the subject of the following subsections.

2.3 Results for the first-order gradient SDE

The stochastic version of (GF) where f is smooth has been well studied and documented in [11]. We recall two main results of that paper on which we will build our study. 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 ([11, Theorem 3.1]). *Consider f and σ that satisfy Assumptions (H_0') and (H). Let $\nu \geq 2$ and consider the SDE:*

$$\begin{cases} dX(t) = -\nabla f(X(t))dt + \sigma(t, X(t))dW(t), \\ X(t_0) = X_0, \end{cases} \tag{2.2}$$

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 A.2.1 for the notation) of (2.2). 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 ([11, Theorem 3.4]). *Let $\nu \geq 2$ and consider the SDE (2.2) with initial data $X_0 \in \mathrm{L}^\nu(\Omega; \mathbb{H})$, where f and σ satisfy Assumptions (H_0) and (H). 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.*.

2.4 Results for the first-order stochastic non-smooth case

The far more intricate non-smooth case has been very recently studied in [13]. Let F, σ satisfy (H_0) and (H). We consider the stochastic differential inclusion

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

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

Definition 2.5. A solution of (SDI) 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 $t > t_0$,

$$\begin{cases} X(t) & = X_0 - \int_{t_0}^t \nabla f(X(s))ds - \eta(t) + \int_{t_0}^t \sigma(s, X(s))dW(s), \\ X(t_0) & = X_0. \end{cases} \tag{2.3}$$

For brevity, we sometimes omit the process η and say that X is a solution of (SDI), 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 A.2.1.

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

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

where g_λ is the Moreau envelope of g with parameter $\lambda > 0$. Under the integrability condition that for every $T > t_0$,

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

it was shown in [62] that there exists a couple (X, η) of stochastic processes which is a solution of (SDI) in the sense of Definition 2.5, and moreover, for every $T > t_0$,

$$\lim_{\lambda \downarrow 0} \mathbb{E}\left(\sup_{t \in [t_0, T]} \|X_\lambda(t) - X(t)\|^2 \right) = 0 \quad \text{and} \quad \lim_{\lambda \downarrow 0} \mathbb{E}\left(\sup_{t \in [t_0, T]} \|\eta_\lambda(t) - \eta(t)\|^2 \right) = 0,$$

where $\eta_\lambda(t) = \int_{t_0}^t \nabla g_\lambda(X_\lambda(s))ds$. In addition, the uniqueness of this solution was proved in [13].

Remark 2.6. Condition (H$_\lambda$) is satisfied under different conditions, some examples are mentioned in [62]. One case where this condition holds is when ∂g is 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.

Now, we can show the main results we have for the dynamic (SDI), this was shown by Maulen-Soto, Fadili, Attouch in [13].

Theorem 2.7. *Consider $F = f + g$ and σ satisfying (H$_0$) and (H). Suppose further that g verifies (H$_\lambda$). Let $\nu \geq 2$, $t_0 \geq 0$, and consider the dynamic (SDI) with initial data $X_0 \in \mathrm{L}^\nu(\Omega; \mathbb{H})$. Then, (SDI) has a unique solution $(X, \eta) \in S^\nu_{\mathbb{H}}[t_0] \times C^1([t_0, +\infty[; \mathbb{H})$.*

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 $\forall x^\star \in \mathcal{S}_F$, $\nabla f(x^\star)$ is constant, 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 w-$\lim_{t \to +\infty} X(t) = X^\star$.

Tikhonov regularization. Let's now turn to a Tikhonov regularization of (SDI), *i.e.*,

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

Solution existence and uniqueness for (SDI $-$ TA) is proved in [13, Theorem 3.3]. We also have the following result from [13].

Theorem 2.8 ([13]). *Let $\nu \geq 2$ and consider the dynamic (SDI $-$ 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). Furthermore, assume that g satisfies (H_λ). Then, there exists a unique solution $X \in S^\nu_{\mathbb{H}}[t_0]$ of (SDI $-$ 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 $F_\varepsilon(x) \overset{\text{def}}{=} F(x) + \frac{\varepsilon}{2}\|x\|^2$. 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 the strong convergence of the trajectory to the minimal norm solution.

We show the following version of Itô's formula for a multi-valued drift, which plays a central role in the study of SDI's.

Proposition 2.9. *Let $\nu \geq 2$, $X_0 \in \mathrm{L}^\nu(\Omega; \mathbb{H})$ and $\mathcal{F}_0$-measurable, $t_0 \geq 0$, and $F = f + g$ that satisfies (H_0), $\sigma : [t_0, +\infty[\times \mathbb{H} \to \mathcal{L}_2(\mathbb{K}; \mathbb{H})$ measurable functions. We consider $(X, \eta) \in S^\nu_{\mathbb{H}}[t_0] \times C^1([t_0, +\infty[; \mathbb{H})$ be the unique solution of*

$$\begin{cases} dX(t) & \in -\partial F(X(t))dt + \sigma(t, X(t))dW(t), \quad t > t_0, \\ X(t_0) & = X_0. \end{cases} \tag{2.4}$$

Let $\phi : [t_0, +\infty[\times \mathbb{H} \to \mathbb{R}$ be such that $\phi(\cdot, x) \in C^1([t_0, +\infty[)$ for every $x \in \mathbb{H}$ and $\phi(t, \cdot) \in C^2(\mathbb{H})$ for every $t \geq t_0$. Then the process

$$\tilde{Y}(t) = \phi(t, X(t)),$$

is an Itô process such that for all $t \geq t_0$

$$\begin{aligned} \tilde{Y}(t) = \tilde{Y}(t_0) &+ \int_{t_0}^t \frac{\partial \phi}{\partial t}(s, X(s))ds - \int_{t_0}^t \left\langle \nabla\phi(s, X(s)), \nabla f(X(s)) + \eta'(s) \right\rangle ds \\ &+ \int_{t_0}^t \langle \sigma^\star(s, X(s)) \nabla\phi(s, X(s)), dW(s) \rangle + \frac{1}{2} \int_{t_0}^t \mathrm{tr}\left(\sigma(s, X(s))\sigma^\star(s, X(s))\nabla^2\phi(s, X(s))\right) ds, \end{aligned} \tag{2.5}$$

where $\eta'(t) \in \partial g(X(t))$ a.s. for almost all $t \geq t_0$. Moreover, if $\mathbb{E}[\tilde{Y}(t_0)] < +\infty$, and if for all $T > t_0$

$$\mathbb{E}\left(\int_{t_0}^T \|\sigma^\star(s, X(s))\nabla\phi(s, X(s))\|^2 ds\right) < +\infty,$$

then $\int_{t_0}^{t} \langle \sigma^\star(s, X(s)) \nabla \phi(s, X(s)), dW(s) \rangle$ *is a square-integrable continuous martingale and*

$$\mathbb{E}[\tilde{Y}(t)] = \mathbb{E}[\tilde{Y}(t_0)] + \mathbb{E}\left(\int_{t_0}^{t} \frac{\partial \phi}{\partial t}(s, X(s)) ds \right) - \mathbb{E}\left(\int_{t_0}^{t} \left\langle \nabla \phi(s, X(s)), \nabla f(X(s)) + \eta'(s) \right\rangle ds \right) + \frac{1}{2} \mathbb{E}\left(\int_{t_0}^{t} \operatorname{tr}\left(\sigma(s, X(s)) \sigma^\star(s, X(s)) \nabla^2 \phi(s, X(s)) \right) ds \right). \quad (2.6)$$

Proof. The existence and uniqueness of a solution $(X, \eta) \in S^{\nu}_{\mathbb{H}}[t_0] \times C^1([t_0, +\infty[; \mathbb{H})$ of (2.4) was proved in [13, Theorem 3.3] following the work of [62]. This solution satisfies (by definition) the following equation:

$$\begin{cases} X(t) & = X_0 + \int_{t_0}^{t} -[\nabla f(X(s)) + \eta'(s)] ds + \int_{t_0}^{t} \sigma(s, X(s)) dW(s), \quad t > t_0, \\ X(t_0) & = X_0. \end{cases} \quad (2.7)$$

and $\eta'(s) \in \partial g(X(s))$ for almost all $t \geq 0$ a.s.. Then, (2.7) is an Itô process with drift $s \mapsto -[\nabla f(X(s)) + \eta'(s)]$ and diffusion $s \mapsto \sigma(s, X(s))$. Consequently, we can apply the classical Itô's formula (see [63, Section 2.3]) to obtain the desired. □

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 (SDI) 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 (SDI) directly carry over to obtain results on the convergence behaviour of ($\mathrm{S-ISIHD_{NS}}$) without passing through a dedicated Lyapunov analysis.

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 (SDI) in s, starting at s_0, with diffusion term σ_1 under hypotheses (H_0) and (H). 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 $\sigma_{1_\infty}(s) \overset{\text{def}}{=} \sup_{x \in \mathbb{H}} \|\sigma_1(s, x)\|_{\mathrm{HS}}$. We rewrite (SDI) adapted to our case,

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

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 $Y(t) \overset{\text{def}}{=} Z(s)$ and t_0 be such that $s_0 = \tau(t_0)$. By the chain rule and [64, Theorem 8.5.7], we have

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

Consider the smooth case, i.e. when $g \equiv 0$ and the hypotheses of Theorem 2.4 ($f \in C^2_L(\mathbb{H})$ 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 [14] the following approach, which is called averaging.

Consider (3.2) and let $X, V : \Omega \times [t_0, +\infty[\to \mathbb{H}$ be two stochastic processes such that:

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

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 $Z_0 \overset{\text{def}}{=} X_0 + \tau'(t_0)V_0$ 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 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 [65, Proposition 2.2]) by (H_γ). For this particular selection of τ, and defining $\tilde{\sigma}_1(t, \cdot) \overset{\text{def}}{=} \frac{\sigma_1(\tau(t), \cdot)}{\sqrt{\Gamma(t)}}$, 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

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

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

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

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.,

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

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 ($\mathrm{S}-\mathrm{ISIHD}_{\mathrm{NS}}$) to a random variable taking values in the set of minimizers of F. When $g \equiv 0$, we also provide convergence rates.

Theorem 3.1. *Let $\nu \geq 2$ and consider the dynamic ($\mathrm{S}-\mathrm{ISIHD}_{\mathrm{NS}}$) 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). 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 ($\mathrm{S}-\mathrm{ISIHD}_{\mathrm{NS}}$). 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 $\theta(t) \overset{\text{def}}{=} \int_{t_0}^t \Gamma(u)du$ and $\tilde{\sigma}(s,\cdot) \overset{\text{def}}{=} \sigma(\theta^{-1}(s),\cdot)\sqrt{\Gamma(\theta^{-1}(s))}$. 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 ($\mathrm{S}-\mathrm{ISIHD}_{\mathrm{NS}}$) 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

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

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

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

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

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

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

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

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:

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

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 A.3, 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 A.3, we utilize Lemma A.2 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, 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.

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), and in the case where γ satisfy (H_γ), $\beta \equiv \Gamma$. Moreover, suppose that either $\mathbb{H}$ is finite dimensional or $f \in C^2(\mathbb{H})$, and*

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

where $\theta(t) \overset{\text{def}}{=} \int_{t_0}^t \Gamma(u)du$. 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 $A(t) \overset{\text{def}}{=} \int_{t_0}^t \frac{du}{\Gamma(u)}$ 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 A.2 to check that $\lim_{t\to\infty} I\left[\frac{1}{\theta}\right](t) = 0$.

Proof. We will utilize the averaging technique used in Theorem 3.1 and Jensens'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 gradient convexity inequality on f to get

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

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 $s \mapsto s\tilde{\sigma}^2_\infty(s) \in \mathrm{L}^1(\mathbb{R}_+)$, 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 3.3. *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), and in the case where γ satisfy (H_γ), $\beta \equiv \Gamma$. Moreover, suppose that $f \in C^2(\mathbb{H})$ and*

$$t \mapsto \theta(t)\Gamma^2(t)\sigma_\infty^2(t) \in \mathrm{L}^1([t_0,+\infty[),$$

where $\theta(t) \overset{\text{def}}{=} \int_{t_0}^t \Gamma(u)du$. 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.

Corollary 3.4 (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). 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 A.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 3.3 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 [66, 67, 68]) and is commonly used to prove linear convergence of gradient descent algorithms.

Definition 3.5 (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 $f \in \mathrm{P}\,\text{Ł}_\mu(\mathcal{S})$.

Theorem 3.6. *Let $\nu \geq 2$ and consider the dynamic (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). Besides, $f \in \mathrm{P}\,\text{Ł}_\mu(\mathcal{S})$ 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 $\hat{t}_\delta > t_0, K_{\mu,\delta}, C_l, C_f > 0$ 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 $f \in \mathrm{P}\,\text{Ł}_\mu(\mathcal{S})$ 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 3.7. If f is $\mu-$strongly convex, then $f \in \mathrm{P}\,\text{Ł}_\mu(\mathcal{S})$ holds on the entire space (i.e. $r = +\infty$).

Proof. Consider the dynamic (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 [11, (i-b), Theorem 4.5] (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 [11, Section 4.2].

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

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

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 3.4.

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

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

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

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

Letting $K_{\mu,\delta} \overset{\text{def}}{=} \sqrt{2\mu}\left(\frac{LC_\infty^2}{2} + C_l C_\infty\sqrt{\delta}\right)e^{\sqrt{2\mu}\left(\frac{t_0+\hat{t}_\delta}{\sqrt{2\mu}}+\hat{s}_\delta\right)} + 2\sqrt{2\mu}C_\infty$, 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 ($S-ISIHD_{NS}$), i.e., for $t > 0$,

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

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). Moreover, suppose that g satisfies (H_λ) and let $\nu \geq 2$. Consider ($\mathrm{S} - \mathrm{ISIHD}_{\mathrm{NS}} - \mathrm{TA}$) 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 ($\mathrm{S} - \mathrm{ISIHD}_{\mathrm{NS}} - \mathrm{TA}$). 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). 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 ($\mathrm{S} - \mathrm{ISIHD}_{\mathrm{NS}} - \mathrm{TA}$) 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 A.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 A.3, we utilize Lemma A.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 Practical situations

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 $\theta(t) \overset{\text{def}}{=} s_0 + \int_{t_0}^{t} \Gamma(s)ds$, 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 [13, 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), 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 $f_\varepsilon(x) \overset{\text{def}}{=} f(x) + \frac{\varepsilon}{2}\|x\|^2$, and let x_ε be the unique minimizer of f_ε. Moreover, let $s_0 > 0$ and for $s_1 > s_0$ consider,

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

where $\theta(t) \overset{\text{def}}{=} s_0 + \int_{t_0}^{t} \Gamma(u)du$. Let also $x^\star \overset{\text{def}}{=} \mathrm{P}_{\mathcal{S}}(0)$, $A(s) \overset{\text{def}}{=} \int_{s_1}^{s} \frac{du}{\Gamma(u)}$, and $t_1 \overset{\text{def}}{=} \theta^{-1}(s_1)$. 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 [13, 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 Auxiliary results

A.1 Deterministic results

Lemma A.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 A.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 $A(t) \overset{\text{def}}{=} \int_{t_0}^t a(u)du$ and $B(t) \overset{\text{def}}{=} e^{-A(t)} \int_{t_0}^t a(u)e^{A(u)}b(u)du$. 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

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

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 A.3. *Under hypothesis* (H_γ)*, then*

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

Proof. Let $q(t) \overset{\text{def}}{=} \int_t^\infty \frac{ds}{p(s)}$, 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

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

□

A.2 Stochastic results

A.2.1 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 $\mathcal{C} \in \mathcal{P}(\Omega)$, we will denote $\sigma(\mathcal{C})$ the $\sigma-$algebra generated by $\mathcal{C}$. We denote $\mathcal{F}_\infty \overset{\text{def}}{=} \sigma\left(\bigcup_{t\geq 0} \mathcal{F}_t\right) \in \mathcal{F}$.

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

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

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

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

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 $X(t) \overset{\text{def}}{=} X(\cdot, t)$. We are going to study SDE's and SDI's, 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 $S^\nu_{\mathbb{H}}[t_0] \stackrel{\text{def}}{=} \bigcap_{T\geq t_0} S^\nu_{\mathbb{H}}[t_0,T]$.

Let $I\subseteq\mathbb{N}$ be a numerable set such that $\{e_i\}_{i\in I}$ is an orthonormal basis of $\mathbb{K}$, and $\{w_i(t)\}_{i\in I,t\geq 0}$ be a sequence of independent Brownian motions defined on the filtered space $(\Omega,\mathcal{F},\mathcal{F}_t,\mathbb{P})$. The process

$$W(t)=\sum_{i\in I} w_i(t)e_i$$

is well-defined (independent from the election of $\{e_i\}_{i\in I}$) and is called a $\mathbb{K}$-valued 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 the application $G\to\int_0^\cdot G(s)dW(s)$ 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 [63, Theorem 2.3]).

Proposition A.4. *(see [69] and [70, Section 1.2]) (Burkholder-Davis-Gundy Inequality) Let $p>0$, W be a $\mathbb{K}$-valued Brownian motion defined over a filtered probability space $(\Omega,\mathcal{F},\{\mathcal{F}_t\}_{t\geq 0},\mathbb{P})$ and $g:\Omega\times\mathbb{R}_+\to\mathbb{K}$ a progressively measurable process (with our usual notation $g(t)\stackrel{\text{def}}{=}g(\cdot,t)$) such that*

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

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

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

Theorem A.5. *[71, Theorem 1.3.9] Let $\{A_t\}_{t\geq 0}$ and $\{U_t\}_{t\geq 0}$ be two continuous adapted increasing processes with $A_0=U_0=0$ a.s.. Let $\{M_t\}_{t\geq 0}$ be a real-valued continuous local martingale with $M_0=0$ a.s.. Let ξ be a nonnegative $\mathcal{F}_0$-measurable random variable. Define*

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

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

References

[1] C. Alecsa, S. László, and T. Pinta. An extension of the second order dynamical system that models Nesterovâs convex gradient method. *Appl. Math. Optim.*, 84:1687–1716, 2021.

[2] M. Muehlebach and M. I. Jordan. A dynamical systems perspective on Nesterov acceleration. In *Proceedings of the 36 th International Conference on Machine Learning*, volume 97. PMLR, 2019.

[3] Hedy Attouch, Jalal Fadili, and Vyacheslav 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.

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

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

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

[7] Bin Shi, Weijie J. Su, and Michael I. Jordan. On learning rates and Schrödinger operators. *Journal of Machine Learning Research*, 24:1–53, 2023.

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

[9] 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.

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

[11] Rodrigo Maulen S., Jalal Fadili, and Hedy Attouch. An SDE perspective on stochastic convex optimization. *arXiv:2207.02750*, 2022.

[12] M. Dambrine, C. Dossal, B. Puig, and A. Rondepierre. Stochastic differential equations for modeling first order optimization methods. *Hal*, 2022.

[13] Rodrigo Maulen-Soto, Jalal Fadili, and Hedy Attouch. Tikhonov regularization for stochastic non-smooth convex optimization in Hilbert spaces. *arXiv:2403.06708*, 2024.

[14] Hedy Attouch, Radu Ioan Bot, and Dang-Khoa Nguyen. Fast convex optimization via time scale and averaging of the steepest descent. *arXiv:2208.08260*, 2022.

[15] Grigorios A. Pavliotis. Stochastic processes and applications. *Springer*, 2014.

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

[17] Alexandre Cabot, Hans Engler, and Gadat Sébastien. 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.

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

[19] Y.E. 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.

[20] 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.

[21] Hedy Attouch and Juan 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.

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

[23] V. Apidopoulos, J.-F. Aujol, and C. 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.

[24] 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.

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

[26] Hedy Attouch, Xavier Goudou, and Patrick Redont. The heavy ball with friction method. i- the continuous dynamical system. *Communications in Contemporary Mathematics*, 2 (1), 2011.

[27] Sébastien Gadat, Fabien Panloup, and Saadane Sofiane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12:461–529, 2018.

[28] Hedy Attouch, Zaki Chbani, Jalal Fadili, and Hassan Riahi. First-order optimization algorithms via inertial systems with Hessian driven damping. *Mathematical Programming*, 193(4), 2020.

[29] Camille Castera, Hedy Attouch, Jalal Fadili, and Peter Ochs. Continuous Newton-like methods featuring inertia and variable mass. *arXiv:2301.08726*, 2023.

[30] Xiang Cheng, Niladri S. Chatterji, Peter L. Bartlett, and Michael I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. *arXiv:1707.03663*, 2017.

[31] Yi-An Ma, Niladri Chatterji, Xiang Cheng, Nicolas Flammarion, Peter Bartlett, and Michael I. Jordan. Is there an analog of Nesterov acceleration for MCMC? *Bernoulli*, 27 (3):1942–1992, 2021.

[32] Arnak S. Dalalyan, Lionel Riou-Durand, and Avetik 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.

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

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

[35] 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.

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

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

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

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

[40] H. Attouch, Z. Chbani, J. Fadili, and H. Riahi. First order optimization algorithms via inertial systems with Hessian driven damping. *Math. Program.*, 2020.

[41] 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.

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

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

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

[45] R. Frostig, S. Kakade R. Ge, 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.

[46] Prateek Jain, Praneeth Netrapalli, Sham M. Kakade, Rahul Kidambi, and Aaron 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.

[47] 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.

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

[49] Bowei 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.

[50] Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electron. J. Stat.*, 12(1):461–529, 2018.

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

[52] Maxime Laborde and Adam Oberman. A lyapunov analysis for accelerated gradient methods: from deterministic to stochastic case. In Silvia 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.

[53] Guanghui Lan. *First-order and stochastic optimization methods for machine learning*. Springer Series in the Data Sciences. Springer, Cham, [2020] ©2020.

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

[55] Derek Driggs, Matthias J. Ehrhardt, and Carola-Bibiane Schönlieb. Accelerating variance-reduced stochastic gradient methods. *Math. Program.*, 191(2, Ser. A):671–715, 2022.

[56] Anis Hamadouche, Yun Wu, Andrew M. Wallace, and João F. C. Mota. Sharper bounds for proximal gradient algorithms with errors. *SIAM Journal on Optimization*, 34(1):278–305, 2024.

[57] Hedy Attouch, Jalal Fadili, and Vyacheslav Kungurtsev. The stochastic ravine accelerated gradient method with general extrapolation coefficients, 2024.

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

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

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

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

[62] Roger Pettersson. Yosida approximations for multivalued stochastic differential equations. *Stochastics and Stochastics reports*, 52:107–120, 1994.

[63] Leszek Gawarecki and Vidyadhar Mandrekar. Stochastic differential equations in infinite dimensions. *Springer*, 2011.

[64] Bernt Öksendal. Stochastic differential equations. *Springer*, 2003.

[65] Hedy Attouch and Alexandre Cabot. Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity. *J. Differential Equations*, 263(9):5412–5458, 2017.

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

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

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

[69] D.L. Burkholder, B. Davis, and R.F. Gundy. Integral inequalities for convex functions of operators on martingales. *Proc. 6th Berkeley Symp. Math. Statistics and Probability*, 2:223–240, 1972.

[70] Aurel Rascanu and Eduard Rotenstein. The fitzpatrick function-a bridge between convex analysis and multivalued stochastic differential equations. *arXiv:0809.4447*, 2009.

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