Stochastic Differential Equations for modeling first order optimization methods

Marc Dambrine, Ch Dossal, Bénédicte Puig, Aude Rondepierre

Stochastic Differential Equations for modeling first order optimization methods

M. Dambrine[*] Ch. Dossal[†] B. Puig[‡] A. Rondepierre[§]

Abstract

In this article, a family of SDEs are derived as a tool to understand the behavior of numerical optimization methods under random evaluations of the gradient. Our objective is to transpose the introduction of continuous version through ODEs to understand the asymptotic behavior of discrete optimization scheme to the stochastic setting. We consider a continuous version of the stochastic gradient scheme and of a stochastic inertial system.

This article first studies the quality of the approximation of the discrete scheme by a SDE when the step size tends to 0. Then, it presents new asymptotic bounds on the values $F(X_t) - F^*$ where X_t is a solution of the SDE and $F^* = \min F$, when F is convex and under integrability conditions on the noise. Results are provided under two sets of hypotheses : first considering $\mathcal{C}^2$ and convex functions and then adding some geometrical properties of F. All these results give an insight on the behavior of these inertial and perturbed algorithms in the setting of stochastic algorithms.

Keywords: Lyapunov functions, rate of convergence, SDEs, optimization, geometrical properties of the objective.

1 Introduction

In the recent literature, continuous time approach to first order minimization of a function F is a fruitful field of research. Indeed under some suitable regularity assumptions the sequence generated by a given optimization scheme may converge when the step size $\boldsymbol{h}$ goes to zero to the solution of an ordinary differential equation (ODE). The most striking result in that direction has been obtained by Su, Boyd and Candes in [25]: let $F : \mathbb{R}^d \to \mathbb{R}$. The Nesterov scheme

$$y_n = x_n + \frac{n-1}{n+2}\left(x_n - x_{n-1}\right) \tag{1}$$

$$x_{n+1} = y_n - \boldsymbol{h}\nabla F(y_n), \tag{2}$$

[*]Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, LMAP, Pau, France (*marc.dambrine@univ-pau.fr*).

[†]IMT, Univ. Toulouse, INSA Toulouse, France (*charles.dossal@insa-toulouse.fr*).

[‡]Universite de Pau et des Pays de l'Adour, E2S UPPA, CNRS, LMAP, Pau, France (*benedicte.puig@univ-pau.fr*).

[§]IMT, University of Toulouse, INSA Toulouse, France & LAAS, University Toulouse, CNRS, Toulouse, France (*Aude.Rondepierre@insa-toulouse.fr*).

can be seen as a special and nonstandard discretization of the second order differential equation

$$\ddot{x}(t) + \frac{3}{t}\,\dot{x}(t) + \nabla F(x(t)) = 0 \text{ with } x(0) = x_0 \text{ and } \dot{x}(0) = 0. \tag{3}$$

The study of such ODE, the behavior of the solution enlighten the properties of the sequence generated by the Nesterov scheme.

In the present work, the same question is addressed in the stochastic case, more precisely when the objective F is defined as an expectation: $F(x) = \mathbb{E}[f(x,.)]$. In general, there is no closed form for this expectation and an approximation is used. A common approximation is then an empirical estimator of the type

$$\hat{F}_N(x) = \frac{1}{N}\sum_{i=1}^{N} f(x, \omega_i),$$

where N is the number of samples. The gradient of F is then approximated by the gradient of the previous estimator so that

$$\nabla F(x) = \nabla \hat{F}_N(x) + G_N,$$

where G_N is a random error. In that context, iterative methods based on gradient evaluation are affected by this error. The gradient scheme of step size $\boldsymbol{h}$

$$x_{n+1} = x_n - \boldsymbol{h}\nabla F(x_n),$$

becomes

$$y_{n+1} = y_n - \boldsymbol{h}\left(\nabla \hat{F}(y_n) + G_{N(n)}\right).$$

Of course, the size of the sampling may vary with the number n of iterations. Therefore, we will consider algorithms of the type

$$y_{n+1} = y_n + \boldsymbol{h}\left(\Phi(t_n, y_n) + \sigma(t_n)G_n\right), \tag{4}$$

where $t_n = n\boldsymbol{h}$ and where Φ and σ are given functions, and the G_n are assumed to be independent Gaussian vectors of dimension d with the same variance I_d. It is known that the sequence of iterates generated by these algorithms converge when h tends to 0 to the solution of the deterministic ODE:

$$du(t) = \Phi(t, u(t))dt \tag{5}$$

but it appears that the sequence is more precisely approximated for a given $h > 0$ by the solution of the following high resolution SDE

$$du(t) = \Phi(t, u(t))dt + \sqrt{\boldsymbol{h}}\;\sigma(t)dB_t. \tag{6}$$

That is why we will study this family of high resolution stochastic differential equation (SDE).

Note that the idea was introduced in the deterministic setting in a recent work [24] by Shi, Su and Jordan. Their idea is to consider an ODE whose coefficients depend on the step size $\boldsymbol{h}$ which is not set to 0 to get a better approximation of the inertial discrete scheme.

Thanks to limiting arguments, the authors proposed the following asymptotic approximation for Nesterov scheme for a μ-strongly convex cost function F:

$$\frac{d}{dt}\begin{pmatrix} x \\ v \end{pmatrix} = \begin{pmatrix} v \\ -\frac{3}{t}\, v - \boldsymbol{h}\ D^2F(x)v - \left(1 + \frac{\sqrt{\boldsymbol{h}}}{3}\right)\ \nabla F(x) \end{pmatrix}, \tag{7}$$

where D^2F stands for the Hessian of F, with the initial conditions

$$x(0) = x_0 \text{ and } v(0) = -\frac{2\boldsymbol{h}\nabla F(x_0)}{1+\sqrt{\mu}\boldsymbol{h}}.$$

The SDE (6) is not classical is the sense that it depends on a parameter h that can go to zero. Actually classical results on the convergence of discretization schemes of SDE do not directly apply to the convergence of the sequence y_n defined in (4) to the solution of (6) since the SDE (6) depends itself on the parameter step h.

The contribution of the following paper is twofold. First we prove that the trajectory of the sequence (y_n) defined in (4) is better approximated by the family of SDE (6), parameterized by h than by the solution of the classical ODE (5), see Proposition 1. The main new feature is the fact that we consider a time dependant diffusion coefficient σ and not a trajectory dependant diffusion as in [14]. We give strong convergence (in the sense that the mean of the error goes to 0) results with an additive noise while to our best knowledge only weak convergence can be found in the literature (see [14]). The purpose of this first analysis is to motivate the study of these high resolution SDE to understand the behavior of the sequence generated by some stochastic algorithms.

Secondly, we propose new convergence rates of the solution of two SDEs, especially a high resolution SDE associated to the Nesterov scheme using some Lyapunov analysis. More precisely we provide convergence rate on the expectation of $F(X_t)-F(X^*)$ when F is convex, see Theorem 3 and when F satisfies various geometrical properties, see for example Theorem 4.

The paper is organized as follows: in the second part we show that high resolution SDE provide better approximations of the scheme (4) than the classical ODE approach giving strong convergence results. In the third part we provide first a general analysis allowing to extend some classical Lyapunov analysis for ODEs to SDEs. We illustrate this approach recovering known bounds on the stochastic gradient flow equation. Then in Part 3.2 three new convergence rates are given for the trajectory of an high resolution SDE associated to the Nesterov scheme depending on the geometry of the function to minimize.

2 On the error of approximation by ODE and SDE

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geqslant 0}, \mathbb{P})$ be a complete filtered probability space. We denote by $(B_t)_{t\geqslant 0}$ a standard d-dimensional Brownian motion defined on this space, considering that the filtration $(\mathcal{F}_t)_{t\geqslant 0}$ is in fact the natural filtration of $(B_t)_{t\geqslant 0}$.

Let $\mathrm{L}^2(\Omega)$ be the usual space of random variables whose square is integrable with respect to the measure $\mathbb{P}$ and $\mathrm{L}^\infty(0,T,\mathrm{L}^2(\Omega))$ the usual Bochner space.

For a given $\boldsymbol{h} > 0$, define the following discrete dynamical system

$$Y_{n+1} = Y_n + \boldsymbol{h}\ \Phi(t_n, Y_n) + \boldsymbol{h}\ \sigma(t_n)G_n, \tag{8}$$

where G_n are independent Gaussian vectors with the same variance I_d. Consider a Brownian motion B_t in $\mathbb{R}^d$ and set $t_n = n\boldsymbol{h}$. Noticing that $\sqrt{\boldsymbol{h}}G_n$ has the same distribution that $B_{t_{n+1}} - B_{t_n}$, we can rewrite the system (8) as

$$X_{n+1} = X_n + \boldsymbol{h}\ \Phi(t_n, X_n) + \sqrt{\boldsymbol{h}}\ \sigma(t_n)(B_{t_{n+1}} - B_{t_n}). \tag{9}$$

Observe that such a scheme is similar to an Euler-Maruyama scheme with an additional $\sqrt{\boldsymbol{h}}$ factor that comes from the $\boldsymbol{h}$ factor in (8).

Proposition 1 states that on a fixed interval $[0, T]$, the limit problem when the time step $\boldsymbol{h}$ goes to 0 is the deterministic ODE $w' = \Phi(t, w)$ and that the error measured in $\mathrm{L}^\infty(0, T, \mathrm{L}^2(\Omega))$ is proportional to $\sqrt{\boldsymbol{h}}$. But the main contribution of Proposition 1 is actually to state that the trajectory X_n is better approximated by the SDE:

$$du(t) = \Phi(t, u(t))dt + \sqrt{\boldsymbol{h}}\ \sigma(t)dB_t,$$

since the error measured in $\mathrm{L}^\infty(0, T, \mathrm{L}^2(\Omega))$, is proportional to $\boldsymbol{h}$ instead of $\sqrt{\boldsymbol{h}}$.

Proposition 1. *Let $\Phi : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ be a $\mathcal{C}^2$ function. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a bounded integrable function. Let u_0 be a given vector in $\mathbb{R}^d$, $\boldsymbol{h} \in [0, 1]$ and T be non negative real number. Fix N an integer and set $T = N\boldsymbol{h}$. Define $t_n = n\boldsymbol{h}$, $0 \leqslant n \leqslant N$ and consider the sequence X_n defined by recursion*

$$X_{n+1} = X_n + \boldsymbol{h}\ \Phi(t_n, X_n) + \sqrt{\boldsymbol{h}}\ \sigma(t_n)(B_{t_{n+1}} - B_{t_n}),$$

starting with $X_0 = u_0$. Let w denotes the solution of the deterministic Cauchy problem

$$w' = \Phi(t, w) \text{ with } w(0) = u_0,$$

and let u be the solution of the stochastic differential equation

$$du(t) = \Phi(t, u(t))dt + \sqrt{\boldsymbol{h}}\sigma(t)dB_t \text{ with } u(0) = u_0.$$

There exist non negative constants C_1 and C_2 independent of $\boldsymbol{h}$ such that

$$\sup_{0 \leqslant n \leqslant N} \|w(t_n) - X_n\|_{\mathrm{L}^2(\Omega)} \leqslant C_1\sqrt{\boldsymbol{h}}. \tag{10}$$

$$\sup_{0 \leqslant n \leqslant N} \|u(t_n) - X_n\|_{\mathrm{L}^2(\Omega)} \leqslant C_2\boldsymbol{h}. \tag{11}$$

The proof of this proposition is postponed to Appendix A.

Notice that these results can be adapted in the case of a variable time step: $\boldsymbol{h}$ is then the supremum of the time steps. Unfortunately it turns out that the Nesterov scheme can not be written using the formulation (8) because the gradient of F (or the function Φ) is evaluated at a point that is not Y_n but slightly shifted. Nevertheless Su, Boyd and Candès in [25] proved that a Lyapunov analysis of the ODE associated to Nesterov scheme may provide an efficient Lyapunov analysis of the discrete scheme. That is why we will consider the SDE associated to this second order ODE, expecting that the analysis will provide tools to analyse the stochastic discrete scheme. Such an analysis would be a natural development of the present work.

3 Convergence rates for stochastic gradient flow and Nesterov ODE

In this section we study the properties of trajectories solution of SDEs of the form:

$$dY_t = \Phi\left(t, Y_t\right) dt + \sqrt{\boldsymbol{h}}\sigma(t) M dB_t, \tag{12}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a diffusion term, $\Phi : \mathbb{R} \times \mathbb{R}^m \mapsto \mathbb{R}^m$ is a given smooth function, M a real matrix of size $m \times d$ and $\boldsymbol{h} > 0$. Throughout the paper, we assume that, for any given initial conditions $y_0 \in \mathbb{R}^m$, the Cauchy problem associated with the SDE (12), admits a unique global solution satisfying $Y_0 = y_0$, see [20, Theorem 5.2.1]. For example since σ does not depend on Y_t, the existence and the uniqueness is ensured if the function Φ is Lipschitz with respect to the second variable.

Our main contribution in this section is to prove that the Lyapunov analysis can be extended from ODEs to SDEs providing new decay rates on the expectation of the values $F - F^*$ along the trajectories solution of the SDE (12). It turns out that our results make assumptions on integrability (in time) of the noise level in the spirit of the perturbation analysis led in the deterministic case by Sebbouh, Dossal and Rondepierre in [22].

A key tool in our analysis is the following proposition ensuring that a Lyapunov function associated to an ODE may provide, not necessarily a non increasing function but at least a bounded function in expectation for the stochastic perturbation of an ODE:

Proposition 2. *Let $t_0 \geqslant 0$. Let $\Phi : \mathbb{R} \times \mathbb{R}^m \mapsto \mathbb{R}^m$ be a C^2 function and M a given matrix in $\mathbb{R}^{m\times d}$. Consider the stochastic differential equation:*

$$dY_t = \Phi\left(t, Y_t\right) dt + \sigma(t) M dB_t \tag{13}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ denotes the noise coefficient. Let $J : \mathbb{R} \times \mathbb{R}^m \mapsto \mathbb{R}$ be a (time dependant) Lyapunov function for the associated deterministic ODE: $\dot{Y} = \Phi(t, Y)$ in the sense that J satisfies:

$$\partial_t J\left(t, Y_t\right) + \nabla J\left(t, Y_t\right) \cdot \Phi\left(t, Y_t\right) \leqslant 0$$

for all $t \geqslant t_0$. Then the process $\mathcal{J}(t)$ defined by: $\mathcal{J}(t) = J\left(t, Y_t\right)$, satisfies:

$$\forall t \geqslant t_0,\ \mathbb{E}\left[\mathcal{J}(t)\right] \leqslant \mathcal{J}(t_0) + \frac{1}{2}\int_{t_0}^{t} \sigma^2(\tau)\mathbb{E}\left[Tr\left\{D^2_{22} J\left(\tau, Y_\tau\right) M M^\top\right\}\right] d\tau,$$

where the notation D^2_{22} denotes the second order derivative with respect to the space variable Y.

Proof. Applying the Itô's formula we get

$$\begin{aligned} d\mathcal{J}(t) = \ & \left[\partial_t J\left(t, Y_t\right) + \nabla J\left(t, Y_t\right) \cdot \Phi(t, Y_t)\right] dt + \sigma(t)\nabla J\left(t, Y_t\right) \cdot M dB_t \\ & + \frac{\sigma(t)^2}{2} Tr\left\{D^2_{22} J\left(t, Y_t\right) M M^\top\right\} dt. \end{aligned}$$

Integrating on $[t_0, t]$, we get

$$\mathcal{J}(t) \leqslant \mathcal{J}(t_0) + \int_{t_0}^{t} \sigma(\tau)\nabla J\left(\tau, Y_\tau\right) \cdot M dB_\tau + \frac{\sigma(\tau)^2}{2} Tr\left\{D^2_{22} J\left(\tau, Y_\tau\right) M M^\top\right\} d\tau.$$

Taking the expectation, we get the announced inequality. □

Our methodology is first illustrated on the family of SDEs associated to stochastic gradient descent methods which are widely studied in the literature. Let us mention in particular the results of Mertikopoulos and Staudigl who prove almost sure convergence of trajectories to a minimizer under the assumption that the noise level decreases at least like $1/\sqrt{\ln t}$ [18, Theorem 4.2]. We then apply the same methodology to the family of SDEs associated to the perturbed Nesterov scheme for the minimization of a given convex function F.

3.1 Convergence rate of the stochastic gradient flow

In this paragraph, we consider the stochastic differential equation classically associated to the stochastic gradient descent:

$$dX_t = -\nabla F(X_t)dt + \sqrt{\boldsymbol{h}}\sigma(t)dB_t, \tag{14}$$

where $F : \mathbb{R}^d \to \mathbb{R}$ denotes the objective function and is assumed to be convex of class C^2. Theorem 1 provides some results for convex functions and Theorem 2 provides more accurate bounds when F is μ-strongly convex.

3.1.1 Convergence rate for C^2 convex functions

Theorem 1. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a $\mathcal{C}^2$ convex function with a bounded Laplacian and admitting at least one minimizer. Let $F^* = \min F$ and $t_0 > 0$. Assume that the diffusion σ satisfies the integrability condition: there exist $C > 0$ and $\varsigma \in [0, 1)$ such that*

$$(D_{1,\varsigma}) \quad \forall t \geqslant t_0, \quad \int_{t_0}^{t} s\sigma^2(s)ds \leqslant Ct^{\varsigma}.$$

Then there exists a constant $C_\varsigma > 0$ (affine in $\boldsymbol{h}$) such that the process solution of (14) with the initial condition $X_{t_0} = x_0$, satisfies:

$$\forall t \geqslant t_0, \ \mathbb{E}[F(X_t) - F^*] \leqslant C_\varsigma t^{\varsigma-1} \tag{15}$$

and for any $\beta > 0$:

$$\forall t \geqslant t_0, \ \mathbb{P}\left(F(X_t) - F^* \geqslant \frac{1}{t^\beta}\right) \leqslant C_\varsigma t^{\varsigma+\beta-1}. \tag{16}$$

Before proving Theorem 1, let us make some comments. First, note that the coefficient ς measures the intensity of the noise. The higher the noise is, the lower the convergence rate becomes. Consider for example the case when $\sigma(t) = t^{-p}$ for some $p > 0$. In that case we have:

$$\int_{t_0}^{t} s\sigma(s)^2 ds = \frac{1}{2(p-1)}\left(t_0^{-2(p-1)} - t^{-2(p-1)}\right).$$

A straightforward computation shows that for $p > 1$, we can set $\varsigma = 0$ which implies a convergence rate in expectation in $\frac{1}{t}$. For $p \in [1/2, 1[$ the parameter ς can be set to $\varsigma = -2p+2$ and the convergence rate in expectation is in $\frac{1}{t^{2p-1}}$ where $0 < 2p - 1 < 1$.

The coefficient β is chosen to control the convergence rate in probability. It can be interpreted as a trade-off between the decrease on the values $F(X_t) - F^*$ and the probability that the bound (16) is satisfied. A natural choice for β is $\beta^* = 1 - \varsigma$ which ensures that the two terms in the convergence error in (16) are of the same order but in that case, the

upper bound in (16) is constant and so does not converge to 0. In order to have a decreasing upper bound, β should be chosen as $1-\varsigma-\epsilon$ for $\epsilon>0$ so that the inequality (16) can be reformulated in terms of ϵ as:

$$\forall t \geqslant t_0,\ \mathbb{P}\left(F(X_t)-F^* \geqslant t^{(\varsigma-1)+\epsilon}\right) \leqslant \tilde{C}_\varsigma t^{-\epsilon}$$

for any $\epsilon \in (0, 1-\varsigma)$.

Proof of Theorem 1. Let us introduce the Lyapunov energy:

$$J(t,x) = t(F(x)-F(x^*)) + \frac{1}{2}\|x-x^*\|^2$$

where $x^* \in \arg\min F$. Let X_t be a trajectory solution of (14) with the initial condition $X_{t_0}=X_0$ and $\mathcal{J}(t)=J(t,X_t)$. Applying Proposition 2 with $M=\sqrt{\boldsymbol{h}}I_d$, we have:

$$\forall t \geqslant t_0,\ \mathbb{E}\left[\mathcal{J}(t)\right] \ \leqslant\ \mathcal{J}(t_0) + \frac{\boldsymbol{h}}{2}\int_{t_0}^t \sigma^2(\tau)\mathbb{E}\left[Tr\left\{D^2_{22}J\left(\tau,X_\tau\right)\right\}\right]d\tau.$$

The Itô's calculus gives: $Tr\left\{D^2_{22}J\left(t,x\right)\right\} = t\Delta F(x)+m$. Assuming now that F has a uniformly bounded Laplacian, we deduce that there exists a constant $A>0$ such that:

$$\forall t \geqslant t_0,\ \mathbb{E}\left[\mathcal{J}(t)\right] \ \leqslant\ \mathcal{J}(t_0) + A\boldsymbol{h}\int_{t_0}^t \tau\sigma^2(\tau)d\tau.$$

Hence using the integrability condition $(D_{1,\varsigma})$:

$$\begin{aligned}\forall t \geqslant t_0,\ t\mathbb{E}\left[F(X_t)-F^*\right] \ &\leqslant\ \mathbb{E}\left[\mathcal{J}(t)\right] \leqslant \mathcal{J}(t_0) + A\boldsymbol{h}\int_{t_0}^t \tau\sigma^2(\tau)d\tau \leqslant \mathcal{J}(t_0) + A\boldsymbol{h}Ct^\varsigma \\ &\leqslant\ \left(\mathcal{J}(t_0)t_0^{-\varsigma} + A\boldsymbol{h}C\right)t^\varsigma\end{aligned}$$

Applying Markov inequality, we finally deduce that:

$$\mathbb{P}\left(F(X_t)-F^* \geqslant \frac{1}{t^\beta}\right) \leqslant t^\beta\mathbb{E}[F(X_t)-F^*] \leqslant \left(\mathcal{J}(t_0)t_0^{-\varsigma} + A\boldsymbol{h}C\right)t^{\varsigma+\beta-1}.$$

□

3.1.2 Convergence rates for μ-strongly convex functions

Let us now consider the class of μ-strongly differentiable convex functions i.e. the class of convex differentiable functions satisfying for any x, y in $\mathbb{R}^d$:

$$\langle \nabla F(x) - \nabla F(y), x-y\rangle \geqslant \mu\|x-y\|^2.$$

In the deterministic framework the convergence for this class of function is exponential, see e.g. [12]. This result can be extended to the stochastic case using a Lyapunov approach:

Theorem 2. *Let $F:\mathbb{R}^d\to\mathbb{R}$ be a $\mathcal{C}^2$ μ-strongly convex function having a minimizer $x^*\in \arg\min\ F$ and $t_0\geqslant 0$. Then the process solution of (14) with $X_{t_0}=x_0$ satisfies*

$$\forall t \geqslant t_0,\ \mathbb{E}[\|X_t-x^*\|^2] \leqslant \left[e^{2\mu t_0}\|x_0-x^*\|^2 + \boldsymbol{h}d\int_{t_0}^t e^{2\mu s}\sigma^2(s)\ ds\right]\ e^{-2\mu t}.$$

Proof. The key quantities in this study are $E(x) = \|x - x^*\|^2$ and $\mathcal{E}(t) = E(X_t)$. Applying Itô's formula to the product $e^{2\mu t}\mathcal{E}(t)$, we get:

$$\begin{aligned} d\left(e^{2\mu t}\mathcal{E}(t)\right) &= \left[2\mu\mathcal{E}(t) - 2\langle \nabla F(X_t), X_t - x^*\rangle\right] e^{2\mu t}\, dt + \boldsymbol{h} d\sigma^2(t) e^{2\mu t} dt \\ &\quad + 2\sqrt{\boldsymbol{h}}\sigma(t) e^{2\mu t}\langle X_t - X^*, dB_t\rangle \end{aligned}$$

that is in integral form:

$$\begin{aligned} \forall t \geqslant t_0,\ e^{2\mu t}\mathcal{E}(t) - e^{2\mu t_0}\mathcal{E}(t_0) &= \int_{t_0}^{t} \left[-2\langle \nabla F(X_s), X_s - x^*\rangle) + 2\mu\mathcal{E}(s) + d\boldsymbol{h}\sigma^2(s)\right] e^{2\mu s} ds \\ &\quad + 2\sqrt{\boldsymbol{h}} \int_{t_0}^{t} \sigma(s) e^{2\mu s}\langle X_s - X^*, dB_s\rangle. \end{aligned}$$

Using now the strong convexity of the objective function F, we get:

$$\forall t \geqslant t_0,\ e^{2\mu t}\mathcal{E}(t) \leqslant e^{2\mu t_0}\mathcal{E}(t_0) + 2\sqrt{\boldsymbol{h}} \int_{t_0}^{t} e^{2\mu s}\sigma(s)\langle X_s - X^*, dB_s\rangle + d\boldsymbol{h} \int_{t_0}^{t} e^{2\mu s}\sigma^2(s) ds,$$

and thus the expected inequality:

$$e^{2\mu t}\mathbb{E}[\mathcal{E}(t)] \leqslant e^{2\mu t_0} E(x_0) + d\boldsymbol{h} \int_{t_0}^{t} e^{2\mu s}\sigma^2(s)\ ds.$$

□

A natural question is then under what condition on the noise the convergence is still exponential with the same rate as in the deterministic case. Theorem 2 imposes that:

$$\int_{t_0}^{+\infty} \sigma^2(s) e^{2\mu s}\ ds$$

is finite to obtain a non degraded rate of convergence. Indeed, keeping the exponential rate of convergence of the deterministic case seems out of reach in practical application.

Consider now the particular case when σ is assumed to be constant. We then have:

$$\int_{t_0}^{t} e^{-2\mu(t-s)}\sigma^2\ ds = \sigma^2 e^{-2\mu t} \int_{t_0}^{t} e^{2\mu s} ds = \frac{\sigma^2(1 - e^{-2\mu(t-t_0)})}{2\mu},$$

so that Theorem 2 gives:

$$\mathbb{E}[\mathcal{E}(t)] \leqslant e^{2\mu t_0}\|x_0 - x^*\|^2\ e^{-2\mu t} + \frac{d\sigma^2\boldsymbol{h}}{2\mu}.$$

In other words, the process reaches with an exponential rate (the one of the deterministic dynamic) a ball whose radius depends on the time step $\boldsymbol{h}$ and on the diffusion coefficient σ. This is a well-known property of the stochastic gradient (see [10] for a review and [13] for an accelerated version).

3.2 Convergence rate of a SDE associated to the Nesterov scheme

In 1983, Nesterov [19] proposes a new inertial optimisation scheme (1), with $\alpha = 3$, to minimize a convex differentiable function F. The sequence $(x_n)_n$ generated by this scheme satisfies $F(x_n) - F(x^*) = \mathcal{O}\left(\frac{1}{n^2}\right)$. Moreover, Nesterov proved that this rate is optimal on the class of convex functions among first order method.

Since the work of Boyd, Candes and Su [25], it is known that the Nesterov scheme corresponds in the limit $\boldsymbol{h} \to 0$ to the second order differential equation

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla F(x(t)) = 0, \tag{17}$$

with initial conditions $x(0) = x_0$ and $\dot{x}(0) = 0$ and where the friction coefficient α is set to 3. In [25], the authors prove that the solution of this ODE satisfies $F(x(t)) - F(x^*) = \mathcal{O}\left(\frac{1}{t^2}\right)$ when $\alpha \geqslant 3$. Note that this rate can be improved under geometrical conditions such as strong convexity or growth conditions, see for example [25, 3, 6].

The purpose of the following analysis is to extend these former results to the continuous stochastic setting. The second order ODE (17) in $\mathbb{R}^d$ is equivalent to the following first order system in the phase space

$$\frac{d}{dt}\begin{pmatrix} x(t) \\ v(t) \end{pmatrix} = \begin{pmatrix} v(t) \\ -\frac{\alpha}{t}v(t) - \nabla F(x(t)) \end{pmatrix}. \tag{18}$$

The corresponding stochastic differential equation is then

$$d\begin{pmatrix} X_t \\ V_t \end{pmatrix} = \begin{pmatrix} V_t \\ -\frac{\alpha}{t}V_t - \nabla F(X_t)) \end{pmatrix} dt + \sqrt{\boldsymbol{h}}\ \sigma(t) M dB_t, \tag{19}$$

where M is the matrix with $m = 2d$ rows and d columns

$$M = \begin{pmatrix} 0_d \\ I_d \end{pmatrix}$$

and $(B_t)_{t\geqslant 0}$ a standard d-dimensional Brownian motion. Observe that the initial speed $v(0) = 0$ is mandatory in Nesterov scheme to absorb the singular term. In order to avoid this difficulty that is rather artificial for deriving an asymptotic behavior at $t \to +\infty$, we choose to translate the origin of time in order to start at $t_0 > 0$ so that we consider Equation (19) for times $t \geqslant t_0$ with the deterministic initial condition $x(t_0) = x_0$ and $v(t_0) = 0$.

3.2.1 Convergence rate for the class of convex functions

Let us first consider the general class of convex functions. Our main result given by Theorem 3, establishes new rates of convergence in expectation and probability under some integrability conditions on the noise:

Theorem 3. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a $\mathcal{C}^2$ convex function and $\sigma : \mathbb{R} \to \mathbb{R}$ the noise level. Assume that σ satisfies the integrability condition: there exist $\varsigma \in [0, 2)$ and $C > 0$ such that*

$$(D_{2,\varsigma}) \quad \forall t \geqslant t_0, \quad \int_{t_0}^{t} s^2\sigma^2(s)ds \leqslant Ct^{\varsigma}.$$

Then there exists a real constant $C_\varsigma > 0$ such that the process solution of (19) with $\alpha \geqslant 3$ satisfies

$$\forall t \geqslant t_0,\ \mathbb{E}[F(X_t) - F^*] \leqslant C_\varsigma t^{\varsigma-2}.$$

Moreover there exists another real constant $\widetilde{C}_\varsigma > 0$ such that for any $\beta > 0$:

$$\forall t \geqslant t_0,\ \mathbb{P}\left(F(X_t) - F^* \geqslant \frac{1}{t^\beta}\right) \leqslant \widetilde{C}_\varsigma t^{2(\varsigma+\beta-2)}. \tag{20}$$

Before proving this result, let us make some comments. As for the gradient flow the parameter ς measures the intensity of the noise. Consider again the case where $\sigma(t) = t^{-p}$ for some $p > 0$. In that case, we have:

$$\int_{t_0}^{t} s^2\sigma(s)^2 ds = \frac{1}{2p-3}\left(t_0^{-2p+3} - t^{-2p+3}\right).$$

Thus for any $p > \frac{3}{2}$, we can set: $\varsigma = 0$ which implies a convergence rate in expectation in $\frac{1}{t^2}$. For $p \leqslant \frac{3}{2}$, then we can choose: $\varsigma = -2p + 3$. Consequently if $p \in]\frac{1}{2}, \frac{3}{2}]$ then the convergence rate in expectation is in $\frac{1}{t^{2p-1}}$. We can thus observe that the convergence rates obtained for the stochasticgradient flow (see Theorem 1) and the one obtained for the Nesterov stochastic differential equation (19) coincide when $p \in]\frac{1}{2}, 1]$. In other words, using the Nesterov SDE presents no interest when the diffusion coefficient σ decreases slowly since the stochastic gradient flow provides the same rate of convergence under less demanding integrability conditions on the noise.

Besides, when $p \in [1, \frac{3}{2}]$, the rate of convergence for the gradient flow saturates to its deterministic value 1 while the rate for the Nesterov SDE continues to increase with p until reaching the deterministic value 2 when $p = \frac{3}{2}$. Thus in that case the Nesterov SDE provides a better convergence rate on the values $F(X_t) - F^*$ than the stochastic gradient.

Finally, as for the gradient flow, the coefficient β is chosen to control the convergence rate in probability. It expresses a trade-off between the the decrease on the values $F(X_t) - F^*$ and the probability that the bound (4) is satisfied. A natural choice for β is $\beta^* = 2 - \varsigma$ so that the two terms in the convergence error are of the same order. However, it that case, the upper bound does not converge to 0. In order to have a decreasing upper bound, the parameter β should be chosen as $2 - \varsigma - \epsilon$ for some $\epsilon > 0$ so that the inequality (20) can be reformulated in terms of ϵ as:

$$\mathbb{P}\left(F(X_t) - F^* \geqslant t^{(\varsigma-2)+\epsilon}\right) \leqslant \widetilde{C}_\varsigma t^{-2\epsilon}.$$

for any $\epsilon \in (0, 2 - \varsigma)$.

Proof of Theorem 3. Let us introduce the following energy function:

$$E(t, (x, v)) = t^2(F(x) - F^*) + \frac{1}{2}\|(\alpha - 1)(x - x^*) + tv\|^2 \tag{21}$$

which is a Lyapunov function as proved in [5, Proposition 1], and: $\mathcal{E}(t) = E(t, (X_t, V_t))$. An elementary computation provides

$$Tr\left\{D^2_{22}E\left(t, (x, v)\right) MM^\top\right\} = \Delta_v E(t, x, v) = t^2 d.$$

Applying Proposition 2 with a diffusion term in $\sqrt{\boldsymbol{h}}\sigma(t)$, provides an upper bound on the expectation of the Lyapunov function evaluated on the random trajectory starting from $(x_0, 0)$ at time $t = t_0$:

$$\forall t \geqslant t_0,\ \mathbb{E}[\mathcal{E}(t)] \leqslant \mathcal{E}(t_0) + \frac{\boldsymbol{h}d}{2}\int_{t_0}^{t} s^2\sigma^2(s)ds, \tag{22}$$

and thus an upper bound for the expectation of the residual:

$$\forall t \geqslant t_0,\ \mathbb{E}[F(X_t) - F^*] \leqslant \frac{1}{t^2}\left(\mathcal{E}(t_0) + \frac{\boldsymbol{h}d}{2}\int_{t_0}^{t} s^2\sigma^2(s)ds\right).$$

Using the integrability condition $(D_{2,\varsigma})$, we then deduce that there exists a constant $C_\varsigma > 0$ such that:

$$\forall t \geqslant t_0,\ \mathbb{E}[F(X_t) - F^*] \leqslant C_\varsigma t^{\varsigma-2}.$$

Observe now that the square of a Lyapunov function is still a Lyapunov function so that we can apply Proposition 2 to $\mathcal{E}^2$ to dominate $\mathbb{E}[\mathcal{E}^2]$. We first compute

$$\mathrm{Tr}\left\{D^2_{22}E(t,(x,v))^2 MM^\top\right\} = \Delta_v(E(t,x,v)^2) = 2\left[E(t,x,v)\Delta_v E(t,x,v) + \|\nabla_v E(t,x,v)\|^2\right]$$

where: $\nabla_v E(t,x,v) = t\left((\alpha-1)(x-x^*) + tv\right)$, hence:

$$\|\nabla_v E(t,x,v)\|^2 = t^2\left\|(\alpha-1)(x-x^*) + tv\right\|^2 \leqslant 2t^2 E(t,x,v)$$

so that

$$0 \leqslant \Delta_v(E(t,x,v)^2) \leqslant 2t^2\,(d+2)\,E(t,x,v).$$

Remembering that (22) still holds, we deduce that there is a constant $A > 0$ such that

$$\mathbb{E}[\Delta_v(\mathcal{E}^2(t))] \leqslant At^{2+\varsigma}.$$

Applying Proposition 2 it follows that:

$$\forall t \geqslant t_0,\ \mathbb{E}[\mathcal{E}^2(t)] \leqslant \mathcal{E}(t_0)^2 + \frac{A}{2}\int_{t_0}^{t} s^{2+\varsigma}\sigma^2(s)ds \leqslant \mathcal{E}(t_0)^2 + \frac{A}{2}t^\varsigma\int_{t_0}^{t} s^2\sigma^2(s)ds.$$

Using now the integrability condition $(D_{2,\varsigma})$, we deduce that there exists a constant $\tilde{C}_\varsigma > 0$ such that for all $t \geqslant t_0$, we have: $\mathbb{E}[\mathcal{E}^2(t)] \leqslant \widetilde{C}_\varsigma t^{2\varsigma}$. After expanding $\mathcal{E}^2$, we obtain the crude upper bound on the expectation of the square of the residual:

$$\forall t \geqslant t_0,\ t^4\mathbb{E}[(F(X_t) - F^*)^2] \leqslant \mathbb{E}[\mathcal{E}^2(t)] \leqslant \tilde{C}_\varsigma t^{2\varsigma}.$$

Applying Markov inequality, we finally get:

$$\mathbb{P}\left(F(X_t) - F^* \geqslant \frac{1}{t^\beta}\right) \leqslant t^{2\beta}\mathbb{E}[(F(X_t) - F^*)^2] \leqslant \tilde{C}_\varsigma t^{2(\varsigma+\beta-2)}.$$

□

3.2.2 Convergence rates for convex functions under flatness and sharpness assumptions

Nesterov proved in [19] that the $\mathcal{O}\left(\frac{1}{t^2}\right)$ rate achieved by its acceleration scheme is optimal in some sense on the set of convex functions. Nevertheless better rates can be achieved if more assumptions are made on the function F to minimize. In [1] authors show that this rate actually depends on geometric assumptions on F and on the parameter α of the numerical scheme. This parameter α, in the associated ODE and SDE is a friction parameter. Indeed the solution of (17) can be seen as the trajectory of a mobile (a ball in the seminal work of Polyak) directed by a potential slow down by a vanishing friction whose intensity is defined by α.

In his seminal work, Polyak [21] considered a constant friction term and quadratic geometries for F and observed that the optimal friction is $2\sqrt{\mu}$ where μ is the minimum eigenvalue of the Hessian of F. If α is smaller than this optimal value, some oscillations may slow down the decay rate, if α is higher, the too large friction slow down the mobile. Hence the optimal friction parameter highly depends on the geometry of F.

In [5] Aujol et al. explain how the friction parameter can be chosen for the Nesterov acceleration scheme to optimize the decay of $F(X_t)-F^*$ depending on geometrical hypotheses on F, more precisely: growth and flatness conditions of F. In this section, these former results are extended to the SDE associated to Nesterov.

Let us consider the subclass of convex functions satisfying both a flatness and a growth condition in the neighborhood of their minimizers:

Definition 1 (Growth condition $\mathcal{G}^r$)**.** *Let $F : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a differentiable convex function such that* $\operatorname{argmin} F \neq \emptyset$*. The function F is said to satisfy a global growth condition $\mathcal{G}^r$ for some exponent $r \geqslant 1$ and some real constant $K_r > 0$ if:*

$$\forall x \in \mathbb{R}^d,\ \frac{K_r}{2} d(x, \operatorname{argmin} F)^r \leqslant F(x) - F^*.$$

Historically, the growth condition $\mathcal{G}^r$ is also called r-conditioning [12] or Hölderian error bounds [9], and is closely related to the Łojasiewicz inequality [15, 16], a key tool in the mathematical analysis of continuous and discrete dynamical systems [7, 8]. Note that in the convex setting, the growth condition $\mathcal{G}^r$ is equivalent to a global version of the Łojasiewicz inequality, with exponent $\theta = 1 - \frac{1}{r} \in (0, 1]$ [2, 12]. The growth condition $\mathcal{G}^r$ can be seen as a sharpness assumption on the function F characterizing functions behaving at least as $\|\cdot\|^r$ in the neighborhood of their minimizers. Observe that $\mathcal{G}^r$ implies the growth condition $\mathcal{G}^{r'}$ for all $r' > r \geqslant 1$.

Definition 2 (Flatness condition $\mathcal{F}_\gamma$)**.** *Let $\gamma \geqslant 1$. The function F is said to satisfy the hypothesis $\mathcal{F}_\gamma$ if for any minimizer $x^* \in \operatorname{argmin} F$, we have:*

$$\forall x \in \mathbb{R}^d, \quad F(x) - F^* \leqslant \frac{1}{\gamma} \langle \nabla F(x), x - x^* \rangle.$$

The hypothesis $\mathcal{F}_\gamma$ is a mild assumption, requesting slightly more than the convexity of F in the neighborhood of its minimizers. Any convex function automatically satisfies $\mathcal{F}_1$ and any differentiable function F for which $(F - F^*)^{\frac{1}{\gamma}}$ is convex for some $\gamma \geqslant 1$, satisfies $\mathcal{F}_\gamma$.

In [6], the authors prove that if F satisfies the condition $\mathcal{F}_\gamma$ for some $\gamma \geqslant 1$ then for any minimizer x^* there exist $M > 0$ and $\eta > 0$ such that:

$$\forall x \in B(x^*, \eta),\ F(x) - F(x^*) \leqslant M\|x - x^*\|^\gamma. \tag{23}$$

In other words, the hypothesis $\mathcal{F}_\gamma$ with $\gamma \geqslant 1$, can be interpreted as a flatness condition: it ensures that the function F is sufficiently flat (at least as flat as $x \mapsto \|x\|^\gamma$) in the neighborhood of its minimizers. We refer the interested reader to [11], [25], [5] or [6] for more details.

Let us first consider the subclass of convex functions with a "sharp" geometry near its set of minimizers, i.e. satisfying a quadratic growth condition $\mathcal{G}^2$ combined with a flatness assumption. Theorem 4 gives conditions on the friction parameter and on the noise to take advantage of the geometry assumptions made on F to improve the $\mathcal{O}\left(\frac{1}{t^2}\right)$ decay rate of Theorem 3:

Theorem 4. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a $\mathcal{C}^2$ convex differentiable function admitting a unique minimizer x^*, and $F^* = \inf F$. Additionally assume that F satisfies both a global quadratic growth condition $\mathcal{G}^2$ and some flatness assumption $\mathcal{F}_\gamma$ for some $\gamma \in [1, 2]$.*

Let $t_0 > 0$. Let (X_t, V_t) be any solution of the SDE (19) with $(X_{t_0}, V_{t_0}) = (x_0, v_0)$. Assume that the noise level σ satisfies the integrability condition: there exist $C > 0$ and $\varsigma \in [0, 2)$ such that

$$(D_{q,\varsigma}) \quad \forall t > 0, \quad \int_{t_0}^{t} s^q \sigma^2(s) ds \leqslant C t^\varsigma \text{ with } q = \frac{2\alpha\gamma}{\gamma + 2}.$$

If $\alpha > 1 + \frac{2}{\gamma}$ then the process solution of (19) satisfies:

$$\forall t \geqslant t_c,\ \mathbb{E}[F(X_t) - F^*] \leqslant 2\left[t_c^{q-\varsigma}\left(1 + \frac{2\alpha}{t_0(\gamma+2)\sqrt{K_2}}\right)^2 \mathcal{E}_M(t_0) + \frac{d}{2}C_\varsigma\right]\frac{e^{\frac{2-\gamma}{\gamma+2}\alpha}}{t^{p+2-\varsigma}}$$

where: $t_c = \max(t_0, \frac{2\alpha\sqrt{\gamma}}{(\gamma+2)\sqrt{K_2}})$ and $\mathcal{E}_M(t) = F(X_t) - F^* + \frac{1}{2}\|V_t\|^2$ denotes the mechanical energy at time t. Moreover, for any $\beta > 0$,

$$\mathbb{P}\left(F(X_t) - F^* \geqslant \frac{1}{t^\beta}\right) \leqslant 2\left[t_c^{q-\varsigma}\left(1 + \frac{2\alpha}{t_0(\gamma+2)\sqrt{K_2}}\right)^2 \mathcal{E}_M(t_0) + \frac{d}{2}C_\varsigma\right]\frac{e^{\frac{2-\gamma}{\gamma+2}\alpha}}{t^{q-\varsigma-\beta}}.$$

First, observe that the convergence rate $q = \frac{2\alpha\gamma}{\gamma+2}$ reached for $\varsigma = 0$, is actually larger than 2 when $\alpha \geqslant 3$ and $\gamma \geqslant 1$. The fact that the decay rate $q = \frac{2\alpha\gamma}{\gamma+2}$ is a growing function of α may be surprising and raises to questions: is it possible to get an arbitrary large decay rate ? Would it be better to choose a very large α ?

The answer to the first question is : yes we can achieve arbitrary large decay rate but the price to pay is twofold:

1. The integrability on the noise is more restrictive for large α.
2. The term
$$\left[\left(\frac{q}{\sqrt{K_2\gamma}}\right)^{q-\varsigma}(1 + \frac{1}{\sqrt{\gamma}})^2\mathcal{E}_M(t_0)t_0^{-\varsigma} + C\frac{d}{2}\right]e^{\frac{2-\gamma}{\gamma+2}\alpha}$$
is an increasing function of α. It follows that for a given t, the optimal choice for α is not to take α as large as possible.

Proof of Theorem 4. Our analysis relies on the same energy $\mathcal{H}$ introduced by Su, Boyd and Candes [25], Attouch, Chbani, Peypouquet and Redont [4] and Aujol, Dossal [5]:

$$\mathcal{H}(t,(x,v)) = t^p \mathcal{E}(t,(x,v)) \tag{24}$$

where

$$\mathcal{E}(t,(x,v)) = t^2(F(x)-F^*) + \frac{1}{2}\|\lambda(x-x^*)+tv\|^2 + \frac{\xi}{2}\|x-x^*\|^2,$$

and x^* is a minimizer of F. Note that the energy $\mathcal{H}$ is not a Lyapunov energy anymore since it is not decreasing along the trajectories of the considered system, but we still have a differential inequation enabling the control on the values. Following the computations detailed in [22, Proof of Theorem 3.1] and applying the Itô's calculus, we have:

$$\begin{aligned} d\mathcal{E}(t,Y_t) &\leqslant (2-\lambda\gamma)t\left(F(X_t)-F^*\right)dt + \left[\xi - \lambda(\lambda+1-\alpha)\right]\langle X_t - x^*, V_t\rangle dt \\ &\quad +(\lambda+1-\alpha)\frac{1}{t}\|\lambda(X_t-x^*)+tV_t\|^2 dt - \frac{\lambda^2}{t}(\lambda+1-\alpha)\|X_t-x^*\|^2 dt \\ &\quad +\frac{d}{2}t^2\sigma(t)^2 dt + t\sigma(t)\langle \lambda(X_t-x^*)+tV_t, dB_t\rangle. \end{aligned}$$

Choosing now: $\xi = \lambda(\lambda+1-\alpha)$, we then get:

$$\begin{aligned} d\mathcal{E}(t,Y_t) &\leqslant (2-\lambda\gamma)t\left(F(X_t)-F^*\right)dt + (\lambda+1-\alpha)\tfrac{1}{t}\|\lambda(X_t-x^*)+tV_t\|^2 dt \\ &\quad -\frac{\lambda^2}{t}(\lambda+1-\alpha)\|X_t-x^*\|^2 dt + \frac{d}{2}t^2\sigma(t)^2 dt + t\sigma(t)\langle\lambda(X_t-x^*)+tV_t, dB_t\rangle \end{aligned}$$

Introducing the energy $\mathcal{H}(t,Y_t) = t^p\mathcal{E}(t,Y_t)$ we then obtain:

$$\begin{aligned} d\mathcal{H}(t,Y_t) &= t^p d\mathcal{E}(t,Y_t) + pt^{p-1}\mathcal{E}(t,Y_t) \\ &\leqslant t^p\left[(2-\lambda\gamma+p)t\left(F(X_t)-F^*\right)dt + \frac{2\lambda+2-2\alpha+p}{2t}\|\lambda(X_t-x^*)+tV_t\|^2 dt\right. \\ &\quad \left. +\frac{\lambda}{2t}(\lambda+1-\alpha)(p-2\lambda)\|X_t-x^*\|^2 + t\sigma(t)\langle\lambda(X_t-x^*)+tV_t, dB_t\rangle + \frac{d}{2}t^2\sigma(t)^2 dt\right] \end{aligned}$$

As in [22, Proof of Theorem 3.1], we choose:

$$\lambda = \frac{2\alpha}{\gamma+2}, \quad p = \frac{2\alpha\gamma}{\gamma+2} - 2 \tag{25}$$

so that: $\xi = \frac{2\alpha}{(\gamma+2)^2}(2+\gamma(1-\alpha))$ and:

$$d\mathcal{H}(t,Y_t) \leqslant t^p\left[\frac{\xi}{2t}(p-2\lambda)\|X_t-x^*\|^2 + t\sigma(t)\langle\lambda(X_t-x^*)+tV_t, dB_t\rangle + \frac{d}{2}t^2\sigma(t)^2 dt\right] \tag{26}$$

Let $A = \xi(p-2\lambda)$. With our choice of parameters, we have: $A = \frac{2\xi}{\gamma+2}\left((\gamma-2)\alpha - (\gamma+2)\right)$. Assuming $\alpha > 1 + \frac{2}{\gamma}$ and $\gamma \leqslant 2$, we necessarily have: $\xi < 0$, and thus $A > 0$. Consequently the energy $\mathcal{E}$ is not a sum of non-negative terms and we need an additional growth condition $\mathcal{G}^2$ to bound the term $\|X_t - x^*\|^2$ as done in [22].

Assuming that F satisfies some quadratic growth condition $\mathcal{G}^2$ and has a unique minimizer, there exists $K_2 > 0$ such that:

$$\forall t \geqslant t_0,\ \|X_t - x^*\|^2 \leqslant \frac{2}{K_2}(F(X_t)-F^*). \tag{27}$$

Hence:

$$\forall t \geqslant t_0,\ \mathcal{E}(t, Y_t) \ \geqslant\ t^2(1 + \frac{\xi}{K_2 t^2})\left(F(X_t) - F^*\right). \tag{28}$$

Observe now that since $\xi = \lambda(\lambda + 1 - \alpha) < 0$, we get:

$$|\xi| = \frac{2\alpha}{\gamma + 2}\left(\frac{\alpha\gamma}{\gamma + 2} - 1\right) \leqslant \frac{2\alpha^2\gamma}{(\gamma + 2)^2}.$$

Let $t_c = \max\left(t_0, \frac{2\alpha\sqrt{\gamma}}{(\gamma+2)\sqrt{K_2}}\right)$. We thus have: $1 + \frac{\xi}{K_2 t^2} \geqslant \frac{1}{2}$ for all $t \geqslant t_c$. Hence:

$$\forall t \geqslant t_c,\ \mathcal{E}(t, Y_t) \ \geqslant\ \frac{t^2}{2}\left(F(X_t) - F^*\right). \tag{29}$$

Combining (26), (27) and (29), we get in expectation:

$$\forall t \geqslant t_c,\ d\mathcal{H}(t, Y_t) \ \leqslant\ \frac{2A}{K_2 t^3}\mathcal{H}(t, Y_t) + t^{p+1}\sigma(t)\langle\lambda(X_t - x^*) + tV_t, dB_t\rangle + \frac{d}{2}t^{p+2}\sigma(t)^2 dt$$

Integrating now between t_c and t, we get:

$$\begin{aligned}\forall t \geqslant t_c,\ \mathbb{E}[\mathcal{H}(t, Y_t)] \ &\leqslant\ \mathbb{E}[\mathcal{H}(t_c, Y_{t_c})] + \frac{2A}{K_2}\int_{t_c}^{t}\frac{\mathbb{E}[\mathcal{H}(s, Y_s)]}{s^3}ds + \frac{d}{2}\int_{t_c}^{t} s^{p+2}\sigma(s)^2 ds \\ &\leqslant\ \mathbb{E}[\mathcal{H}(t_c, Y_{t_c})] + \frac{2A}{K_2}\int_{t_c}^{t}\frac{\mathbb{E}[\mathcal{H}(s, Y_s)]}{s^3}ds + C_\varsigma\frac{d}{2}t^\varsigma\end{aligned}$$

According to the Grönwall lemma, we then have:

$$\begin{aligned}\forall t \geqslant t_c,\ \mathbb{E}[\mathcal{H}(t, Y_t)] \ &\leqslant\ \left(\mathbb{E}[\mathcal{H}(t_c, Y_{t_c})] + \frac{d}{2}\int_{t_c}^{t} s^{p+2}\sigma(s)^2 ds\right)\exp\left(\int_{t_c}^{t}\frac{2A}{K_2 s^3}ds\right) \\ &\leqslant\ \left(\mathbb{E}[\mathcal{H}(t_c, Y_{t_c})] + \frac{d}{2}C_\varsigma t^\varsigma\right)\exp\left(\frac{A}{K_2 t_c^2}\right).\end{aligned}$$

Let us now prove that the energy $\mathcal{H}(t_c, Y_{t_c})$ is uniformly bounded by the mechanical energy $\mathcal{E}_M(t) = F(X_t) - F^* + \frac{1}{2}\|V_t\|^2$ at time t_c:

$$\begin{aligned}\mathcal{E}(t_c, Y_{t_c}) \ &=\ t_c^2(F(X_{t_c}) - F^*) + \frac{1}{2}\|\lambda(X_{t_c} - x^*) + t_c V_{t_c}\|^2 + \frac{\xi}{2}\|X_{t_c} - x^*\|^2 \\ &=\ t_c^2\mathcal{E}_M(t_c) + \frac{\lambda^2}{2}\|X_{t_c} - x^*\|^2 + \lambda t_c\langle X_{t_c} - x^*, V_{t_c}\rangle + \frac{\xi}{2}\|X_{t_c} - x^*\|^2 \\ &\leqslant\ t_c^2\mathcal{E}_M(t_c) + \frac{\lambda^2}{2}\|X_{t_c} - x^*\|^2 + \lambda t_c\langle X_{t_c} - x^*, V_{t_c}\rangle.\end{aligned}$$

Using the quadratic growth condition $\mathcal{G}^2$ and the following inequality:

$$2|\langle X_{t_c} - x^*, v_{t_c}\rangle| \leqslant \sqrt{K_2}\|X_{t_c} - x^*\|^2 + \frac{1}{\sqrt{K_2}}\|V_{t_c}\|^2, \tag{30}$$

we finally get:

$$\mathcal{E}(t_c, Y_{t_c}) \ \leqslant\ \left(t_c^2 + 2\frac{\lambda t_c}{\sqrt{K_2}} + \frac{\lambda^2}{K_2}\right)\mathcal{E}_M(t_c) = \left(t_c + \frac{\lambda}{\sqrt{K_2}}\right)^2\mathcal{E}_M(t_c).$$

Hence:

$$\begin{aligned}\forall t \geqslant t_c,\ \mathbb{E}[\mathcal{H}(t,Y_t)] &\leqslant \left[t_c^p\left(t_c+\frac{\lambda}{\sqrt{K_2}}\right)^2\mathbb{E}[\mathcal{E}_M(t_c)]+\frac{d}{2}C_\varsigma t^\varsigma\right]\exp(\frac{A}{K_2t_c^2})\\ &\leqslant \left[t_c^{p+2}\left(1+\frac{\lambda}{t_c\sqrt{K_2}}\right)^2\mathbb{E}[\mathcal{E}_M(t_c)]+\frac{d}{2}C_\varsigma t^\varsigma\right]\exp(\frac{2-\gamma}{\gamma+2}\alpha)\end{aligned}$$

since by construction: $\frac{A}{K_2t_c^2}\leqslant\frac{A}{2|\xi|}=\frac{2-\gamma}{\gamma+2}\alpha$. Observe now that the mean of the mechanical energy is non-increasing. Indeed:

$$\forall t\geqslant t_0,\ \mathcal{E}_M(t) = \mathcal{E}_M(t_0)+\int_{t_0}^t d\mathcal{E}_M(s)ds=\mathcal{E}_M(t_0)+\int_{t_0}^t\left(-\frac{\alpha}{s}\|V_s\|^2ds+\sigma(s)\langle V_s,dB_s\rangle\right)$$

Hence: $\forall t\geqslant t_0,\ \mathbb{E}[\mathcal{E}_M(t)]\leqslant\mathcal{E}_M(t_0)$ and:

$$\begin{aligned}\forall t \geqslant t_c,\ \mathbb{E}[\mathcal{H}(t,Y_t)] &\leqslant \left[t_c^{p+2}\left(1+\frac{\lambda}{t_c\sqrt{K_2}}\right)^2\mathcal{E}_M(t_0)+\frac{d}{2}C_\varsigma t^\varsigma\right]\exp(\frac{2-\gamma}{\gamma+2}\alpha)\\ &\leqslant \left[t_c^{p+2}\left(1+\frac{\lambda}{t_0\sqrt{K_2}}\right)^2\mathcal{E}_M(t_0)+\frac{d}{2}C_\varsigma t^\varsigma\right]\exp(\frac{2-\gamma}{\gamma+2}\alpha)\end{aligned}$$

Remember now that the values of F along the trajectories of the SDE (19) are controlled by the energy $\mathcal{E}$ as stated in (29). Thus:

$$\begin{aligned}\forall t \geqslant t_c,\ \mathbb{E}[F(X_t)-F^*] &\leqslant 2\left[t_c^{p+2}\left(1+\frac{\lambda}{t_0\sqrt{K_2}}\right)^2\mathcal{E}_M(t_0)+\frac{d}{2}C_\varsigma t^\varsigma\right]\frac{e^{\frac{2-\gamma}{(\gamma+2)\alpha}}}{t^{p+2}}\\ &\leqslant 2\left[t_c^{p+2-\varsigma}\left(1+\frac{\lambda}{t_0\sqrt{K_2}}\right)^2\mathcal{E}_M(t_0)+\frac{d}{2}C_\varsigma\right]\frac{e^{\frac{2-\gamma}{(\gamma+2)\alpha}}}{t^{p+2-\varsigma}}\end{aligned}$$

Finally observe that if $t_c=\frac{2\alpha\sqrt{\gamma}}{(\gamma+2)\sqrt{K_2}}>t_0$, we then get:

$$\forall t\geqslant t_c,\ \mathbb{E}[F(X_t)-F^*] \leqslant 2\left[\left(\frac{2\alpha\sqrt{\gamma}}{(\gamma+2)\sqrt{K_2}}\right)^{p+2-\varsigma}\left(1+\frac{\lambda}{t_0\sqrt{K_2}}\right)^2\mathcal{E}_M(t_0)+\frac{d}{2}C_\varsigma\right]\frac{e^{\frac{2-\gamma}{\gamma+2}\alpha}}{t^{p+2-\varsigma}}.$$

as expected. □

The second Theorem considers the case of convex functions with a "flat" geometry near its set of minimizers, i.e. roughly speaking functions behaving like $\|\cdot\|^\gamma$ with $\gamma>2$.

Theorem 5. *Let $F:\mathbb{R}^d\to\mathbb{R}$ be a $\mathcal{C}^2$ convex differentiable function admitting a unique minimizer x^*, and $F^*=\inf F$. Additionally assume that F satisfies both a global growth condition $\mathcal{G}^\gamma$ and some flatness assumption $\mathcal{F}_\gamma$ for some $\gamma>2$.*

Let $t_0>0$. Let (X_t,V_t) be any solution of the SDE (19) with $(X_{t_0},V_{t_0})=(x_0,v_0)$. Assume that the noise level σ satisfies the integrability condition: there exist $C>0$ and $\varsigma\in[0,2)$ such that

$$(D_{p,\varsigma})\qquad \int_{t_0}^t s^{p+2}\sigma(s)^2ds\leqslant Ct^\varsigma\quad \text{where } p=\frac{4}{\gamma-2}. \tag{31}$$

If $\alpha \geqslant \frac{\gamma+2}{\gamma-2}$ then for all $t \geqslant t_0$ we have:

$$\mathbb{E}[F(X_t) - F^*] \leqslant \left(\left(t_0^{-\varsigma} \mathcal{H}(t_0, Y_{t_0}) + \frac{d}{2} C \right)^{\frac{\gamma-2}{\gamma}} + \lambda\alpha \frac{K_\gamma^{-\frac{2}{\gamma}}}{2} t_0^{-\varsigma\frac{\gamma-2}{\gamma}} \right)^{\frac{\gamma}{\gamma-2}} t^{-\frac{2\gamma}{\gamma-2}+\varsigma}. \tag{32}$$

Moreover, for any $\beta > 0$,

$$\mathbb{P}\left(F(X_t) - F^* \geqslant \frac{1}{t^\beta} \right) \leqslant \left(\left(t_0^{-\varsigma} \mathcal{H}(t_0, Y_{t_0}) + \frac{d}{2} C \right)^{\frac{\gamma-2}{\gamma}} + \lambda\alpha \frac{K_\gamma^{-\frac{2}{\gamma}}}{2} t_0^{-\varsigma\frac{\gamma-2}{\gamma}} \right)^{\frac{\gamma}{\gamma-2}} t^{-\frac{2\gamma}{\gamma-2}+\varsigma+\beta}.$$

We can notice here that for any $\gamma > 2$ the decay rate $\frac{2\gamma}{\gamma-2}$ (corresponding to $\varsigma = 0$) is greater than 2 when γ is large enough. Indeed this Theorem give conditions on the friction parameter α and on the noise to take advantage of the geometry of F. If α is chosen too small, the trajectories will oscillate and the decay rate will be smaller. The critical value for α is $\frac{\gamma+2}{\gamma-2}$. For all values beyond this one, the asymptotic rate is the same : $\frac{2\gamma}{\gamma-2} - \varsigma$, but the bound is a growing function of α and thus, the optimal value of α seems to be $\frac{\gamma+2}{\gamma-2}$.

Proof of Theorem 5. In this proof we still consider the energy (24) defined by:

$$\mathcal{H}(t, (x, v)) = t^p \mathcal{E}(t, (x, v))$$

where $\mathcal{E}(t, (x, v)) = t\left(a(t, (x, v)) + b(t, (x, v)) + \xi c(t, (x, v))\right)$ using the following notations:

$$a(t, (x, v)) = t(F(x) - F^*), \quad b(t, (x, v)) = \frac{1}{2t}\|\lambda(x - x^*) + tv)\|^2, \quad c(t, (x, v)) = \frac{1}{2t}\|x - x^*\|^2.$$

But the choice of parameters is slightly different. Here we choose:

$$\lambda = \frac{2}{\gamma - 2}, \ p = 2\lambda, \ \xi = \lambda(\lambda + 1 - \alpha) \tag{33}$$

Note that with this choice, we have: $\xi < 0$ since $\alpha \geqslant \frac{\gamma+2}{\gamma-2}$.

Let $Y_t = (X_t, V_t)$. Applying the Itô calculus to $\mathcal{H}(t, Y_t)$ and following the same steps as in [22, Proof of Theorem 3.1], we can prove that:

$$d\mathcal{H}(t, Y_t) \quad \leqslant \quad t^p \left[2(\frac{\gamma+2}{\gamma-2} - \alpha) b(t, Y_t) dt + t\sigma(t) \langle \lambda(X_t - x^*) + tV_t, dB_t \rangle + \frac{d}{2} t^2 \sigma(t)^2 dt \right]$$

which implies that for all $t \geqslant t_0$

$$\mathbb{E}[\mathcal{H}(t, Y_t)] \leqslant \mathbb{E}[\mathcal{H}(t_0, Y_{t_0})] + 2\int_{t_0}^t (\frac{\gamma+2}{\gamma-2} - \alpha) s^p \mathbb{E}[b(s, Y_s)] ds + \frac{d}{2}\int_{t_0}^t s^{p+2}\sigma(s)^2 ds \tag{34}$$

Hence for any $\alpha \geqslant \frac{\gamma+2}{\gamma-2}$,

$$\forall t \geqslant t_0, \ \mathbb{E}[\mathcal{H}(t, Y_t)] \leqslant \mathcal{H}(t_0, Y_{t_0}) + \frac{d}{2}\int_{t_0}^t s^{p+2}\sigma(s)^2 ds. \tag{35}$$

We now need the control on the values $F(X_t)-F^*$ from the energy. Using $|\xi| = \lambda(\alpha-\lambda-1) \leqslant \lambda\alpha$, we have:

$$\mathcal{H}(t,Y_t) \quad = \quad t^{p+1}(a(t,Y_t)+b(t,Y_t)+\xi c(t,Y_t) \geqslant t^{p+1}a(t,Y_t) - \lambda\alpha t^{p+1}c(t,Y_t).$$

Hence:

$$\forall t \geqslant t_0,\ t^{p+1}\mathbb{E}[a(t,Y_t)] - \lambda\alpha t^{p+1}\mathbb{E}[c(t,Y_t)] \leqslant \mathbb{E}[\mathcal{H}(t_0,Y_{t_0})] + \frac{d}{2}\int_{t_0}^t s^{p+2}\sigma(s)^2 ds.$$

Assume now that F has a unique minimizer and satisfies the growth condition $\mathcal{G}^\gamma$ with $\gamma > 2$. We then have:

$$c(t,Y_t) \leqslant \frac{K_\gamma^{-\frac{2}{\gamma}}}{2^{1-\frac{2}{\gamma}}} t^{-\frac{2}{\gamma}-1} a(t,Y_t)^{\frac{2}{\gamma}}.$$

Noticing that: $\frac{2}{\gamma}(p+2)+1 = p+1$, we have:

$$t^{p+1}c(t,Y_t) \leqslant \frac{K_\gamma^{-\frac{2}{\gamma}}}{2^{1-\frac{2}{\gamma}}} \left(t^{p+1}a(t,Y_t)\right)^{\frac{2}{\gamma}}.$$

Hence:

$$\begin{aligned}\mathbb{E}[t^{p+1}a(t,Y_t) - \lambda\alpha\frac{K_\gamma^{-\frac{2}{\gamma}}}{2^{1-\frac{2}{\gamma}}}\left(t^{p+1}a(t,Y_t)\right)^{\frac{2}{\gamma}}] &\leqslant t^{p+1}\mathbb{E}[a(t,Y_t)] - \lambda\alpha t^{p+1}\mathbb{E}[c(t,Y_t)] \\ &\leqslant \mathbb{E}[\mathcal{H}(t_0,Y_{t_0})] + \frac{d}{2}\int_{t_0}^t s^{p+2}\sigma(s)^2 ds.\end{aligned}$$

Applying the Jensen inequality to the convex function $x \mapsto x^{\frac{\gamma}{2}}$ we get

$$\mathbb{E}[t^{p+1}a(t,Y_t)] - \lambda\alpha\frac{K_\gamma^{-\frac{2}{\gamma}}}{2}\mathbb{E}[\left(t^{p+1}a(t,Y_t)\right)]^{\frac{2}{\gamma}} \leqslant \mathbb{E}[\mathcal{H}(t_0,Y_{t_0})] + \frac{d}{2}\int_{t_0}^t s^{p+2}\sigma(s)^2 ds$$

Let us apply then the following lemma with $\delta = \frac{2}{\gamma}$, $m = \lambda\alpha\frac{K_\gamma^{-\frac{2}{\gamma}}}{2}$ and $M = \mathbb{E}[\mathcal{H}(t_0,Y_{t_0})] + \frac{d}{2}\int_{t_0}^t s^{p+2}\sigma(s)^2 ds$ for any $t \geqslant t_0$:

Lemma 1. *Let $\delta \in (0,1)$, $m > 0$ and $M > 0$*

$$x - mx^\delta \leqslant M \Rightarrow x \leqslant (M^{1-\delta}+m)^{\frac{1}{1-\delta}}. \tag{36}$$

We then deduce that:

$$\mathbb{E}[t^{p+1}a(t)] \leqslant \left(\left(\mathbb{E}[\mathcal{H}(t_0,Y_{t_0})] + \frac{d}{2}\int_{t_0}^t s^{p+2}\sigma(s)^2 ds\right)^{\frac{\gamma-2}{\gamma}} + \lambda\alpha\frac{K_\gamma^{-\frac{2}{\gamma}}}{2}\right)^{\frac{\gamma}{\gamma-2}} \tag{37}$$

Remembering that: $a(t) = t(F(x(t)) - F^*)$, we finally deduce that for all $t \geqslant t_0$,

$$\begin{aligned} t^{p+2}\mathbb{E}[F(x(t)) - F^*] &\leqslant \left(\left(\mathbb{E}[\mathcal{H}(t_0, Y_{t_0})] + \frac{d}{2} \int_{t_0}^{t} s^{p+2}\sigma(s)^2 ds \right)^{\frac{\gamma-2}{\gamma}} + \lambda\alpha \frac{K_\gamma^{-\frac{2}{\gamma}}}{2} \right)^{\frac{\gamma}{\gamma-2}} \\ &\leqslant \left(\left(\mathbb{E}[\mathcal{H}(t_0, Y_{t_0})] + \frac{d}{2} C t^{\varsigma} \right)^{\frac{\gamma-2}{\gamma}} + \lambda\alpha \frac{K_\gamma^{-\frac{2}{\gamma}}}{2} \right)^{\frac{\gamma}{\gamma-2}} \\ &\leqslant \left(\left(t_0^{-\varsigma}\mathcal{H}(t_0, Y_{t_0}) + \frac{d}{2} C \right)^{\frac{\gamma-2}{\gamma}} + \lambda\alpha \frac{K_\gamma^{-\frac{2}{\gamma}}}{2} t_0^{-\varsigma\frac{\gamma-2}{\gamma}} \right)^{\frac{\gamma}{\gamma-2}} t^{-\frac{2\gamma}{\gamma-2}+\varsigma}. \end{aligned}$$

□

4 Conclusion

In many practical applications, the number of samples cannot be arbitrary large and σ has a nonnegative limit when t goes to $+\infty$. In that case, the stochastic gradient method is well-known to reach a ball around the minimizer with a radius depending on this limit. In particular it does not converges to the minimizer even in the very favorable case of a μ-strictly convex objective. Therefore, a variable time step is usually used: for example $\boldsymbol{h}_n = \boldsymbol{h}\ n^{-a}$ with $a \in (0, 1)$. We are considering the variable time stepping case in a on going work.

In [23] the authors proposed new optimization algorithms based on the symplectic Euler method classically used for the numerical approximation of Hamiltonian ODE. Under the strong convexity assumption of the cost function F, the authors manage to recover the acceleration phenomenon with rather large step size. Their conclusion is *high-resolution ODEs and symplectic schemes are critical to achieve acceleration using numerical discretization... Lyapunov functions play a key role in such analyses, and also allow aspects of the continuous-time analysis to be transferred to discrete time.* A natural extension of this work is to study the stochastic version of the high resolution ODE (7).

Thanks. The first author is indebted to Pr. G. Vallet for instructive discussions on the numerical discretizations of SDEs and for pointing to us the reference [17] and the basis of the analysis of Section 2.

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-PRC-CE23 Masdol.

References

[1] V. Apidopoulos, Jean-François Aujol, Charles Dossal, and Aude Rondepierre. Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions. *Mathematical Programming*, 187(1):151–193, 2021.

[2] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-ł ojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[3] H. Attouch, A. Cabot, Z. Chbani, and H. Riahi. Inertial forward–backward algorithms with perturbations: Application to tikhonov regularization. *Journal of Optimization Theory and Applications*, 179(1):1–36, 2018.

[4] H. Attouch, Z. Chbani, J. Peypouquet, and P. Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming*, 168(1-2):123–175, 2018.

[5] J-F. Aujol and Ch. Dossal. Optimal rate of convergence of an ODE associated to the fast gradient descent schemes for $b > 0$. *Hal Preprint hal-01547251*, June 2017.

[6] J.-F. Aujol, Ch. Dossal, and A. Rondepierre. Optimal convergence rates for Nesterov acceleration. *SIAM Journal on Optimization*, 29(4):3131–3153, 2019.

[7] J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[8] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

[9] J. Bolte, T.P. Nguyen, J. Peypouquet, and B.W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[10] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[11] A. Cabot, H. Engler, and S. Gadat. 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.

[12] G. Garrigos, L. Rosasco, and S. Villa. Convergence of the forward-backward algorithm: Beyond the worst case with the help of geometry. *arXiv preprint arXiv:1703.09477*, 2017.

[13] M. Laborde and A. M. Oberman. Nesterov's method with decreasing learning rate leads to accelerated stochastic gradient descent, 2020.

[14] Q. Li, C. Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2101–2110. PMLR, 06–11 Aug 2017.

[15] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles (Paris, 1962)*, pages 87–89. Éditions du Centre National de la Recherche Scientifique, Paris, 1963.

[16] S. Łojasiewicz. Sur la géométrie semi- et sous-analytique. *Annales de l'Institut Fourier. Université de Grenoble*, 43(5):1575–1595, 1993.

[17] G. J. Lord, C. E. Powell, and T. Shardlow. *An introduction to computational stochastic PDEs*. Cambridge Texts in Applied Mathematics. Cambridge University Press, New York, 2014.

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

[19] Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(\frac{1}{k^2})$. In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

[20] B. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer-Verlag, Berlin, Heidelberg, 1992.

[21] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[22] O. Sebbouh, Ch. Dossal, and A. Rondepierre. Convergence rates of damped inertial dynamics under geometric conditions and perturbations. *SIAM Journal on Optimization*, 30(3):1850–1877, 2020.

[23] B. Shi, S.S. Du, W.J. Su, and M.I. Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[24] B. Shi, S.S. Du, W.J. Su, and M.I. Jordan. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pages 1–70, 2021.

[25] W. Su, S. Boyd, and E. J. Candes. A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

A Proofs of Proposition 1

Let us now detail the proofs of Proposition 1. The two statements (10) and (11) will be proven independently. For those proofs, we will need the following inequality: for any collection of K vectors $x_1, \dots, x_K$, it holds

$$\left| \sum_{i=1}^{K} x_i \right|^2 \leqslant K \sum_{i=1}^{K} |x_i|^2 \tag{38}$$

which is a consequence of the Jensen inequality, and the discrete Grönwall lemma:

Lemma 2 (Discrete Grönwall Lemma)**.** *Let $(z_n)_{n\in\mathbb{N}}$ be a sequence of positive real numbers with $z_0 = 0$. If there are nonnegative constants C, τ and L such that, for any $n \geqslant 0$ it holds*

$$z_{n+1} \leqslant C + L\tau \sum_{k=0}^{n} z_k,$$

then for all $n \geqslant 0$

$$z_n \leqslant C e^{Ln\tau}.$$

The error analysis for the deterministic limit stated in (10) is a direct consequence of the following result.

Proposition 3. *Let $\Phi : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ be a Lipschitz function.Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a bounded integrable function. Let u_0 be a given vector in $\mathbb{R}^d$, $\boldsymbol{h} \in [0,1]$ and T be non negative real number. Fix N an integer and set $T = N\boldsymbol{h}$. Define $t_n = n\boldsymbol{h}$, $0 \leqslant n \leqslant N$ and consider the sequences X_n and Y_n defined by recursion*

$$X_{n+1} = X_n + \boldsymbol{h}\ \Phi(t_n, X_n) + \sqrt{\boldsymbol{h}}\ \sigma(t_n)(B_{t_{n+1}} - B_{t_n}),$$

$$Y_{n+1} = Y_n + \boldsymbol{h}\ \Phi(t_n, Y_n),$$

starting with $X_0 = Y_0 = u_0$. Then, there exists a non negative constant C_1 such that the following error estimate holds

$$\sup_{0 \leqslant n \leqslant N} \|Y_n - X_n\|_{\mathrm{L}^2(\Omega)} \leqslant C_1 \sqrt{\boldsymbol{h}}. \tag{39}$$

Proof of Proposition 3 Consider the misfit $e_k = X_k - Y_k$, $k \in \mathbb{N}$. By construction, it satisfies

$$e_{k+1} = e_k + \boldsymbol{h}\,(\Phi(t_k, X_k) - \Phi(t_k, Y_k)) - \sqrt{\boldsymbol{h}}\sigma(t_k)(B_{t_{k+1}} - B_{t_k}).$$

Let $n \in \mathbb{N}$. Summing these relations from $k = 0$ to $k = n$ and remembering that $X_0 = Y_0$, we get:

$$e_{n+1} \quad = \quad \boldsymbol{h} \sum_{k=1}^{n} (\Phi(t_k, X_k) - \Phi(t_k, Y_k)) - \sqrt{\boldsymbol{h}} \sum_{k=0}^{n} \sigma(t_k)(B_{t_{k+1}} - B_{t_k})$$

which implies:

$$\mathbb{E}[|e_{n+1}|^2] \leqslant 2\boldsymbol{h}^2 \mathbb{E}\left[\left(\sum_{k=1}^{n} (\Phi(t_k, X_k) - \Phi(t_k, Y_k))\right)^2\right] + 2\boldsymbol{h}\mathbb{E}\left[\left(\sum_{k=0}^{n} \sigma(t_k)(B_{t_{k+1}} - B_{t_k})\right)^2\right].$$

Observe now that using the inequality (38), we have:

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{k=1}^{n} (\Phi(t_k, X_k) - \Phi(t_k, Y_k))\right)^2\right] \quad &\leqslant n \textstyle\sum_{k=1}^{n} \mathbb{E}\left[(\Phi(t_k, X_k) - \Phi(t_k, Y_k))^2\right] \\ &\leqslant n Lip(\Phi)^2 \textstyle\sum_{k=1}^{n} \mathbb{E}[|e_k|^2],\end{aligned}$$

where $Lip(\Phi)$ denotes the Lipschitz constant of the function Φ. For the second term, we use that the Brownian increments are both independent and centered to get

$$\mathbb{E}\left[\left(\sum_{k=0}^{n} \sigma(t_k)(B_{t_{k+1}} - B_{t_k})\right)^2\right] = \sum_{k=0}^{n} \sigma(t_k)^2 \mathbb{E}\left[\left(B_{t_{k+1}} - B_{t_k}\right)^2\right] \leqslant (n+1)\|\sigma\|_\infty^2 \boldsymbol{h}.$$

Combining the last two inequalities we then get:

$$\begin{aligned}\mathbb{E}[|e_{n+1}|^2] \quad &\leq \quad 2n\boldsymbol{h}^2 \left(Lip(\Phi)^2 \sum_{k=0}^{n} \mathbb{E}[|e_k|^2] + \frac{n+1}{n}\|\sigma\|_\infty^2\right) \\ &\leqslant \quad 2\boldsymbol{h}T \left(Lip(\Phi)^2 \sum_{k=0}^{n} \mathbb{E}[|e_k|^2] + 2\|\sigma\|_\infty^2\right)\end{aligned}$$

Finally, applying the discrete Grönwall Lemma we obtain the inequality (39) as expected:

$$\forall n \leqslant N,\ \mathbb{E}[|e_n|^2] \leqslant 4T\|\sigma\|_\infty^2 e^{2TLip(\Phi)^2}\ \boldsymbol{h}.$$

The second inequality (10) is a direct consequence of the previous one and of the classical error estimate for the explicit Euler scheme.

The key of the proof of second result is mainly a precise error estimate for the Euler-Maruyama method in the case of an additive noise with a precise computation of the constant. This requires more regularity on the function Φ. We use the notation $Lip(\sigma)$ to denote the (best) Lipschitz constant of σ and where $\Delta_2\Phi$ to denote the Laplacian of Φ with respect to the second variable.

Proposition 4. *Let $\Phi : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ be a $\mathcal{C}^2$ Lipschitz function with a bounded second derivative. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a bounded Lipschitz function. Let u_0 be a given vector in $\mathbb{R}^d$ and let $h, \lambda \in [0,1]$ and T be non negative real number. Fix N an integer and set $T = \boldsymbol{h}N$. Let u be the solution of the stochastic differential equation:*

$$du = \Phi(t, u(t))dt + \lambda\sigma(t)dB_t,$$

with the initial condition $u(0) = u_0$. Set $t_n = n\boldsymbol{h}$ and consider the sequence X_n defined by recursion

$$X_{n+1} = X_n + \boldsymbol{h}\ \Phi(t_n, X_n) + \lambda\sigma(t_n)(B_{t_{n+1}} - B_{t_n}).$$

Then

$$\sup_{0\leqslant n\leqslant N} \|u(t_n) - X_n\|_{\mathrm{L}^2(\Omega)} \leqslant C\boldsymbol{h},$$

where the non negative constant C can be chosen as

$$C = e^{T^2Lip(\Phi)^2}\Big(\frac{2}{3}TLip(\sigma)^2 + 6T^2Lip(\Phi)^2\|\Phi\|_\infty^2 + 9TLip(\Phi)^2\|\sigma\|_\infty^2 + \frac{3}{2}T^2\|\Delta_2\Phi\|_\infty^2\|\sigma\|_\infty^4 + 2T^2Lip(\Phi)^2\Big)$$

independent of λ and $\boldsymbol{h}$.

In this work, we are deeply interested in the particular case $\lambda = \sqrt{\boldsymbol{h}}$ corresponding to (X_n) defined in (8), we obtain the error analysis for the approximation by the SDE stated in (11).

Proof of Proposition 4 Consider the error $e_n = u(t_n) - X_n$ at step n. By definition, it satisfies the recursion formula :

$$e_{n+1} = e_n + \int_{t_n}^{t_{n+1}} (\Phi(s, u(s)) - \Phi(t_n, X_n))\, ds + \lambda \int_{t_n}^{t_{n+1}} (\sigma(s) - \sigma(t_n))\, dB_s.$$

Summing from 0 to n, we get

$$e_{n+1} = \int_0^{t_{n+1}} \big(\Phi(s, u(s)) - \Phi(t_{n(s)}, X_{n(s)})\big)\, ds + \lambda \int_0^{t_{n+1}} \big(\sigma(s) - \sigma(t_{n(s)})\big)\, dB_s$$

where $n(s)$ denotes the entire part: $n(s) = \lfloor s/\boldsymbol{h} \rfloor$ for any $0 \leqslant s \leqslant T$ so that: $n(s) = k$ when $t_k \leqslant s < t_{k+1}$. It follows that

$$\begin{aligned}\mathbb{E}[|e_{n+1}|^2] \leqslant 2 \quad & \mathbb{E}\left[\left(\int_0^{t_{n+1}} \left(\Phi(s,u(s)) - \Phi(t_{n(s)}, X_{n(s)})\right) ds\right)^2\right] \\ & +2\lambda^2 \mathbb{E}\left[\left(\int_0^{t_{n+1}} \left(\sigma(s) - \sigma(t_{n(s)})\right) dB_s\right)^2\right].\end{aligned}$$

The second term is estimated using the Itô isometry:

$$\mathbb{E}\left[\left(\int_0^{t_{n+1}} \left(\sigma(s) - \sigma(t_{n(s)})\right) dB_s\right)^2\right] = \int_0^{t_{n+1}} \left(\sigma(s) - \sigma(t_{n(s)})\right)^2 dt \leqslant \frac{1}{3} T Lip(\sigma)^2 \boldsymbol{h}^2.$$

The first term is split to get a suitable bound

$$\begin{aligned}\Phi(s,u(s)) - & \Phi(t_{n(s)}, X_{n(s)}) = \Phi(s,u(s)) - \Phi(s,u(t_{n(s)})) \\ + \quad & \Phi(s,u(t_{n(s)})) - \Phi(t_{n(s)},u(t_{n(s)})) + \Phi(t_{n(s)},u(t_{n(s)})) - \Phi(t_{n(s)},X_{n(s)}),\end{aligned}$$

so that:

$$\mathbb{E}\left[\left(\int_0^{t_{n+1}} \left(\Phi(s,u(s)) - \Phi(t_{n(s)}, X_{n(s)})\right) ds\right)^2\right] \leqslant 3(E_1 + E_2 + E_3)$$

where

$$\begin{aligned}E_1 &= \mathbb{E}\left[\left(\int_0^{t_{n+1}} \left(\Phi(s,u(s)) - \Phi(s,u(t_{n(s)}))\right) ds\right)^2\right], \\ E_2 &= \mathbb{E}\left[\left(\int_0^{t_{n+1}} \left(\Phi(s,u(t_{n(s)})) - \Phi(t_{n(s)},u(t_{n(s)}))\right) ds\right)^2\right], \\ E_3 &= \mathbb{E}\left[\left(\int_0^{t_{n+1}} \left(\Phi(t_{n(s)},u(t_{n(s)})) - \Phi(t_{n(s)},X_{n(s)})\right) ds\right)^2\right].\end{aligned}$$

The E_2 term is bounded using Lipschitz properties of Φ

$$E_2 \leqslant \frac{1}{3} T^2 Lip(\Phi)^2 \boldsymbol{h}^2,$$

and that by Cauchy-Schwarz inequality

$$\begin{aligned}\left(\int_0^{t_{n+1}} \left(\Phi(t_{n(s)},u(t_{n(s)})) - \Phi(t_{n(s)},X_{n(s)})\right) ds\right)^2 &\leqslant Lip(\Phi)^2 \left(\int_0^{t_{n+1}} |u(t_{n(s)}) - X_{n(s)}| ds\right)^2 \\ &\leqslant Lip(\Phi)^2 t_{n+1} \int_0^{t_{n+1}} |u(t_{n(s)}) - X_{n(s)}|^2 ds = Lip(\Phi)^2 t_{n+1} \sum_{k=1}^n |e_k|^2 \boldsymbol{h}\end{aligned}$$

so that

$$E_3 \leqslant \boldsymbol{h} Lip(\Phi)^2 t_{n+1} \sum_{k=1}^n \mathbb{E}[|e_k|^2].$$

To bound E_1, we first use Itô formula to the process $\tau \mapsto \Phi(s,u(\tau))$ to get

$$\begin{aligned}\Phi(s,u(s)) - \Phi(s,u(t_{n(s)})) = & \int_{t_{n(s)}}^s \partial_2 \Phi(s,u(\tau)) \Phi(\tau,u(\tau)) d\tau \\ & + \int_{t_{n(s)}}^s \partial_2 \Phi(s,u(\tau)) \lambda \sigma(\tau) dB_\tau + \frac{1}{2} \int_{t_{n(s)}}^s \Delta_2 \Phi(s,u(\tau)) \lambda^2 \sigma^2(\tau) d\tau.\end{aligned}$$

Remembering that $n(s) = k$ when $s \in [t_k, t_{k+1}]$, we have:

$$\begin{aligned}
E_1 &= \mathbb{E}\left[\left(\int_0^{t_{n+1}} \left(\Phi(s,u(s)) - \Phi(s,u(t_{n(s)}))\right) ds\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{k=0}^n \int_{t_k}^{t_{k+1}} \left(\int_{t_k}^s \partial_2\Phi(s,u(\tau))\Phi(\tau,u(\tau))d\tau + \int_{t_k}^s \partial_2\Phi(s,u(\tau))\lambda\sigma(\tau)dB_\tau \right.\right.\right. \\
&\qquad \left.\left.\left. + \frac{1}{2}\int_{t_k}^s \Delta_2\Phi(s,u(\tau))\lambda^2\sigma^2(\tau)d\tau\right) ds\right)^2\right] \\
&\leqslant 3\mathbb{E}\left[\left(\sum_{k=0}^n \int_{t_k}^{t_{k+1}} \int_{t_k}^s \partial_2\Phi(s,u(\tau))\Phi(\tau,u(\tau))d\tau ds\right)^2\right] \\
&\quad + 3\mathbb{E}\left[\left(\sum_{k=0}^n \int_{t_k}^{t_{k+1}} \int_{t_k}^s \partial_2\Phi(s,u(\tau))\lambda\sigma(\tau)dB_\tau ds\right)^2\right] \\
&\quad + \frac{3}{4}\mathbb{E}\left[\left(\sum_{k=0}^n \int_{t_k}^{t_{k+1}} \int_{t_k}^s \Delta_2\Phi(s,u(\tau))\lambda^2\sigma^2(\tau)d\tau ds\right)^2\right]
\end{aligned}$$

using the convexity inequality (38). We now study each term in the last sum. Using (38) again, the Cauchy-Schwarz inequality and the properties of Φ, we get

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{k=0}^n \int_{t_k}^{t_{k+1}} \int_{t_k}^s \partial_2\Phi(s,u(\tau))\Phi(\tau,u(\tau))d\tau ds\right)^2\right] \\
&\leqslant (n+1)\sum_{k=0}^n \mathbb{E}\left[\left(\int_{t_k}^{t_{k+1}} \int_{t_k}^s \partial_2\Phi(s,u(\tau))\Phi(\tau,u(\tau))d\tau ds\right)^2\right] \\
&\leqslant (n+1)\sum_{k=0}^n \boldsymbol{h}\ \mathbb{E}\left[\int_{t_k}^{t_{k+1}} \left(\int_{t_k}^s \partial_2\Phi(s,u(\tau))\Phi(\tau,u(\tau))d\tau\right)^2 ds\right] \\
&\leqslant t_{n+1}\mathbb{E}\left[\int_0^{t_{n+1}} \left(\int_{t_{n(s)}}^s \partial_2\Phi(s,u(\tau))\Phi(\tau,u(\tau))d\tau\right)^2 ds\right] \leqslant \frac{1}{3}\, t_{n+1}^2 Lip(\Phi)^2\|\Phi\|_\infty^2 \boldsymbol{h}^2.
\end{aligned}$$

The same method holds for the third term

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{k=0}^n \int_{t_k}^{t_{k+1}} \int_{t_k}^s \Delta_2\Phi(s,u(\tau))\lambda^2\sigma^2(\tau)d\tau ds\right)^2\right] \\
&\leqslant t_{n+1}\mathbb{E}\left[\int_0^{t_{n+1}} \left(\int_{t_{n(s)}}^s \Delta_2\Phi(s,u(\tau))\lambda^2\sigma^2(\tau)d\tau\right)^2 ds\right] \leqslant \frac{1}{3}\, t_{n+1}^2\lambda^4\|\Delta_2\Phi\|_\infty^2\|\sigma\|_\infty^4\boldsymbol{h}^2.
\end{aligned}$$

For the second term, we first expand the square, then use the fact that the increments of the noise are independent and centered, finally the Cauchy-Schwarz inequality and Itô isometry

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_{k=0}^n \int_{t_k}^{t_{k+1}} \int_{t_k}^s \partial_2\Phi(s,u(\tau))\lambda\sigma(\tau)dB_\tau ds\right)^2\right] \\
&= \sum_{k=0}^n\sum_{l=0}^n \mathbb{E}\left[\int_{t_k}^{t_{k+1}} \int_{t_k}^s \partial_2\Phi(s,u(\tau))\lambda\sigma(\tau)dB_\tau ds \int_{t_l}^{t_{l+1}} \int_{t_l}^s \partial_2\Phi(s,u(\tau))\lambda\sigma(\tau)dB_\tau ds\right] \\
&= \sum_{k=0}^n \sum_{\substack{l=0 \\ l\neq k}}^n \mathbb{E}\left[\int_{t_k}^{t_{k+1}} \int_{t_k}^s \partial_2\Phi(s,u(\tau))\lambda\sigma(\tau)dB_\tau ds\right]\mathbb{E}\left[\int_{t_l}^{t_{l+1}} \int_{t_l}^s \partial_2\Phi(s,u(\tau))\lambda\sigma(\tau)dB_\tau ds\right] \\
&\qquad + \sum_{k=0}^n \mathbb{E}\left[\left(\int_{t_k}^{t_{k+1}} \int_{t_k}^s \partial_2\Phi(s,u(\tau))\lambda\sigma(\tau)dB_\tau ds\right)^2\right] \\
&\leq \sum_{k=0}^n \mathbb{E}\left[\left(\int_{t_k}^{t_{k+1}} ds\right)\right] \int_{t_k}^{t_{k+1}} \mathbb{E}\left[\left(\int_{t_k}^s \partial_2\Phi(s,u(\tau))\lambda\sigma(\tau)dB_\tau\right)^2 ds\right] \\
&\leqslant \boldsymbol{h}\sum_{k=0}^n \int_{t_k}^{t_{k+1}} \mathbb{E}\left[\int_{t_k}^s \left(\partial_2\Phi(s,u(\tau))\lambda\sigma(\tau)\right)^2 d\tau\right] ds \leqslant \frac{1}{2}\, t_{n+1}\lambda^2 Lip(\Phi)^2\|\sigma\|_\infty^2\boldsymbol{h}^2.
\end{aligned}$$

Gathering all the previous estimations, we obtain that

$$\mathbb{E}\left[|e_{n+1}|^2\right] \leqslant C\boldsymbol{h}^2 + 6\boldsymbol{h}TLip(\Phi)^2 \sum_{k=1}^{n} \mathbb{E}\left[|e_k|^2\right].$$

where the constant C is explicitly given by

$$C = \frac{2}{3}\lambda^2 TLip(\sigma)^2 + 6T^2 Lip(\Phi)^2\|\Phi\|_\infty^2 \qquad +9T\lambda^2 Lip(\Phi)^2\|\sigma\|_\infty^2$$
$$+\frac{3}{2}T^2\lambda^4\|\Delta_2\Phi\|_\infty^2\|\sigma\|_\infty^4 + 2T^2 Lip(\Phi)^2.$$

As $\lambda \in [0,1]$,

$$C \leqslant \tilde{C} = \frac{2}{3}TLip(\sigma)^2 + 6T^2 Lip(\Phi)^2\|\Phi\|_\infty^2 \qquad +9TLip(\Phi)^2\|\sigma\|_\infty^2$$
$$+\frac{3}{2}T^2\|\Delta_2\Phi\|_\infty^2\|\sigma\|_\infty^4 + 2T^2 Lip(\Phi)^2.$$

We apply Grönwall lemma to get that for any $n \leqslant N$

$$\mathbb{E}\left[|e_n|^2\right] \leqslant \tilde{C}\boldsymbol{h}^2 e^{n\boldsymbol{h}TLip(\Phi)^2} \leqslant \tilde{C}\boldsymbol{h}^2 e^{T^2 Lip(\Phi)^2}.$$