Proceedings of Machine Learning Research vol 134:1–94, 2021 34rd Annual Conference on Learning Theory

Convergence rates and approximation results for SGD and its continuous-time counterpart

Xavier Fontaine XAVIER.FONTAINE@POLYTECHNIQUE.EDU
Centre Borelli, ENS Paris-Saclay

Valentin De Bortoli VALENTIN.DEBORTOLI@GMAIL.COM
Department of Statistics, University of Oxford

Alain Durmus ALAIN.DURMUS@ENS-PARIS-SACLAY.FR
Centre Borelli, ENS Paris-Saclay

Editors: Mikhail Belkin and Samory Kpotufe

Abstract

This paper proposes a thorough theoretical analysis of Stochastic Gradient Descent (SGD) with non-increasing step sizes. First, we show that the recursion defining SGD can be provably approximated by solutions of a time inhomogeneous Stochastic Differential Equation (SDE) using an appropriate coupling. In the specific case of a batch noise we refine our results using recent advances in Stein's method. Then, motivated by recent analyses of deterministic and stochastic optimization methods by their continuous counterpart, we study the long-time behavior of the continuous processes at hand and establish non-asymptotic bounds. To that purpose, we develop new comparison techniques which are of independent interest. Adapting these techniques to the discrete setting, we show that the same results hold for the corresponding SGD sequences. In our analysis, we notably improve non-asymptotic bounds in the convex setting for SGD under weaker assumptions than the ones considered in previous works. Finally, we also establish finite-time convergence results under various conditions, including relaxations of the famous Łojasiewicz inequality, which can be applied to a class of non-convex functions.

Keywords: Stochastic Gradient Descent, Stochastic Differential Equations, approximation results, convergence rates

1. Introduction

Recently, first-order optimization methods (Su et al., 2016) have been shown to share similar long-time behavior with solutions of certain Ordinary Differential Equations (ODE). One starting point of this analysis is to remark that most of these algorithms can be regarded as discretization schemes. For instance, gradient descent (GD) can be seen as the Euler discretization of the gradient flow corresponding to the objective function f, *i.e.*, the ODE $\mathrm{d}x(t)/\mathrm{d}t = -\nabla f(x(t))$. The analysis of the long-time behavior of solutions of this gradient flow equation provides fruitful insights on the convergence of GD. This idea has been adapted to the Nesterov acceleration scheme (Nesterov, 1983) by Su et al. (2016), and in this case the limiting continuous flow is associated with a second-order ODE. This result then allows for a much more intuitive analysis of this scheme and the technique has been subsequently extended to derive tighter estimates (Shi et al., 2018) or to analyze different settings (Krichene et al., 2015; Aujol et al., 2018; Apidopoulos et al., 2019).

Following this approach this paper proposes a new analysis of the Stochastic Gradient Descent (SGD) algorithm to optimize a continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ given stochastic estimates of its gradient in convex and non-convex settings. Using ODEs, and in particular the gradient flow equation, to study SGD and derive non-asymptotic convergence bounds has already been applied in numerous papers (Ljung, 1977; Kushner and Clark, 1978; Métivier and Priouret, 1984, 1987; Benveniste et al., 1990; Benaim, 1996; Borkar, 2009; Tadić and Doucet, 2017), see also (Hauer and Mazón, 2019) for an extension to metric spaces. However, to take into account more precisely the noisy nature of SGD, it has been recently suggested to use Stochastic Differential Equations (SDE) as continuous-time models for the analysis of SGD. Li et al. (2017) introduced Stochastic Modified Equations and established weak approximations theorems, gaining more intuition on SGD, in particular to obtain new hyper-parameter adjustment policies. In another line of work, Feng et al. (2019) derived uniform in time approximation bounds using ergodic properties of SDEs.

The first contribution of this paper is to show that SDEs can also be used as continuous-time processes properly modeling SGD with non-increasing stepsizes. In Section 2, we show that SGD with non-increasing stepsizes is a discretization of a certain class of stochastic continuous processes $(\mathbf{X}_t)_{t\geq 0}$ solutions of time inhomogeneous SDEs. More precisely, we derive strong and weak approximation estimates between the discrete and continuous processes. Our strong approximation results are new and rely on some appropriate coupling between SGD and the associated SDE. These new estimates highlight the advantages and limitations of the analysis of an SDE as a continuous-time proxy for SGD. In the specific case of a batch noise we can sharpen our analysis using recent advances in Stein's method.

However, in general, these approximation bounds between solutions of SDEs and recursions defined by SGD are derived under a finite time horizon $T \geq 0$ and the error between the discrete and the continuous-time processes does not go to zero as T goes to infinity, which is a strong limitation to study the long-time behavior of SGD, see (Li et al., 2017, 2019). We emphasize that our goal is not to address this problem here by showing uniform in time bounds between the two processes. Hence, as a second distinct contribution, we highlight how the long-time behavior of the continuous process related to SGD can be used to gain insight on the convergence of SGD itself. In that sense our work follows the same lines as (Su et al., 2016; Krichene et al., 2015; Aujol et al., 2018) which use continuous-time approaches to provide intuitive ways of deriving convergence results. More precisely, in the rest of the paper, we first study the behavior of $(t \mapsto \mathbb{E}[f(\mathbf{X}_t)] - \min_{\mathbb{R}^d} f)$ which can be analyzed under different sets of assumptions on f, including a convex and weakly quasi-convex setting. Then, we propose an adaptation of the main arguments of this analysis to the discrete setting. This allows us to show, under the same conditions, that $(\mathbb{E}[f(X_n)] - \min_{\mathbb{R}^d} f)_{n\in\mathbb{N}}$ also converges to 0 with the same rates, where $(X_n)_{n\in\mathbb{N}}$ is the recursion defined by SGD.

Based on this interpretation, we provide much simpler proofs of existing results and obtain sharper convergence rates for SGD than the ones derived in previous works in the convex and the weakly quasi-convex settings (Bach and Moulines, 2011; Taylor and Bach, 2019; Orvieto and Lucchi, 2019). In the convex setting, we prove for the first time that the convergence rates of SGD match the minimax lower-bounds (Agarwal et al., 2012) under the same assumptions as (Bach and Moulines, 2011), *i.e.*, for a convex objective function with Lipschitz gradient. Finally, we consider a relaxation of the weakly quasi-convex setting introduced in (Hardt et al., 2018). Recent works (Orvieto and Lucchi, 2019) use SDEs to analyze SGD and derive convergence rates in the weakly quasi-convex setting. However the rates they obtain are not optimal and we show that our analysis leads to better rates under weaker assumptions. To summarize, our contributions are as follows:

(i) We derive strong approximation results between the discrete-time and the continuous-time processes in Section 2. Our strong approximation results are new and rely on a specific coupling between SGD and the associated SDE. Contrary to other works our bounds cover the case of non-increasing stepsizes.

(ii) We introduce our main tools for the analysis of discrete and continuous-time processes and apply them in the context of strongly-convex functions to give intuition on our approach in Section 3. Then, we use them to study SGD for the minimization of convex functionals in Section 4. We show for the first time that the convergence rate is at least of order $\mathcal{O}(n^{-1/2})$, with stepsize $\gamma_n = \mathcal{O}(n^{-1/2})$, without bounded gradient assumptions (both for the continuous-time and discrete-time processes). This disproves a conjecture of (Bach and Moulines, 2011).

(iii) In Section 5, we relax the convexity assumption (weakly quasi-convex assumption). In this framework we improve on recent bounds by Orvieto and Lucchi (2019) and derive new convergence results under general Łojasiewicz-type assumptions.

2. SGD with Non-Increasing Stepsizes as a Time Inhomogeneous Diffusion Process

2.1. Problem Setting and Main Assumptions

Throughout this paper we consider the problem of the unconstrained minimization of $f \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R})$, an objective function satisfying the following regularity condition.

A1 *For any $x, y \in \mathbb{R}^d$, $\|\nabla f(x) - \nabla f(y)\| \leq \mathtt{L} \|x - y\|$, with $\mathtt{L} \geq 0$, i.e., f is $\mathtt{L}$-smooth.*

We consider the general case where we do not have access to ∇f but only to unbiased estimates. There are classically two ways to handle this and we will treat both of them in this paper.

A2 *There exists a Polish probability space $(\mathsf{Z}, \mathcal{Z}, \pi^Z)$ and $\eta \geq 0$ such that one of the following conditions holds:*
(a) There exists a function $H : \mathbb{R}^d \times \mathsf{Z} \to \mathbb{R}^d$ such that for any $x \in \mathbb{R}^d$,

$$\int_{\mathsf{Z}} H(x,z) \mathrm{d}\pi^Z(z) = \nabla f(x) \ , \qquad \int_{\mathsf{Z}} \|H(x,z) - \nabla f(x)\|^2 \, \mathrm{d}\pi^Z(z) \leq \eta \ .$$

(b) There exists a function $\tilde{f} : \mathbb{R}^d \times \mathsf{Z} \to \mathbb{R}$ such that for all $z \in \mathsf{Z}$, $\tilde{f}(\cdot, z) \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R})$ is $\mathtt{L}$-smooth. In addition, there exists $x^\star \in \mathbb{R}^d$ such that for any $x \in \mathbb{R}^d$

$$\int_{\mathsf{Z}} \tilde{f}(x,z) \mathrm{d}\pi^Z(z) = f(x) \ , \quad \int_{\mathsf{Z}} \nabla \tilde{f}(x,z) \mathrm{d}\pi^Z(z) = \nabla f(x) \ , \quad \int_{\mathsf{Z}} \|\nabla \tilde{f}(x^\star, z)\|^2 \mathrm{d}\pi^Z(z) \leq \eta \ .$$

In this case, we define $H = \nabla \tilde{f}$.

The first setting **A**2-(a) corresponds to the stochastic approximation setting with a square-integrable noise term and has been studied in (Robbins and Monro, 1951; Bach and Moulines, 2011; Orvieto and Lucchi, 2019). This is a weaker assumption than the bounded gradient assumption considered in (Kingma and Ba, 2014; Shamir and Zhang, 2013; Feng et al., 2019; Rakhlin et al., 2012). The second setting **A**2-(b) relaxes the square-integrability condition, which is often not satisfied in classical machine learning problems (logistic regression or smooth Support Vector Machines) at the cost of imposing the Lipschitz regularity of $H(\cdot, z)$ for all $z \in \mathsf{Z}$. We also point out that the

Polish assumption (*i.e.*, the space Z is metric, complete and separable) is only used in the proof of Theorem 1 and can be avoided in the rest of the paper.

Under **A**1 and **A**2, we introduce the sequence $(X_n)_{n\in\mathbb{N}}$ starting from $X_0 \in \mathbb{R}^d$ corresponding to SGD with non-increasing stepsizes and defined for any $n \in \mathbb{N}$ by

$$X_{n+1} = X_n - \gamma(n+1)^{-\alpha} H(X_n, Z_{n+1}) \ , \tag{1}$$

where $\gamma > 0$, $\alpha \in [0,1]$ and $(Z_n)_{n\in\mathbb{N}}$ is a sequence of independent random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ valued in $(\mathsf{Z}, \mathcal{Z})$ such that for any $n \in \mathbb{N}$, Z_n is distributed according to π^Z. We now turn to the continuous counterpart of (1). Define for any $x \in \mathbb{R}^d$, the semi-definite positive matrix $\Sigma(x) = \pi^Z(\{H(x,\cdot) - \nabla f(x)\}\{H(x,\cdot) - \nabla f(x)\}^\top)$ and, for $\alpha \in [0,1)$, consider the time inhomogeneous SDE,

$$\mathrm{d}\mathbf{X}_t = -(\gamma_\alpha + t)^{-\alpha}\{\nabla f(\mathbf{X}_t)\mathrm{d}t + \gamma_\alpha^{1/2}\Sigma(\mathbf{X}_t)^{1/2}\mathrm{d}\mathbf{B}_t\} \ , \tag{2}$$

where $\gamma_\alpha = \gamma^{1/(1-\alpha)}$ and $(\mathbf{B}_t)_{t\geq 0}$ is a d-dimensional Brownian motion. For solutions of this SDE to exist in a strong sense, we consider the following assumption on $x \mapsto \Sigma(x)^{1/2}$.

A3 *There exists* $\mathtt{M} \geq 0$ *such that for any* $x, y \in \mathbb{R}^d$, $\|\Sigma(x)^{1/2} - \Sigma(y)^{1/2}\| \leq \mathtt{M}\|x - y\|$.

Indeed, using (Karatzas and Shreve, 1991, Chapter 5, Theorem 2.5), strong solutions $(\mathbf{X}_t)_{t\geq 0}$ exist if **A**1 and **A**3 hold. Condition **A**3 can be hard to check in practice and can be replaced by the following stronger (but easier to verify) assumption: $\Sigma \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R}^{d\times d})$ with bounded Hessian, see (Stroock and Varadhan, 2007, Theorem 5.2.3). In the sequel, $(\mathbf{X}_t)_{t\geq 0}$ is referred to as the *continuous* SGD process in contrast to $(X_n)_{n\in\mathbb{N}}$ which is referred to as the *discrete* SGD process.

2.2. Approximations Results

In this section, we prove that $(\mathbf{X}_t)_{t\geq 0}$ solution of (2) is indeed, under some conditions, a continuous counterpart of $(X_n)_{n\in\mathbb{N}}$ given by (1). First, we informally derive the form of (2). Let $(\tilde{\mathbf{X}}_t)_{t\geq 0}$ be the linear interpolation of $(X_n)_{n\in\mathbb{N}}$, *i.e.*, for any $t \in [n\gamma_\alpha, (n+1)\gamma_\alpha]$, $n \in \mathbb{N}$, $\tilde{\mathbf{X}}_t = ((t - n\gamma_\alpha)X_{n+1} + ((n+1)\gamma_\alpha - t)X_n)/\gamma_\alpha$, with $\gamma_\alpha = \gamma^{1/(1-\alpha)}$. Using a first-order Taylor expansion and assuming that the noise is roughly Gaussian with zero-mean and covariance matrix $\Sigma(\tilde{\mathbf{X}}_{n\gamma_\alpha})$, we have the following approximation,

$$\begin{aligned}\tilde{\mathbf{X}}_{(n+1)\gamma_\alpha} - \tilde{\mathbf{X}}_{n\gamma_\alpha} &= X_{n+1} - X_n \approx -\gamma(n+1)^{-\alpha} H(\tilde{\mathbf{X}}_{n\gamma_\alpha}, Z_{n+1}) \\ &\approx -\gamma_\alpha(n\gamma_\alpha + \gamma_\alpha)^{-\alpha}\{\nabla f(\tilde{\mathbf{X}}_{n\gamma_\alpha}) + \Sigma(\tilde{\mathbf{X}}_{n\gamma_\alpha})^{1/2} G_{n+1}\} \\ &\approx -\int_{n\gamma_\alpha}^{(n+1)\gamma_\alpha}(s+\gamma_\alpha)^{-\alpha}\nabla f(\tilde{\mathbf{X}}_s)\mathrm{d}s - \gamma_\alpha^{1/2}\int_{n\gamma_\alpha}^{(n+1)\gamma_\alpha}(s+\gamma_\alpha)^{-\alpha}\Sigma(\tilde{\mathbf{X}}_s)^{1/2}\mathrm{d}\mathbf{B}_s \ ,\end{aligned} \tag{3}$$

where for any $n \in \mathbb{N}$, G_n is a d-dimensional standard Gaussian random variable. The next result justifies the *ansatz* (3) and establishes strong approximation bounds for SGD. We recall the definition of the Wasserstein (extended) distance of order 2, denoted $\mathbf{W}_2 : \mathscr{P}(\mathbb{R}^d) \times \mathscr{P}(\mathbb{R}^d) \to [0, +\infty]$ (where $\mathscr{P}(\mathbb{R}^d)$ is the set of probability measures over $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and given for any $\mu_1, \mu_2 \in \mathscr{P}(\mathbb{R}^d)$ by $\mathbf{W}_2^2(\mu_1, \mu_2) = \inf_{\Lambda\in\Gamma(\mu_1,\mu_2)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|v_1 - v_2\|^2 \mathrm{d}\Lambda(v_1, v_2)$, where $\Gamma(\mu_1, \mu_2) \subset \mathscr{P}(\mathbb{R}^{2d})$ is the set of transference plans between μ_1 and μ_2, *i.e.*, $\Lambda \in \Gamma(\mu_1, \mu_2)$ if for any $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, $\Lambda(\mathsf{A} \times \mathbb{R}^d) = \mu_1(\mathsf{A})$ and $\Lambda(\mathbb{R}^d \times \mathsf{A}) = \mu_2(\mathsf{A})$.

Theorem 1 *Let $\bar{\gamma} > 0$ and $\alpha \in [0,1)$. Assume* **A**1*,* **A**2-(b) *and* **A**3*. Given $(Z_n)_{n\in\mathbb{N}}$ a sequence of independent random variables such that for any $n \in \mathbb{N}$, Z_n is distributed according to π^Z, there exists $(\mathbf{B}_t)_{t\geq 0}$ such that the following hold:*

(a) $(\mathbf{B}_t)_{t\geq 0}$ is a d-dimensional Brownian motion.

(b) For any $T \geq 0$, there exists $C \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $n \in \mathbb{N}$ with $n \leq n_T = \lfloor T/\gamma_\alpha \rfloor$ and $\gamma_\alpha = \gamma^{1/(1-\alpha)}$ we have

$$\mathbb{E}^{1/2}\left[\|\mathbf{X}_{n\gamma_\alpha} - X_n\|^2\right] \leq C(\gamma^\delta \varepsilon + \gamma)(1 + \log(\gamma^{-1})) \ , \quad \textit{with } \delta = \min(1, (2-2\alpha)^{-1}) \ ,$$

where $(\mathbf{X}_t)_{t\geq 0}$ is solution of (2), $(X_n)_{n\in\mathbb{N}}$ is defined by (1) with $\mathbf{X}_0 = X_0 \in \mathbb{R}^d$ and

$$\varepsilon^2 = \sup_{n\in\{0,\dots,n_T\}} \mathbb{E}[\mathbf{W}_2^2(\nu^{\mathrm{d}}(\mathbf{X}_{n\gamma_\alpha}), \nu^{\mathrm{c}}(\mathbf{X}_{n\gamma_\alpha}))] \ , \tag{4}$$

where for any $\tilde{x} \in \mathbb{R}^d$, $\nu^{\mathrm{d}}(\tilde{x})$ is the distribution of $H(\tilde{x}, Z_0)$ and $\nu^{\mathrm{c}}(\tilde{x})$ is the distribution of $\nabla f(\tilde{x}) + \Sigma^{1/2}(\tilde{x})G$, with G a standard Gaussian random variable.

The proof is postponed to Appendix B.4. Note that the term $\log(\gamma^{-1})$ can be avoided if $\alpha \neq 1/2$. However for the sake of simplicity we include it for all values of $\alpha \in [0,1)$. Our proof relies on a coupling argument which is made explicit in Appendix B.2 and uses tools from the optimal transport theory. The rest of the proof extends approximation results from Milstein (1995) to our coupled setting. To the best of our knowledge, this strong approximation result is new. A few remarks are in order:

(a) This result illustrates the fundamental difference between SGD and discretization of SDEs such as the Euler-Maruyama (EM) discretization. In the fixed stepsize setting, *i.e.*, $\alpha = 0$, consider $(\mathbf{Y}_t)_{t\geq 0}$ and its EM discretization $(Y_n)_{n\in\mathbb{N}}$ given by $\mathbf{Y}_0 = Y_0 \in \mathbb{R}^d$ and for any $n \in \mathbb{N}$

$$\mathrm{d}\mathbf{Y}_t = \mathrm{b}(\mathbf{Y}_t)\mathrm{d}t + \sigma(\mathbf{Y}_t)\mathrm{d}\mathbf{B}_t \ , \qquad Y_{n+1} = Y_n + \gamma \mathrm{b}(Y_n) + \sqrt{\gamma}\sigma(Y_n)G_{n+1} \ , \tag{5}$$

with $b : \mathbb{R}^d \to \mathbb{R}^d$, $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$, and $(G_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables such that for any $n \in \mathbb{N}$, $\mathbb{E}[G_n] = 0$ and $\mathbb{E}[G_n G_n^\top] = \mathrm{Id}$. Recall that using Theorem 1, we have that in the Gaussian case the strong approximation bound for SGD is at least of order 1[1]. For SDE, this depends on the structure of σ. If σ is constant then the strong approximation is of order 1, otherwise it is of order $1/2$, see *e.g.*, (Kloeden and Platen, 2011; Milstein, 1995). In addition, it can be shown that if $(G_n)_{n\in\mathbb{N}}$ is no longer a sequence of Gaussian random variables then for $\mathrm{b} = 0$, $\sigma = \mathrm{Id}$, (but it holds under mild conditions on b and σ), there exists $C \geq 0$ such that for any $T \geq 0$, $\gamma > 0$, $n \in \mathbb{N}$, $n\gamma \leq T$, $\mathbb{E}^{1/2}[\|\mathbf{Y}_{n\gamma} - Y_n\|^2] \geq C\sqrt{T}$, *i.e.*, no strong approximation holds. The behavior is different for SGD for which we obtain a strong approximation of order $\mathcal{O}(\gamma^{1/2}\varepsilon)$, regardless the structure of the noise.

(b) We remark that a strong approximation of order $\mathcal{O}(\gamma^{1/2})$ can also be derived for the error between SGD and the associated gradient flow ODE. Replacing the gradient flow by a stochastic continuous-time process improves this error bound up to $\mathcal{O}(\gamma^{1/2}\varepsilon)$, where ε is a measure of the distance between the noise and some Gaussian distribution in $\mathbf{W}_2$. This highlights the fact that the SDE (2) is well-suited to model SGD (1) in the case of a noise which is close to a Gaussian, but might not be better than a classical ODE approach for a more general noise. In this case, we conjecture that an appropriate Lévy process would further improve these bounds.

[1] A method is of order $p > 0$ if $\mathbb{E}[\|\mathbf{Y}_{n\gamma} - Y_n\|^2] = \mathcal{O}(\gamma^p)$

(c) We highlight that Theorem 1 can be improved to obtain functional strong approximation bounds using Doob's inequality. Note also that we derive our results under the regularity assumption **A**2-(b) which implies that for any $z \in \mathsf{Z}$, $x \mapsto H(x, z)$ is Lipschitz continuous. It is not clear if our results can be extended to **A**2-(a). We postpone these investigations to future work.

(d) Our current results do not cover the case $\alpha = 1$. However, changing the sequence of step sizes to $\gamma_n = \gamma/(\gamma n + \delta)$ where $\delta > 0$ is some parameter a modified *ansatz* holds and our strong approximation results can be adapted to this setting.

We now present a refinement of Theorem 1 in the case of batch noise. We begin by recalling the batch noise setting. Assume that $(\mathsf{Z}, \mathcal{Z}) = (\mathsf{Y}^M, \mathcal{Y}^{\otimes M})$, $\pi^Z = \pi^{\otimes M}$ with $M \in \mathbb{N}$, π a probability measure on the Polish space $(\mathsf{Y}, \mathcal{Y})$ and for any $x \in \mathbb{R}^d$, $z = \{y_i\}_{i=1}^M$, let

$$H(x, z) = (1/M) \textstyle\sum_{i=1}^M \nabla \tilde{f}(x, y_i) \ . \tag{6}$$

Note that $\Sigma = (1/M)\Sigma_f$, where for any $x \in \mathbb{R}^d$, $\Sigma_f(x) = \pi[(\nabla \tilde{f} - \nabla f(x))(\nabla \tilde{f} - \nabla f(x))^\top]$.

Corollary 2 *Let $\bar{\gamma} > 0$ and $\alpha \in [0, 1)$. Assume* **A**1*,* **A**2-(b) *and* **A**3 *(with respect to $(\mathsf{Y}, \mathcal{Y}, \pi)$). Let H be given by (6). Assume that there exists $x^\star \in \mathbb{R}^d$, $C, p \geq 0$ such that for any $x \in \mathbb{R}^d$ and $y \in \mathsf{Y}$*

$$\textstyle\int_{\mathsf{Y}} \|\nabla \tilde{f}(x^\star, y)\|^4 \mathrm{d}\pi(y) < +\infty \ , \quad \|\Sigma_f(x)^{-1/2}\| \leq C(1 + \|x\|^p) \ .$$

Then, there exists a random variable $((\mathbf{B}_t)_{t \geq 0}, (Z_n)_{n \in \mathbb{N}})$ such that for any $T \geq 0$, there exists $C \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $n \in \mathbb{N}$ with $n\gamma_\alpha \leq T$ $\gamma_\alpha = \gamma^{1/(1-\alpha)}$ we have

$$\mathbb{E}^{1/2}\left[\|\mathbf{X}_{n\gamma_\alpha} - X_n\|^2\right] \leq C(\gamma^\delta M^{-1} + \gamma)(1 + \log(\gamma^{-1})) \ , \quad \textit{with } \delta = \min(1, (2 - 2\alpha)^{-1}) \ .$$

The proof is postponed to Appendix B.5 and heavily relies on new quantitative bounds for the Central Limit Theorem established using Stein's method in Bonis (2020). Corollary 2 shows that in the presence of batch noise (and in the fixed stepsize setting), choosing a batch size $M = \mathcal{O}(\gamma^{-1/2})$ is enough to obtain a linear approximation between the continuous-time process and SGD. In Appendix B.5, we also show that a batch-size of order $M = \mathcal{O}(\gamma^{-1})$ is necessary to obtain a linear approximation between the deterministic gradient flow and SGD. Finally, we also establish weak approximation errors between continuous and discrete versions of SGD but due to space constraints, they are stated and proved in Appendix B.6.

3. Convergence of the Continuous and Discrete SGD Processes

3.1. Two Basic Comparison Lemmas

We now turn to the convergence of SGD. Our general strategy is as follows: in the continuous-time setting, in order to derive sharp convergence rates for (2), we consider appropriate energy functions $\mathscr{V} : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}_+$ which depend on the conditions imposed on the function f. Then, we show that $t \mapsto v(t) = \mathbb{E}[\mathscr{V}(t, \mathbf{X}_t)]$ satisfies an ODE and prove that it is bounded using the following simple lemma.

Lemma 3 *Let $F \in \mathrm{C}^1(\mathbb{R}_+ \times \mathbb{R}, \mathbb{R})$ and $v \in \mathrm{C}^1(\mathbb{R}_+, \mathbb{R}_+)$ such that for all $t \geq 0$, $\mathrm{d}v(t)/\mathrm{d}t \leq F(t, v(t))$. If there exists $t_0 > 0$ and $A > 0$ such that for all $t \geq t_0$ and for all $u \geq A$, $F(t, u) < 0$, then there exists $B > 0$ such that for all $t \geq 0$, $v(t) \leq B$, with $B = \max(\max_{t \in [0, t_0]} v(t), A)$.*

Proof Assume that there exists $t \geq 0$ such that $v(t) > B$, and let $t_1 = \inf\{t \geq 0 : v(t) > B\}$. By definition of B, $t_1 \geq t_0$, and by continuity of v, $v(t_1) = B$. By assumption, $F(t_1, v(t_1)) < 0$. Then $\mathrm{d}v(t_1)/\mathrm{d}t < 0$ and there exists $t_2 < t_1$ such that $v(t_2) > v(t_1) = B$, hence the contradiction. ■

Considering discrete analogues of the energy functions and ODEs found in the study of the continuous process solution of (2), we derive explicit convergence bounds for the discrete SGD process. To that purpose, we establish a discrete analog of Lemma 3 whose proof is postponed to Appendix C.

Lemma 4 *Let $F : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ satisfying for any $n \in \mathbb{N}$, $F(n, \cdot) \in \mathrm{C}^1(\mathbb{R}, \mathbb{R})$. Let $(u_n)_{n\in\mathbb{N}}$ be a sequence of non-negative numbers satisfying for all $n \in \mathbb{N}$, $u_{n+1} - u_n \leq F(n, u_n)$. Assume that there exist $n_0 \in \mathbb{N}$ and $A_1 > 0$ such that for all $n \geq n_0$ and for all $x \geq A_1$, $F(n, x) < 0$. In addition, assume that there exists $A_2 > 0$ such that for all $n \geq n_0$ and for all $x \geq 0$, $F(n, x) \leq A_2$. Then, there exists $B > 0$ such that for all $n \in \mathbb{N}$, $u_n \leq B$ with $B = \max(\max_{n\leq n_0+1} u_n, A_1) + A_2$.*

3.2. Strongly-Convex Case

First, we illustrate the simplicity and effectiveness of our approach by recovering optimal convergence rates if the objective function is strongly convex. Due to the two settings associated with **A**2, we consider two versions of the strong convexity hypothesis, either directly on f if **A**2-(a) holds or on $\tilde{f}$ if **A**2-(b) holds.

F1 *Either one of the following conditions holds:*
(a) Case **A**2-(a)*: f is μ-strongly convex with $\mu > 0$, i.e., for any $x, y \in \mathbb{R}^d$, $\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \mu \|x - y\|^2$.*
(b) Case **A**2-(b)*: for all $z \in \mathsf{Z}$, $\tilde{f}(\cdot, z)$ is μ-strongly convex.*

Note that **F**1-(b) implies directly the strong convexity of f. The results presented below are not new, see (Bach and Moulines, 2011) for the discrete case and (Orvieto and Lucchi, 2019) for the continuous one, but they can be obtained very easily within our framework. We only derive our results in the continuous-time setting for pedagogical purposes, and gather their discrete counterparts in Appendix C. First, we derive convergence rates on the last iterates. Denote by $x^\star$ the unique minimizer of f (which exists under **F**1).

Theorem 5 *Let $\alpha, \gamma \in (0, 1)$ and $(\mathbf{X}_t)_{t\geq 0}$ be given by (2). Assume* **A**1*,* **A**2*,* **A**3 *and* **F**1*. Then there exists $C \geq 0$ (explicit in the proof) such that for any $T \geq 1$, $\mathbb{E}[\|\mathbf{X}_T - x^\star\|^2] \leq CT^{-\alpha}$.*

This result holds for both versions of **F**1 and we present below a proof under **F**1-(a). The result under **F**1-(b) is stated and proved in Appendix D.

Proof Let $\alpha, \gamma \in (0, 1)$ and consider $\mathcal{E} : \mathbb{R}_+ \to \mathbb{R}_+$ defined for $t \geq 0$ by $\mathcal{E}(t) = \mathbb{E}[(t+\gamma_\alpha)^\alpha \|\mathbf{X}_t - x^\star\|^2]$, with $\gamma_\alpha = \gamma^{1/(1-\alpha)}$. Using Dynkin's formula, see Lemma 48, we have for any $t \geq 0$,

$$\mathcal{E}(t) = \mathcal{E}(0) + \alpha \int_0^t \frac{\mathcal{E}(s)}{s + \gamma_\alpha} \mathrm{d}s + \int_0^t \gamma_\alpha \frac{\mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_s))\right]}{(s + \gamma_\alpha)^\alpha} \mathrm{d}s - 2\int_0^t \mathbb{E}\left[\langle \nabla f(\mathbf{X}_s), \mathbf{X}_s - x^\star\rangle\right] \mathrm{d}s \ .$$

We now differentiate this expression with respect to t and using **F**1-(a) and **A**2-(a), we get for any $t > 0$,

$$\begin{aligned}
\mathrm{d}\mathcal{E}(t)/\mathrm{d}t &= \alpha\mathcal{E}(t)(t + \gamma_\alpha)^{-1} - 2\mathbb{E}\left[\langle \nabla f(\mathbf{X}_t), \mathbf{X}_t - x^\star\rangle\right] + \gamma_\alpha \mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_t))\right](t + \gamma_\alpha)^{-\alpha} \\
&\leq \alpha\mathcal{E}(t)/(t + \gamma_\alpha) - 2\mu\mathbb{E}[\|\mathbf{X}_t - x^\star\|^2] + \gamma_\alpha\eta/(t + \gamma_\alpha)^\alpha \\
&\leq F(t, \mathcal{E}(t)) = \alpha\mathcal{E}(t)(t + \gamma_\alpha)^{-1} - 2\mu\mathcal{E}(t)(t + \gamma_\alpha)^{-\alpha} + \gamma_\alpha\eta(t + \gamma_\alpha)^{-\alpha} \ ,
\end{aligned}$$

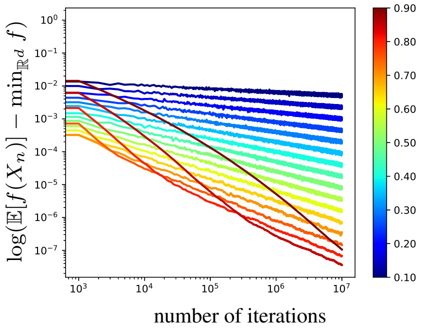

Figure 1: Evolution of $(\log(\mathbb{E}[f(X_n)] - \min_{\mathbb{R}^d} f))_{n\in\mathbb{N}}$

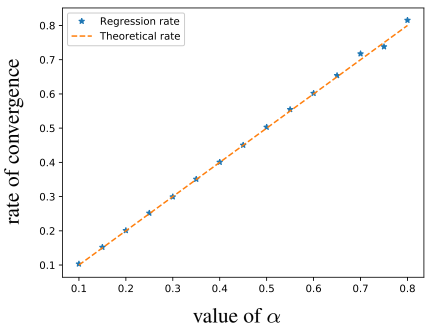

Figure 2: Empirical rates match theoretical rates for different values of α.

where we have used that $\mathrm{Tr}(\Sigma(x)) \leq \eta$ for any $x \in \mathbb{R}^d$ by **A**2-(a). Hence, since F satisfies the conditions of Lemma 3 with $t_0 = (\alpha/\mu)^{1/(1-\alpha)}$ and $A = 2\gamma_\alpha \eta/\mu$, applying this result we get, for any $t \geq 0$, $\mathcal{E}(t) \leq B$ with $B = \max(\max_{s\in[0,t_0]} \mathcal{E}(s), A)$ which concludes the proof. ■

Due to space constraints and to avoid over-complicated propositions, we do not precise the dependency of C with respect to the parameters μ, η and the initial condition. However, in Theorem 34 we obtain that (i) the constant in front of the asymptotic term $T^{-\alpha}$ scales as η/μ and (ii) the initial condition is forgotten exponentially fast.

In Theorem 31, we extend this result to the discrete setting using Lemma 4 and recover the rates obtained in (Bach and Moulines, 2011, Theorem 1) in the case where $\alpha \in (0, 1]$. In particular, if $\alpha = 1$, we obtain a convergence rate of order $\mathcal{O}(T^{-1})$ which matches the minimax lower-bounds established in (Nemirovsky and Yudin, 1983; Agarwal et al., 2012). In Figure 1 and Figure 2, we experimentally verify that the results we obtain are tight in the simple case where $f(x) = \|x\|^2$.

We emphasize that the strong convexity assumption can be relaxed if we only assume that f is weakly μ-strongly convex, *i.e.*, for any $x \in \mathbb{R}^d$, $\langle \nabla f(x), x - x^\star \rangle \geq \mu \|x - x^\star\|^2$. In (Kleinberg et al., 2018) the authors experimentally show that modern neural networks satisfy a relaxation of this last condition and it was proved in (Li and Yuan, 2017) that two-layer neural networks with ReLU activation functions are weakly μ-strongly convex if the inputs are Gaussian. Finally, we show in Corollary 30 and Corollary 32 that Theorem 5 also implies convergence rates for the process $(\mathbb{E}[f(\mathbf{X}_t)] - \min_{\mathbb{R}^d} f)_{t\geq 0}$ and its discrete counterpart.

4. Convex Case

In this section, we relax the strong convexity condition. Again we need to consider two different settings depending on the version of **A**2 we consider.

F2 *Either one of the following conditions holds:*
(a) Case **A**2-(a)*: f is convex, i.e., for any $x, y \in \mathbb{R}^d$, $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$, and there exists a minimizer $x^\star \in \arg\min_{\mathbb{R}^d} f$.*
(b) Case **A**2-(b)*: for all $z \in \mathsf{Z}$, $\tilde{f}(\cdot, z)$ is convex and there exists a minimizer $x^\star \in \arg\min_{\mathbb{R}^d} f$.*

Similarly to the strongly-convex case, we start by studying the continuous process. The discrete analog of the following result is given in Theorem 8.

Theorem 6 *Let $\alpha, \gamma \in (0,1)$ and $(\mathbf{X}_t)_{t\geq 0}$ be given by (2). Assume $f \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R})$,* **A**1*,* **A**2*,* **A**3 *and* **F**2*. Then, there exists $C \geq 0$ (explicit and given in the proof) such that for any $T \geq 1$*

$$\mathbb{E}\left[f(\mathbf{X}_T)\right] - \min_{\mathbb{R}^d} f \leq C(1+\log(T))^2/T^{\alpha\wedge(1-\alpha)} \; .$$

To the best of our knowledge, these non-asymptotic results are new for the continuous process $(\mathbf{X}_t)_{t\geq 0}$ defined by (2). Note that for $\alpha = 1/2$ the convergence rate is of order $\mathcal{O}(T^{-1/2}\log^2(T))$ which matches (up to a logarithmic term) the minimax lower-bound for the discrete-time process (Agarwal et al., 2012) and is in accordance with the tight bounds derived in the discrete case under additional assumptions (Shamir and Zhang, 2013). The general proof is postponed to Appendix E.2. The main strategy to prove Theorem 6 is to carefully analyze a continuous version of the suffix averaging (Shamir and Zhang, 2013; Harvey et al., 2019), introduced in the discrete case by Zhang (2004). We can relax the assumption $f \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R})$ assuming that the set $\arg\min_{\mathbb{R}^d} f$ is bounded.

Corollary 7 *Let $\alpha, \gamma \in (0,1)$ and $(\mathbf{X}_t)_{t\geq 0}$ be given by (2). Assume that $\arg\min_{\mathbb{R}^d} f$ is bounded,* **A**1*,* **A**2*,* **A**3 *and* **F**2*. Then, there exists $C \geq 0$ (explicit and given in the proof) such that for any $T \geq 1$,*

$$\mathbb{E}\left[f(\mathbf{X}_T)\right] - \min_{\mathbb{R}^d} f \leq C(1+\log(T))^2/T^{\alpha\wedge(1-\alpha)} \; .$$

The proof is postponed to Appendix E.2 and relies on the fact that if f is convex then for any $\varepsilon > 0$, $f * g_\varepsilon$ is also convex, where $(g_\varepsilon)_{\varepsilon>0}$ is a family of non-negative mollifiers. We now turn to the discrete counterpart of Theorem 6.

Theorem 8 *Let $\gamma, \alpha \in (0,1)$ and $(X_n)_{n\geq 0}$ be given by (1). Assume* **A**1*,* **A**2 *and* **F**2*. Then, there exists $C \geq 0$ (explicit and given in the proof) such that for any $N \geq 1$,*

$$\mathbb{E}\left[f(X_N)\right] - \min_{\mathbb{R}^d} f \leq C(1+\log(N+1))^2/(N+1)^{\alpha\wedge(1-\alpha)} \; .$$

The proof is postponed to Appendix E.3 and takes its inspiration from the proof of the continuous counterpart Theorem 6. Note that in the case $\alpha = 1/2$ we recover (up to a logarithmic term) the rate $\mathcal{O}(N^{-1/2}\log(N+1))$ derived in (Shamir and Zhang, 2013, Theorem 2) which matches the minimax lower-bound Agarwal et al. (2012), up to a logarithmic term. We also extend this result to the case $\alpha \neq 1/2$. Note however that our setting differs from the one of (Shamir and Zhang, 2013). Indeed, (Shamir and Zhang, 2013, Theorem 2) established the optimal convergence rate for a projected version of SGD onto a convex compact set of $\mathbb{R}^d$ under the assumption that f is convex (possibly non-smooth) and $(\mathbb{E}[\|H(X_n, Z_{n+1})\|^2])_{n\in\mathbb{N}}$ is bounded. Our result avoids the boundedness assumption and the projection step of (Shamir and Zhang, 2013), since in Theorem 8, we replace the boundedness condition by the regularity condition **A**1 (actually our proof can be very easily adapted to the setting of (Shamir and Zhang, 2013), see Corollary 61). Our main contributions in the convex setting are summarized in Table 1 and Figure 4.

On the other hand, the framework we consider is the same as (Bach and Moulines, 2011), but we always obtain better convergence rates and in particular we get an optimal choice for α ($\alpha = 1/2$) different from theirs ($\alpha = 2/3$), see Table 1. Hence, we disprove the conjecture formulated in (Bach and Moulines, 2011) which asserts that the minimax rate for SGD in this setting is $1/3$.

In Figure 3, we experimentally assess the results of Theorem 8. We apply SGD on the family of functions $(\varphi_p)_{p\in\mathbb{N}^\star}$, where for any $x \in \mathbb{R}$, $p \in \mathbb{N}^\star$,

$$\varphi_p(x) = x^{2p} \text{ , if } x \in [-1,1] \text{ , } \varphi_p(x) = 2p(|x|-1)+1 \text{ , otherwise .}$$

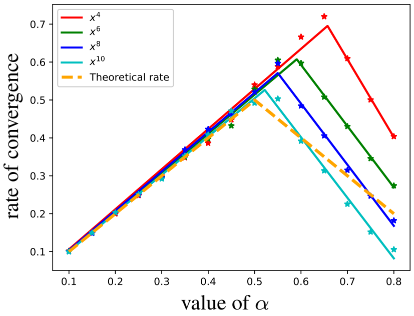

Figure 3: Convergence rates for φ_p match the theoretical results of Theorem 8 asymptotically.

Reference	Thm.8 (L)	(BM'11) (B, L)	(BM'11) (L)
$\alpha \in (0, 1/3)$	α	$\times$	$\times$
$(1/3, 1/2)$	α	$(3\alpha - 1)/2$	$\times$
$(1/2, 2/3)$	$1 - \alpha$	$\alpha/2$	$\alpha/2$
$(2/3, 1)$	$1 - \alpha$	$1 - \alpha$	$1 - \alpha$

Table 1: Convergence rates for convex SGD (B: Bounded gradients, L: Lipschitz gradient).

For any $p \in \mathbb{N}$, φ_p satisfies and **A**1 and **F**2. Denoting $\alpha_p^\star$ the non-increasing rate α for which the convergence rate $r_p^\star$ is maximum, we experimentally check that $\lim_{p\to+\infty} r_p^\star = 1/2$ and $\lim_{p\to+\infty} \alpha_p^\star = 1/2$. Note also that $\alpha_p^\star$ decreases as p grows, which is in accordance with the deterministic setting where the optimal rate in this case is given by $p/(p-2)$, see (Bolte et al., 2017; Frankel et al., 2015). As an immediate consequence of Theorem 8, we can show that $(\mathbb{E}[\|\nabla f(X_n)\|^2])_{n\in\mathbb{N}}$ enjoys the same rates of convergence as $(\mathbb{E}[f(X_n)] - \min_{\mathbb{R}^d} f)_{n\in\mathbb{N}}$, using that f is smooth.

Corollary 9 *Let $\gamma, \alpha \in (0, 1)$ and $(X_n)_{n\geq 0}$ be given by (1). Assume* **A**1*,* **A**2 *and* **F**2*. Then, there exists $C \geq 0$ (explicit and given in the proof) such that for any $N \geq 1$,*

$$\mathbb{E}[\|\nabla f(X_N)\|^2] \leq C(1 + \log(N+1))^2/(N+1)^{\alpha\wedge(1-\alpha)} .$$

In particular, $(\mathbb{E}[\|\nabla f(X_n)\|^2])_{n\in\mathbb{N}}$ is bounded which is often found as an assumption for the study of the convergence of SGD in the convex setting (Shalev-Shwartz et al., 2011; Nemirovski et al., 2009; Hazan and Kale, 2014; Shamir and Zhang, 2013; Recht et al., 2011). Our result shows that this assumption is unnecessary.

5. Weakly Quasi-Convex Case

In this section, we no longer consider that f is convex but a relaxation of this condition.

F3 *There exist $r_1 \in (0, 2)$, $r_2 \geq 0$, $\tau > 0$ such that for any $x \in \mathbb{R}^d$*

$$\|\nabla f(x)\|^{r_1} \|x - x^\star\|^{r_2} \geq \tau(f(x) - f(x^\star)) , \quad \textit{where } x^\star \in \arg\min_{\mathbb{R}^d} f \neq \emptyset .$$

This setting is a generalization of the weakly quasi-convex assumption considered in (Orvieto and Lucchi, 2019) and introduced in (Hardt et al., 2018) as follows.

F3b *The function f is weakly quasi-convex if there exists $\tau > 0$ such that for any $x \in \mathbb{R}^d$*

$$\langle \nabla f(x), x - x^\star \rangle \geq \tau(f(x) - f(x^\star)) , \quad \textit{where } x^\star \in \arg\min_{\mathbb{R}^d} f \neq \emptyset .$$

This last condition itself is a modification of the quasi-convexity assumption (Hazan et al., 2015). It was shown in (Hardt et al., 2018) that an idealized risk for linear dynamical system identification is weakly quasi-convex, and in (Yuan et al., 2019) the authors experimentally check that a residual network (ResNet20) used on CIFAR-10 (with differentiable activation units) satisfy the weakly quasi-convex assumption.

The assumption **F**3 also encompasses the setting where f satisfies some Kurdyka-Łojasiewicz condition (Bolte et al., 2017), *i.e.*, if there exist $r \in (0,2)$ and $\tilde{\tau} > 0$ such that for any $x \in \mathbb{R}^d$,

$$\|\nabla f(x)\|^r \geq \tilde{\tau}(f(x) - f(x^\star)) \ , \qquad \text{where } x^\star \in \arg\min_{\mathbb{R}^d} f \neq \emptyset \ . \tag{7}$$

In this case, **F**3 is satisfied with $r_1 = r$, $r_2 = 0$ and $\tau = \tilde{\tau}$. Kurdyka-Łojasiewicz conditions have often been used in the context of non-convex minimization (Attouch et al., 2010; Noll, 2014). Even though the case $r_1 = 2$ and $r_2 = 0$ is not considered in **F**3, one can still derive convergence of order α for $\alpha \in (0,1)$, see Proposition 36, extending the results obtained in the strongly convex setting. We now state the main theorem of this section.

Theorem 10 *Let $\alpha, \gamma \in (0,1)$ and $(\mathbf{X}_t)_{t\geq 0}$ be given by (2). Assume $f \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R})$,* **A**1*,* **A**2-(a)*,* **A**3 *and* **F**3*. In addition, assume that there exist $\beta, \varepsilon \geq 0$ and $C_{\beta,\varepsilon} \geq 0$ such that for any $t \geq 0$,*

$$\mathbb{E}[\|\mathbf{X}_t - x^\star\|^{r_2 r_3}] \leq C_{\beta,\varepsilon}(\gamma_\alpha + t)^\beta (1 + \log(1 + \gamma_\alpha^{-1} t))^\varepsilon \ ,$$

where $\gamma_\alpha = \gamma^{1/(1-\alpha)}$ and $r_3 = (1 - r_1/2)^{-1}$. Then, there exists $C \geq 0$ (explicit and given in the proof) such that for any $T \geq 1$

$$\mathbb{E}\left[f(\mathbf{X}_T)\right] - \min_{\mathbb{R}^d} f \leq C T^{-\delta}[1 + \log(1 + \gamma_\alpha^{-1} T)]^\varepsilon \ , \tag{8}$$

where $\delta_1 \wedge \delta_2$, $\quad \delta_1 = (r_1/2)(1 - r_1/2)^{-1}(1-\alpha) - \beta \quad$ and $\quad \delta_2 = (r_1/2)\alpha - \beta(1 - r_1/2)$.

The proof is postponed to Appendix G. First, note that if f satisfies a Kurdyka-Łojasiewicz condition of type (7) then **F**3 is satisfied with $r_1 = r$ and $r_2 = 0$ and the rates in Theorem 10 simplify and we obtain that $\delta = \min((r/2)(1-r/2)^{-1}(1-\alpha), (r/2)\alpha)$. The rate is maximized for $\alpha = (2-r/2)^{-1}$ and in this case, $\delta = r/(4-r)$. Therefore, if $r \to 2$, then $\delta \to 1$ and we obtain at the limit the same convergence rate that the case where f is strongly convex **F**1.

In the general case $r_2 \neq 0$, the convergence rates obtained in Theorem 10 depend on β where $(\mathbb{E}[\|\mathbf{X}_t - x^\star\|^{r_2 r_3}](\gamma_\alpha + t)^{-\beta})_{t\geq 0}$ has at most logarithmic growth. If $\beta \neq 0$, then the convergence rates deteriorate. In what follows, we shall consider different scenarios under which β can be explicitly controlled and Theorem 10 improved.

Corollary 11 *Let $\alpha, \gamma \in (0,1)$ and $(\mathbf{X}_t)_{t\geq 0}$ given by (2). Assume $f \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R})$,* **A**1*,* **A**2-(a)*,* **A**3*.*
(a) If **F**3b *holds, then there exists $C \geq 0$ such that for any $T \geq 1$*

$$\mathbb{E}\left[f(\mathbf{X}_T)\right] - \min_{\mathbb{R}^d} f \leq C[T^{(1-3\alpha)/2} + T^{-\alpha/2} + T^{\alpha-1}] \ .$$

(b) If **F**3b *holds and there exist $R \geq 0$ and $c > 0$ such that for any $x \in \mathbb{R}^d$ with $\|x - x^\star\| \geq R$, $f(x) - f(x^\star) \geq c\|x - x^\star\|$ then there exists $C \geq 0$ such that for any $T \geq 1$*

$$\mathbb{E}\left[f(\mathbf{X}_T)\right] - \min_{\mathbb{R}^d} f \leq C[T^{-\alpha/2} + T^{\alpha-1}] \ . \tag{9}$$

Reference	Corollary 11-(a)	Corollary 11-(b)	(OL'19)
$\alpha \in (0, 1/3)$	$\times$	$\alpha/2$	$\times$
$\alpha \in (1/3, 1/2)$	$(3\alpha - 1)/2$	$\alpha/2$	$\times$
$\alpha = 1/2$	$1/4 + \log.$	$1/4 + \log.$	$\times$
$\alpha \in (1/2, 2/3)$	$\alpha/2$	$1 - \alpha$	$2\alpha - 1$
$\alpha \in (2/3, 1)$	$1 - \alpha$	$1 - \alpha$	$1 - \alpha$

Table 2: Rates for continuous SGD with non-convex assumptions

(c) If **F**3 *holds and if there exist $R \geq 0$ and $\mathtt{m} > 0$ such that for any $x \in \mathbb{R}^d$ with $\|x - x^\star\| \geq R$, $\langle \nabla f(x), x - x^\star \rangle \geq \mathtt{m} \|x - x^\star\|^2$, then there exists $C \geq 0$ such that for any $T \geq 1$,* (9) *holds.*

The proof is postponed to Appendix G. The main ingredient of the proof is to control the growth of $t \mapsto \mathbb{E}[\|\mathbf{X}_t - x^\star\|^2]$ using either the SDE satisfied by $(\|\mathbf{X}_t - x^\star\|^2)_{t \geq 0}$ in the case of (a) and (c), or the SDE satisfied by $(f(\mathbf{X}_t) - \min_{\mathbb{R}^d} f)_{t \geq 0}$ in the case of (b).

Under **F**3b, we compare the rates we obtain using Corollary 11-(a) with the ones derived by (Orvieto and Lucchi, 2019) in Table 2 and Figure 4. Note that compared to (Orvieto and Lucchi, 2019), we establish that SGD converges as soon as $\alpha > 1/3$ and not $\alpha > 1/2$. In addition, the convergence rates we obtain are always better than the ones of (Orvieto and Lucchi, 2019). However, note that in both cases, the optimal convergence rate is $1/3$ obtained using $\alpha = 2/3$. In addition, under additional growth conditions on the function f, and using Corollary 11-(b)-(c) we show that the convergence of SGD in the weak quasi-convex case occurs as soon as $\alpha > 0$. Finally, as in the previous sections, we extend our main result to the discrete setting.

Theorem 12 *Let $\alpha, \gamma \in (0, 1)$ and $(X_n)_{n \in \mathbb{N}}$ be given by* (1)*. Assume* **A**1*,* **A**2-(a) *and* **F**3*. In addition, assume that there exist $\beta, \varepsilon, C_{\beta,\varepsilon} \geq 0$ such that for any $n \in \mathbb{N}$, $\mathbb{E}\left[\|X_n - x^\star\|^{r_2 r_3}\right] \leq C_{\beta,\varepsilon}(n+1)^\beta \{1 + \log(1+n)\}^\varepsilon$, where $r_3 = (1 - r_1/2)^{-1}$. Then, there exists $C \geq 0$ (explicit and given in the proof) such that for any $N \geq 1$*

$$\mathbb{E}\left[f(X_N)\right] - \min_{\mathbb{R}^d} f \leq C N^{-\delta_1 \wedge \delta_2} \left(1 + \log(1 + N)\right)^\varepsilon ,$$

where δ_1, δ_2 are given in (8)*.*

The proof is postponed to Appendix G. We can formulate the same remarks as the ones after Theorem 10, and Corollary 11 can be extended to the discrete case, see Corollary 80 in Appendix G.

6. Conclusion

In this paper we investigated the connection between SGD and solutions of appropriate time inhomogenuous SDEs. We first proved approximation bounds between these two processes motivating convergence analysis of continuous SGD. Then, we turned to the convergence behavior of SGD and showed how the continuous process can provide a better understanding of SGD using tools from ODE analysis and stochastic calculus. In particular, we obtained optimal convergence rates in the strongly convex case and new optimal convergence rates in the convex case. In the non-convex setting, we considered a relaxation of the weakly quasi-convex condition and improved the state-of-the art convergence rates in both the continuous and discrete-time setting.

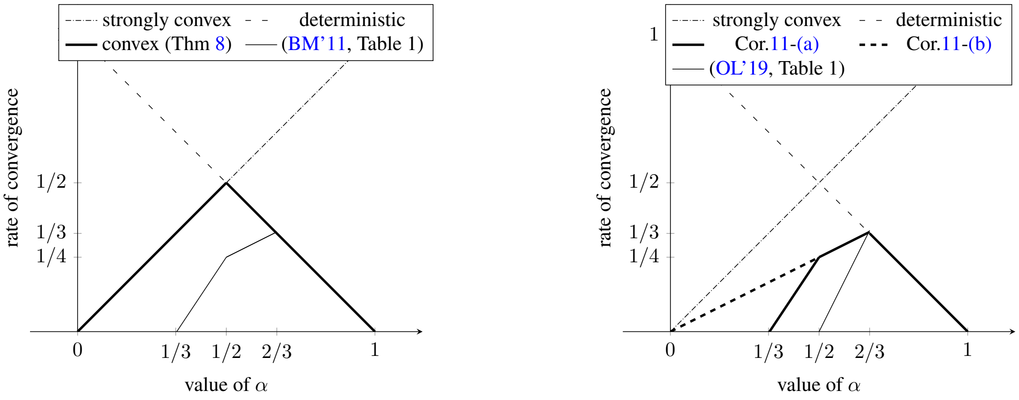

Figure 4: Comparison of convergence rates in convex (left) and weakly quasi-convex (right) settings.

7. Acknowledgement

Alain Durmus acknowledges support of the Lagrange Mathematics and Computing Research Center. Valentin De Bortoli is supported by the EPSRC CoSInES (COmputational Statistical INference for Engineering and Security) grant EP/R034710/1. Xavier Fontaine acknowledges support from Région Île-de-France.

References

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Trans. Information Theory*, 58(5):3235–3249, 2012. doi: 10.1109/TIT.2011.2182178. URL https://doi.org/10.1109/TIT.2011.2182178.

Charalambos D. Aliprantis and Kim C. Border. *Infinite dimensional analysis*. Springer, Berlin, third edition, 2006. ISBN 978-3-540-32696-0; 3-540-32696-0. A hitchhiker's guide.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008. ISBN 978-3-7643-8721-1.

Vassilis Apidopoulos, Jean-Franccois Aujol, Charles Dossal, and Aude Rondepierre. Convergence rates of an inertial gradient descent algorithm under growth and flatness conditions. 2019.

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

Jean Franccois Aujol, Aude Rondepierre, and Charles Dossal. Optimal convergence rates for nesterov acceleration. *arXiv preprint arXiv:1805.05719*, 2018.

Francis R. Bach and Eric Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011.*

Proceedings of a meeting held 12-14 December 2011, Granada, Spain, pages 451–459, 2011. URL http://papers.nips.cc/paper/4316-non-asymptotic-analysis-of-stochastic-approximation-algorithms-for-machine-learning.

Michel Benaim. A dynamical system approach to stochastic approximations. *SIAM J. Control Optim.*, 34(2):437–472, 1996. ISSN 0363-0129. doi: 10.1137/S0363012993253534. URL https://doi.org/10.1137/S0363012993253534.

Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990. ISBN 3-540-52894-6. doi: 10.1007/978-3-642-75894-2. URL https://doi.org/10.1007/978-3-642-75894-2. Translated from the French by Stephen S. Wilson.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334, 1997.

Ju. N. Blagovescenskii and M. I. Freidlin. Some properties of diffusion processes depending on a parameter. *Dokl. Akad. Nauk SSSR*, 138:508–511, 1961. ISSN 0002-3264.

Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

Thomas Bonis. Stein's method for normal approximation in Wasserstein distances with application to the multivariate central limit theorem. *Probab. Theory Related Fields*, 178(3-4):827–860, 2020. ISSN 0178-8051. doi: 10.1007/s00440-020-00989-4. URL https://doi.org/10.1007/s00440-020-00989-4.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Yuanyuan Feng, Tingran Gao, Lei Li, Jian-Guo Liu, and Yulong Lu. Uniform-in-time weak error analysis for stochastic gradient descent algorithms via diffusion approximation. *CoRR*, abs/1902.00635, 2019. URL http://arxiv.org/abs/1902.00635.

Pierre Frankel, Guillaume Garrigos, and Juan Peypouquet. Splitting methods with variable metric for kurdyka–łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2015.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *J. Mach. Learn. Res.*, 19:29:1–29:44, 2018. URL http://jmlr.org/papers/v19/16-465.html.

Nicholas J. A. Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 1579–1613. PMLR, 2019. URL http://proceedings.mlr.press/v99/harvey19a.html.

Daniel Hauer and José Mazón. Kurdyka–łojasiewicz–simon inequality for gradient flows in metric spaces. *Transactions of the American Mathematical Society*, 372(7):4917–4976, 2019.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15:2489–2512, 2014. URL http://jmlr.org/papers/v15/hazan14a.html.

Elad Hazan, Kfir Y. Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1594–1602, 2015. URL http://papers.nips.cc/paper/5718-beyond-convexity-stochastic-quasi-convex-optimization.

Ioannis Karatzas and Steven E. Shreve. *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1991. ISBN 0-387-97655-8. doi: 10.1007/978-1-4612-0949-2. URL https://doi.org/10.1007/978-1-4612-0949-2.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition. In *European Conference on Machine Learning and Knowledge Discovery in Databases - Volume 9851*, ECML PKDD 2016, pages 795–811, Berlin, Heidelberg, 2016. Springer-Verlag. ISBN 9783319461274. doi: 10.1007/978-3-319-46128-1_50. URL https://doi.org/10.1007/978-3-319-46128-1_50.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Robert Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2703–2712, 2018. URL http://proceedings.mlr.press/v80/kleinberg18a.html.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2011. ISBN 9783540540625. URL https://books.google.fr/books?id=BCvtssom1CMC.

Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2845–2853. Curran Associates, Inc., 2015. URL http://papers.nips.cc/paper/5843-accelerated-mirror-descent-in-continuous-and-discrete-time.pdf.

Hiroshi Kunita. On the decomposition of solutions of stochastic differential equations. In *Stochastic integrals (Proc. Sympos., Univ. Durham, Durham, 1980)*, volume 851 of *Lecture Notes in Math.*, pages 213–255. Springer, Berlin-New York, 1981.

Harold J. Kushner and Dean S. Clark. *Stochastic approximation methods for constrained and unconstrained systems*, volume 26 of *Applied Mathematical Sciences*. Springer-Verlag, New York-Berlin, 1978. ISBN 0-387-90341-0.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 2101–2110, 2017. URL http://proceedings.mlr.press/v70/li17f.html.

Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: mathematical foundations. *J. Mach. Learn. Res.*, 20:40:1–40:47, 2019. URL http://jmlr.org/papers/v20/17-526.html.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 597–607, 2017. URL http://papers.nips.cc/paper/6662-convergence-analysis-of-two-layer-neural-networks-with-relu-activation.

Lennart Ljung. Analysis of recursive stochastic algorithms. *IEEE Trans. Automatic Control*, AC-22 (4):551–575, 1977. ISSN 0018-9286.

Michel Métivier and Pierre Priouret. Applications of a Kushner and Clark lemma to general classes of stochastic algorithms. *IEEE Trans. Inform. Theory*, 30(2, part 1):140–151, 1984. ISSN 0018-9448. doi: 10.1109/TIT.1984.1056894. URL https://doi.org/10.1109/TIT.1984.1056894.

Michel Métivier and Pierre Priouret. Théorèmes de convergence presque sure pour une classe d'algorithmes stochastiques à pas décroissant. *Probab. Theory Related Fields*, 74(3):403–428, 1987. ISSN 0178-8051. doi: 10.1007/BF00699098. URL https://doi.org/10.1007/BF00699098.

Grigori N. Milstein. *Numerical integration of stochastic differential equations*, volume 313 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1995. ISBN 0-7923-3213-X. doi: 10.1007/978-94-015-8455-5. URL https://doi.org/10.1007/978-94-015-8455-5. Translated and revised from the 1988 Russian original.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Arkadi S. Nemirovsky and David B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. ISBN 0-471-10345-4. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics.

Yurii E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

Yurii E. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004. ISBN 1-4020-7553-7. doi: 10.1007/978-1-4419-8853-9. URL https://doi.org/10.1007/978-1-4419-8853-9. A basic course.

Dominikus Noll. Convergence of non-smooth descent methods using the Kurdyka-Łojasiewicz inequality. *Journal of Optimization, Theory and Applications*, 2014. URL https://hal.archives-ouvertes.fr/hal-01868363.

Antonio Orvieto and Aurélien Lucchi. Continuous-time models for stochastic optimization algorithms. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 12589–12601, 2019. URL http://papers.nips.cc/paper/9424-continuous-time-models-for-stochastic-optimization-algorithms.

Baburao G. Pachpatte. *Inequalities for differential and integral equations*, volume 197 of *Mathematics in Science and Engineering*. Academic Press, Inc., San Diego, CA, 1998. ISBN 0-12-543430-8.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012. URL http://icml.cc/2012/papers/261.pdf.

Benjamin Recht, Christopher Ré, Stephen J. Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 693–701, 2011. URL http://papers.nips.cc/paper/4390-hogwild-a-lock-free-approach-to-parallelizing-stochastic-gradient-descent.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Chris Rogers and David Williams. *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2000. ISBN 0-521-77593-0. doi: 10.1017/CBO9781107590120. URL https://doi.org/10.1017/CBO9781107590120. Itô calculus, Reprint of the second (1994) edition.

Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: primal estimated sub-gradient solver for SVM. *Math. Program.*, 127(1):3–30, 2011. doi: 10.1007/s10107-010-0420-4.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

Bin Shi, Simon S. Du, Michael I. Jordan, and Weijie J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *CoRR*, abs/1810.08907, 2018. URL http://arxiv.org/abs/1810.08907.

Daniel W Stroock and SR Srinivasa Varadhan. *Multidimensional diffusion processes*. Springer, 2007.

Weijie Su, Stephen P. Boyd, and Emmanuel J. Candès. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *J. Mach. Learn. Res.*, 17:153:1–153:43, 2016. URL http://jmlr.org/papers/v17/15-084.html.

V. B. Tadić and A. Doucet. Asymptotic bias of stochastic gradient search. *Ann. Appl. Probab.*, 27(6): 3255–3304, 2017. ISSN 1050-5164. doi: 10.1214/16-AAP1272. URL https://doi.org/10.1214/16-AAP1272.

Denis Talay and Luciano Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic analysis and applications*, 8(4):483–509, 1990.

Adrien Taylor and Francis Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2934–2992, Phoenix, USA, 25–28 Jun 2019. PMLR. URL http://proceedings.mlr.press/v99/taylor19a.html.

Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9. URL https://doi.org/10.1007/978-3-540-71050-9. Old and new.

Zhuoning Yuan, Yan Yan, Rong Jin, and Tianbao Yang. Stagewise training accelerates convergence of testing error over SGD. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 2604–2614, 2019. URL http://papers.nips.cc/paper/8529-stagewise-training-accelerates-convergence-of-testing-error-over-sgd.

Tong Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, 2004. doi: 10.1145/1015330.1015332. URL https://doi.org/10.1145/1015330.1015332.

Appendix A. Organization of the appendix

In these appendices we gather the proofs of our results. We start by deriving strong and weak approximation bounds in Appendix B. We then turn to the study of the long-time behavior of SGD and its continuous-time counterpart for the minimization of strongly convex functions in Appendix C under **A**2-(a). The counterpart of these results in the case where **A**2-(b) holds is presented in Appendix D. In Appendix E, we analyze the convex setting under **A**2-(a). Again, the counterpart of these results in the case where **A**2-(b) holds is given in Appendix F. We conclude with the proofs of the weakly quasi-convex setting in Appendix G.

Contents

Appendix B. Approximation Results

In this section, we present the proof of our strong and weak approximation results. In Appendix B.1, we gather technical lemmas which will be of use throughout the section. Our coupling construction between the discrete-time and continuous processes is presented in Appendix B.2. In Appendix B.3 we provide moment bounds which constitute the first step towards deriving the strong approximation bounds in Appendix B.4. The refinement of our theorem in the presence of batch-noise is considered in Appendix B.5. We also derive weak approximation bounds in Appendix B.6. Throughout this section we will denote all the constants by the letter $\mathtt{A}$ followed by some subscript.

B.1. Technical Lemmas

In order to derive the finite-time strong approximations from the one-step approximations we will make use of the following version of the discrete Grönwall's lemma.

Lemma 13 *Let $(u_n)_{n\in\mathbb{N}}$, $(v_n)_{n\in\mathbb{N}}$ and $(w_n)_{n\in\mathbb{N}}$ such that for any $n \in \mathbb{N}$, $u_n, v_n, w_n \geq 0$ and $u_{n+1} \leq (1+v_n)u_n + w_n$. Then for any $n \in \mathbb{N}$*

$$u_n \leq \exp\left[\sum_{k=0}^{n-1} v_k\right]\left(u_0 + \sum_{k=0}^{n-1} w_k\right) .$$

Proof The proof is a straightforward consequence of the discrete Grönwall's lemma. ■

The sums appearing in Lemma 13 will be controlled with the following lemma.

Lemma 14 *Let $r > 0$, $\gamma > 0$, $\alpha \in [0,1)$ and $\gamma_\alpha = \gamma^{1/(1-\alpha)}$. Then for any $T \geq 0$, there exists $\mathtt{A}_{\alpha,r} \geq 0$ such that for any $N \in \mathbb{N}$ with $N\gamma_\alpha \leq T$ we have*

$$\gamma^r \sum_{k=0}^{N-1} (k+1)^{-\alpha r} \leq \begin{cases} \mathtt{A}_{\alpha,r}\gamma^r(1+\log(\gamma^{-1}))(1+\log(T)) , & \textit{if } \alpha \geq 1/r , \\ \mathtt{A}_{\alpha,r}\gamma^r\gamma_\alpha^{\alpha r-1}T^{1-\alpha r} , & \textit{otherwise} . \end{cases}$$

Proof Let $r > 0$, $\gamma > 0$ and $\alpha \in [0,1)$. If $\alpha > 1/r$ then there exists $\mathtt{A}_{\alpha,r} \geq 0$ such that

$$\gamma^r \sum_{k=0}^{N-1} (k+1)^{-\alpha r} \leq \mathtt{A}_{\alpha,r}\gamma^r .$$

If $\alpha < 1/r$ then there exists $\mathtt{A}_{\alpha,r} \geq 0$ such that

$$\gamma^r \sum_{k=0}^{N-1} (k+1)^{-\alpha r} \leq \mathtt{A}_{\alpha,r}\gamma^r N^{-\alpha r+1} \leq \mathtt{A}_{\alpha,r}\gamma^r\gamma_\alpha^{\alpha r-1}T^{1-\alpha r} .$$

if $\alpha = 1/r$ then there exists $\mathtt{A}_{\alpha,r} \geq 0$ such that

$$\gamma^r \sum_{k=0}^{N-1} (k+1)^{-\alpha r} \leq \gamma^r(1+\log(N)) \leq \mathtt{A}_{\alpha,r}\gamma^r(1+\log(T))(1+\log(\gamma^{-1})) .$$

■

Note that if $r = 1$ then $\gamma^r \sum_{k=0}^{N-1}(k+1)^{-\alpha r} \leq \mathtt{A}_{\alpha,1}T^{1-\alpha}$. Using a slight modification of Lemma 14 we also obtain that there exists $\tilde{\mathtt{A}}$ such that if $r = 1$ then $\gamma^r \sum_{k=0}^{N-1}(k+1)^{-\alpha r} \leq T^{1-\alpha} + \tilde{\mathtt{A}}$.

The following lemma derives upper-bound from the regularity assumption **A**1 and **A**3.

Lemma 15 *Assume **A**1 and **A**3. Then there exists $C \geq 0$ such that for any $x \in \mathbb{R}^d$,*

$$\|\nabla f(x)\| \leq C(1+\|x\|) , \qquad \|\Sigma^{1/2}(x)\| \leq C(1+\|x\|) , \qquad \|\Sigma(x)\| \leq C(1+\|x\|^2) .$$

Proof First, we have for any $x \in \mathbb{R}^d$ using **A**1

$$\|\nabla f(x)\| \leq \|\nabla f(0)\| + \mathtt{L}\|x\| \leq (\|\nabla f(0)\| + \mathtt{L})(1 + \|x\|) .$$

Similarly, we have for any $x \in \mathbb{R}^d$ using **A**3,

$$\|\Sigma^{1/2}(x)\| \leq (\|\nabla f(0)\| + \mathtt{M})(1 + \|x\|) . \tag{10}$$

Denote for any $x \in \mathbb{R}^d$, $(a_{i,j}(x))_{1 \leq i,j \leq d} = \Sigma(x)$ and $(b_{i,j}(x))_{1 \leq i,j \leq d} = \Sigma(x)^{1/2}$. Using the fact that for any $u, v \in \mathbb{R}$, $2uv \leq u^2 + v^2$ we get that for any $x \in \mathbb{R}^d$

$$\|\Sigma(x)\| \leq \sum_{i,j=1}^{d} |a_{i,j}(x)| \leq \sum_{i,j,k=1}^{d} |b_{i,j}(x) b_{j,k}(x)| \leq 2d\|\Sigma^{1/2}(x)\|^2 .$$

We conclude the proof upon combining this result and (10). ■

B.2. Construction of the coupling

In this section, we describe and prove the existence of an appropriate coupling between the discrete-time and continuous-time process. In the following sections, we always assume that $(\mathbf{B}_t)_{t \geq 0}$ and $(Z_n)_{n \in \mathbb{N}}$ are given by Theorem 16. The proof of Theorem 16 is based on an abstract construction of an appropriate measure on a joint space. In order to construct such a measure we use the gluing lemma (Ambrosio et al., 2008, Lemma 5.3.2) and tools from the optimal transport theory to impose the desired properties on the marginals.

Theorem 16 *Assume* **A**1 *and* **A**3*. Let $\alpha \in [0, 1)$, $\bar{\gamma} > 0$ and $\gamma \in (0, \bar{\gamma}]$. Then, there exists $(\mathbf{B}_t)_{t \geq 0}$ a d-dimensional Brownian motion and $(Z_n)_{n \in \mathbb{N}}$ such that the following hold:*

(a) For any $k \in \mathbb{N}$, Z_{k+1} has distribution π^Z and is independent from $\mathcal{K}_k$, where for any $k \in \mathbb{N}$

$$\mathcal{K}_k = \sigma(\{\mathbf{B}_t, Z_j \,:\, t \in [0, k\gamma_\alpha] \,,\, j \in \{1, \dots, k\}\}) ,$$

with $\mathcal{K}_0 = \{\emptyset, \Omega\}$. Similarly, for any $k \in \mathbb{N}$, $(\mathbf{B}_t - \mathbf{B}_{k\gamma_\alpha})_{t \geq 0}$ is independent from $\mathcal{K}_k$.

(b) For any $k \in \mathbb{N}$, there exists $\mathsf{A}_k \in \mathcal{B}(\mathbb{R}^d)$ such that $\mathbb{P}(\mathbf{X}_{k\gamma_\alpha} \in \mathsf{A}_k) = 1$ and for any $\tilde{x} \in \mathsf{A}_k$.

$$\mathbf{W}_2^2(\nu_k^{\mathrm{d}}(\tilde{x}), \nu_k^{\mathrm{c}}(\tilde{x})) = \mathbb{E}\left[\left\| H(\tilde{x}, Z_{k+1}) - \nabla f(\tilde{x}) - \gamma_\alpha^{-1/2}\Sigma^{1/2}(\tilde{x}) \int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha} \mathrm{d}\mathbf{B}_s \right\|^2\right] ,$$

where $(\mathbf{X}_t)_{t \geq 0}$ is a unique strong solution of (2) starting from $\mathbf{X}_0 = X_0 \in \mathbb{R}^d$, $\nu^{\mathrm{d}}(\tilde{x})$ is the distribution of $H(\tilde{x}, Z_0)$ and $\nu^{\mathrm{c}}(\tilde{x})$ is the distribution of $\nabla f(\tilde{x}) + \Sigma^{1/2}(\tilde{x})G$ with G a Gaussian random variable with zero mean and identity covariance matrix.

Proof Let $\alpha \in [0, 1)$, $\bar{\gamma} > 0$ and $\gamma \in (0, \bar{\gamma}]$. By recursion, we show that there exists $((\mathbf{B}_t^k)_{t \in [0, \gamma_\alpha]}, Z_k)_{k \in \mathbb{N}}$ such that for any $k \in \mathbb{N}$, the following assertion **H**1(k) is true.

H1 **(k)** *We have that* $(\mathbf{B}_t^{k+1})_{t\in[0,\gamma_\alpha]}$ *and* Z_{k+1} *are independent from* $\mathcal{H}_k = \sigma(\{\mathbf{B}_t^j, Z_j \;:\; t \in [0,\gamma_\alpha]\,,\; j \in \{1,\ldots,k-1\}\})$ *(with* $\mathcal{H}_0 = \{\emptyset, \Omega\}$*) and there exists* $\mathsf{A}_k \in \mathcal{B}(\mathbb{R}^d)$ *such that* $\mathbb{P}\left(\mathbf{Y}_{k\gamma_\alpha} \in \mathsf{A}_k\right) = 1$ *and for any* $\tilde{x} \in \mathsf{A}_k$*.*

$$\mathbf{W}_2(\nu^{\mathrm{d}}(\tilde{x}), \nu^{\mathrm{c}}(\tilde{x})) = \mathbb{E}\left[\left\| H(\tilde{x}, Z_{k+1}) - \nabla f(\tilde{x}) - \gamma_\alpha^{-1/2}\Sigma^{1/2}(\tilde{x}) \int_0^{\gamma_\alpha} \mathrm{d}\mathbf{B}_s^{k+1} \right\|^2\right] , \tag{11}$$

where $\nu^{\mathrm{d}}(\tilde{x})$ *is the distribution of* $H(\tilde{x}, Z_0)$ *and* $\nu^{\mathrm{c}}(\tilde{x})$ *is the distribution of* $\nabla f(\tilde{x}) + \Sigma^{1/2}(\tilde{x})G$ *with* G *a Gaussian random variable with zero mean and identity covariance matrix, and for any* $t \in [0, k\gamma_\alpha]$

$$\begin{aligned}
\mathbf{Y}_t = \mathbf{X}_0 + \sum_{j=0}^{n_t-1} \Big(& -\int_0^{\gamma_\alpha} ((j+1)\gamma_\alpha + s)^{-\alpha} \nabla f(\mathbf{Y}_{j\gamma_\alpha+s})\mathrm{d}s \\
& \qquad - \gamma_\alpha^{1/2} \int_0^{\gamma_\alpha} ((j+1)\gamma_\alpha + s)^{-\alpha} \Sigma^{1/2}(\mathbf{Y}_{j\gamma_\alpha+s})\mathrm{d}\mathbf{B}_s^{j+1} \Big) \\
& \qquad\quad - \int_0^{t-n_t\gamma_\alpha} ((n_t+1)\gamma_\alpha + s)^{-\alpha} \nabla f(\mathbf{Y}_{n_t\gamma_\alpha+s})\mathrm{d}s \\
& \qquad\qquad - \gamma_\alpha^{1/2} \int_0^{t-n_t\gamma_\alpha} ((n_t+1)\gamma_\alpha + s)^{-\alpha} \Sigma^{1/2}(\mathbf{Y}_{n_t\gamma_\alpha+s})\mathrm{d}\mathbf{B}_s^{n_t+1} ,
\end{aligned} \tag{12}$$

where $n_t = \lfloor t/\gamma_\alpha \rfloor$*. Denote* μ_k *the distribution of* $((\mathbf{B}_t^j)_{t\in[0,\gamma_\alpha]}, Z_j)_{j\in\{0,\ldots,k\}}$*.*

We denote $\pi^B \in \mathscr{P}(\mathrm{C}_{\gamma_\alpha}) = \mathscr{P}(\mathrm{C}([0,\gamma_\alpha]\,, \mathbb{R}^d))$ the distribution of the Brownian motion up to time γ_α. For any $k \in \mathbb{N}$ we denote $\mathsf{E}^k = (\mathrm{C}_{\gamma_\alpha} \times \mathsf{Z})^k$ and $\mathsf{E} = \mathsf{E}^1$. Similarly, we denote $\pi^Z \in \mathscr{P}(\mathsf{Z})$ the distribution of Z. For any $k \in \mathbb{N}$, let $F :\; \mathbb{R}^d \times \mathrm{C}_{\gamma_\alpha}$ such that for any $x \in \mathbb{R}^d$ and π^B-almost every $\mathrm{w} \in \mathrm{C}_{\gamma_\alpha}$ we have

$$F_k(x, \mathrm{w}) = \nabla f(x) + \gamma_\alpha^{-1/2}\Sigma^{1/2}(x) \int_0^{\gamma_\alpha} \mathrm{d}\mathrm{w}_s \;.$$

Since $x \mapsto F(x, \mathrm{w})$ is continuous for π^B-almost every $\mathrm{w} \in \mathrm{C}_{\gamma_\alpha}$ by **A**1 and **A**3, and for any $x \in \mathbb{R}^d$, $\mathrm{w} \mapsto F(x, \mathrm{w})$ is measurable, we get that F is measurable using (Aliprantis and Border, 2006, Lemma 4.51). In addition, using **A**1, **A**3 and (Rogers and Williams, 2000, Theorem 10.4), for any $k \in \mathbb{N}$, there exists a measurable mapping $\tilde{S}_{k+1} :\; \mathbb{R}^d \times \mathrm{C}_{\gamma_\alpha} \to \mathbb{R}^d$ such that for any Brownian motion $(\mathbf{B}_t)_{t\in[0,\gamma_\alpha]}$, $\tilde{S}_{k+1}(\tilde{x}, (\mathbf{B}_t)_{t\in[0,\gamma_\alpha]}) = \mathbf{Y}_{\gamma_\alpha}^k$, where $(\mathbf{Y}_t^k)_{t\in[0,\gamma_\alpha]}$ is the unique strong solution to the following SDE: for any $t \in [0, \gamma_\alpha]$

$$\mathbf{Y}_t^k = \tilde{x} - \int_0^t ((k+1)\gamma_\alpha + s)^{-\alpha} \nabla f(\mathbf{Y}_s^k)\mathrm{d}s - \gamma_\alpha^{1/2} \int_0^t ((k+1)\gamma_\alpha + s)^{-\alpha} \Sigma^{1/2}(\mathbf{Y}_s^k)\mathrm{d}\mathbf{B}_s \;.$$

In addition, let $\tilde{S}_0 :\; \mathbb{R}^d \times \mathrm{C}_{\gamma_\alpha} \to \mathbb{R}^d$ such that for any $\mathrm{w} \in \mathrm{C}_{\gamma_\alpha}$, $\tilde{S}_0(\tilde{x}, \mathrm{w}) = \tilde{x}$. For any $k \in \mathbb{N}$, denote $S_k :\; \mathsf{E}^{k+1} \to \mathbb{R}^d$ such that for any $\{(\mathrm{w}_t^j)_{t\in[0,\gamma_\alpha]}, Z_j\}_{j=0}^k \in \mathsf{E}^k$, we have

$$\begin{aligned}
& S_k(\{(\mathrm{w}_t^j)_{t\in[0,\gamma_\alpha]}, Z_j\}_{j=1}^k) \\
& \quad = \tilde{S}_k(\tilde{S}_{k-1}(\ldots(\tilde{S}_1(\tilde{S}_0(\mathbf{X}_0, (\mathrm{w}_t^0)_{t\in[0,\gamma_\alpha]}), (\mathrm{w}_t^1)_{t\in[0,\gamma_\alpha]}))\ldots, (\mathrm{w}_t^{k-1})_{t\in[0,\gamma_\alpha]}), (\mathrm{w}_t^k)_{t\in[0,\gamma_\alpha]})) \;.
\end{aligned}$$

Finally, for any $k \in \mathbb{N}$, let $\mathscr{F}_k : \mathsf{E}^{k+1} \times \mathrm{C}_{\gamma_\alpha} \to \mathbb{R}^d$ and $\mathscr{H}_k : \mathsf{E}^{k+1} \times \mathsf{Z} \to \mathbb{R}^d$ such that for any $k \in \mathbb{N}$, $\{(\mathrm{w}_t^j)_{t\in[0,\gamma_\alpha]}, Z_j\}_{j=0}^k \in \mathsf{E}^{k+1}$, $(\mathrm{w}_t)_{t\in[0,\gamma_\alpha]} \in \mathrm{C}_{\gamma_\alpha}$ and $z \in \mathsf{Z}$

$$\mathscr{F}_k(\{(\mathrm{w}_t^j)_{t\in[0,\gamma_\alpha]}, Z_j\}_{j=0}^k, (\mathrm{w}_t)_{t\in[0,\gamma_\alpha]}) = F(S_k(\{(\mathrm{w}_t^j)_{t\in[0,\gamma_\alpha]}, Z_j\}_{j=0}^k), (\mathrm{w}_t)_{t\in[0,\gamma_\alpha]}) ,$$
$$\mathscr{H}_k(\{(\mathrm{w}_t^j)_{t\in[0,\gamma_\alpha]}, Z_j\}_{j=0}^k, z) = H(S_k(\{(\mathrm{w}_t^j)_{t\in[0,\gamma_\alpha]}, Z_j\}_{j=0}^k), z) .$$

Note that for any $k \in \mathbb{N}$, $\mathscr{F}_k$ and $\mathscr{H}_k$ are measurable. For any $k \in \mathbb{N}$, let $\mathrm{Q}_k : \mathsf{E} \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$, the Markov kernel given for any $u \in \mathsf{E}^{k+1}$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d)$ by

$$\mathrm{Q}_k(u, \mathsf{A}) = \mathtt{Opt}(\mathscr{F}_k(u,\cdot)_\# \pi^B, \mathscr{H}_k(u,\cdot)_\# \pi^Z)(\mathsf{A}) ,$$

where for all $u \in \mathsf{E}^{k+1}$, $\mathtt{Opt}(\mathscr{F}_k(u,\cdot)_\# \pi^B, \mathscr{H}_k(u,\cdot)_\# \pi^Z)$ is the optimal transference plan between $\mathscr{F}_k(u,\cdot)_\# \pi^B$ and $\mathscr{H}_k(u,\cdot)_\# \pi^Z$ w.r.t. to the $\mathbf{W}_2$, which exists by (Villani, 2009, Theorem 4.1). Note that for any $k \in \mathbb{N}$, Q_k is well-defined since $u \mapsto \mathtt{Opt}(\mathscr{F}_k(u,\cdot)_\# \pi^B, \mathscr{H}_k(u,\cdot)_\# \pi^Z)$ is measurable, (Villani, 2009, Corollary 5.22).

We divide the rest of the proof into two parts. First, we show by recursion that for any $k \in \mathbb{N}$ the assertion $\mathbf{H}1(k)$ is true. Second, we show that we can construct a Brownian motion from the random variables introduced in $\mathbf{H}1(k)$ for any $k \in \mathbb{N}$ such that the proposition holds.

(a) We start by proving that $\mathbf{H}1(0)$ holds. Let $\mu_0 \in \mathscr{P}(\mathrm{C}_{\gamma_\alpha} \times \mathsf{Z})$ be any coupling between π^B and π^Z. Let $\eta^0 \in \mathscr{P}(\mathsf{E} \times \mathrm{C}_{\gamma_\alpha} \times \mathbb{R}^d \times \mathsf{E})$, $\eta^1 \in \mathscr{P}(\mathbb{R}^d \times \mathsf{E} \times \mathbb{R}^d)$ and $\eta^2 \in \mathscr{P}(\mathsf{E} \times \mathbb{R}^d \times \mathsf{E} \times \mathsf{Z})$ such that

$$\eta^0 = (\mathrm{Id}, (\mathscr{F}_0, \Pi_1))_\#(\mu_0 \otimes \pi^B) , \qquad \eta^2 = ((\Pi_1, \mathscr{H}_0), \mathrm{Id})_\#(\mu_0 \otimes \pi^Z) ,$$

where Π_1 is the projection on the first variable. In addition, let $\eta^1 = \mu_0 \otimes \mathrm{Q}_0$, *i.e.*, for any $\mathsf{A} \in \mathcal{B}(E)$ and $\mathsf{B}_1, \mathsf{B}_2 \in \mathcal{B}(\mathbb{R}^d)$ we have

$$\eta^1(\mathsf{B}_1 \times \mathsf{A} \times \mathsf{B}_2) = \int_{\mathsf{A}} \mathrm{Q}_0(x, \mathsf{B}_1 \times \mathsf{B}_2) \mathrm{d}\mu_0(x) .$$

Note that $\eta^0_1 = (\mathscr{F}_0, \mathrm{Id})_\#(\mu_0 \otimes \pi^B) = \eta^1_{12}$. Therefore, using the gluing lemma (Ambrosio et al., 2008, Lemma 5.3.2) (which is valid since E, $\mathrm{C}_{\gamma_\alpha}$ and Z are Polish spaces), there exists a probability measure $\tilde{\eta} \in \mathscr{P}(\mathsf{E} \times \mathrm{C}_{\gamma_\alpha} \times \mathbb{R}^d \times \mathsf{E} \times \mathbb{R}^d)$ such that $\tilde{\eta}_{1234} = \eta^0$ and $\tilde{\eta}_{345} = \eta^1$. In addition note that $\tilde{\eta}_{45} = (\mathrm{Id}, \mathscr{H}_0)_\#(\mu_0 \otimes \pi^Z) = \eta^2_{12}$. Therefore, using the gluing lemma, there exists a probability measure $\eta \in \mathscr{P}(\mathsf{E} \times \mathrm{C}_{\gamma_\alpha} \times \mathbb{R}^d \times \mathsf{E} \times \mathbb{R}^d \times \mathsf{Z} \times \mathsf{E})$ such that $\eta_{12345} = \tilde{\eta}$ and $\eta_{4567} = \eta^1$. In particular, $\eta^{1234} = \eta^0$ and $\eta^{4567} = \eta^1$. Let $(U_i)_{i\in\{1,\dots,7\}}$ be a random variable with distribution η. Then, using that $\eta^{1234} = \eta^0$ and $\eta^{4567} = \eta^1$, we have almost surely

$$U_3 = \mathscr{F}_0(U_1, U_2) , \qquad U_4 = U_1 , \qquad U_4 = U_7 , \qquad U_5 = \mathscr{H}_0(U_7, U_6) .$$

Therefore, we get that

$$(U_1, \dots, U_7) = (U_1, U_2, \mathscr{F}_0(U_1, U_2), U_1, \mathscr{H}_0(U_1, U_6), U_6, U_1) .$$

Since U_2 is independent from U_1 and U_6 is independent from U_7, we get that U_6 is independent from U_1. Hence, there exists $\mu_1 \in \mathscr{P}(\mathsf{E} \times \mathrm{C}_{\gamma_\alpha} \times \mathsf{Z})$ such that

$$\eta = (\Pi_1, \Pi_2, \mathscr{F}_0(\Pi_1, \Pi_2), \Pi_1, \mathscr{H}_0(\Pi_1, \Pi_3), \Pi_3, \Pi_1)_\# \mu_1 .$$

Let $((\mathbf{B}_t^0)_{t\in[0,\gamma_\alpha]}, Z_0, (\mathbf{B}_t^1)_{t\in[0,\gamma_\alpha]}, Z_1)$ be a random variable with distribution μ_1. Then $(\mathbf{B}_t^1)_{t\in[0,\gamma_\alpha]}$ and Z_1 are independent from $((\mathbf{B}_t^0)_{t\in[0,\gamma_\alpha]}, Z_0)$, $(\mathbf{B}_t^1)_{t\in[0,\gamma_\alpha]}$ has distribution π^B and Z_1 has distribution π^Z. Hence, $(\mathbf{B}_t^1)_{t\in[0,\gamma_\alpha]}$ and Z_1 are independent from $\mathcal{H}_0$. Finally, we show that (11) holds. Denote by R_0 the Markov kernel given for any $u \in \mathsf{E}$ and $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d)$ by

$$\mathrm{R}_0(u, \mathsf{A}) = \int_{\mathbb{R}^d} (\mathscr{F}_0(u, \Pi_2(\cdot)), \mathscr{H}_0(u, \Pi_3(\cdot))))_{\#}\mu_1 \mathrm{d}\mu_0(u) ,$$

Note that $\mu_1 = \mu_0 \otimes \mathrm{R}_0$. But by definition of μ_1 we also have that $\mu_1 = \mu_0 \otimes \mathrm{Q}_0$. Therefore, μ_0 almost surely we have $\mathrm{R}_0(x, \cdot) = \mathrm{Q}_0(x, \cdot)$, which concludes the proof of (11). Therefore $\mathbf{H}1(0)$ holds. Assume that $\mathbf{H}1(k)$ is true with $k \in \mathbb{N}$. Then $\mathbf{H}1(k+1)$ holds. The proof is similar to the one for $\mathbf{H}1(0)$ upon replacing μ_0 by μ_k, $\mathscr{F}_0$ by $\mathscr{F}_k$, $\mathscr{H}_0$ by $\mathscr{H}_k$ and Q_0 by Q_k. We conclude by recursion.

(b) Finally, it remains to define a Brownian motion $(\mathbf{B}_t)_{t\geq 0}$ such that for any $k \in \mathbb{N}$, $\mathcal{K}_k = \mathcal{H}_k$ and $(\mathbf{X}_t)_{t\geq 0} = (\mathbf{Y}_t)_{t\geq 0}$. For any $t \geq 0$, let $n_t = \lfloor t/\gamma_\alpha \rfloor$ and $(\mathbf{B}_t)_{t\geq 0}$ such that for any $t \geq 0$

$$\mathbf{B}_t = \mathbf{B}^{n_t}_{(t-n_t)\gamma_\alpha} + \sum_{k=0}^{n_t - 1} \mathbf{B}^k_{\gamma_\alpha} .$$

Since $((\mathbf{B}_t^k)_{t\in[0,\gamma_\alpha]})_{k\in\mathbb{N}}$ is a sequence of independent Brownian motion, we get that $(\mathbf{B}_t)_{t\geq 0}$ is a Brownian motion. In addition, there exists a measurable bijection mapping $(\mathbf{B}_t)_{t\geq 0}$ to $((\mathbf{B}_t^k)_{t\in[0,\gamma_\alpha]})_{k\in\mathbb{N}}$ and therefore for any $k \in \mathbb{N}$, $\mathcal{K}_k = \mathcal{H}_k$. Finally, we have that $(\mathbf{Y}_t)_{t\geq 0}$ solution to (12) is a solution to (2) with initial condition $\mathbf{Y}_0 = \mathbf{X}_0$, *i.e.*, $(\mathbf{Y}_t)_{t\geq 0} = (\mathbf{X}_t)_{t\geq 0}$ which concludes the proof.

■

B.3. Moment bounds and one-step approximation

The following result is well-known in the field of SDE but its proof is given for completeness. For any $t \geq 0$ and $k \in \mathbb{N}$, denote $\mathcal{F}_t = \sigma(\{\mathbf{X}_s \ : \ s \in [0, t]\})$ and $\mathcal{G}_k = \sigma(\{Z_j \ : \ j \in \{0, \dots, k\}\})$. We derive classical moment bounds in Lemma 17. This bounds are then used in Lemma 18 in order to provide one-step approximations. The proof of these lemmas is based on the repeated application of the Grönwall's lemma (both discrete and continuous) and Itô's formula.

Lemma 17 *Let $p \in \mathbb{N}$, $\bar{\gamma} > 0$ and $\alpha \in [0, 1)$. Assume* **A**1*,* **A**2 *and* **A**3*. Then for any $T \geq 0$, there exists $\mathtt{A}_{T,1} \geq 0$, such that for any $s \geq 0$ and $t \in [s, s+T]$, $\gamma \in (0, \bar{\gamma}]$, we have*

$$\mathbb{E}[1 + \|\mathbf{X}_t\|^{2p} | \mathcal{F}_s] \leq \mathtt{A}_{T,1}(1 + \|\mathbf{X}_s\|^{2p}) ,$$

where $(\mathbf{X}_t)_{t\geq 0}$ is the solution of (2). In addition, if there exists $x^\star$ such that $\int_{\mathsf{Z}} \|H(x^\star, z)\|^{2p} \, \mathrm{d}\pi^Z(z) < +\infty$, then for any $T \geq 0$, there exists $\tilde{\mathtt{A}}_{T,1} \geq 0$, such that for any $k_0 \geq 0$, $\gamma \in (0, \bar{\gamma}]$ and $k \in \{k_0, \dots, k_0 + N\}$ with $N\gamma_\alpha \leq T$, we have

$$\mathbb{E}\left[1 + \|X_k\|^{2p} \middle| \mathcal{G}_{k_0}\right] \leq \tilde{\mathtt{A}}_{T,1}(1 + \|X_{k_0}\|^{2p}) ,$$

where $(X_k)_{k\in\mathbb{N}}$ satisfies the recursion (1).

Proof We prove the result under **A**2-(b). The proof under **A**2-(a) is similar and left to the reader. Let $p \in \mathbb{N}$, $\alpha \in [0,1)$, $s, T \in [0,+\infty)$, $t \in [s, s+T]$, and $g_p \in \mathrm{C}^2(\mathbb{R}^d, [0,+\infty))$ such that for any $x \in \mathbb{R}^d$, $g_p(x) = 1 + \|x\|^{2p}$. Let $\bar{\gamma} > 0$ and $\gamma \in (0, \bar{\gamma}]$.

We divide the proof into two parts.

(a) Let $(\mathbf{X}_t)_{t\geq 0}$ be a solution to (2). We have for any $x \in \mathbb{R}^d$

$$\nabla g_p(x) = 2p \left\|x\right\|^{2(p-1)} x \ , \qquad \nabla^2 g_p(x) = 4p(p-1) \left\|x\right\|^{2(p-2)} xx^\top + 2p \left\|x\right\|^{2(p-1)} \mathrm{Id} \ . \quad (13)$$

Let $n \in \mathbb{N}$, and set $\tau_n = \inf\{u \geq 0 \ : \ g_p(\mathbf{X}_u) > n\}$. Applying Itô's lemma and using (2) and (13) we get

$$\begin{aligned}
&\mathbb{E}\left[g_p(\mathbf{X}_{t\wedge\tau_n})|\mathcal{F}_s\right] - \mathbb{E}\left[g_p(\mathbf{X}_{s\wedge\tau_n})|\mathcal{F}_s\right] \\
&\qquad = \mathbb{E}\left[\int_{s\wedge\tau_n}^{t\wedge\tau_n} -(\gamma_\alpha + u)^{-\alpha}\langle \nabla f(\mathbf{X}_u), \nabla g_p(\mathbf{X}_u)\rangle \middle| \mathcal{F}_s\right] \mathrm{d}u \\
&\qquad\quad + (\gamma_\alpha/2)\mathbb{E}\left[\int_{s\wedge\tau_n}^{t\wedge\tau_n} (\gamma_\alpha + u)^{-2\alpha}\langle \Sigma(\mathbf{X}_u), \nabla^2 g_p(\mathbf{X}_u)\rangle \mathrm{d}u \middle| \mathcal{F}_s\right] \ . \quad (14)
\end{aligned}$$

Using **A**1, (13) and the Cauchy-Schwarz inequality we get that for any $u \in [s, s+T]$

$$\begin{aligned}
|\langle \nabla f(\mathbf{X}_u), \nabla g_p(\mathbf{X}_u)\rangle| &\leq 2p\|\mathbf{X}_u\|^{2(p-1)} \{|\langle \nabla f(\mathbf{X}_u) - \nabla f(0), \mathbf{X}_u\rangle| + \|\nabla f(0)\|\|\mathbf{X}_u\|\} \\
&\leq 2p(\mathtt{L} + \|\nabla f(0)\|) g_p(\mathbf{X}_u) \ . \quad (15)
\end{aligned}$$

In addition, using **A**1, Lemma 15, (13) and the Cauchy-Schwarz inequality we get that for any $u \in [s, s+T]$

$$\begin{aligned}
\left|\langle \Sigma(\mathbf{X}_u), \nabla^2 g_p(\mathbf{X}_u)\rangle\right| &\leq C(1 + \|\mathbf{X}_u\|^2)\left\|\nabla^2 g_p(\mathbf{X}_u)\right\| \quad (16)\\
&\leq C(1 + \|\mathbf{X}_u\|^2)(8p(p-1)d + 2pd)\left\|\mathbf{X}_u\right\|^{2(p-1)} \\
&\leq 4Cdp(4(p-1)+1) g_p(\mathbf{X}_u) \ .
\end{aligned}$$

Combining (15) and (16) in (14) we get for large enough $n \in \mathbb{N}$

$$\begin{aligned}
&\mathbb{E}\left[g_p(\mathbf{X}_{t\wedge\tau_n})|\mathcal{F}_s\right] - g_p(\mathbf{X}_s) \\
&\quad \leq 2p(\mathtt{L} + \|\nabla f(0)\|)\mathbb{E}\left[\int_s^{t\wedge\tau_n} g_p(\mathbf{X}_u)\mathrm{d}u \middle| \mathcal{F}_s\right] + \bar{\gamma}_\alpha p(2p-1)\mathbb{E}\left[\int_s^{t\wedge\tau_n} g_p(\mathbf{X}_u)\mathrm{d}u \middle| \mathcal{F}_s\right] \\
&\quad \leq \{2p(\mathtt{L} + \|\nabla f(0)\|) + 4\gamma_\alpha Cdp(4(p-1)+1)\} \int_s^t \mathbb{E}\left[g_p(\mathbf{X}_{\wedge\tau_n})|\mathcal{F}_s\right] \mathrm{d}u \ .
\end{aligned}$$

Using Grönwall's lemma we obtain

$$\mathbb{E}\left[g_p(\mathbf{X}_{t\wedge\tau_n})|\mathcal{F}_s\right] \leq g_p(\mathbf{X}_s) \exp\left[T\left\{2p(\mathtt{L} + \|\nabla f(0)\|) + 4\gamma_\alpha Cdp(4(p-1)+1)\right\}\right] \ .$$

We conclude upon using Fatou's lemma and remarking that $\lim_n \tau_n = +\infty$, since $\mathbf{X}_t$ is well-defined for any $t \geq 0$.

(b) Let $(X_k)_{k\in\mathbb{N}}$ be a sequence which satisfies the recursion (1). Let $A_k = X_k\gamma(k+1)^{-\alpha}$ and $B_k = -\gamma(k+1)^{-\alpha}\{\nabla f(X_k) - H(X_k, Z_{k+1})\}$. We have, using Cauchy-Schwarz inequality and the binomial formula,

$$\begin{aligned}\|X_{k+1}\|^{2p} &= \|A_k + B_k\|^{2p} = \left\{\|A_k\|^2 + 2\langle A_k, B_k\rangle + \|B_k\|^2\right\}^p \\ &\leq \sum_{i=0}^{p}\sum_{j=0}^{i}\binom{p}{i}\binom{i}{j}\|A_k\|^{2(p-i)+j}\|B_k\|^{2i-j}\times 2^j \\ &\leq \|A_k\|^{2p} + 2^p\sum_{i=1}^{p}\sum_{j=0}^{i}\binom{p}{i}\binom{i}{j}\|A_k\|^{2(p-i)+j}\|B_k\|^{2i-j}\ .\end{aligned} \tag{17}$$

Using **A**1, there exists $\tilde{\mathtt{A}}_{T,1}^{(a)}, \tilde{\mathtt{A}}_{T,1}^{(b)}, \tilde{\mathtt{A}}_{T,1}^{(c)} \geq 0$ such that for any $\ell \in \{0, \dots, 2p\}$

$$\begin{aligned}\|A_k\|^{\ell} &\leq \sum_{m=0}^{\ell}\binom{\ell}{m}(1+\gamma(k+1)^{-\alpha}\mathtt{L})^m\|X_k\|^m\left(\gamma(k+1)^{-\alpha}\|\nabla f(0)\|\right)^{\ell-m} \\ &\leq (1+\gamma(k+1)^{-\alpha}\tilde{\mathtt{A}}_{T,1}^{(a)})\|X_k\|^{\ell} + \gamma(k+1)^{-\alpha}\tilde{\mathtt{A}}_{T,1}^{(b)}(1+\|X_k\|^{2p}) \\ &\leq (1+\gamma(k+1)^{-\alpha}\tilde{\mathtt{A}}_{T,1}^{(c)})(1+\|X_k\|^{2p})\ .\end{aligned}$$

In addition, we have that for any $\ell \in \{1, \dots, p\}$, $x \in \mathbb{R}^d$ and $z \in \mathsf{Z}$,

$$\|H(x,z)\|^{\ell} \leq (\|H(x,z)\| + \mathtt{L}\|x\|)^{2\ell} \leq 2^{2\ell-1}\|H(x,z)\|^{2\ell} + 2^{2\ell-1}\mathtt{L}^{2\ell}\|x\|^{2\ell}\ .$$

Therefore, there exists $\eta_\ell > 0$ such that for any $\ell \in \mathbb{N}$, $\mathbb{E}\left[\|B_k\|^{2\ell}\middle|\mathcal{G}_{k_0}\right] \leq \gamma^{2\ell}(k+1)^{-2\alpha\ell}\eta_\ell(1+\|X\|^{2\ell})$. Combining this result, (17), Jensen's inequality and that f we have

$$\begin{aligned}\mathbb{E}\left[\|X_{k+1}\|^{2p}\middle|\mathcal{G}_k\right] &\leq \|X_k\|^{2p} + \gamma(k+1)^{-\alpha}(\tilde{\mathtt{A}}_{T,1}^{(a)} + \tilde{\mathtt{A}}_{T,1}^{(b)})(1+\|X_k\|^{2p}) \\ &+ 2^{p+1}(1+\gamma(k+1)^{-\alpha}\tilde{\mathtt{A}}_{T,1}^{(c)})(1+\|X_k\|^{2p})\sum_{i=1}^{p}\sum_{j=0}^{i}\binom{p}{i}\binom{i}{j}\eta_{2i-j}^{1/2}\gamma^{2i-j}(k+1)^{-\alpha(2i-j)}\ .\end{aligned}$$

Therefore, there exists $\tilde{\mathtt{A}}_{T,1}^{(d)} \geq 0$ such that

$$\mathbb{E}\left[1+\|X_{k+1}\|^{2p}\middle|\mathcal{G}_{k_0}\right] \leq (1+\tilde{\mathtt{A}}_{T,1}^{(d)}\gamma(k+1)^{-\alpha})\mathbb{E}\left[1+\|X_k\|^{2p}\middle|\mathcal{G}_{k_0}\right] + \tilde{\mathtt{A}}_{T,1}^{(d)}\gamma(k+1)^{-\alpha}\ .$$

We conclude combining this result, Lemma 13 and Lemma 14.

■

Lemma 18 *Let $p \in \mathbb{N}$, $\bar{\gamma} > 0$ and $\alpha \in [0,1)$. Assume* **A**1, **A**2-(b), **A**3 *and that there exists $x^\star$ such that $\int_{\mathsf{Z}}\|H(x^\star, z)\|^{2p}\,\mathrm{d}\pi^Z(z) < +\infty$. Then for any $T \geq 0$, there exists $\mathtt{A}_{T,2} \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $k \in \mathbb{N}$ with $(k+1)\gamma_\alpha \leq T$, $t \in [k\gamma_\alpha, (k+1)\gamma_\alpha]$, we have*

$$\begin{aligned}&\mathbb{E}\left[\|X_{k+1} - X_k\|^{2p}\middle|\mathcal{G}_k\right] \leq \mathtt{A}_{T,2}(k+1)^{-2\alpha p}\gamma^{2p}(1+\|X_k\|^{2p})\ , \\ &\mathbb{E}\left[\|\mathbf{X}_t - \mathbf{X}_{k\gamma_\alpha}\|^{2p}\middle|\mathcal{F}_{k\gamma_\alpha}\right] \leq \mathtt{A}_{T,2}(k+1)^{-2\alpha p}\gamma^{2p}(1+\|\mathbf{X}_{k\gamma_\alpha}\|^{2p}\ .\end{aligned}$$

where $(X_k)_{k\in\mathbb{N}}$ satisfies the recursion and $(\mathbf{X}_t)_{t\geq 0}$ is the solution of (2).

Proof Let $p \in \mathbb{N}$, $\alpha \in [0,1)$, $\bar{\gamma} > 0$, $\gamma \in (0, \bar{\gamma}]$, $k \in \mathbb{N}$, $t \in [k\gamma_\alpha, (k+1)\gamma_\alpha]$. We divide the rest of the proof into two parts.

(a) Let $(\mathbf{X}_s)_{s \geq 0}$ be a solution to (2). Using **A**1, **A**2, Jensen's inequality, Burkholder-Davis-Gundy's inequality (Rogers and Williams, 2000, Theorem 42.1) and Lemma 17 there exists $B_p \geq 0$ such that

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{X}_t - \mathbf{X}_{k\gamma_\alpha}\|^{2p} \middle| \mathcal{F}_{k\gamma_\alpha}\right] \\
&\leq 2^{2p-1}\mathbb{E}\left[\left\|\int_{k\gamma_\alpha}^t (\gamma_\alpha + s)^{-\alpha}\nabla f(\mathbf{X}_s)\mathrm{d}s\right\|^{2p} \middle| \mathcal{F}_{k\gamma_\alpha}\right] \\
&\quad + 2^{2p-1}\gamma_\alpha^p\mathbb{E}\left[\left\|\int_{k\gamma_\alpha}^t (\gamma_\alpha + s)^{-\alpha}\Sigma(\mathbf{X}_s)^{1/2}\mathrm{d}\mathbf{B}_s\right\|^{2p} \middle| \mathcal{F}_{k\gamma_\alpha}\right] \\
&\leq 2^{2p-1}\gamma_\alpha^{2p-1}\int_{k\gamma_\alpha}^t (\gamma_\alpha + s)^{-2\alpha p}\mathbb{E}\left[\|\nabla f(\mathbf{X}_s)\|^{2p} \middle| \mathcal{F}_{k\gamma_\alpha}\right]\mathrm{d}s \\
&\quad + B_p 2^{2p-1}\gamma_\alpha^p\left(\int_{k\gamma_\alpha}^t (\gamma_\alpha + s)^{-2\alpha}\mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_s))|\mathcal{F}_{k\gamma_\alpha}\right]\mathrm{d}s\right)\mathcal{F}_{k\gamma_\alpha}{}^p \\
&\leq 2^{2p-1}\gamma_\alpha^{2p-1-2\alpha p}(k+1)^{-2\alpha p}(B_p+1)\left\{\int_{k\gamma_\alpha}^t \mathbb{E}\left[\|\nabla f(\mathbf{X}_s)\|^{2p} \middle| \mathcal{F}_{k\gamma_\alpha}\right]\mathrm{d}s\right. \\
&\quad \left. + \int_{k\gamma_\alpha}^t \mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_s))|\mathcal{F}_{k\gamma_\alpha}\right]^p\mathrm{d}s\right\} \\
&\leq 2^{4p}(1+\mathtt{L}^{2p})\gamma^{2p}\gamma_\alpha^{-1}(k+1)^{-2\alpha p}(B_p+1)\int_{k\gamma_\alpha}^t C^{2p}\left(1 + \mathbb{E}\left[\|\mathbf{X}_s\|^{2p} \middle| \mathcal{F}_{k\gamma_\alpha}\right]\right)\mathrm{d}s \\
&\leq 2^{4p}(1+\mathtt{L}^{2p})\gamma^{2p}(k+1)^{-2\alpha p}(B_p+1)\left(1 + \sup_{s\in[k\gamma_\alpha,t]}\mathbb{E}\left[\|\mathbf{X}_s\|^{2p} \middle| \mathcal{F}_{k\gamma_\alpha}\right]\right) \\
&\leq 2^{4p}(1+\mathtt{L}^{2p})\gamma^{2p}(k+1)^{-2\alpha p}(B_p+1)\left(1+\mathtt{A}_{T,1}\right)g_p(\mathbf{X}_{k\gamma_\alpha}) \ .
\end{aligned}
$$

(b) Let $(X_n)_{n\in\mathbb{N}}$ which satisfies the recursion (1). Using **A**1 and **A**2-(b) we get that

$$
\begin{aligned}
&\mathbb{E}\left[\|X_{k+1} - X_k\|^{2p} \middle| \mathcal{G}_k\right] \\
&\quad = \mathbb{E}\left[\left\|-\gamma(k+1)^{-\alpha}(H(X_k, Z_{k+1}) - H(x^\star, Z_{k+1}) + H(x^\star, Z_{k+1}))\right\|^{2p} \middle| \mathcal{G}_k\right] \\
&\quad \leq 2^{2p}\gamma^{2p}(k+1)^{-2\alpha p}\left(\mathtt{L}^{2p}\|X_k - x^\star\|^{2p} + \mathbb{E}\left[\|H(x^\star, Z_{k+1})\|^{2p} \middle| \mathcal{G}_k\right]\right) \\
&\quad \leq 2^{4p}\gamma^{2p}(k+1)^{-2\alpha p}\left(\mathtt{L}^{2p}\|X_k\|^{2p} + \mathtt{L}^{2p}\|x^\star\|^{2p} + \mathbb{E}\left[\|H(x^\star, Z_{k+1})\|^{2p} \middle| \mathcal{G}_k\right]\right) \ ,
\end{aligned}
$$

which concludes the proof.

■

B.4. Mean-square approximation

In this section, we introduce an auxiliary process $(\overline{\mathbf{X}}_t)_{t\geq 0}$. This process is a continuous interpolation of the discrete-time process $(\bar{X}_k)_{k\in\mathbb{N}}$ such that for any $k \in \mathbb{N}$,

$$
\bar{X}_{k+1} = \bar{X}_k - \gamma(1+k)^{-\alpha}\left\{\nabla f(\bar{X}_k) + \gamma_\alpha^{1/2}\Sigma^{1/2}(\bar{X}_k)G_{k+1}\right\} \ ,
$$

where $(G_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. Gaussian random variables with zero mean. For any $x \in \mathbb{R}^d$ and $k \in \mathbb{N}$, $\nabla f(x) + \gamma_\alpha^{1/2}\Sigma^{1/2}(x)G_{k+1}$ is a Gaussian approximation of the true noise term $H(x, Z_{k+1})$. Using Theorem 16, G_{k+1} and Z_{k+1} will be coupled in order to minimize the distance between the two discrete-time processes.

We now introduce the continuous-time process $(\overline{\mathbf{X}}_t)_{t\geq 0}$. Consider the stochastic process $(\overline{\mathbf{X}}_t)_{t\geq 0}$ defined by $\overline{\mathbf{X}}_0 = X_0$ and solution of the following SDE

$$\mathrm{d}\overline{\mathbf{X}}_t = -\gamma_\alpha^{-1}\sum_{k=0}^{+\infty}\mathbb{1}_{[k\gamma_\alpha,(k+1)\gamma_\alpha)}(t)(1+k)^{-\alpha}\gamma\left\{\nabla f(\overline{\mathbf{X}}_{k\gamma_\alpha})\mathrm{d}t + \gamma_\alpha^{1/2}\Sigma(\overline{\mathbf{X}}_{k\gamma_\alpha})^{1/2}\mathrm{d}\mathbf{B}_t\right\} .$$

Note that for any $k \in \mathbb{N}$, we have

$$\overline{\mathbf{X}}_{(k+1)\gamma_\alpha} = \overline{\mathbf{X}}_{k\gamma_\alpha} - \gamma(k+1)^{-\alpha}\left\{\nabla f(\overline{\mathbf{X}}_{k\gamma_\alpha}) + \Sigma(\overline{\mathbf{X}}_{k\gamma_\alpha})^{1/2}G_k\right\} ,$$

with $G_k = \gamma_\alpha^{-1/2}\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha}\mathrm{d}\mathbf{B}_s$. Hence, for any $k \in \mathbb{N}$, $\overline{\mathbf{X}}_{k\gamma_\alpha}$ has the same distribution as X_k given by (1) with $H(x,z) = \nabla f(x) + \Sigma(x)^{1/2}z$, $(\mathsf{Z}, \mathcal{Z}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and π^Z the Gaussian probability distribution with zero mean and covariance matrix identity. In our proof, we will not consider this process but a similar version whose initial point is given by the continuous-time process, see for instance. However, we found that introducing $(\overline{\mathbf{X}}_t)_{t\geq 0}$ and its discrete-time counterpart provides intuition for our derivation.

In Lemma 19, we bound the one-step difference between the continuous-time auxiliary process $(\overline{\mathbf{X}}_t)_{t\geq 0}$ and the discrete-time process $(X_k)_{k\in\mathbb{N}}$. In Lemma 20, we bound the one-step difference between the continuous-time auxiliary process $(\overline{\mathbf{X}}_t)_{t\geq 0}$ and the continuous-time process $(\mathbf{X}_t)_{t\geq 0}$. We combine these estimates in Proposition 21. We conclude by proving Proposition 22 which is a restatement of Theorem 1.

Recall that $n_T = \lfloor T/\gamma_\alpha \rfloor$. In what follows, we denote

$$\varepsilon^2 = \sup_{k\in\{0,\dots,n_T\}} \mathbb{E}\left[\mathbf{W}_2^2(\nu_k^{\mathrm{d}}(\mathbf{X}_{k\gamma_\alpha}), \nu_k^{\mathrm{c}}(\mathbf{X}_{k\gamma_\alpha}))\right] , \tag{18}$$

where for any $\tilde{x} \in \mathbb{R}^d$, $\nu_k^{\mathrm{d}}(\tilde{x})$ is the distribution of $H(\tilde{x}, Z_{n+1})$, $\nu_k^{\mathrm{c}}(\tilde{x})$ is the distribution of $\nabla f(\tilde{x}) + \gamma_\alpha^{-1/2}\Sigma^{1/2}(\tilde{x})\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha}(\gamma_\alpha + s)^{-\alpha}\mathrm{d}\mathbf{B}_s$.

Lemma 19 *Assume* **A**1 *and* **A**3. *Let $\bar{\gamma} > 0$ and $\alpha \in [0,1)$. Then for any $T \geq 0$, there exists $\mathtt{A}_{T,3} \geq 0$ such that for any $\gamma \in (0,\bar{\gamma}]$, $k \in \mathbb{N}$ with $(k+1)\gamma_\alpha \leq T$ and $X_0 \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[\|\tilde{\mathbf{X}}_{(k+1)\gamma_\alpha}^k - \tilde{X}_{k+1}\|^2\right] \leq \mathtt{A}_{T,3}\gamma^2(k+1)^{-2\alpha}\varepsilon^2 ,$$

where for any $k \in \mathbb{N}$ and $t \in [k\gamma_\alpha, (k+1)\gamma_\alpha]$

$$\tilde{X}_{k+1} = \mathbf{X}_{k\gamma_\alpha} - \gamma(k+1)^{-\alpha}H(\mathbf{X}_{k\gamma_\alpha}, Z_{k+1}) ,$$
$$\tilde{\mathbf{X}}_t = \mathbf{X}_{k\gamma_\alpha} - \gamma_\alpha^{-1}\gamma(1+k)^{-\alpha}\left\{(t-k\gamma_\alpha)\nabla f(\mathbf{X}_{k\gamma_\alpha}) + \gamma_\alpha^{1/2}\Sigma^{1/2}(\mathbf{X}_{k\gamma_\alpha})\int_{k\gamma_\alpha}^t \mathrm{d}\mathbf{B}_s\right\} ,$$

We recall that for any $\tilde{x} \in \mathbb{R}^d$, $\nu_k^{\mathrm{d}}(\tilde{x})$ is the distribution of $H(\tilde{x}, Z_{n+1})$, $\nu_k^{\mathrm{c}}(\tilde{x})$ is the distribution of $\nabla f(\tilde{x}) + \gamma_\alpha^{-1/2}\Sigma^{1/2}(\tilde{x})\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha}(\gamma_\alpha + s)^{-\alpha}\mathrm{d}\mathbf{B}_s$.

Proof Let $\alpha \in [0,1)$, $\bar{\gamma} > 0$, $\gamma \in (0, \bar{\gamma}]$, $k \in \mathbb{N}$, $t \in [k\gamma_\alpha, (k+1)\gamma_\alpha]$ and $X_0 \in \mathbb{R}^d$. Using Theorem 16 we have

$$\begin{aligned}
&\mathbb{E}\left[\|\overline{\mathbf{X}}_{(k+1)\gamma_\alpha} - X_{k+1}\|^2\right] \\
&\quad = \gamma^2 (k+1)^{-2\alpha} \mathbb{E}\left[\left\|\nabla f(\mathbf{X}_{k\gamma_\alpha}) + \Sigma^{1/2}(X_0) G_k - H(\mathbf{X}_{k\gamma_\alpha}, Z_k)\right\|^2\right] \\
&\quad \leq 2\gamma^2 (k+1)^{-2\alpha} \mathbb{E}\left[\mathbf{W}_2^2(\nu_k^{\mathrm{d}}(\mathbf{X}_{k\gamma_\alpha}), \nu_k^{\mathrm{c}}(\mathbf{X}_{k\gamma_\alpha}))\right] \ .
\end{aligned}$$

which concludes the proof upon using (33). ■

Lemma 20 *Let $\bar{\gamma} > 0$ and $\alpha \in [0,1)$. Assume* **A**1 *and* **A**3*. Then for any $T \geq 0$, there exists $\mathtt{A}_{T,4} \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $k \in \mathbb{N}$ with $(k+1)\gamma_\alpha \leq T$ and $X_0 \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[\|\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{\mathbf{X}}_{(k+1)\gamma_\alpha}\|^2 \middle| \mathcal{F}_{k\gamma_\alpha}\right] \leq \mathtt{A}_{T,4}\left\{\gamma^4 (k+1)^{-4\alpha} + \gamma^2 (k+1)^{-2(1+\alpha)}\right\}(1 + \|\mathbf{X}_{k\gamma_\alpha}\|^2) \ ,$$

where $(\mathbf{X}_t)_{t \geq 0}$ be the solution of (2) and for any $t \in [k\gamma_\alpha, (k+1)\gamma_\alpha]$ we have

$$\tilde{\mathbf{X}}_t = \mathbf{X}_{k\gamma_\alpha} - \gamma_\alpha^{-1}\gamma(1+k)^{-\alpha}\left\{(t - k\gamma_\alpha)\nabla f(\mathbf{X}_{k\gamma_\alpha}) + \gamma_\alpha^{1/2}\Sigma^{1/2}(\mathbf{X}_{k\gamma_\alpha})\int_{k\gamma_\alpha}^{t} \mathrm{d}\mathbf{B}_s\right\} \ . \tag{19}$$

Proof Let $\alpha \in [0,1)$, $\bar{\gamma} > 0$, $\gamma \in (0, \bar{\gamma}]$, $k \in \mathbb{N}$ and $t \in [k\gamma_\alpha, (k+1)\gamma_\alpha]$. Let $(\mathbf{X}_t)_{t \geq 0}$ is the solution of (2) and $(\tilde{\mathbf{X}}_t)_{t \in [k\gamma_\alpha, (k+1)\gamma_\alpha]}$ given by (19). Using Jensen's inequality and that $\gamma_\alpha \gamma^{-1} = \gamma_\alpha^\alpha$ we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{\mathbf{X}}_{(k+1)\gamma_\alpha}\right\|^2 \middle| \mathcal{F}_{k\gamma_\alpha}\right] \qquad (20) \\
&\leq \mathbb{E}\left[\left\| - \int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha} (\gamma_\alpha + s)^{-\alpha}\nabla f(\mathbf{X}_s)\mathrm{d}s - \gamma_\alpha^{1/2}\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha} (\gamma_\alpha + s)^{-\alpha}\Sigma(\mathbf{X}_s)^{1/2}\mathrm{d}\mathbf{B}_s \right.\right. \\
&\qquad \left.\left. + \gamma(k+1)^{-\alpha}\nabla f(\mathbf{X}_{k\gamma_\alpha}) + \gamma\gamma_\alpha^{-1/2}(k+1)^{-\alpha}\Sigma(\mathbf{X}_{k\gamma_\alpha})^{1/2}\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha}\mathrm{d}\mathbf{B}_s\right\|^2 \middle| \mathcal{F}_{k\gamma_\alpha}\right] \\
&\leq 2\mathbb{E}\left[\left\| -\gamma_\alpha^{-\alpha}\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha} (1 + \gamma_\alpha^{-1}s)^{-\alpha}\nabla f(\mathbf{X}_s)\mathrm{d}s + \gamma(k+1)^{-\alpha}\nabla f(\mathbf{X}_{k\gamma_\alpha})\right\|^2 \middle| \mathcal{F}_{k\gamma_\alpha}\right] \\
&\quad + 2\mathbb{E}\left[\left\| -\gamma_\alpha^{1/2-\alpha}\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha} (1 + \gamma_\alpha^{-1}s)^{-\alpha}\Sigma(\mathbf{X}_s)^{1/2}\mathrm{d}\mathbf{B}_s \right.\right. \\
&\qquad \left.\left. + \gamma\gamma_\alpha^{-1/2}(k+1)^{-\alpha}\Sigma(\mathbf{X}_{k\gamma_\alpha})^{1/2}\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha}\mathrm{d}\mathbf{B}_s\right\|^2 \middle| \mathcal{F}_{k\gamma_\alpha}\right] \\
&\leq 2\gamma_\alpha^{-2\alpha}\mathbb{E}\left[\left\|\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha}\left\{(k+1)^{-\alpha}\nabla f(\mathbf{X}_{k\gamma_\alpha}) - (1 + \gamma_\alpha^{-1}s)^{-\alpha}\nabla f(\mathbf{X}_s)\right\}\mathrm{d}s\right\|^2 \middle| \mathcal{F}_{k\gamma_\alpha}\right] \\
&\quad + 2\gamma_\alpha^{1-2\alpha}\mathbb{E}\left[\left\|\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha}\left\{(k+1)^{-\alpha}\Sigma(\mathbf{X}_{k\gamma_\alpha})^{1/2} - (1 + \gamma_\alpha^{-1}s)^{-\alpha}\Sigma(\mathbf{X}_s)^{1/2}\right\}\mathrm{d}\mathbf{B}_s\right\|^2 \middle| \mathcal{F}_{k\gamma_\alpha}\right] \ .
\end{aligned}$$

We now treat each term separately.

Using Jensen's inequality, Itô isometry, Fubini-Tonelli's theorem, **A**1, **A**3 and Lemma 18 we have

$$\mathbb{E}\left[\left\|\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha}\left\{(k+1)^{-\alpha}\Sigma(\mathbf{X}_{k\gamma_\alpha})^{1/2}-(1+\gamma_\alpha^{-1}s)^{-\alpha}\Sigma(\mathbf{X}_s)^{1/2}\right\}\mathrm{d}\mathbf{B}_s\right\|^2\middle|\mathcal{F}_{k\gamma_\alpha}\right] \tag{21}$$

$$\begin{aligned}
&\leq 2\left[(k+1)^{-2\alpha}\left|\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha}\mathbb{E}\left[\|\Sigma(\mathbf{X}_{k\gamma_\alpha})^{1/2}-\Sigma(\mathbf{X}_s)^{1/2}\|^2\middle|\mathcal{F}_{k\gamma_\alpha}\right]\mathrm{d}s\right|\right.\\
&\left.+\eta\left|\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha}\{(k+1)^{-\alpha}-(1+\gamma_\alpha^{-1}s)\}^2\mathrm{d}s\right|\right]\\
&\leq 2\gamma_\alpha\left[(k+1)^{-2\alpha}\mathtt{M}^2\sup_{s\in[k\gamma_\alpha,(k+1)\gamma_\alpha]}\mathbb{E}\left[\|\mathbf{X}_s-\mathbf{X}_{k\gamma_\alpha}\|^2\middle|\mathcal{F}_{k\gamma_\alpha}\right]+\eta\alpha^2(k+1)^{-2(1+\alpha)}\right]\\
&\leq 2\gamma_\alpha\left[(k+1)^{-4\alpha}\mathtt{M}^2\mathtt{A}_{T,2}\gamma^2+\eta\alpha^2(k+1)^{-2(1+\alpha)}\right](1+\|\mathbf{X}_{k\gamma_\alpha}\|^2)\ .
\end{aligned}$$

Using Jensen's inequality, Fubini-Tonelli's theorem, the fact that for any $u>0$, $u^{-\alpha}-(u+1)^{-\alpha}\leq\alpha u^{-(\alpha+1)}$, **A**1 and Lemma 18 we get that

$$\mathbb{E}\left[\left\|\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha}\left\{(k+1)^{-\alpha}\nabla f(\mathbf{X}_{k\gamma_\alpha})-(1+\gamma_\alpha^{-1}s)^{-\alpha}\nabla f(\mathbf{X}_s)\right\}\mathrm{d}s\right\|^2\middle|\mathcal{F}_{k\gamma_\alpha}\right] \tag{22}$$

$$\begin{aligned}
&\leq\gamma_\alpha^2\sup_{s\in[k\gamma_\alpha,(k+1)\gamma_\alpha]}\left\{\mathbb{E}\left[\left\|(k+1)^{-\alpha}\nabla f(\mathbf{X}_{k\gamma_\alpha})-(1+\gamma_\alpha^{-1}s)^{-\alpha}\nabla f(\mathbf{X}_s)\right\|^2\right]\mathcal{F}_{k\gamma_\alpha}\mathrm{d}s\right\}\\
&\leq 2\gamma_\alpha^2\sup_{s\in[k\gamma_\alpha,(k+1)\gamma_\alpha]}\left\{\|\nabla f(\mathbf{X}_{k\gamma_\alpha})\|^2|(k+1)^{-\alpha}-(1+\gamma_\alpha^{-1}s)^{-\alpha}|^2\right.\\
&\qquad\left.+(1+\gamma_\alpha s^{-1})^{-2\alpha}\mathbb{E}\left[\|\nabla f(\mathbf{X}_s)-\nabla f(\mathbf{X}_{k\gamma_\alpha})\|^2\right]\mathcal{F}_{k\gamma_\alpha}\right\}\\
&\leq 2\gamma_\alpha^2\left(\alpha^2\|\nabla f(\mathbf{X}_{k\gamma_\alpha})\|^2(k+1)^{-2(1+\alpha)}\right.\\
&\qquad\left.+(k+1)^{-2\alpha}\mathtt{L}^2\sup_{s\in[k\gamma_\alpha,(k+1)\gamma_\alpha]}\mathbb{E}\left[\|\mathbf{X}_s-\mathbf{X}_{k\gamma_\alpha}\|^2\middle|\mathcal{F}_{k\gamma_\alpha}\right]\right)\\
&\leq 2\gamma_\alpha^2\left[\alpha^2\|\nabla f(\mathbf{X}_{k\gamma_\alpha})\|^2(k+1)^{-2(1+\alpha)}+(k+1)^{-4\alpha}\mathtt{L}^2\mathtt{A}_{T,2}\gamma^2(1+\|\mathbf{X}_{k\gamma_\alpha}\|^2)\right]\\
&\leq 2\gamma_\alpha^2\left[\alpha^2(\|\nabla f(0)\|^2+\mathtt{L}^2)(k+1)^{-2(1+\alpha)}+(k+1)^{-4\alpha}\mathtt{L}^2\mathtt{A}_{T,2}\gamma^2\right](1+\|\mathbf{X}_{k\gamma_\alpha}\|^2)\ .
\end{aligned}$$

Combining (20), (22) and (21) concludes the proof upon setting

$$\mathtt{A}_{T,4}=4\left[\mathtt{M}^2\mathtt{A}_{T,2}+\eta\alpha^2+\alpha^2(\|\nabla f(0)\|^2+\mathtt{L}^2)+\mathtt{L}^2\mathtt{A}_{T,2}\right]\ .$$

■

Proposition 21 *Let $\bar{\gamma} > 0$ and $\alpha \in [0, 1)$. Assume* **A**1, **A**2-(b) *and* **A**3. *Then for any $T \geq 0$, there exists $\mathtt{A}_{T,5} \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $k \in \mathbb{N}$ with $(k+1)\gamma_\alpha \leq T$ and $X_0 \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[\|\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{X}_{k+1}\|^2 \middle| \mathcal{F}_{k\gamma_\alpha}\right] \leq \mathtt{A}_{T,5}\left\{\gamma^4(k+1)^{-4\alpha} + \gamma^2(k+1)^{-2\alpha}\varepsilon^2\right\}(1 + \|\mathbf{X}_{k\gamma_\alpha}\|^2) ,$$

where $(\mathbf{X}_t)_{t\geq 0}$ is the solution of (2) *and for any $k \in \mathbb{N}$*

$$\tilde{X}_{k+1} = \mathbf{X}_{k\gamma_\alpha} - \gamma(k+1)^{-\alpha} H(\mathbf{X}_{k\gamma_\alpha}, Z_{k+1}) .$$

Proof The proof is straightforward upon combining Lemma 19 and Lemma 20. ■

Proposition 22 *Let $\bar{\gamma} > 0$, $\alpha \in [0, 1)$ and $\gamma \in (0, \bar{\gamma}]$. Assume* **A**1, **A**2-(b) *and* **A**3. *Then there exists a coupling $((\mathbf{B}_t)_{t\geq 0}, (Z_n)_{n\in\mathbb{N}})$. such that the following hold:*

(a) $(Z_n)_{n\in\mathbb{N}}$ is a sequence of independent random variables such that for any $n \in \mathbb{N}$, Z_n is distributed according to π^Z.

(b) For any $T \geq 0$, there exists $C \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $n \in \mathbb{N}$ with $\gamma_\alpha = \gamma^{1/(1-\alpha)}$, $n\gamma_\alpha \leq T$ we have

$$\mathbb{E}^{1/2}\left[\|\mathbf{X}_{n\gamma_\alpha} - X_n\|^2\right] \leq C(\gamma^\delta\varepsilon + \gamma)(1 + \log(\gamma^{-1})) , \quad \textit{with } \delta = \min(1, (2-2\alpha)^{-1}),$$

where $(\mathbf{X}_t)_{t\geq 0}$ is solution of (2), *$(X_n)_{n\in\mathbb{N}}$ is defined by* (1) *with $\mathbf{X}_0 = X_0 \in \mathbb{R}^d$ and*

$$\varepsilon^2 = \sup_{k\in\{0,\dots,n_T\}} \mathbb{E}\left[\mathbf{W}_2^2(\nu_k^{\mathrm{d}}(\mathbf{X}_{k\gamma_\alpha}), \nu_k^{\mathrm{c}}(\mathbf{X}_{k\gamma_\alpha}))\right] , \tag{23}$$

where for any $\tilde{x} \in \mathbb{R}^d$, $\nu_k^{\mathrm{d}}(\tilde{x})$ is the distribution of $H(\tilde{x}, Z_{n+1})$, $\nu_k^{\mathrm{c}}(\tilde{x})$ is the distribution of $\nabla f(\tilde{x}) + \gamma_\alpha^{-1/2}\Sigma^{1/2}(\tilde{x})\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha}(\gamma_\alpha + s)^{-\alpha}\mathrm{d}\mathbf{B}_s$.

Proof Let $\alpha \in [0, 1)$, $\bar{\gamma} > 0$, $\gamma \in (0, \bar{\gamma}]$, $k \in \mathbb{N}$, and $X_0 \in \mathbb{R}^d$. The first part of the proof is a direct consequence of Theorem 16. We now turn to the second part of the proof. Let $(E_k)_{k\in\mathbb{N}}$ such that for any $k \in \mathbb{N}$, $E_k = \mathbb{E}[\|\mathbf{X}_{k\gamma_\alpha} - X_k\|^2]$. Note that $E_0 = 0$. Let $\tilde{X}_{k+1} = \mathbf{X}_{k\gamma_\alpha} - \gamma(k+1)^{-\alpha}H(\mathbf{X}_{k\gamma_\alpha}, Z_{k+1})$. We have

$$\begin{aligned}
E_{k+1} &= \mathbb{E}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - X_{k+1}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{X}_{k+1} + \tilde{X}_{k+1} - X_{k+1}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{X}_{k+1}\right\|^2\right] + 2\mathbb{E}\left[\langle \mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{X}_{k+1}, \tilde{X}_{k+1} - X_{k+1}\rangle\right] \\
&\quad + \mathbb{E}\left[\left\|\tilde{X}_{k+1} - X_{k+1}\right\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{X}_{k+1}\right\|^2\right] + \mathbb{E}\left[\left\|\tilde{X}_{k+1} - X_{k+1}\right\|^2\right] \\
&\quad + 2\mathbb{E}\left[\langle \mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{X}_{k+1}, \mathbf{X}_{k\gamma_\alpha} - X_k\right] \\
&\quad + 2\gamma(k+1)^{-\alpha}\mathbb{E}\left[\langle \mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{X}_{k+1}, H(X_k, Z_{k+1}) - H(\mathbf{X}_{k\gamma_\alpha}, Z_{k+1})\rangle\right] .
\end{aligned} \tag{24}$$

Let $a_k = \gamma^4(k+1)^{-4\alpha} + \gamma^2(k+1)^{-2\alpha}$ and $a_k^{\varepsilon} = \gamma^4(k+1)^{-4\alpha} + \varepsilon^2\gamma^2(k+1)^{-2\alpha}$. We now bound each of the four terms appearing in (24)

(a) First, using Proposition 21 and Lemma 17 we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{X}_{k+1}\right\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{X}_{k+1}\right\|^2 \middle| \mathcal{F}_{k\gamma_\alpha}\right]\right] \\
&\leq \mathbb{E}\left[\mathtt{A}_{T,5}(\gamma^4(k+1)^{-4\alpha} + \varepsilon^2\gamma^2(k+1)^{-2\alpha})\left(1 + \|\mathbf{X}_{k\gamma_\alpha}\|^2\right)\right] \\
&\leq \mathtt{A}_{T,1}\mathtt{A}_{T,5}(\gamma^4(k+1)^{-4\alpha} + \varepsilon^2\gamma^2(k+1)^{-2\alpha})\left(1 + \|X_0\|^2\right) \leq \mathtt{A}_{T,6}^{(a)} a_k^{\varepsilon} \ ,
\end{aligned} \tag{25}$$

with $\mathtt{A}_{T,6}^{(a)} \geq 0$ which does not depend on γ and k.

(b) Second, using **A**1, **A**2-(a) and that for any $a, b \geq 0$, $(a+b)^2 \leq 2a^2 + 2b^2$ we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\tilde{X}_{k+1} - X_{k+1}\right\|^2\right] &= \mathbb{E}\left[\left\|\mathbf{X}_{k\gamma_\alpha} - X_k - \gamma(k+1)^{-\alpha}(H(\mathbf{X}_{k\gamma_\alpha}, Z_{k+1}) - H(X_k, Z_{k+1}))\right\|^2\right] \\
&\leq (1 + \gamma\mathtt{L}(k+1)^{-\alpha})^2\mathbb{E}[\|\mathbf{X}_{k\gamma_\alpha} - X_k\|^2] \\
&\leq (1 + 2\gamma\mathtt{L}(k+1)^{-\alpha} + \gamma^2\mathtt{L}^2(k+1)^{-2\alpha})E_k \leq (1 + \mathtt{A}_{T,6}^{(b)} a_k^{1/2})E_k \ ,
\end{aligned} \tag{26}$$

with $\mathtt{A}_{T,6}^{(b)} \geq 0$ which does not depend on γ and k.

(c) In what follows, let $\tilde{\mathbf{X}}_{(k+1)\gamma_\alpha} = \mathbf{X}_{k\gamma_\alpha} - \gamma(k+1)^{-\alpha}\left\{\nabla f(\mathbf{X}_{k\gamma_\alpha}) + \Sigma(\mathbf{X}_{k\gamma_\alpha})^{1/2}G_k\right\}$, with $G_k = \gamma_\alpha^{-1/2}\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha} \mathrm{d}\mathbf{B}_s$. Let $b_k = \gamma^3(k+1)^{-3\alpha} + \gamma(k+1)^{-2(1+\alpha/2)}$.

Using **A**2 we have $\mathbb{E}[\tilde{\mathbf{X}}_{(k+1)\gamma_\alpha}|\mathcal{K}_k] = \mathbb{E}[\tilde{X}_{k+1}|\mathcal{K}_k]$. Combining this result, the Cauchy-Schwarz inequality, Lemma 20, Lemma 17 and that for any $a, b \geq 0$, $(a+b)^{1/2} \leq a^{1/2} + b^{1/2}$ and $2ab \leq a^2 + b^2$ we obtain

$$\begin{aligned}
&\mathbb{E}\left[\langle\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{X}_{k+1}, \mathbf{X}_{k\gamma_\alpha} - X_k\rangle\right] \\
&\quad = \mathbb{E}\left[\langle\mathbb{E}[\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{X}_{k+1}|\mathcal{K}_k], \mathbf{X}_{k\gamma_\alpha} - X_k\rangle\right] \\
&\quad = \mathbb{E}\left[\langle\mathbb{E}[\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{\mathbf{X}}_{(k+1)\gamma_\alpha}|\mathcal{K}_k], \mathbf{X}_{k\gamma_\alpha} - X_k\rangle\right] \\
&\quad \leq \mathbb{E}\left[\mathbb{E}^{1/2}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{\mathbf{X}}_{(k+1)\gamma_\alpha}\right\|^2 \middle| \mathcal{K}_k\right]\|\mathbf{X}_{k\gamma_\alpha} - X_k\|\right] \\
&\quad \leq \mathbb{E}^{1/2}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{\mathbf{X}}_{(k+1)\gamma_\alpha}\right\|^2\right]\mathbb{E}^{1/2}\left[\|\mathbf{X}_{k\gamma_\alpha} - X_k\|^2\right] \\
&\quad \leq \mathtt{A}_{T,1}^{1/2}\mathtt{A}_{T,4}^{1/2}\left\{\gamma^4(k+1)^{-4\alpha} + \gamma^2(k+1)^{-2(1+\alpha)}\right\}^{1/2}(1 + \|X_0\|^2)E_k^{1/2} \\
&\quad \leq \mathtt{A}_{T,1}^{1/2}\mathtt{A}_{T,4}^{1/2}\left\{\gamma^{3/2}(k+1)^{-3\alpha/2} + \gamma^{1/2}(k+1)^{-(1+\alpha/2)}\right\}(1 + \|X_0\|^2)\gamma^{1/2}(k+1)^{-\alpha/2}E_k^{1/2} \\
&\quad \leq \mathtt{A}_{T,6}^{(c)}\left\{\gamma^3(k+1)^{-3\alpha} + \gamma(k+1)^{-2(1+\alpha/2)}\right\}/2 + a_k^{1/2}E_k/2 \leq \mathtt{A}_{T,6}^{(c)}b_k + a_k^{1/2}E_k \ .
\end{aligned} \tag{27}$$

with $\mathtt{A}_{T,6}^{(c)} \geq 0$ which does not depend on γ and k.

(d) Finally, using the Cauchy-Schwarz inequality, (25), **A**2 and **A**1 and that for any $a, b \geq 0$, $(a+b)^{1/2} \leq a^{1/2} + b^{1/2}$, we have

$$\begin{aligned}
&\gamma(k+1)^{-\alpha}\mathbb{E}\left[\langle \mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{X}_{k+1}, H(X_k, Z_{k+1}) - H(\mathbf{X}_{k\gamma_\alpha}, Z_{k+1})\rangle\right] \qquad (28)\\
&\leq \gamma(k+1)^{-\alpha}\mathbb{E}^{1/2}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - \tilde{X}_{k+1}\right\|^2\right]\mathbb{E}^{1/2}\left[\|H(X_k, Z_{k+1}) - H(\mathbf{X}_{k\gamma_\alpha}, Z_{k+1})\|^2\right]\\
&\leq (\mathtt{A}^{(a)}_{T,6})^{1/2}\gamma(k+1)^{-\alpha}a_k^{1/2}\mathtt{L}E_k \leq \mathtt{A}^{(d)}_{T,6}a_kE_k\ .
\end{aligned}$$

with $\mathtt{A}^{(d)}_{T,6} \geq 0$ which does not depend on γ and k.

Let $\bar{\mathtt{A}}_{T,6} = \mathtt{A}^{(a)}_{T,6} + \mathtt{A}^{(b)}_{T,6} + \mathtt{A}^{(c)}_{T,6} + \mathtt{A}^{(d)}_{T,6}$. Finally, we have using (25), (26), (27) and (28) in (24)

$$E_{k+1} \leq \bar{\mathtt{A}}_{T,6}(a_k + a_k^{1/2})E_k + \bar{\mathtt{A}}_{T,6}(a_k^\varepsilon + b_k)\ . \qquad (29)$$

We denote $v_k = \bar{\mathtt{A}}_{T,6}(a_k^{1/2} + a_k)$ and $w_k = \bar{\mathtt{A}}_{T,6}(a_k^\varepsilon + b_k)$. Using Lemma 14 and that $a_k^{1/2} \leq \gamma(k+1)^{-\alpha} + \gamma^2(k+1)^{-2\alpha}$, there exists $\mathtt{A}^{(e)}_{T,6} \geq 0$ which does not depend on γ and k such that

$$\sum_{k=0}^{N-1} v_k \leq \mathtt{A}^{(e)}_{T,6}\ . \qquad (30)$$

In addition, we have that for any $k \in \mathbb{N}$,

$$v_k \leq \bar{\mathtt{A}}_{T,6}(\gamma^2(k+1)^{-2\alpha}\varepsilon^2 + \gamma^3(k+1)^{-3\alpha} + \gamma^4(k+1)^{-4\alpha} + \gamma(k+1)^{-2(1+\alpha/2)})\ .$$

Using that $\gamma\gamma_\alpha^\alpha = \gamma_\alpha$ and Lemma 14 there exists $\mathtt{A}^{(f)}_{T,6} \geq 0$ which does not depend on γ and k such that

$$\sum_{k=0}^{N-1} v_k \leq \begin{cases} \mathtt{A}^{(f)}_{T,6}\gamma^2(1 + \log(\gamma^{-1})) & \text{if } \alpha \geq 1/2\ , \\ \mathtt{A}^{(f)}_{T,6}\gamma_\alpha\varepsilon^2 & \text{if } \alpha < 1/2\ . \end{cases} \qquad (31)$$

Using (29) and Lemma 13 we obtain that

$$\begin{aligned}
E_k &\leq \sum_{k=0}^{N-1} w_k + \exp\left[\sum_{k=0}^{N-1} v_k\right]\sum_{k=0}^{N-1} v_kw_k\\
&\leq \sum_{k=0}^{N-1} w_k + \exp\left[\sum_{k=0}^{N-1} v_k\right]\left(\sum_{k=0}^{N-1} v_k\right)\left(\sum_{k=0}^{N-1} w_k\right)\ . \qquad (32)
\end{aligned}$$

Combining (30), (31) and (32) concludes the first part of the proof. ■

B.5. The case of batch noise

In this section, we refine our results in the specific case of a batch noise. We recall our main result in this setting in Corollary 23. The proof is based on quantitative bounds in the CLT w.r.t. to $\mathbf{W}_2$, see Bonis (2020). In Proposition 24, we show that contrary to the SDE setting the gradient flow has an error of order at least $\mathcal{O}(M^{-1})$.

Corollary 23 *Let $\bar{\gamma} > 0$ and $\alpha \in [0,1)$. Assume* **A**1, **A**2-(b) *and* **A**3 *(with respect to $(\mathsf{Y}, \mathcal{Y}, \pi)$). Let H be given by (6). Assume that there exists $x^\star \in \mathbb{R}^d$, $C, p \geq 0$ such that for any $x \in \mathbb{R}^d$ and $y \in \mathsf{Y}$*

$$\int_{\mathsf{Y}} \|\nabla \tilde{f}(x^\star, y)\|^4 \mathrm{d}\pi(y) < +\infty \ , \quad \|\Sigma_f(x)^{-1/2}\| \leq C(1 + \|x\|^p) \ .$$

Then, there exists a random variable $((\mathbf{B}_t)_{t\geq 0}, (Z_n)_{n\in\mathbb{N}})$ such that for any $T \geq 0$, there exists $C \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $n \in \mathbb{N}$ with $n\gamma_\alpha \leq T$ $\gamma_\alpha = \gamma^{1/(1-\alpha)}$ we have

$$\mathbb{E}^{1/2}\left[\|\mathbf{X}_{n\gamma_\alpha} - X_n\|^2\right] \leq C(\gamma^\delta M^{-1} + \gamma)(1 + \log(\gamma^{-1})) \ , \quad \textit{with } \delta = \min(1, (2-2\alpha)^{-1}) \ .$$

Proof Let $\bar{\gamma} > 0$, $\alpha \in [0,1)$ and $M \in \mathbb{N}$. Applying Theorem 1, there exists a random variable $((\mathbf{B}_t)_{t\geq 0}, (Z_n)_{n\in\mathbb{N}})$ such that or any $T \geq 0$, there exists $C \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $n \in \mathbb{N}$ with $n\gamma_\alpha \leq T$ $\gamma_\alpha = \gamma^{1/(1-\alpha)}$ we have

$$\mathbb{E}^{1/2}\left[\|\mathbf{X}_{n\gamma_\alpha} - X_n\|^2\right] \leq C(\gamma^\delta \varepsilon + \gamma)(1 + \log(\gamma^{-1})) \ , \quad \text{with } \delta = \min(1, (2-2\alpha)^{-1}) \ ,$$

where $(\mathbf{X}_t)_{t\geq 0}$ is solution of (2), $(X_n)_{n\in\mathbb{N}}$ is defined by (1) with $\mathbf{X}_0 = X_0 \in \mathbb{R}^d$ and

$$\varepsilon^2 = \sup_{k\in\{0,\ldots,n_T\}} \mathbb{E}\left[\mathbf{W}_2^2(\nu^{\mathrm{d}}(\mathbf{X}_{k\gamma_\alpha}), \nu^{\mathrm{c}}(\mathbf{X}_{k\gamma_\alpha}))\right] \ , \tag{33}$$

where for any $\tilde{x} \in \mathbb{R}^d$, $\nu^{\mathrm{d}}(\tilde{x})$ is the distribution of $H(\tilde{x}, Z_{n+1})$ and $\nu^{\mathrm{c}}(\tilde{x})$ is the distribution of $\nabla f(\tilde{x}) + \gamma_\alpha^{-1/2}\Sigma^{1/2}(\tilde{x})\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha}(\gamma_\alpha + s)^{-\alpha}\mathrm{d}\mathbf{B}_s$. Our goal is now to control ε in this specific setting. Let $x \in \mathbb{R}^d$, $k \in \mathbb{N}$ and (X_1^x, X_2^x) be an optimal coupling between ν_k^{d} and ν_k^{c}. Note that X_2^x is a Gaussian random variable with mean $\nabla f(x)$ and covariance matrix $\Sigma(x)$, where $\Sigma(x) = (1/M)\Sigma_f(x)$ with $\Sigma_f = \pi[(\tilde{\nabla} f(x,\cdot) - \nabla f(x))(\tilde{\nabla} f(x,\cdot) - \nabla f(x))^\top]$. In particular, we get that

$$\mathbf{W}_2^2(\nu^{\mathrm{d}}(x), \nu^{\mathrm{c}}(x)) \leq M^{-1}\|\Sigma_f^{1/2}(x)\|^2 \mathbf{W}_2^2(\tilde{\nu}(x), \nu(x)) \ , \tag{34}$$

where $\tilde{\nu}(x)$ is the distribution of $M^{-1}\sum_{k=1}^M \Sigma_f(x)^{-1/2}\{\nabla\tilde{f}(x, Y_i) - \nabla f(x)\}$ with $\{Y_i\}_{i=1}^M$ distributed according to $\pi^{\otimes M}$ and ν the distribution of a Gaussian random variable with zero mean and identity covariance matrix. Denote $\{Y_i\}_{i=1}^M = \{\Sigma_f(x)^{-1/2}(\nabla\tilde{f}(x, Y_i) - \nabla f(x))\}_{i=1}^M$. The random variables $\{Y_i\}_{i=1}^M$ are i.i.d., $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}[Y_i Y_i^\top] = \mathrm{Id}$. In addition, using **A**1 and **A**2-(b) we have that

$$\begin{aligned}
\|Y_1\| &\leq 8\|\Sigma_f(x)^{-1/2}\|^4(\|\nabla f(x, U_1)\|^4 + \|\nabla f(x)\|^4) \\
&\leq 216C^4(1 + \|x\|^p)^4(\|\nabla f(x^\star, U_1)\|^4 + \mathtt{L}^4\|x\|^4 + \mathtt{L}^4\|x^\star\|^4 + \|\nabla f(0)\|^4 + \mathtt{L}^4\|x\|^4) \ .
\end{aligned}$$

Combining this result and the fact $\int_{\mathsf{Z}} \|\nabla\tilde{f}(x^\star, y)\|^4\mathrm{d}\pi(y) < +\infty$, there exists $q \in \mathbb{N}$ and $C \geq 0$ such that

$$\|\Sigma_f^{1/2}(x)\|^2\mathbb{E}[\|Y_1\|^4] \leq C(1 + \|x\|^{2q}) \ .$$

Therefore combining (Bonis, 2020, Theorem 1) and (34), there exists $C \geq 0$ such that for any $x \in \mathbb{R}^d$

$$\mathbf{W}_2^2(\nu_k^{\mathrm{d}}(x), \nu_k^{\mathrm{c}}(x)) \leq CM^{-1/2}(1 + \|x\|^{2q}) \ .$$

Using Lemma 17, there exists $C \geq 0$ such that $\varepsilon \leq CM^{-1}$, which concludes the proof. ■

Proposition 24 *Let $\alpha \in [0, 1/2)$, $T \geq 0$ and $\bar{\gamma} > 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $(T - \gamma_\alpha)^{1-2\alpha} - \gamma_\alpha^{1-2\alpha} \geq (T/2)^{1-2\alpha}$. Let $f = 0$ and $\tilde{f} : \mathbb{R}^d \to \mathbb{R}$ such that for any $x \in \mathbb{R}^d$ and $z \in \mathbb{R}^d$, $\tilde{f}(x, z) = \langle x, z \rangle$, $(\mathsf{Z}, \mathcal{Z}) = ((\mathbb{R}^d)^M, \mathcal{B}(\mathbb{R}^d)^{\otimes M})$, $\pi^Z = \pi^{\otimes M}$ with $M \in \mathbb{N}$, π a Gaussian distribution with zero mean and identity covariance matrix. In this case, for any $\gamma \in (0, \bar{\gamma}]$ we have for $n = \lfloor T/\gamma_\alpha \rfloor$*

$$\mathbb{E}^{1/2}\left[\|\mathbf{X}_{n\gamma_\alpha} - X_n\|^2\right] \geq M^{-1/2}\gamma^\delta(1 - 2\alpha)^{-1/2}(T/2)^{1/2-\alpha} , \quad \textit{with } \delta = \min(1, (2 - 2\alpha)^{-1}) ,$$

where $(\mathbf{X}_t)_{t\geq 0}$ is the solution of $\mathrm{d}\mathbf{X}_t = -\nabla f(\mathbf{X}_t)\mathrm{d}t$ and $(X_n)_{n\in\mathbb{N}}$ is a solution of (1).

Proof Note that for any $t \geq 0$, $\mathbf{X}_t = 0$. In addition, for any $n \in \mathbb{N}$, we have $X_n = (1/M)\gamma\sum_{k=0}^{n-1}(k+1)^{-\alpha}\sum_{m=1}^{M} Z^{k,m}$, where $\left\{Z^{k,m} : k, m \in \mathbb{N}\right\}$ is a collection of independent Gaussian random variables with zero mean and identity covariance matrix. Therefore we get that

$$\begin{aligned}\mathbb{E}\left[\|X_n\|^2\right] &= (1/M)\gamma^2\sum_{k=0}^{n-1}(k+1)^{-2\alpha} \\ &\geq (1/M)\gamma^2\int_1^n t^{-2\alpha}\mathrm{d}t \geq M^{-1}\gamma^2\gamma_\alpha^{2\alpha-1}(1 - 2\alpha)^{-1/2}((T - \gamma_\alpha)^{1-2\alpha} - \gamma_\alpha^{1-2\alpha}) ,\end{aligned}$$

which concludes the proof. ■

B.6. Weak approximation

We also derive weak approximation estimates of order 1. Note that in the case where $\alpha \geq 1/2$, these weak results are a direct consequence of Theorem 1. Denote by $\mathbb{G}_{p,k}$ the set of k-times continuously differentiable functions g such that there exists $\mathtt{K} \geq 0$ such that for any $x \in \mathbb{R}^d$, $\max(\|\nabla g(x)\|, \dots, \|\nabla^k g(x)\|) \leq \mathtt{K}(1 + \|x\|^p)$. We state our main result in Proposition 25.

Proposition 25 *Let $\bar{\gamma} > 0$, $\alpha \in [0, 1)$ and $p \in \mathbb{N}$. Assume that $f \in \mathbb{G}_{p,4}$, $\Sigma^{1/2} \in \mathbb{G}_{p,3}$,* **A**1, **A**2-(b) *and* **A**3. *Let $g \in \mathbb{G}_{p,2}$. In addition, assume that for any $m \in \mathbb{N}$ there exists $x^\star \in \mathbb{R}^d$ such that $\int_{\mathsf{Z}} \|H(x^\star, z)\|^{2m} \mathrm{d}\pi^Z(z) < +\infty$. Then for any $T \geq 0$, there exists $C \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $n \in \mathbb{N}$ with $\gamma_\alpha = \gamma^{1/(1-\alpha)}$, $n\gamma_\alpha \leq T$ we have*

$$|\mathbb{E}\left[g(\mathbf{X}_{n\gamma_\alpha}) - g(X_n)\right]| \leq C\gamma(1 + \log(\gamma^{-1})) .$$

These results extend (Li et al., 2017, Theorem 1.1 (a)) to the non-increasing stepsize case. Once again, the result obtained in Proposition 25 must be compared to similar weak error controls for SDEs. For example, under appropriate conditions, (Talay and Tubaro, 1990) shows that the EM discretization $Y_{n+1} = Y_n + \gamma\mathrm{b}(Y_n) + \sqrt{\gamma}\sigma(Y_n)G_{n+1}$ is a weak approximation of order 1 of (5).

We now turn to the proof of Proposition 25. We start with a useful technical lemma in Lemma 26. Then, before giving the proof of Proposition 25, we highlight that the result is straightforward for $\alpha \in [1/2, 1)$ in Proposition 27. We provide a one-step approximation error bound in Proposition 28 and conclude in Proposition 29. We recall that $\mathbb{G}_p$ is the set of twice continuously differentiable functions from $\mathbb{R}^d$ to $\mathbb{R}$ such that for any $g \in \mathbb{G}_p$, there exists $\mathtt{K} \geq 0$ such that for any $x \in \mathbb{R}^d$

$$\max\left\{\|\nabla g(x)\|, \left\|\nabla^2 g(x)\right\|\right\} \leq \mathtt{K}(1 + \|x\|^p) , \tag{35}$$

with $p \in \mathbb{N}$.

Lemma 26 *Let $p \in \mathbb{N}$, $g \in \mathbb{G}_p$ and let $\mathtt{K} \geq 0$ as in (35). Then, for any $x, y \in \mathbb{R}^d$*

$$|g(y) - g(x) - \langle \nabla g(x), y - x \rangle| \leq \mathtt{K}(1 + \|x\|^p + \|y\|^p) \|x - y\|^2 \ .$$

Proof Using that for any $x \mapsto \|x\|^p$ is convex, and Cauchy-Schwarz inequality we get for any $x, y \in \mathbb{R}^d$

$$\begin{aligned} |g(x) - g(y) - \langle \nabla g(x), y - x \rangle| &\leq \int_0^1 |\nabla^2 g(x + t(y-x))(y-x)^{\otimes 2}| \mathrm{d}t \\ &\leq \|x - y\|^2 \int_0^1 |\nabla^2 g(x + t(y-x))(y-x)^{\otimes 2}| \mathrm{d}t \\ &\leq \mathtt{K}(1 + \|x\|^p + \|y\|^p) \|x - y\|^2 \ . \end{aligned}$$

■

Proposition 27 *Let $\bar{\gamma} > 0$ and $\alpha \in [1/2, 1)$ and $p \in \mathbb{N}$. Assume **A**1, **A**2-(b) and **A**3. In addition, assume that for any $m \in \mathbb{N}$ there exists $x^\star \in \mathbb{R}^d$ such that $\int_{\mathsf{Z}} \|H(x^\star, z)\|^{2m} \mathrm{d}\pi^Z(z) < +\infty$. Then for any $T \geq 0$ and $g \in \mathbb{G}_p$, there exists $\mathtt{A}_{T,7} \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $k \in \mathbb{N}$ with $k\gamma_\alpha \leq T$ and $X_0 \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[|g(\mathbf{X}_{k\gamma_\alpha}) - g(X_k)| | \mathcal{F}_k\right] \leq \mathtt{A}_{T,7} \gamma (1 + \log(\gamma^{-1})) \ ,$$

where $(X_k)_{k \in \mathbb{N}}$ satisfies the recursion (1) and $(\mathbf{X}_t)_{t \geq 0}$ is the solution of (2) with $\mathbf{X}_0 = X_0$

Proof Let $p \in \mathbb{N}$, $g \in \mathbb{G}_p$, $\alpha \in [1/2, 1)$, $\bar{\gamma} > 0$, $\gamma \in (0, \bar{\gamma}]$, $k \in \mathbb{N}$, and $X_0 \in \mathbb{R}^d$. Using that for any $x \mapsto \|x\|^p$ is convex, for any $x, y \in \mathbb{R}^d$ we get

$$\begin{aligned} |g(x) - g(y)| &\leq \int_0^1 |\langle \nabla g(x + t(y-x)), y - x \rangle| \mathrm{d}t \leq \|y - x\| \int_0^1 \|\nabla g(x + t(y-x))\| \mathrm{d}t \\ &\leq \|y - x\| \mathtt{K}(1 + \|x\|^p + \|y\|^p) \ . \end{aligned}$$

Combining this result, Proposition 22, Lemma 17 and the Cauchy-Schwarz inequality we get that

$$\mathbb{E}\left[|g(\mathbf{X}_{k\gamma_\alpha}) - g(X_k)|\right] \leq \mathtt{K}\mathtt{A}_{T,6} \gamma (1 + \log(\gamma^{-1})) (\mathtt{A}_{T,1} + \tilde{\mathtt{A}}_{T,1})^{1/2} (1 + \|X_0\|^{2p})^{1/2} \ ,$$

which concludes the proof. ■

Proposition 28 *Let $p \in \mathbb{N}$ and $g \in \mathbb{G}_p$. Let $\bar{\gamma} > 0$ and $\alpha \in [0, 1)$. Assume **A**1, **A**2-(b), **A**3 and that for any $m \in \mathbb{N}$ there exists $x^\star \in \mathbb{R}^d$ such that $\int_{\mathsf{Z}} \|H(x^\star, z)\|^{2m} \mathrm{d}\pi^Z(z) < +\infty$. Then for any $T \geq 0$, there exists $\mathtt{A}_{T,8} \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $k \in \mathbb{N}$ with $(k+1)\gamma_\alpha \leq T$ and $X_0 \in \mathbb{R}^d$ we have*

$$\left|\mathbb{E}\left[g(\mathbf{X}_{(k+1)\gamma_\alpha}) - g(X_{k+1}) \middle| \mathcal{G}_k\right]\right| \leq \mathtt{A}_{T,8} \left\{\gamma^2 (k+1)^{-2\alpha} + \gamma (k+1)^{-(1+\alpha)}\right\} (1 + \|\mathbf{X}_{k\gamma_\alpha}\|^{p+2}) \ ,$$

where (X_k) is the solution of (1) and $\mathbf{X}_t$ is the solution of

$$\mathbf{X}_t = X_k - \int_{k\gamma_\alpha}^t (s + \gamma_\alpha)^{-\alpha} \nabla f(\mathbf{X}_s) \mathrm{d}s + \gamma_\alpha^{1/2} \int_{k\gamma_\alpha}^t (s + \gamma_\alpha)^{-\alpha} \Sigma(\mathbf{X}_s)^{1/2} \mathrm{d}\mathbf{B}_s \ .$$

Proof Let $\overline{\mathbf{X}}_{(k+1)\gamma_\alpha} = X_k - \gamma(k+1)^{-\alpha}\left\{\nabla f(\mathbf{X}_{k\gamma_\alpha}) + \Sigma(X_k)^{1/2} G_k\right\}$, with $G_k = \gamma_\alpha^{-1/2}\int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha} \mathrm{d}\mathbf{B}_s$. Using **A**2 we have $\mathbb{E}\left[\overline{\mathbf{X}}_{(k+1)\gamma_\alpha} \middle| \mathcal{G}_k\right] = \mathbb{E}\left[X_{k+1} | \mathcal{G}_k\right]$. Using Lemma 17, Lemma 18, Lemma 20, Lemma 26 and the Cauchy-Schwarz inequality we have

$$\begin{aligned}
&\left|\mathbb{E}\left[g(\mathbf{X}_{(k+1)\gamma_\alpha}) - g(X_{k+1}) \middle| \mathcal{G}_k\right]\right| \\
&\quad \leq \left|\mathbb{E}\left[\langle \nabla g(X_k), \mathbf{X}_{(k+1)\gamma_\alpha} - X_{k+1}\rangle \middle| \mathcal{G}_k\right]\right| \\
&\qquad + \mathtt{K}\mathbb{E}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - X_k\right\|^2 (1 + \|X_k\|^p + \left\|\mathbf{X}_{(k+1)\gamma_\alpha}\right\|^p) \middle| \mathcal{G}_k\right] \\
&\qquad + \mathtt{K}\mathbb{E}\left[\|X_{k+1} - X_k\|^2 (1 + \|X_k\|^p + \|X_{k+1}\|^p) \middle| \mathcal{G}_k\right] \\
&\quad \leq \left|\langle \nabla g(X_k), \mathbb{E}\left[\mathbf{X}_{(k+1)\gamma_\alpha} - \overline{\mathbf{X}}_{k+1} \middle| \rangle\right]\right| \mathcal{G}_k \\
&\qquad + 3^{1/2}\mathtt{K}\mathbb{E}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - X_k\right\|^4 \middle| \mathcal{G}_k\right]^{1/2} \mathbb{E}\left[(1 + \|X_k\|^{2p} + \|X_{k+1}\|^{2p}) \middle| \mathcal{G}_k\right]^{1/2} \\
&\qquad + 3^{1/2}\mathtt{K}\mathbb{E}\left[\|X_{k+1} - X_k\|^4 \middle| \mathcal{G}_k\right]^{1/2} \mathbb{E}\left[(1 + \|X_k\|^{2p} + \left\|\mathbf{X}_{(k+1)\gamma_\alpha}\right\|^{2p}) \middle| \mathcal{G}_k\right]^{1/2} \\
&\quad \leq \mathtt{K}(1 + \|X_k\|^p)\mathbb{E}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - \overline{\mathbf{X}}_{k+1}\right\|^2 \middle| \mathcal{G}_k\right]^{1/2} \\
&\qquad + 3^{1/2}\mathtt{K}\mathbb{E}\left[\left\|\mathbf{X}_{(k+1)\gamma_\alpha} - X_k\right\|^4 \middle| \mathcal{G}_k\right]^{1/2} (1 + \mathtt{A}_{T,1})^{1/2}(1 + \|X_k\|^p) \\
&\qquad + 3^{1/2}\mathtt{K}\mathbb{E}\left[\|X_{k+1} - X_k\|^4 \middle| \mathcal{G}_k\right]^{1/2} (1 + \tilde{\mathtt{A}}_{T,1})^{1/2}(1 + \|X_k\|^p) \\
&\quad \leq \mathtt{K}(1 + \|X_k\|^p)\mathtt{A}_{T,4}^{1/2}\left\{\gamma^2(k+1)^{-2\alpha} + \gamma(k+1)^{-(1+\alpha)}\right\}(1 + \|X_k\|) \\
&\qquad + 3^{1/2}\mathtt{K}\mathtt{A}_{T,2}^{1/2}\gamma^2(k+1)^{-2\alpha}(1 + \|X_k\|^2)(1 + \mathtt{A}_{T,1})^{1/2}(1 + \|X_k\|^p) \\
&\qquad + 3^{1/2}\mathtt{K}\mathtt{A}_{T,2}^{1/2}\gamma^2(k+1)^{-2\alpha}(1 + \|X_k\|^2)(1 + \tilde{\mathtt{A}}_{T,1})^{1/2}(1 + \|X_k\|^p) \ ,
\end{aligned}$$

which concludes the proof. ■

Proposition 29 *Let $\bar{\gamma} > 0$ and $\alpha \in [0, 1)$. Assume that $f \in \mathbb{G}_{p,4}$, $\Sigma^{1/2} \in \mathbb{G}_{p,3}$* **A**1*,* **A**2-(b) *and* **A**3*. Let $p \in \mathbb{N}$ and $g \in \mathbb{G}_{p,2}$. In addition, assume that for any $m \in \mathbb{N}$ there exists $x^\star \in \mathbb{R}^d$ such that $\int_{\mathsf{Z}} \|H(x^\star, z)\|^{2m} \mathrm{d}\pi^Z(z) < +\infty$. Then for any $T \geq 0$, there exists $\mathtt{A}_{T,9} \geq 0$ such that for any $\gamma \in (0, \bar{\gamma}]$, $k \in \mathbb{N}$ with $k\gamma_\alpha \leq T$ and $X_0 \in \mathbb{R}^d$ we have*

$$\left|\mathbb{E}\left[g(\mathbf{X}_{k\gamma_\alpha}) - g(X_k)\right]\right| \leq \mathtt{A}_{T,9}\gamma(1 + \log(\gamma^{-1})) \ ,$$

where $(X_k)_{k\in\mathbb{N}}$ satisfies the recursion (1) and $(\mathbf{X}_t)_{t\geq 0}$ is the solution of (2) with $\mathbf{X}_0 = X_0$.

Proof For any $k \in \mathbb{N}$ with $k\gamma_\alpha \leq T$, let $g_k(x) = \mathbb{E}\left[g(\mathbf{X}_{k\gamma_\alpha})\right]$ with $\mathbf{X}_0 = x$. Since $f \in \mathbb{G}_{p,4}$, $\Sigma^{1/2} \in \mathbb{G}_{p,3}$ and $g \in \mathbb{G}_{p,2}$ one can show, see (Blagovescenskii and Freidlin, 1961) or (Kunita, 1981, Proposition 2.1), that there exists $m \in \mathbb{N}$ and $\mathtt{K} \geq 0$ such that for any $k \in \mathbb{N}$ $g_k \in \mathrm{C}^m(\mathbb{R}^d, \mathbb{R})$ and

$$\max\left\{\|g_k(x)\|, \ldots, \|\nabla^m g_k(x)\|\right\} \leq \mathtt{K}(1 + \|x\|^p) \ .$$

Therefore, $g_k \in \mathbb{G}_{p,m}$ with constants uniform in $k \in \mathbb{N}$. In addition, for any $k \in \mathbb{N}$ with $k\gamma_\alpha \leq T$, let $h_k^{(1)}(x) = \mathbb{E}\left[g_k(X_{k+1})\right]$ with $X_k = x$ and $h_k^{(2)}(x) = \mathbb{E}\left[g_k(\mathbf{X}_{(k+1)\gamma_\alpha})\right]$ with $\mathbf{X}_{k\gamma_\alpha} = x$. Using Proposition 28 we have for any $k \in \mathbb{N}$, $k\gamma_\alpha \leq T$

$$\left|h_k^{(1)}(x) - h_k^{(2)}(x)\right| \leq \mathtt{A}_{T,8}\left\{\gamma^2(k+1)^{-2\alpha} + \gamma(k+1)^{-(1+\alpha)}\right\}(1 + \|x\|^{m+2}) \ .$$

Therefore, using Lemma 17 we have for any $k \in \mathbb{N}$ with $k\gamma_\alpha \leq T$ and $j \leq k$,

$$\left|\mathbb{E}\left[h^{(1)}_{k-j-1}(X_j) - h^{(2)}_{k-j-1}(X_j)\right]\right| \leq \tilde{\mathtt{A}}_{T,1}\mathtt{A}_{T,8}\left\{\gamma^2(k+1)^{-2\alpha} + \gamma(k+1)^{-(1+\alpha)}\right\}(1+\|X_0\|^{m+2}) \,. \tag{36}$$

Now, let $k \in \mathbb{N}$ with $k\gamma_\alpha \leq T$ and consider the family $\{(X^j_\ell)_{\ell\in\mathbb{N}} \,:\, j = 0, \dots, N\}$, defined by the following recursion: for any $j \in \{0, \dots, N\}$ $X^j_0 = X_0$ and for any $\ell \in \mathbb{N}$:

(a) if $\ell \geq j$,

$$X^j_{\ell+1} = X^j_\ell - \gamma(k+1)^{-\alpha}H(X^j_\ell, Z_{\ell+1}) \,,$$

(b) if $\ell < j$, $X^j_{\ell+1} = \mathbf{X}^j_{(\ell+1)\gamma_\alpha}$, where $\mathbf{X}^j_{\ell\gamma_\alpha} = X^j_\ell$ and for any $t \in [\ell\gamma_\alpha, (\ell+1)\gamma_\alpha]$ we have

$$\mathbf{X}^j_t = X^j_\ell - \int_{\ell\gamma_\alpha}^t (\gamma_\alpha + s)^{-\alpha}\nabla f(\mathbf{X}^j_s)\mathrm{d}s - \gamma_\alpha^{1/2}\int_{\ell\gamma_\alpha}^t (\gamma_\alpha + s)^{-\alpha}\Sigma^{1/2}(\mathbf{X}^j_s)\mathrm{d}\mathbf{B}_s \,.$$

We have

$$|\mathbb{E}\left[g(\mathbf{X}_{k\gamma_\alpha}) - g(X_k)\right]| = \left|\mathbb{E}\left[g(X^k_k) - g(X^0_k)\right]\right| = \sum_{j=0}^{k-1}\left|\mathbb{E}\left[g(X^{j+1}_k) - g(X^j_k)\right]\right| \,.$$

Using (36) we get

$$\begin{aligned}\left|\mathbb{E}\left[g(X^{j+1}_k) - g(X^j_k)\right]\right| &= \left|\mathbb{E}\left[\mathbb{E}\left[g(X^j_k) - g(X^{j+1}_k)\middle| X^j_k\right]\right]\right| \\ &= \left|\mathbb{E}\left[h^{(1)}_{k-j-1}(X_j) - h^{(2)}_{k-j-1}(X_j)\right]\right| \\ &\leq \tilde{\mathtt{A}}_{T,1}\mathtt{A}_{T,8}\left\{\gamma^2(k+1)^{-2\alpha} + \gamma(k+1)^{-(1+\alpha)}\right\}(1 + \|X_0\|^{m+2}) \\ &\leq \mathtt{A}^{(a)}_{T,9}\gamma^2(k+1)^{-2\alpha} + \gamma(k+1)^{-(1+\alpha)} \,,\end{aligned}$$

with $\mathtt{A}^{(a)}_{T,9} \geq 0$ which does not depend on k or γ In addition, using Lemma 14 there exists $\mathtt{A}^{(b)}_{T,9} \geq 0$ such that

$$\sum_{k=0}^{N-1}\left\{\gamma^2(k+1)^{-2\alpha} + \gamma(k+1)^{-(1+\alpha)}\right\} \leq \mathtt{A}^{(b)}_{T,9}\gamma \,.$$

Combining these last two results concludes the proof. ■

Appendix C. Strongly-Convex case (under A2-(a))

In this section, we gather the proofs for the study of the long-time behavior of SGD in the strongly convex case. Note that all of our proofs are derived under **A**2-(a). We refer to Appendix D for similar results under **A**2-(b). First, we start by deriving and proving our main results in the strongly convex case both for the continuous-time and the discrete-time dynamics in Appendix C.1. Then, we refine our study to explicit the dependency of the constant w.r.t. to the parameters of the problem in Appendix C.2. Finally, we show that our results can be extended to cover the case where the strongly convex assumption is replaced by a weaker Kurdyka-Łojasiewicz condition, in Appendix C.3.

C.1. Convergence results in the strongly convex case

First, we begin by deriving Corollary 30 which is a consequence of Theorem 5 and provides convergence rates for $(\mathbb{E}\left[f(\mathbf{X}_t)\right] - \min_{\mathbb{R}^d} f)_{t \geq 0}$. Then, we turn to the study of the discrete-time setting. We start by giving the proof of Lemma 4. The discrete analogous of Theorem 5 is given in Theorem 31. Similarly the discrete-time counterpart to Corollary 30 is given in Corollary 32.

Corollary 30 *Let $\alpha, \gamma \in (0,1)$ and $(\mathbf{X}_t)_{t \geq 0}$ be given by (2). Assume* **A**1,**A**2-(a)*,* **A**3 *and* **F**1-(a)*. Then there exists $C \geq 0$ such that for any $T > 0$, $\mathbb{E}\left[f(\mathbf{X}_T)\right] - \min_{\mathbb{R}^d} f \leq CT^{-\alpha}$.*

Proof The proof is a direct consequence of **A**1, (Nesterov, 2004, Lemma 1.2.3) and Theorem 5. ■

Proof [Proof of Lemma 4] Assume that there exists $n \in \mathbb{N}$ such that $u_n > B$, and let $n_1 = \inf\{n \geq 0 \, : \, u_n > B\}$. By definition of B we have $n_1 \geq n_0 + 1$. Moreover we have $u_{n_1} - u_{n_1 - 1} \leq F(n_1 - 1, u_{n_1 - 1})$. Since $n_1 - 1 \geq n_0$ we get that $u_{n_1} - u_{n_1-1} \leq A_2$ and $u_{n_1 - 1} \geq u_{n_1} - A_2 \geq A_1$. Consequently, $F(n_1 - 1, u_{n_1-1}) < 0$ and $u_{n_1} < u_{n_1 - 1}$, which is a contradiction. ■

We state a discrete analogous of Theorem 5. Note that the proof is considerably simpler than the one of (Bach and Moulines, 2011).

Theorem 31 *Let $\gamma \in (0,1)$ and $\alpha \in (0,1]$. Let $(X_n)_{n \geq 0}$ be given by (1). Assume* **A**2-(a) *and* **F**1-(a)*. Then there exists $\mathtt{B}_3 > 0$ such that for all $N \geq 1$,*

$$\mathbb{E}[\|X_N - x^\star\|^2] \leq \mathtt{B}_3 N^{-\alpha} \; .$$

In the case where $\alpha = 1$ we have to assume additionally that $\gamma > 1/(2\mu)$.

Proof Let $\gamma \in (0,1)$ and $\alpha \in (0,1]$. Let $(X_n)_{n \geq 0}$ be given by (1). Using **F**1-(a) we get for all $n \geq 0$,

$$\begin{aligned}
\mathbb{E}\left[\|X_{n+1} - x^\star\|^2 \middle| \mathcal{F}_n\right] &= \mathbb{E}\left[\left\|X_n - x^\star - \gamma(n+1)^{-\alpha} H(X_n, Z_{n+1})\right\|^2\right] \\
&= \|X_n - x^\star\|^2 + \gamma^2 (n+1)^{-2\alpha} \mathbb{E}\left[\|H(X_n, Z_{n+1})\|^2 \middle| \mathcal{F}_n\right] \\
&\quad - 2\gamma(n+1)^{-\alpha} \mathbb{E}\left[\langle X_n - x^\star, H(X_n, Z_{n+1})\rangle | \mathcal{F}_n\right] \\
&\leq \|X_n - x^\star\|^2 + \gamma^2(n+1)^{-2\alpha}\left[\eta + \|\nabla f(X_n)\|^2\right] \\
&\quad - 2\gamma(n+1)^{-\alpha} \langle X_n - x^\star, \nabla f(X_n)\rangle \; .
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\|X_{n+1} - x^\star\|^2\right] \\
&\qquad \leq \mathbb{E}\left[\|X_n - x^\star\|^2\right]\left[1 - 2\gamma(n+1)^{-\alpha}\mu + \gamma^2 (n+1)^{-2\alpha}\mathtt{L}^2\right] + \eta\gamma^2(n+1)^{-2\alpha} \; . \qquad (37)
\end{aligned}$$

We note now $u_n = \mathbb{E}\left[\|X_n - x^\star\|^2\right]$ and $v_n = n^\alpha u_n$. Using (37) and Bernoulli's inequality we have, for all $n \geq 0$

$$\begin{aligned}
v_{n+1} - v_n &= (n+1)^\alpha u_{n+1} - n^\alpha u_n \\
&= (n+1)^\alpha (u_{n+1} - u_n)) + u_n((n+1)^\alpha - n^\alpha) \\
&\leq \left[-2\gamma\mu + \gamma^2 \mathtt{L}^2 (n+1)^{-\alpha}\right] u_n + \eta\gamma^2(n+1)^{-\alpha} + u_n n^\alpha \left[(1 + 1/n)^\alpha - 1\right] \\
&\leq \left[-2\gamma\mu + \gamma^2\mathtt{L}^2(n+1)^{-\alpha} + \alpha n^{\alpha - 1}\right] u_n + \eta\gamma^2 (n+1)^{-\alpha} \; .
\end{aligned}$$

Therefore, in the case where $\alpha < 1$, there exists $n_0 \geq 0$ such that for all $n \geq n_0$,

$$\begin{aligned} v_{n+1} - v_n &\leq -\gamma\mu u_n + \eta\gamma^2(n+1)^{-\alpha} \\ &\leq -\gamma\mu n^{-\alpha} v_n + \eta\gamma^2(n+1)^{-\alpha} \leq (n+1)^{-\alpha}(-\gamma\mu v_n + \eta\gamma^2) \ . \end{aligned}$$

And in the case where $\alpha = 1$, if $\gamma > 1/(2\mu)$ we have the existence of $n_1 \geq 0$ such that for all $n \geq n_1$,

$$v_{n+1} - v_n \leq \left[(1/2 - \gamma\mu) + \gamma^2 \mathtt{L}^2 (n+1)^{-\alpha} + \alpha n^{\alpha-1}\right] u_n + \eta\gamma^2(n+1)^{-\alpha} \ .$$

Using Lemma 4 this shows that, for $\alpha \in (0,1]$, there exists a constant $\mathtt{B}_3 > 0$ such that for all $n \geq 0$, $v_n \leq \mathtt{B}_3$. This proves the result. ■

Using **A**1 and the descent lemma (Nesterov, 2004, Lemma 1.2.3) we have the immediate corollary

Corollary 32 *Let $\alpha \in (0,1]$ and $\gamma \in (0,1)$. Let $(X_n)_{n\geq 0}$ be given by (1). Assume **A**1,**A**2-(a) and **F**1-(a). Then there exists $\mathtt{B}_4 > 0$ such that for all $N \geq 1$,*

$$\mathbb{E}\left[f(X_N) - f^\star\right] \leq \mathtt{B}_4 N^{-\alpha} \ .$$

If $\alpha = 1$ we have also assumed that $\gamma > 1/(2\mu)$.

C.2. Quantitative constants in the strongly convex setting

We first state in Lemma 33 a specific version of Lemma 3 in the case where there exists $t_0 > 0$ such that for any $t \geq 0$ and $F(t,x) \geq -\mathrm{f}(x)\mathrm{g}(t)$ with f superlinear. In particular, this lemma allows to obtain (i) an exponential forgetting of the initial conditions, (ii) a more explicit expression of the constant appearing in Lemma 3. The improved version of Theorem 5 with explicit constants is stated Theorem 34.

Lemma 33 *Let $F \in \mathrm{C}^1(\mathbb{R}_+ \times \mathbb{R}, \mathbb{R})$ and $v \in \mathrm{C}^1(\mathbb{R}_+, \mathbb{R}_+)$ such that for all $t \geq 0$, $\mathrm{d}v(t)/\mathrm{d}t \leq F(t, v(t))$. Assume that there exist $\mathrm{f} : \mathbb{R} \to \mathbb{R}$, $\mathrm{g} \in \mathrm{C}(\mathbb{R}_+, \mathbb{R}_+)$, $t_0 > 0$, $A \geq 0$ and $\upbeta > 0$ such that the following conditions hold.*

(a) For any $t \geq t_0$, $r \in (0,1]$ and $u \geq 0$, $rF(t,u) \leq F(t, ru)$.

(b) For any $t \geq t_0$ and $u \geq 0$, $F(t,u) \leq -\mathrm{f}(u)\mathrm{g}(t)$.

(c) For any $u \geq A$, $\mathrm{f}(u) > \upbeta u$.

Then, for any $t \geq 0$,

$$v(t) \leq \max\{A, \exp[\upbeta(G(t_0) - G(t))] \max_{s\in[0,t_0]} v(s)\} \ ,$$

with $G(t) = \int_0^t \mathrm{g}(s)\mathrm{d}s$.

Proof Let $T \geq 0$ and $v_T(t) = v(t)\exp[\beta(G(t) - G(T))]$. Using condition (a) and that G is non-decreasing since for any $t \geq 0$, $\mathrm{g}(t) \geq 0$, we have for any $t \in (0, T]$

$$\mathrm{d}v_T(t)/\mathrm{d}t \leq \exp[\beta(G(t) - G(T))]F(t, v(t)) + \beta\mathrm{g}(t)v_T(t) \leq F(t, v_T(t)) + \beta\mathrm{g}(t)v_T(t) \ .$$

Using this result and conditions (b)-(c), we have for any $t \geq t_0$ such that $v_T(t) \geq A$

$$\mathrm{d}v_T(t)/\mathrm{d}t \leq -\mathrm{f}(v_T(t))\mathrm{g}(t) + \beta v_T(t)\mathrm{g}(t) < 0 \ . \tag{38}$$

Let $B = \max(A, \max_{s \in [0,t_0]} v_T(s))$. Assume that $\mathsf{A} = \{t \in [0, T] \ : \ v_T(t) > B\} \neq \emptyset$ and let $t_1 = \inf \mathsf{A}$. Note that $t_1 \geq t_0$ and $v_T(t_1) \geq A$. Therefore, using (38) we have $\mathrm{d}v_T(t_1)/\mathrm{d}t < 0$ and therefore, there exists $0 < t_2 < t_1$ such that $v_T(t_2) > v_T(t_1)$ but then $t_2 \in \mathsf{A}$ and $t_2 < \inf \mathsf{A}$. Hence, $\mathsf{A} = \emptyset$ and we get that for any $t \in [0, T]$, $v_T(t) \leq B$. Therefore, we get that for any $t \geq 0$,

$$v(t) = v_t(t) \leq \max\{A, \exp[\beta(G(t_0) - G(t))] \max_{s \in [0,t_0]} v(s)\} \ ,$$

which concludes the proof. ■

Theorem 34 *Let $\alpha, \gamma \in (0, 1)$ and $(\mathbf{X}_t)_{t \geq 0}$ be given by (2). Assume* **A**1*,* **A**2-(a)*,* **A**3 *and* **F**1-(a)*. Then for any $T \geq 0$,*

$$\mathbb{E}[\|\mathbf{X}_T - x^\star\|^2] \leq \max\left\{4\gamma_\alpha\eta/\mu, C\mathbb{E}[\|\mathbf{X}_0 - x^\star\|^2]\exp[-\mu(\gamma_\alpha + T)^{1-\alpha}/(2 - 2\alpha)]\right\}(\gamma_\alpha + T)^{-\alpha} \ ,$$

with

$$C = (1 + \eta\Psi(\alpha, t_0))\exp[\mu(\gamma_\alpha + t_0)^{1-\alpha}/(2 - 2\alpha)](\gamma_\alpha + t_0)^\alpha \ .$$

Proof Let $\alpha, \gamma \in (0, 1]$ and consider $\mathcal{E} : \mathbb{R}_+ \to \mathbb{R}_+$ defined for $t \geq 0$ by $\mathcal{E}(t) = \mathbb{E}[(t + \gamma_\alpha)^\alpha\|\mathbf{X}_t - x^\star\|^2]$, with $\gamma_\alpha = \gamma^{1/(1-\alpha)}$. Using Dynkin's formula, see Lemma 48, we have for any $t \geq 0$,

$$\mathcal{E}(t) = \mathcal{E}(0) + \alpha\int_0^t \frac{\mathcal{E}(s)}{s + \gamma_\alpha}\mathrm{d}s + \int_0^t \gamma_\alpha\frac{\mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_s))\right]}{(s + \gamma_\alpha)^\alpha}\mathrm{d}s - 2\int_0^t \mathbb{E}\left[\langle\nabla f(\mathbf{X}_s), \mathbf{X}_s - x^\star\rangle\right]\mathrm{d}s \ .$$

We now differentiate this expression with respect to t and using **F**1 and **A**2, we get for any $t > 0$,

$$\begin{aligned}
\mathrm{d}\mathcal{E}(t)/\mathrm{d}t &= \alpha\mathcal{E}(t)(t + \gamma_\alpha)^{-1} - 2\mathbb{E}\left[\langle\nabla f(\mathbf{X}_t), \mathbf{X}_t - x^\star\rangle\right] + \gamma_\alpha\mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_t))\right](t + \gamma_\alpha)^{-\alpha} \\
&\leq \alpha\mathcal{E}(t)/(t + \gamma_\alpha) - 2\mu\mathbb{E}[\|\mathbf{X}_t - x^\star\|^2] + \gamma_\alpha\eta/(t + \gamma_\alpha)^\alpha \\
&\leq F(t, \mathcal{E}(t)) = \alpha\mathcal{E}(t)(t + \gamma_\alpha)^{-1} - 2\mu\mathcal{E}(t)(t + \gamma_\alpha)^{-\alpha} + \gamma_\alpha\eta(t + \gamma_\alpha)^{-\alpha} \ ,
\end{aligned}$$

where we have used in the penultimate line that $\mathrm{Tr}(\Sigma(x)) \leq \eta$ for any $x \in \mathbb{R}^d$ by **A**2. Let $t_0 = \max((\alpha/\mu)^{1/(1-\alpha)} - \gamma_\alpha, \gamma_\alpha)$. We have for any $t \geq t_0$, and $u \geq 0$

$$F(t, u) \leq -\mathrm{f}(u)\mathrm{g}(t) \ , \qquad \mathrm{g}(t) = (t + \gamma_\alpha)^{-\alpha} \ , \qquad \mathrm{f}(u) = \mu u - \gamma_\alpha\eta \ .$$

Hence the conditions (a) and (b) of Lemma 33 are satisfied. Let $\beta = \mu/2$ and $A = 4\gamma_\alpha\eta/\mu$. We obtain that for any $t \geq t_0$ and $u \geq A$, $\mathrm{f}(u) > \mu u/2$ and therefore condition (c) of Lemma 33 is satisfied. Applying Lemma 33, we obtain that for any $t \geq 0$

$$\mathcal{E}(t) \leq \max(4\gamma_\alpha\eta/\mu, \exp[-\mu(\gamma_\alpha + t)^{1-\alpha}/(2 - 2\alpha)]B) \ ,$$

with $B = \exp[\mu(\gamma_\alpha + t_0)^{1-\alpha}/(2-2\alpha)] \max_{s\in[0,t_0]} \mathcal{E}(s)$. We have that $\max_{s\in[0,t_0]} \mathcal{E}(s) \leq (t_0 + \gamma_\alpha)^\alpha \max_{s\in[0,t_0]} \mathbb{E}[\|\mathbf{X}_s - x^\star\|]^2$. Using Dynkin's formula, see Lemma 48, we have for any $t \geq 0$,

$$\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|\right]^2 \leq \mathbb{E}\left[\|\mathbf{X}_0 - x^\star\|\right]^2 + \eta\Psi(\alpha, t_0) ,$$

with

$$\Psi(\alpha, t_0) = \begin{cases} \gamma^2/(2\alpha-1) & \text{if } 2\alpha > 1 \ , \\ \gamma_\alpha \log(\gamma_\alpha^{-1}(t_0+\gamma_\alpha)) & \text{if } 2\alpha = 1 \ , \\ \gamma_\alpha(t_0+\gamma_\alpha)^{1-2\alpha}/(1-2\alpha) & \text{otherwise} \ . \end{cases}$$

We conclude the proof upon setting $C = (1+\eta\Psi(\alpha,t_0))\exp[\mu(\gamma_\alpha+t_0)^{1-\alpha}/(2-2\alpha)](\gamma_\alpha+t_0)^\alpha$. ■

C.3. Convergence results under Kurdyka-Łojasiewicz conditions

We state now an equivalent result of Corollary 30 under weaker assumptions, namely the Łojasiewicz inequality with $r = 2$, that we restate as it is usually given, with $c > 0$, *i.e.*, for any $x \in \mathbb{R}^d$,

$$f(x) - f(x^\star) \leq c\|\nabla f(x)\|^2 \ . \tag{39}$$

Note that (39) is verified for all strongly convex functions (Karimi et al., 2016). The equivalent of Corollary 30 is stated in Proposition 35 (for the continuous-time process). The equivalent of Corollary 32 is given in Proposition 36 (for the discrete-time process).

Proposition 35 *Let $\alpha, \gamma \in (0,1)$ and $(\mathbf{X}_t)_{t\geq 0}$ be given by (2). Assume* **A**1, **A**2-(a), **A**3 *and that f verifies (39). Then there exists $\mathtt{B}_5 > 0$ such that for any $T > 0$,*

$$\mathbb{E}\left[f(\mathbf{X}_T) - f^\star\right] \leq \mathtt{B}_5 T^{-\alpha} \ .$$

Proof Let $\alpha, \gamma \in (0,1)$ and $(\mathbf{X}_t)_{t\geq 0}$ be given by (2). Without loss of generality we can assume that $f^\star = \min_{x\in\mathbb{R}^d} f(x) = 0$. We note $\mathcal{E}(t) = (t+\gamma_\alpha)^\alpha \mathbb{E}\left[f(\mathbf{X}_t)\right]$ and we apply Lemma 48 to the stochastic process $((t+\gamma_\alpha)^\alpha f(\mathbf{X}_t))_{t\geq 0}$, and using **A**1, **A**2-(a), **A**3, (39) and Lemma 47 this gives, for all $t > 0$,

$$\begin{aligned} \mathcal{E}(t) - \mathcal{E}(0) = \int_0^t \alpha(s+\gamma_\alpha)^{\alpha-1}\mathbb{E}\left[f(\mathbf{X}_s)\right] \mathrm{d}s - \int_0^t \mathbb{E}\left[\|\nabla f(\mathbf{X}_s)\|^2\right] \mathrm{d}s \\ + (\gamma_\alpha/2)\int_0^t (s+\gamma_\alpha)^{-\alpha}\mathbb{E}\left[\mathrm{Tr}(\nabla^2 f(\mathbf{X}_s)\Sigma(\mathbf{X}_s))\right] \mathrm{d}s \\ \mathrm{d}\mathcal{E}(t)/\mathrm{d}t \leq \alpha\mathcal{E}(t)(t+\gamma_\alpha)^{-1} - (1/c)\mathcal{E}(t)(t+\gamma_\alpha)^{-\alpha} + (\gamma_\alpha/2)\mathtt{L}\eta(t+\gamma_\alpha)^{-\alpha} \ . \end{aligned}$$

We can now apply Lemma 3 to $F(t,x) = \alpha x(t+\gamma_\alpha)^{-1} - (1/c)x(t+\gamma_\alpha)^{-\alpha} + (\gamma_\alpha/2)\mathtt{L}\eta(t+\gamma_\alpha)^{-\alpha}$ with $t_0 = (2c\alpha)^{1/(1-\alpha)}$ and $A = 2\gamma_\alpha c\mathtt{L}\eta$, which shows the existence of $\mathtt{B}_5 > 0$ such that for all $t > 0$, $\mathcal{E}(t) \leq \mathtt{B}_5$, concluding the proof. ■

And we now state its discrete counterpart, which is an equivalent of Corollary 32.

Proposition 36 *Let $\alpha \in (0,1]$ and $\gamma \in (0,1)$. Let $(X_n)_{n \geq 0}$ be given by (1). Assume* **A**1*,* **A**2-(a) *and that f verifies (39). Then there exists $\mathtt{B}_6 > 0$ such that for all $N \geq 1$,*

$$\mathbb{E}\left[f(X_N) - f^\star\right] \leq \mathtt{B}_6 N^{-\alpha} .$$

In the case where $\alpha = 1$ we have to assume additionally that $\gamma > 2/c$.

Proof Let $\alpha \in (0,1]$ and $\gamma \in (0,1)$. Let $(X_n)_{n \geq 0}$ be given by (1). Let $n \geq 0$. Applying the descent lemma (Nesterov, 2004, Lemma 1.2.3) (using **A**1) we get

$$\begin{aligned}
\mathbb{E}\left[f(X_{n+1})|\mathcal{F}_n\right] &= \mathbb{E}\left[f(X_n - \gamma/(n+1)^\alpha H(X_n, Z_{n+1}))|\mathcal{F}_n\right] \\
&\leq f(X_n) - \gamma/(n+1)^\alpha \mathbb{E}\left[\langle \nabla f(X_n), H(X_n, Z_{n+1})\rangle|\mathcal{F}_n\right] \\
&\quad + \gamma^2/(n+1)^{2\alpha}(\mathtt{L}/2)\mathbb{E}[\|H(X_n, Z_{n+1})\|^2 \,|\mathcal{F}_n] \\
&\leq f(X_n) - \gamma/(n+1)^\alpha \|\nabla f(X_n)\|^2 + (\mathtt{L}\gamma^2/2)(n+1)^{-2\alpha}\left[\eta + \|\nabla f(X_n)\|^2\right] \\
\mathbb{E}\left[f(X_{n+1})\right] - f^\star &\leq \mathbb{E}\left[f(X_n)\right] - f^\star + \gamma(n+1)^{-\alpha}\mathbb{E}[\|\nabla f(X_n)\|^2]\left[-1 + (\mathtt{L}\gamma/2)(n+1)^{-\alpha}\right] \\
&\quad + (\mathtt{L}\gamma^2/2)(n+1)^{-2\alpha}\eta .
\end{aligned}$$

This shows the existence of $n_2 \geq 0$ such that using (39) we have for all $n \geq n_2$,

$$\begin{aligned}
\mathbb{E}\left[f(X_{n+1})\right] - f^\star &\leq \mathbb{E}\left[f(X_n)\right] - f^\star - (\gamma/2)(n+1)^{-\alpha}\mathbb{E}\left[\|\nabla f(X_n)\|^2\right] + (\mathtt{L}\gamma^2/2)(n+1)^{-2\alpha}\eta \\
&\leq \left(\mathbb{E}\left[f(X_n)\right] - f^\star\right)\left[1 - (\gamma c^{-1}/2)(n+1)^{-\alpha}\right] + (\mathtt{L}\gamma^2/2)(n+1)^{-2\alpha}\eta .
\end{aligned}$$

We note now for all $n \geq 0$, $u_n = \mathbb{E}\left[f(X_n)\right] - f^\star$ and $v_n = n^\alpha u_n$. We have

$$\begin{aligned}
v_{n+1} - v_n &= (n+1)^\alpha u_{n+1} - n^\alpha u_n \\
&= (n+1)^\alpha (u_{n+1} - u_n)) + u_n((n+1)^\alpha - n^\alpha) \\
&\leq -(\gamma c^{-1}/2)u_n + (\mathtt{L}\gamma^2\eta/2)(n+1)^{-\alpha} + u_n n^\alpha\left[(1+1/n)^\alpha - 1\right] \\
&\leq u_n(-(\gamma c^{-1}/2) + \alpha n^{\alpha-1}) + (\mathtt{L}\gamma^2\eta/2)(n+1)^{-\alpha} .
\end{aligned}$$

If $\alpha < 1$, or if $1 - \gamma c^{-1}/2 < 0$ we have the existence of $n_3 \geq n_2$ and $\tilde{\mathtt{B}} > 0$ such that for all $n \geq n_3$,

$$\begin{aligned}
v_{n+1} - v_n &\leq -\tilde{\mathtt{B}} u_n + (\mathtt{L}\gamma^2\eta/2)(n+1)^{-\alpha} \\
&\leq \left\{-\tilde{\mathtt{B}} v_n + (\mathtt{L}\gamma^2\eta/2)\right\}(n+1)^{-\alpha}
\end{aligned}$$

This proves the existence of $\mathtt{B}_6 > 0$ such that for all $n \geq 0$, $v_n \leq \mathtt{B}_6$, which concludes the proof. ■

Appendix D. Strongly convex case (under A2-(b))

This section gather the proofs for the study of the strongly convex case under **A**2-(b). It is the counterpart of Appendix C. We start by establishing useful lemmas under **A**2-(b) in Appendix D.1. Then we present the counterpart of the results obtained in Appendix C.1 in Appendix D.3.

D.1. Technical results

We begin by several lemmas to control $\operatorname{Tr}\Sigma$ and $\mathbb{E}[\|\nabla\tilde{f}\|^2]$. We will note $\mathtt{L_T} = 6\mathtt{L}^2 + 4\mathtt{L} + 3\eta$.

Lemma 37 *Assume* **A**2-(b). *Then, for all $x \in \mathbb{R}^d$, we have*

$$\int_{\mathsf{Z}} \|\nabla\tilde{f}(x,z)\|^2 \mathrm{d}\pi^Z(z) \le \mathtt{L_T}(\|x - x^\star\|^2 + 1) \ .$$

Proof Let $x \in \mathbb{R}^d$. Using **A**2-(b) we have

$$\begin{aligned}\int_{\mathsf{Z}} \|\nabla\tilde{f}(x,z)\|^2 \mathrm{d}\pi^Z(z) &= \int_{\mathsf{Z}} \|\nabla\tilde{f}(x,z) - \nabla\tilde{f}(x^\star,z) + \nabla\tilde{f}(x^\star,z)\|^2 \mathrm{d}\pi^Z(z)\\ &\le 2\int_{\mathsf{Z}} \|\nabla\tilde{f}(x,z) - \nabla\tilde{f}(x^\star,z)\|^2 + \int_{\mathsf{Z}} \|\nabla\tilde{f}(x^\star,z)\|^2 \mathrm{d}\pi^Z(z)\mathrm{d}\pi^Z(z)\\ &\le 2\mathtt{L}^2\|x - x^\star\|^2 + 2\eta \ ,\end{aligned}$$

which concludes the proof. ■

Lemma 38 *Assume* **A**2-(b). *Then, for all $x \in \mathbb{R}^d$, we have*

$$\operatorname{Tr}(\Sigma(x)) \le \mathtt{L_T}\left(1 + \|x - x^\star\|^2\right) \ .$$

Proof Let $x \in \mathbb{R}^d$. Using **A**2-(b) we have

$$\begin{aligned}\operatorname{Tr}(\Sigma(x)) &= \operatorname{Tr}\left(\int_{\mathsf{Z}} (\nabla\tilde{f}(x,z) - \nabla f(x))(\nabla\tilde{f}(x,z) - \nabla f(x))^\top \mathrm{d}\pi^Z(z)\right)\\ &= \int_{\mathsf{Z}} \|\nabla\tilde{f}(x,z) - \nabla f(x)\|^2 \mathrm{d}\pi^Z(z)\\ &= \int_{\mathsf{Z}} \|\nabla\tilde{f}(x,z) - \nabla\tilde{f}(x^\star,z) + \nabla\tilde{f}(x^\star,z) - \nabla f(x)\|^2 \mathrm{d}\pi^Z(z)\\ &\le 3\int_{\mathsf{Z}} \left(\|\nabla\tilde{f}(x,z) - \nabla\tilde{f}(x^\star,z)\|^2 + \|\nabla\tilde{f}(x^\star,z)\|^2 + \|\nabla f(x)\|^2\right) \mathrm{d}\pi^Z(z)\\ &\le 6\mathtt{L}^2\|x - x^\star\|^2 + 3\eta \ \le \mathtt{L_T}\left(1 + \|x - x^\star\|^2\right) \ ,\end{aligned}$$

which concludes the proof. ■

Lemma 39 *Let $a, b \in \mathbb{R}^d$. Then $\|a+b\|^2 \ge \|a\|^2/2 - \|b\|^2$.*

Proof Let $a, b \in \mathbb{R}^d$. Using the fact that $2xy \le 2y^2 + x^2/2$ for all $x, y \in \mathbb{R}$, we have

$$\begin{aligned}\|a+b\|^2 &= \|a\|^2 + \|b\|^2 + 2\langle a, b\rangle\\ &\ge \|a\|^2 + \|b\|^2 - 2\|a\|\|b\| \ge \|a\|^2 + \|b\|^2 - 2\|b\|^2 - \|a\|^2/2 \ge \|a\|^2/2 - \|b\|^2 \ .\end{aligned}$$

■

Lemma 40 *Let $f \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R})$. Assume that there exists $\mathtt{L} \geq 0$ such that for any $x, y \in \mathbb{R}^d$, ∇f is $\mathtt{L}$-Lipschitz. Then for any $x \in \mathbb{R}^d$*

$$\|\nabla f(x)\|^2 \leq 2\mathtt{L}(f(x) - \inf_{\mathbb{R}^d} f) \ . \tag{40}$$

Proof Using (Nesterov, 2004, Lemma 1.2.3), we have for any $x, y \in \mathbb{R}^d$

$$f(y) - f(x) \leq \langle \nabla f(x), y - x \rangle + (\mathtt{L}/2) \|y - x\|^2 \ .$$

We obtain (40) by minimizing both side of the previous inequality w.r.t. y. ■

Lemma 41 *Assume* **A**2-(b). *In addition, assume that $\tilde{f}(\cdot, z)$ is convex for all $z \in \mathsf{Z}$. For all $x \in \mathbb{R}^d$, we have*

$$\int_{\mathsf{Z}} \|\nabla \tilde{f}(x, z)\|^2 \mathrm{d}\pi^Z(z) \leq \mathtt{L}_{\mathtt{T}} \left(f(x) - f(x^\star) + 1 \right) \ .$$

Proof Let $x \in \mathbb{R}^d$ and $z \in \mathsf{Z}$. Using the smoothness and convexity of $\tilde{f}(\cdot, z)$, taking the expectation and using Lemma 39 we have

$$\begin{aligned} \tilde{f}(x, z) - \tilde{f}(x^\star, z) &\geq \langle \nabla \tilde{f}(x^\star, z), x - x^\star \rangle + (1/2\mathtt{L}) \|\nabla \tilde{f}(x, z) - \nabla \tilde{f}(x^\star, z)\|^2 \\ f(x) - f(x^\star) &\geq (1/2\mathtt{L}) \mathbb{E}[\|\nabla \tilde{f}(x, z) - \nabla \tilde{f}(x^\star, z)\|^2] \\ &\geq (1/4\mathtt{L}) \mathbb{E}[\|\nabla \tilde{f}(x, z)\|^2] - (1/2\mathtt{L}) \mathbb{E}[\|\nabla f(x)\|^2] \ , \end{aligned}$$

We conclude upon combining this result with Lemma 40. ■

Lemma 42 *Assume* **A**1, **A**2-(b) *and* **A**3. *Assume additionally that $\tilde{f}(\cdot, z)$ is convex for all $z \in \mathsf{Z}$. For all $x \in \mathbb{R}^d$, we have*

$$\mathrm{Tr}(\Sigma(x)) \leq \mathtt{L}_{\mathtt{T}} \left(f(x) - f(x^\star) + 1 \right) \ .$$

Proof Let $x \in \mathbb{R}^d$. Then using **A**2-(b) and Lemma 41 we have

$$\begin{aligned} \mathrm{Tr}\left(\Sigma(x)\right) &= \int_{\mathsf{Z}} \|\nabla \tilde{f}(x, z) - \nabla f(x)\|^2 \mathrm{d}\pi^Z(z) \\ &= \int_{\mathsf{Z}} \|\nabla \tilde{f}(x, z)\|^2 + \|\nabla f(x)\|^2 - 2\langle \nabla \tilde{f}(x, z), \nabla f(x) \rangle \mathrm{d}\pi^Z(z) \\ &= \int_{\mathsf{Z}} \|\nabla \tilde{f}(x, z)\|^2 \mathrm{d}\pi^Z(z) - \|\nabla f(x)\|^2 \\ &\leq \int_{\mathsf{Z}} \|\nabla \tilde{f}(x, z)\|^2 \mathrm{d}\pi^Z(z) \leq \mathtt{L}_{\mathtt{T}} \left(f(x) - f(x^\star) + 1 \right) \ , \end{aligned}$$

which concludes the proof. ■

D.2. Equivalent to Appendix C.1

The equivalent of Theorem 5 and Theorem 31 are given in Theorem 43 and Theorem 44 respectively.

Theorem 43 *Let $\alpha, \gamma \in (0,1)$ and $(\mathbf{X}_t)_{t \geq 0}$ be given by (2). Assume* **A**1*,* **A**2-(b)*,* **A**3 *and* **F**1-(b)*. Then there exists $C \geq 0$ (explicit in the proof) such that for any $T \geq 1$, $\mathbb{E}[\|\mathbf{X}_T - x^\star\|^2] \leq CT^{-\alpha}$.*

Proof Let $\alpha, \gamma \in (0,1]$ and consider $\mathcal{E} : \mathbb{R}_+ \to \mathbb{R}_+$ defined for $t \geq 0$ by $\mathcal{E}(t) = \mathbb{E}[(t+\gamma_\alpha)^\alpha \|\mathbf{X}_t - x^\star\|^2]$, with $\gamma_\alpha = \gamma^{1/(1-\alpha)}$. Using Dynkin's formula, see Lemma 48, we have for any $t \geq 0$,

$$\mathcal{E}(t) = \mathcal{E}(0) + \alpha \int_0^t \frac{\mathcal{E}(s)}{s+\gamma_\alpha} \mathrm{d}s + \int_0^t \gamma_\alpha \frac{\mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_s))\right]}{(s+\gamma_\alpha)^\alpha} \mathrm{d}s - 2\int_0^t \mathbb{E}\left[\langle \nabla f(\mathbf{X}_s), \mathbf{X}_s - x^\star \rangle\right] \mathrm{d}s \ .$$

We now differentiate this expression with respect to t and using **F**1, **A**2 and Lemma 38, we get for any $t > 0$,

$$\begin{aligned}
\mathrm{d}\mathcal{E}(t)/\mathrm{d}t &= \alpha\mathcal{E}(t)(t+\gamma_\alpha)^{-1} - 2\mathbb{E}\left[\langle \nabla f(\mathbf{X}_t), \mathbf{X}_t - x^\star \rangle\right] + \gamma_\alpha \mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_t))\right](t+\gamma_\alpha)^{-\alpha} \\
&\leq \alpha\mathcal{E}(t)/(t+\gamma_\alpha) - 2\mu\mathbb{E}[\|\mathbf{X}_t - x^\star\|^2] + \gamma_\alpha \mathtt{L}_\mathtt{T}/(t+\gamma_\alpha)^\alpha + \gamma_\alpha \mathtt{L}_\mathtt{T}\mathbb{E}[\|\mathbf{X}_t - x^\star\|^2](t+\gamma_\alpha)^{-\alpha} \\
&\leq \alpha\mathcal{E}(t)(t+\gamma_\alpha)^{-1} - 2\mu\mathcal{E}(t)(t+\gamma_\alpha)^{-\alpha} + \gamma_\alpha \mathtt{L}_\mathtt{T}(t+\gamma_\alpha)^{-\alpha} + \gamma_\alpha \mathtt{L}_\mathtt{T}\mathcal{E}(t)(t+\gamma_\alpha)^{-2\alpha} \ .
\end{aligned}$$

Hence, using Lemma 3 we get, for any $t \geq 0$, $\mathcal{E}(t) \leq B$, which concludes the proof. ■

Theorem 44 *Let $\gamma \in (0,1)$ and $\alpha \in (0,1]$. Let $(X_n)_{n \geq 0}$ be given by (1). Assume* **A**2-(b) *and* **F**1-(b)*. Then there exists $\mathtt{B}_3 > 0$ such that for all $N \geq 1$,*

$$\mathbb{E}[\|X_N - x^\star\|^2] \leq \mathtt{B}_3 N^{-\alpha} \ .$$

In the case where $\alpha = 1$ we have to assume additionally that $\gamma > 1/(2\mu)$.

Proof Let $\gamma \in (0,1)$ and $\alpha \in (0,1]$. Let $(X_n)_{n \geq 0}$ be given by (1). Using **F**1-(b) and Lemma 37 we get for all $n \geq 0$,

$$\begin{aligned}
\mathbb{E}\left[\|X_{n+1} - x^\star\|^2 \middle| \mathcal{F}_n\right] &= \mathbb{E}\left[\left\|X_n - x^\star - \gamma(n+1)^{-\alpha}\nabla\tilde{f}(X_n, Z_{n+1})\right\|^2 \middle| \mathcal{F}_n\right] \\
&= \|X_n - x^\star\|^2 + \gamma^2(n+1)^{-2\alpha}\mathbb{E}\left[\left\|\nabla\tilde{f}(X_n, Z_{n+1})\right\|^2 \middle| \mathcal{F}_n\right] \\
&\quad - 2\gamma(n+1)^{-\alpha}\mathbb{E}\left[\langle X_n - x^\star, \nabla\tilde{f}(X_n, Z_{n+1})\rangle \middle| \mathcal{F}_n\right] \\
&\leq \|X_n - x^\star\|^2 + \mathtt{L}_\mathtt{T}\gamma^2(n+1)^{-2\alpha}\left[1 + \|X_n - x^\star\|^2\right] \\
&\quad - 2\gamma(n+1)^{-\alpha}\langle X_n - x^\star, \nabla f(X_n)\rangle \ .
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\left[\|X_{n+1} - x^\star\|^2\right] \\
&\qquad \leq \mathbb{E}\left[\|X_n - x^\star\|^2\right]\left[1 - 2\gamma(n+1)^{-\alpha}\mu + \gamma^2(n+1)^{-2\alpha}\mathtt{L}_\mathtt{T}\right] + \mathtt{L}_\mathtt{T}\gamma^2(n+1)^{-2\alpha} \ . \qquad (41)
\end{aligned}$$

We note now $u_n = \mathbb{E}\left[\|X_n - x^\star\|^2\right]$ and $v_n = n^\alpha u_n$. Using (41) and Bernoulli's inequality we have, for all $n \geq 0$

$$\begin{aligned} v_{n+1} - v_n &= (n+1)^\alpha u_{n+1} - n^\alpha u_n \\ &= (n+1)^\alpha (u_{n+1} - u_n)) + u_n((n+1)^\alpha - n^\alpha) \\ &\leq \left[-2\gamma\mu + \gamma^2 \mathtt{L}_\mathtt{T}(n+1)^{-\alpha}\right] u_n + \mathtt{L}_\mathtt{T}\gamma^2(n+1)^{-\alpha} + u_n n^\alpha \left[(1+1/n)^\alpha - 1\right] \\ &\leq \left[-2\gamma\mu + \gamma^2 \mathtt{L}_\mathtt{T}(n+1)^{-\alpha} + \alpha n^{\alpha-1}\right] u_n + \mathtt{L}_\mathtt{T}\gamma^2(n+1)^{-\alpha} . \end{aligned}$$

Therefore, in the case where $\alpha < 1$, there exists $n_0 \geq 0$ such that for all $n \geq n_0$,

$$\begin{aligned} v_{n+1} - v_n &\leq -\gamma\mu u_n + \mathtt{L}_\mathtt{T}\gamma^2(n+1)^{-\alpha} \\ &\leq -\gamma\mu n^{-\alpha} v_n + \mathtt{L}_\mathtt{T}\gamma^2(n+1)^{-\alpha} \\ &\leq (n+1)^{-\alpha}(-\gamma\mu v_n + \mathtt{L}_\mathtt{T}\gamma^2) . \end{aligned}$$

And in the case where $\alpha = 1$, if $\gamma > 1/(2\mu)$ we have the existence of $n_1 \geq 0$ such that for all $n \geq n_1$,

$$v_{n+1} - v_n \leq \left[(1/2 - \gamma\mu) + \gamma^2 \mathtt{L}_\mathtt{T}(n+1)^{-\alpha} + \alpha n^{\alpha-1}\right] u_n + \mathtt{L}_\mathtt{T}\gamma^2(n+1)^{-\alpha} .$$

Using Lemma 4 this shows that, for $\alpha \in (0,1]$, there exists a constant $\mathtt{B}_3 > 0$ such that for all $n \geq 0$, $v_n \leq \mathtt{B}_3$. This proves the result. ■

D.3. Equivalent to Appendix C.3

The equivalent of Proposition 35 and Proposition 36 are given in Proposition 45 and Proposition 46 respectively.

Proposition 45 *Let $\alpha, \gamma \in (0,1)$ and $(\mathbf{X}_t)_{t\geq 0}$ be given by* (2)*. Assume* **A**1*,* **A**2-(b)*,* **A**3 *and that f verifies* (39)*. Then there exists $\mathtt{B}_5 > 0$ such that for any $T > 0$,*

$$\mathbb{E}\left[f(\mathbf{X}_T) - f^\star\right] \leq \mathtt{B}_5 T^{-\alpha} .$$

Proof Let $\alpha, \gamma \in (0,1)$ and $(\mathbf{X}_t)_{t\geq 0}$ be given by (2). Without loss of generality we can assume that $f^\star = \min_{x\in\mathbb{R}^d} f(x) = 0$. We note $\mathcal{E}(t) = (t+\gamma_\alpha)^\alpha \mathbb{E}\left[f(\mathbf{X}_t)\right]$ and we apply Lemma 48 to the stochastic process $((t+\gamma_\alpha)^\alpha f(\mathbf{X}_t))_{t\geq 0}$, and using **A**1, **A**2-(b), (39) and Lemma 42 this gives, for all $t > 0$,

$$\begin{aligned} \mathcal{E}(t) - \mathcal{E}(0) = \int_0^t \alpha(s+\gamma_\alpha)^{\alpha-1}\mathbb{E}\left[f(\mathbf{X}_s)\right] \mathrm{d}s - \int_0^t \mathbb{E}\left[\|\nabla f(\mathbf{X}_s)\|^2\right] \mathrm{d}s \\ + (\gamma_\alpha/2)\int_0^t (s+\gamma_\alpha)^{-\alpha}\mathbb{E}\left[\mathrm{Tr}(\nabla^2 f(\mathbf{X}_s)\Sigma(\mathbf{X}_s))\right] \mathrm{d}s \\ \mathrm{d}\mathcal{E}(t)/\mathrm{d}t \leq \alpha\mathcal{E}(t)(t+\gamma_\alpha)^{-1} - (1/c)\mathcal{E}(t)(t+\gamma_\alpha)^{-\alpha} + (\gamma_\alpha/2)\mathtt{L}\mathtt{L}_\mathtt{T}\left(1 + \mathbb{E}\left[f(\mathbf{X}_t)\right]\right)(t+\gamma_\alpha)^{-\alpha} \\ \leq \alpha\mathcal{E}(t)(t+\gamma_\alpha)^{-1} - (1/c)\mathcal{E}(t)(t+\gamma_\alpha)^{-\alpha} + (\gamma_\alpha/2)\mathtt{L}\mathtt{L}_\mathtt{T}\mathcal{E}(t)(t+\gamma_\alpha)^{-2\alpha} \\ + (\gamma_\alpha/2)\mathtt{L}\mathtt{L}_\mathtt{T}(t+\gamma_\alpha)^{-\alpha} . \end{aligned}$$

We can now apply Lemma 3, concluding the proof. ■

Proposition 46 *Let $\alpha \in (0,1]$ and $\gamma \in (0,1)$. Let $(X_n)_{n \geq 0}$ be given by (1). Assume* **A**1*,* **A**2*-(b) and that f verifies (39). Then there exists $\mathtt{B}_6 > 0$ such that for all $N \geq 1$,*

$$\mathbb{E}\left[f(X_N) - f^\star\right] \leq \mathtt{B}_6 N^{-\alpha} \ .$$

In the case where $\alpha = 1$ we have to assume additionally that $\gamma > 2/c$.

Proof Let $\alpha \in (0,1]$ and $\gamma \in (0,1)$. Let $(X_n)_{n \geq 0}$ be given by (1). Let $n \geq 0$. Applying the descent lemma (using **A**1) and Lemma 41 gives

$$\begin{aligned}
\mathbb{E}\left[f(X_{n+1}) | \mathcal{F}_n\right] &= \mathbb{E}\left[f(X_n - \gamma/(n+1)^\alpha \nabla \tilde{f}(X_n, Z_{n+1}) \middle| \mathcal{F}_n\right] \\
&\leq f(X_n) - \gamma/(n+1)^\alpha \mathbb{E}\left[\langle \nabla f(X_n), \nabla \tilde{f}(X_n, Z_{n+1})\rangle \middle| \mathcal{F}_n\right] \\
&\quad + \gamma^2/(n+1)^{2\alpha}(\mathtt{L}/2)\mathbb{E}\left[\left\|\nabla \tilde{f}(X_n, Z_{n+1})\right\|^2 \middle| \mathcal{F}_n\right] \\
&\leq f(X_n) - \gamma/(n+1)^\alpha \left\|\nabla f(X_n)\right\|^2 + (\mathtt{L}\gamma^2/2)(n+1)^{-2\alpha}\mathtt{L}_\mathtt{T}\left[1 + f(X_n)\right] \\
\mathbb{E}\left[f(X_{n+1})\right] - f^\star &\leq \mathbb{E}\left[f(X_n)\right] - f^\star - \gamma(n+1)^{-\alpha}\mathbb{E}\left[\left\|\nabla f(X_n)\right\|^2\right] \\
&\quad + (\mathtt{L}\mathtt{L}_\mathtt{T}\gamma^2/2)(n+1)^{-2\alpha} + (\mathtt{L}\mathtt{L}_\mathtt{T}\gamma^2/2)(n+1)^{-2\alpha}\left(\mathbb{E}\left[f(X_n)\right] - f^\star\right) \ .
\end{aligned}$$

This shows the existence of $n_2 \geq 0$ such that using (39) we have for all $n \geq n_2$,

$$\mathbb{E}\left[f(X_{n+1})\right] - f^\star \leq \left(\mathbb{E}\left[f(X_n)\right] - f^\star\right)\left[1 - (\gamma c^{-1}/2)(n+1)^{-\alpha}\right] + (\mathtt{L}\mathtt{L}_\mathtt{T}\gamma^2/2)(n+1)^{-2\alpha}\eta \ .$$

We note now for all $n \geq 0$, $u_n = \mathbb{E}\left[f(X_n)\right] - f^\star$ and $v_n = n^\alpha u_n$. We have

$$\begin{aligned}
v_{n+1} - v_n &= (n+1)^\alpha u_{n+1} - n^\alpha u_n \\
&= (n+1)^\alpha(u_{n+1} - u_n)) + u_n((n+1)^\alpha - n^\alpha) \\
&\leq -(\gamma c^{-1}/2)u_n + (\mathtt{L}\gamma^2\mathtt{L}_\mathtt{T}/2)(n+1)^{-\alpha} + u_n n^\alpha\left[(1 + 1/n)^\alpha - 1\right] \\
&\leq u_n(-(\gamma c^{-1}/2) + \alpha n^{\alpha-1}) + (\mathtt{L}\gamma^2\mathtt{L}_\mathtt{T}/2)(n+1)^{-\alpha} \ .
\end{aligned}$$

If $\alpha < 1$, or if $1 - \gamma c^{-1}/2 < 0$ we have the existence of $n_3 \geq n_2$ and $\tilde{\mathtt{B}} > 0$ such that for all $n \geq n_3$,

$$\begin{aligned}
v_{n+1} - v_n &\leq -\tilde{\mathtt{B}} u_n + (\mathtt{L}\gamma^2\mathtt{L}_\mathtt{T}/2)(n+1)^{-\alpha} \\
&\leq \left\{-\tilde{\mathtt{B}} v_n + (\mathtt{L}\gamma^2\mathtt{L}_\mathtt{T}/2)\right\}(n+1)^{-\alpha}
\end{aligned}$$

This proves the existence of $\mathtt{B}_6 > 0$ such that for all $n \geq 0$,

$$v_n \leq \mathtt{B}_6 \ ,$$

concluding the proof. ■

Appendix E. Convex case (under A2-(a))

In this section we gather our results about the long-time behavior of SGD and its continuous-time counterpart in **A**2-(a). In Appendix E.1, we derive technical results. In Appendix E.2 we provide the proof of Theorem 6 (continuous-time setting). In Appendix E.3, we give the proof of Theorem 8 (discrete-time setting).

E.1. Technical Results

Lemma 47 *Let $f \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R})$. Assume* **A**1 *and* **A**2-(a)*. Then for any $x \in \mathbb{R}^d$ we have*

$$\left|\langle \nabla^2 f(x), \Sigma(x) \rangle\right| \leq \mathtt{L}\eta \;, \qquad \left|\langle \nabla f(x) \nabla f(x)^\top, \Sigma(x) \rangle\right| \leq \eta^2 \left\|\nabla f(x)\right\|^2 \;.$$

Similarly, assume **A**1 *and* **A**2-(b)*. Then there exists $C \geq 0$ such tha for any $x \in \mathbb{R}^d$ we have*

$$\left|\langle \nabla^2 f(x), \Sigma(x) \rangle\right| C(1 + \|x\|^2) \;, \qquad \left|\langle \nabla f(x) \nabla f(x)^\top, \Sigma(x) \rangle\right| \leq C \left\|\nabla f(x)\right\|^2 (1 + \|x\|^2) \;.$$

Proof Let $x \in \mathbb{R}^d$. Using Cauchy-Schwarz's inequality, we have $\left|\langle \nabla^2 f(x), \Sigma(x) \rangle\right| \leq \|\nabla^2 f(x)\| \|\Sigma(x)\|_*$, where $\|\cdot\|$ is the operator norm and $\|\cdot\|_*$ is the nuclear norm. Using **A**1 we have $\|\nabla^2 f(x)\| \leq \mathtt{L}$ for all $x \in \mathbb{R}^d$. In addition, denoting $(\lambda_i)_{i \in \{1,\dots,d\}}$ the eigenvalues of $\Sigma(x)$, using that Σ is positive semi-definite and **A**2 we have

$$\|\Sigma(x)\|_* = \textstyle\sum_{i=1}^d |\lambda_i| = \sum_{i=1}^d \lambda_i = \mathrm{Tr}(\Sigma(x)) \leq \eta \;.$$

This concludes the first part of the proof. For the second part we have

$$\left|\langle \nabla f(x) \nabla f(x)^\top, \Sigma(x) \rangle\right| \leq \sup_{i \in \{1,\dots,d\}} \lambda_i \left\|\nabla f(x)\right\|^2 \leq \eta^2 \left\|\nabla f(x)\right\|^2 \;,$$

which concludes the first part of the proof. The last part of the proof is an immediate consequence of Lemma 15. ■

The following lemma consists into taking the expectation in Itô's formula.

Lemma 48 *Let $\alpha \in [0, 1)$ and $\gamma > 0$. Assume $f, g \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R})$,* **A**1*,* **A**2 *and* **A**3 *and let $(\mathbf{X}_t)_{t \geq 0}$ solution of* (2)*. Then for any $\varphi \in \mathrm{C}^1([0, +\infty), \mathbb{R})$, $Y \in \mathcal{F}_0$ and $\mathbb{E}\left[\|Y\|^2 + |g(Y)|\right] < +\infty$, we have the following results:*

(a) For any $t \geq 0$,

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{X}_t - Y\|^2 \varphi(t)\right] &= \mathbb{E}\left[\|\mathbf{X}_0 - Y\|^2 \varphi(0)\right] \\ &- 2 \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \mathbb{E}\left[\langle \nabla f(\mathbf{X}_s), \mathbf{X}_s - Y \rangle\right] \mathrm{d}s \\ &+ \gamma_\alpha \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_s))\right] \mathrm{d}s + \int_0^t \varphi'(s) \mathbb{E}[\|\mathbf{X}_s - Y\|^2] \mathrm{d}s \;. \end{aligned} \tag{42}$$

(b) For any $t \geq 0$

$$\begin{aligned}\mathbb{E}\left[(f(\mathbf{X}_t) - g(Y))\varphi(t)\right] &= \mathbb{E}\left[(f(\mathbf{X}_0) - g(Y))\varphi(0)\right] \\ &- \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \mathbb{E}[\|\nabla f(\mathbf{X}_s)\|^2] \mathrm{d}s \\ &+ (\gamma_\alpha/2) \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \mathbb{E}\left[\langle \nabla^2 f(\mathbf{X}_s), \Sigma(\mathbf{X}_s) \rangle\right] \mathrm{d}s \\ &+ \int_0^t \varphi'(s) \mathbb{E}\left[f(\mathbf{X}_s) - g(Y)\right] \mathrm{d}s \;. \end{aligned}$$

(c) If $\mathbb{E}[\|Y\|^{2p}] < +\infty$, then for any $t \geq 0$

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{X}_t - Y\|^{2p}\varphi(t)\right] = \mathbb{E}\left[\|\mathbf{X}_0 - Y\|^{2p}\varphi(0)\right] \\
&- 2p\int_0^t (\gamma_\alpha + s)^{-\alpha}\varphi(s)\mathbb{E}\left[\langle\nabla f(\mathbf{X}_s), \mathbf{X}_s - Y\rangle\|\mathbf{X}_t - Y\|^{2(p-1)}\right]\mathrm{d}s \\
&+ \gamma_\alpha p\int_0^t (\gamma_\alpha + s)^{-2\alpha}\varphi(s)\mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_s))\|\mathbf{X}_s - Y\|^{2(p-1)}\right]\mathrm{d}s \\
&+ \gamma_\alpha 2p(p-1)\int_0^t (\gamma_\alpha + s)^{-2\alpha}\varphi(s)\mathbb{E}\left[\langle\Sigma(\mathbf{X}_s), (\mathbf{X}_t - Y)(\mathbf{X}_t - Y)^\top\rangle\|\mathbf{X}_s - Y\|^{2(p-2)}\right]\mathrm{d}s \\
&+ \int_0^t \varphi'(s)\mathbb{E}\left[(f(\mathbf{X}_s) - g(Y))^{2p}\right]\mathrm{d}s\ .
\end{aligned}$$

(d) If $\mathbb{E}[|g(Y)|^p] < +\infty$, then for any $t \geq 0$

$$\begin{aligned}
&\mathbb{E}\left[(f(\mathbf{X}_t) - g(Y))^p\varphi(t)\right] = \mathbb{E}\left[(f(\mathbf{X}_0) - g(Y))^p\varphi(0)\right] \\
&- p\int_0^t (\gamma_\alpha + s)^{-\alpha}\varphi(s)\mathbb{E}\left[\|\nabla f(\mathbf{X}_s)\|^2(f(\mathbf{X}_s) - g(Y))^{p-1}\right]\mathrm{d}s \\
&+ \gamma_\alpha(p/2)\int_0^t (\gamma_\alpha + s)^{-2\alpha}\varphi(s)\mathbb{E}\left[\langle\nabla^2 f(\mathbf{X}_s), \Sigma(\mathbf{X}_s)\rangle(f(\mathbf{X}_s) - g(Y))^{p-2}\right]\mathrm{d}s \\
&+ \gamma_\alpha p(p-1)/2\int_0^t (\gamma_\alpha + s)^{-2\alpha}\varphi(s)\mathbb{E}\left[\langle\nabla f(\mathbf{X}_s)\nabla f(\mathbf{X}_s)^\top, \Sigma(\mathbf{X}_s)\rangle(f(\mathbf{X}_s) - g(Y))^{p-2}\right] \\
&+ \int_0^t \varphi'(s)\mathbb{E}\left[(f(\mathbf{X}_s) - g(Y))^p\right]\mathrm{d}s\ .
\end{aligned}$$

Proof Let $\alpha \in [0, 1)$, $\gamma > 0$ and $(\mathbf{X}_t)_{t\geq 0}$ the solution of (2). Note that for any $t \geq 0$, we have

$$\langle\mathbf{X}\rangle_t = \gamma_\alpha\int_0^t (\gamma_\alpha + s)^{-2\alpha}\,\mathrm{Tr}(\Sigma(\mathbf{X}_s))\mathrm{d}s\ .$$

We divide the rest of the proof into our parts.

(a) First, let $y \in \mathbb{R}^d$ and $F_y : [0, +\infty) \times \mathbb{R}^d$ such that for any $t \in [0, +\infty)$, $x \in \mathbb{R}^d$, $F_y(t, x) = \varphi(t)\|x - y\|^2$. Since $(\mathbf{X}_t)_{t\geq 0}$ is a strong solution of (2) we have that $(\mathbf{X}_t)_{t\geq 0}$ is a continuous semimartingale. Using this result, the fact that $F \in \mathrm{C}^{1,2}([0, +\infty), \mathbb{R}^d)$ and Itô's lemma (Karatzas and

Shreve, 1991, Chapter 3, Theorem 3.6) we obtain that for any $t \geq 0$ almost surely

$$\begin{aligned}
F_y(t, \mathbf{X}_t) &= F_y(0, \mathbf{X}_0) + \int_0^t \partial_1 F_y(s, \mathbf{X}_s) \mathrm{d}s + \int_0^t \langle \partial_2 F_y(s, \mathbf{X}_s), \mathrm{d}\mathbf{X}_s \rangle \\
&\quad + (1/2) \int_0^t \langle \partial_{2,2} F_y(s, \mathbf{X}_s), \mathrm{d}\langle \mathbf{X} \rangle_s \rangle \\
&= F_y(0, \mathbf{X}_0) + \int_0^t \varphi'(s) \left\| \mathbf{X}_s - y \right\|^2 \mathrm{d}s + \int_0^t \langle \partial_2 F_y(s, \mathbf{X}_s), \mathrm{d}\mathbf{X}_s \rangle \\
&\quad + (1/2) \int_0^t \langle \partial_{2,2} F_y(s, \mathbf{X}_s), \mathrm{d}\langle \mathbf{X} \rangle_s \rangle \\
&= F_y(0, \mathbf{X}_0) + \int_0^t \varphi'(s) \left\| \mathbf{X}_s - y \right\|^2 \mathrm{d}s - 2 \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \langle \nabla f(\mathbf{X}_s), \mathbf{X}_s - y \rangle \mathrm{d}s \\
&\quad + 2\gamma_\alpha^{1/2} \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \langle \mathbf{X}_s - y, \Sigma(\mathbf{X}_s)^{1/2} \mathrm{d}\mathbf{B}_s \rangle \\
&\quad + \gamma_\alpha \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \operatorname{Tr}(\Sigma(\mathbf{X}_s)) \mathrm{d}s \; . \qquad (43)
\end{aligned}$$

Using **A**1 have for any $x \in \mathbb{R}^d$,

$$|\langle \nabla f(x), x - y \rangle| \leq \|\nabla f(0)\| \, \|x - y\| + \mathtt{L} \, \|x\| \, \|x - y\| \; .$$

Therefore, using this result Lemma 17, Cauchy-Schwarz's inequality and that $\mathbb{E}[\|Y\|^2] < +\infty$, we obtain that for any $t \geq 0$ there exists $\bar{\mathtt{A}} \geq 0$ such that

$$\sup_{s \in [0,t]} \mathbb{E}[\|\mathbf{X}_s - Y\|^2] \leq \bar{\mathtt{A}} \; , \qquad \sup_{s \in [0,t]} \mathbb{E}\left[|\langle \nabla f(\mathbf{X}_s), \mathbf{X}_s - Y \rangle|\right] \leq \bar{\mathtt{A}} \; . \qquad (44)$$

In addition, we have using Lemma 15 that for any $t \geq 0$, $\mathbb{E}[|\operatorname{Tr}(\Sigma(\mathbf{X}_s))|] = \mathbb{E}[\operatorname{Tr}(\Sigma(\mathbf{X}_s))] \leq C(1 + \bar{\mathtt{A}})$ if **A**2-(b) holds or $\mathbb{E}[|\operatorname{Tr}(\Sigma(\mathbf{X}_s))|] = \mathbb{E}[\operatorname{Tr}(\Sigma(\mathbf{X}_s))] \leq \eta$ if **A**2-(a) holds. Combining these results, (44), (43), that $(\int_0^t (\gamma_\alpha + t)^{-\alpha} \varphi(t) \langle \mathbf{X}_t - Y, \Sigma(\mathbf{X}_t)^{1/2} \mathrm{d}\mathbf{B}_t \rangle)_{t \geq 0}$ is a martingale and Fubini-Lebesgue's theorem we obtain for any $t \geq 0$

$$\begin{aligned}
\mathbb{E}\left[\varphi(t) \left\| \mathbf{X}_t - Y \right\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[F_Y(t, \mathbf{X}_t) | \mathcal{F}_0\right]\right] \\
&= \mathbb{E}\left[\varphi(0) \left\| \mathbf{X}_0 - Y \right\|^2\right] + \int_0^t \varphi'(s) \mathbb{E}\left[\left\| \mathbf{X}_s - Y \right\|^2\right] \mathrm{d}s \\
&\quad - 2 \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \mathbb{E}\left[\langle \nabla f(\mathbf{X}_s), \mathbf{X}_s - Y \rangle\right] \mathrm{d}s \\
&\quad + \gamma_\alpha \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \mathbb{E}\left[\operatorname{Tr}(\Sigma(\mathbf{X}_s))\right] \mathrm{d}s \; ,
\end{aligned}$$

which concludes the proof of (42).

(b) Second, let $y \in \mathbb{R}^d$ and $F : [0, +\infty) \times \mathbb{R}^d$ such that for any $t \in [0, +\infty)$, $x \in \mathbb{R}^d$, $F_y(t, x) = \varphi(t)(f(x) - g(y))$. Using that $(\mathbf{X}_t)_{t \geq 0}$ is a continuous semi-martingale, the fact that $F \in \mathrm{C}^{1,2}([0, +\infty), \mathbb{R}^d)$

and Itô's lemma (Karatzas and Shreve, 1991, Chapter 3, Theorem 3.6) we obtain that for any $t \geq 0$ almost surely

$$\begin{aligned}
F_y(t, \mathbf{X}_t) &= F_y(0, \mathbf{X}_0) + \int_0^t \partial_1 F_y(s, \mathbf{X}_s) \mathrm{d}s + \int_0^t \langle \partial_2 F_y(s, \mathbf{X}_s), \mathrm{d}\mathbf{X}_s \rangle \\
&\quad + (1/2) \int_0^t \langle \partial_{2,2} F_y(s, \mathbf{X}_s), \mathrm{d}\langle \mathbf{X} \rangle_s \rangle \\
&= F_y(0, \mathbf{X}_0) + \int_0^t \varphi'(s)(f(\mathbf{X}_s) - g(y)) \mathrm{d}s + \int_0^t \langle \partial_2 F_y(s, \mathbf{X}_s), \mathrm{d}\mathbf{X}_s \rangle \\
&\quad + (1/2) \int_0^t \langle \partial_{2,2} F_y(s, \mathbf{X}_s), \mathrm{d}\langle \mathbf{X} \rangle_s \rangle \\
&= F_y(0, \mathbf{X}_0) + \int_0^t \varphi'(s)(f(\mathbf{X}_s) - g(y)) \mathrm{d}s - \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \left\| \nabla f(\mathbf{X}_s) \right\|^2 \mathrm{d}s \\
&\quad + \gamma_\alpha^{1/2} \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \langle \nabla f(\mathbf{X}_s), \Sigma(\mathbf{X}_s)^{1/2} \mathrm{d}\mathbf{B}_s \rangle \\
&\quad + (\gamma_\alpha/2) \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \langle \nabla^2 f(\mathbf{X}_s), \Sigma(\mathbf{X}_s) \rangle \mathrm{d}s \ .
\end{aligned}$$

Using **A**1 and that for any $a, b \geq 0$, $(a+b)^2 \leq 2(a^2 + b^2)$ we have for any $x, y \in \mathbb{R}^d$,

$$|f(x) - g(y)| \leq |f(0)| + \|\nabla f(0)\| \|x\| + (\mathtt{L}/2) \|x\|^2 + |g(y)| \ , \quad \|\nabla f(x)\|^2 \leq 2 \left\| \nabla f(0) \right\|^2 + 2\mathtt{L}^2 \left\| x \right\|^2 \ .$$

Therefore, using this result Lemma 17, Cauchy-Schwarz's inequality and that $\mathbb{E}[g(Y)^2] < +\infty$, we obtain that for any $t \geq 0$ there exists $\bar{\mathtt{A}} \geq 0$ such that

$$\sup_{s \in [0,t]} \mathbb{E}\left[|f(\mathbf{X}_s) - g(Y)| \right] \leq \bar{\mathtt{A}} \ , \quad \sup_{s \in [0,t]} \mathbb{E}[\|\nabla f(\mathbf{X}_s)\|^2] \leq \bar{\mathtt{A}} \ , \quad \sup_{s \in [0,t]} \mathbb{E}[\left| \langle \nabla^2 f(\mathbf{X}_s), \Sigma(\mathbf{X}_s) \rangle \right|] \leq \bar{\mathtt{A}} \ .$$

Combining this result, Lemma 47, the fact that $(\int_0^t \varphi(s) \langle \nabla f(\mathbf{X}_s), \Sigma(\mathbf{X}_s)^{1/2} \mathrm{d}\mathbf{B}_s \rangle)_{t \geq 0}$ is a martingale and Fubini-Lebesgue's theorem we obtain that for any $t \geq 0$

$$\begin{aligned}
\mathbb{E}\left[F_y(t, \mathbf{X}_t) \right] &= \mathbb{E}\left[\mathbb{E}\left[F_Y(t, \mathbf{X}_t) | \mathcal{F}_0 \right] \right] \\
&= \mathbb{E}\left[\varphi(0)(f(\mathbf{X}_0) - g(Y)) \right] + \int_0^t \varphi'(s) \mathbb{E}\left[(f(\mathbf{X}_s) - g(Y)) \right] \mathrm{d}s \\
&\quad - \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \mathbb{E}\left[\|\nabla f(\mathbf{X}_s)\|^2 \right] \mathrm{d}s \\
&\quad + (\gamma_\alpha/2) \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \mathbb{E}\left[\langle \nabla^2 f(\mathbf{X}_s), \Sigma(\mathbf{X}_s) \rangle \right] \mathrm{d}s \ .
\end{aligned}$$

(c) Let $y \in \mathbb{R}^d$ and $F_y : [0, +\infty) \times \mathbb{R}^d$ such that for any $t \in [0, +\infty)$, $x, y \in \mathbb{R}^d$, $F_y(t, x) = \varphi(t) \left\| x - y \right\|^{2p}$. Using that $(\mathbf{X}_t)_{t \geq 0}$ is a continuous semi-martingale, that $F_y \in \mathrm{C}^{1,2}([0, +\infty), \mathbb{R}^d)$ and Itô's lemma (Karatzas and Shreve, 1991, Chapter 3, Theorem 3.6) we obtain that for any $t \geq 0$

almost surely

$$
\begin{aligned}
F_y(t,\mathbf{X}_t) = F_y(0,\mathbf{X}_0) &+ \int_0^t \partial_1 F_y(s,\mathbf{X}_s)\mathrm{d}s + \int_0^t \langle \partial_2 F_y(s,\mathbf{X}_s), \mathrm{d}\mathbf{X}_s\rangle \\
&+ (1/2)\int_0^t \langle \partial_{2,2} F_y(s,\mathbf{X}_s), \mathrm{d}\langle \mathbf{X}\rangle_s\rangle \\
= F_y(0,\mathbf{X}_0) &+ \int_0^t \varphi'(s)\left\|\mathbf{X}_s - y\right\|^{2p}\mathrm{d}s + \int_0^t \langle \partial_2 F_y(s,\mathbf{X}_s), \mathrm{d}\mathbf{X}_s\rangle \\
&+ (1/2)\int_0^t \langle \partial_{2,2} F_y(s,\mathbf{X}_s), \mathrm{d}\langle \mathbf{X}\rangle_s\rangle \\
= F_y(0,\mathbf{X}_0) &+ \int_0^t \varphi'(s)\left\|\mathbf{X}_s - y\right\|^{2p}\mathrm{d}s \\
&- 2p\int_0^t (\gamma_\alpha + s)^{-\alpha}\varphi(s)\langle \nabla f(\mathbf{X}_s), \mathbf{X}_s - y\rangle \left\|\mathbf{X}_s) - y\right\|^{2(p-1)}\mathrm{d}s \\
&+ 2p\gamma_\alpha^{1/2}\int_0^t (\gamma_\alpha + s)^{-\alpha}\varphi(s)\langle \mathbf{X}_s - y, \Sigma(\mathbf{X}_s)^{1/2}\left\|\mathbf{X}_s - y\right\|^{2(p-1)}\mathrm{d}\mathbf{B}_s\rangle \\
&+ p\gamma_\alpha\int_0^t (\gamma_\alpha + s)^{-2\alpha}\varphi(s)\operatorname{Tr}(\Sigma(\mathbf{X}_s))\left\|\mathbf{X}_s - y\right\|^{2(p-1)}\mathrm{d}s \\
&+ 2p(p-1)\int_0^t (\gamma_\alpha + s)^{-2\alpha}\varphi(s)\langle (\mathbf{X}_s - y)(\mathbf{X}_s - y)^\top, \Sigma(\mathbf{X}_s)\rangle \left\|\mathbf{X}_s - y\right\|^{2(p-2)}\mathrm{d}s \; .
\end{aligned}
$$

Using **A**1 and that for any $a, b \geq 0$, $(a+b)^2 \leq 2(a^2+b^2)$ we have for any $x, y \in \mathbb{R}^d$, Therefore, using this result Lemma 17, Cauchy-Schwarz's inequality and that $\mathbb{E}[\|Y\|^2] < +\infty$, we obtain that for any $t \geq 0$ there exists $\bar{\mathtt{A}} \geq 0$ such that

$$
\sup_{s\in[0,t]} \mathbb{E}\left[\left\|\mathbf{X}_s - Y\right\|^{2p}\right] \leq \bar{\mathtt{A}} \; , \qquad \sup_{s\in[0,t]} \mathbb{E}\left[\left|\langle \nabla f(\mathbf{X}_s), \mathbf{X}_s - Y\rangle \left\|\mathbf{X}_s - Y\right\|^{2(p-1)}\right|\right] \leq \bar{\mathtt{A}} \; ,
$$

and

$$
\sup_{s\in[0,t]} \mathbb{E}\left[\left\|\Sigma^{1/2}(\mathbf{X}_s)\right\|\left\|\mathbf{X}_s - Y\right\|^{2p-1}\right] \leq \bar{\mathtt{A}} \; , \quad \sup_{s\in[0,t]} \mathbb{E}\left[\left\|\Sigma(\mathbf{X}_s)\right\|\left\|\mathbf{X}_s - y\right\|^{2(p-1)}\right] \leq \bar{\mathtt{A}} \; .
$$

Combining these results, Lemma 47, that $(\int_0^t \varphi(s)\langle \nabla f(\mathbf{X}_s), \Sigma(\mathbf{X}_s)^{1/2}(f(\mathbf{X}_s) - g(Y))^{p-1}\mathrm{d}\mathbf{B}_s\rangle)_{t\geq 0}$ is a martingale and Fubini-Lebesgue's theorem we obtain that for any $t \geq 0$

$$
\begin{aligned}
&\mathbb{E}\left[F_y(t,\mathbf{X}_t)\right] = \mathbb{E}\left[\mathbb{E}\left[F_Y(t,\mathbf{X}_t)|\mathcal{F}_0\right]\right] \\
&\quad = \mathbb{E}\left[\varphi(0)\left\|\mathbf{X}_0 - Y\right\|^{2p}\right] + \int_0^t \varphi'(s)\mathbb{E}\left[\left\|\mathbf{X}_s - Y\right\|^{2p}\right]\mathrm{d}s \\
&\quad - 2p\int_0^t (\gamma_\alpha + s)^{-\alpha}\varphi(s)\mathbb{E}\left[\langle \nabla f(\mathbf{X}_s), \mathbf{X}_s - y\rangle \left\|\mathbf{X}_s) - y\right\|^{2(p-1)}\right]\mathrm{d}s \\
&\quad + \gamma_\alpha p\int_0^t (\gamma_\alpha + s)^{-2\alpha}\varphi(s)\mathbb{E}\left[\operatorname{Tr}(\Sigma(\mathbf{X}_s))\left\|\mathbf{X}_s - y\right\|^{2(p-1)}\right]\mathrm{d}s \\
&\quad + 2\gamma_\alpha p(p-1)\int_0^t (\gamma_\alpha + s)^{-2\alpha}\varphi(s)\mathbb{E}\left[\langle (\mathbf{X}_s - y)\nabla(\mathbf{X}_s - y)^\top, \Sigma(\mathbf{X}_s)\rangle \left\|\mathbf{X}_s - y\right\|^{2(p-2)}\right]\mathrm{d}s \; .
\end{aligned}
$$

(d) Let $y \in \mathbb{R}^d$ and $F : [0,+\infty) \times \mathbb{R}^d$ such that for any $t \in [0,+\infty)$, $x, y \in \mathbb{R}^d$, $F_y(t,x) = \varphi(t)(f(x) - g(y))^{2p}$. Using that $(\mathbf{X}_t)_{t \geq 0}$ is a continuous semi-martingale, the fact that $F \in \mathrm{C}^{1,2}([0,+\infty), \mathbb{R}^d)$ and Itô's lemma (Karatzas and Shreve, 1991, Chapter 3, Theorem 3.6) we obtain that for any $t \geq 0$ almost surely

$$
\begin{aligned}
F_y(t,\mathbf{X}_t) &= F_y(0,\mathbf{X}_0) + \int_0^t \partial_1 F_y(s,\mathbf{X}_s)\mathrm{d}s + \int_0^t \langle \partial_2 F_y(s,\mathbf{X}_s), \mathrm{d}\mathbf{X}_s\rangle \\
&+ (1/2)\int_0^t \langle \partial_{2,2} F_y(s,\mathbf{X}_s), \mathrm{d}\langle \mathbf{X}\rangle_s\rangle \\
&= F_y(0,\mathbf{X}_0) + \int_0^t \varphi'(s)(f(\mathbf{X}_s) - g(y))^{2p}\mathrm{d}s \\
&+ \int_0^t \langle \partial_2 F_y(s,\mathbf{X}_s), \mathrm{d}\mathbf{X}_s\rangle + (1/2)\int_0^t \langle \partial_{2,2} F_y(s,\mathbf{X}_s), \mathrm{d}\langle \mathbf{X}\rangle_s\rangle \\
&= F_y(0,\mathbf{X}_0) + \int_0^t \varphi'(s)(f(\mathbf{X}_s) - g(y))^{2p}\mathrm{d}s \\
&- 2p\int_0^t (\gamma_\alpha + s)^{-\alpha}\varphi(s)\left\|\nabla f(\mathbf{X}_s)\right\|^2 (f(\mathbf{X}_s) - g(y))^{2(p-1)}\mathrm{d}s \\
&+ 2p\gamma_\alpha^{1/2}\int_0^t (\gamma_\alpha + s)^{-\alpha}\varphi(s)\langle \nabla f(\mathbf{X}_s), \Sigma(\mathbf{X}_s)^{1/2}(f(\mathbf{X}_s) - g(y))^{2(p-1)}\mathrm{d}\mathbf{B}_s\rangle \\
&+ p\gamma_\alpha\int_0^t (\gamma_\alpha + s)^{-2\alpha}\varphi(s)\langle \nabla^2 f(\mathbf{X}_s), \Sigma(\mathbf{X}_s)\rangle(f(\mathbf{X}_s) - g(y))^{2(p-1)}\mathrm{d}s \\
&+ 2p(p-1)\int_0^t (\gamma_\alpha + s)^{-2\alpha}\varphi(s)\langle \nabla f(\mathbf{X}_s)\nabla f(\mathbf{X}_s)^\top, \Sigma(\mathbf{X}_s)\rangle(f(\mathbf{X}_s) - g(y))^{2(p-2)}\mathrm{d}s
\end{aligned}
$$

Using **A**1 and that for any $a, b \geq 0$, $(a+b)^2 \leq 2(a^2 + b^2)$ we have for any $x, y \in \mathbb{R}^d$,

$$
\begin{aligned}
&|f(x) - g(y)|^{2p} \leq 4^{2p-1}|f(0)|^{2p} + 4^{2p-1}\|\nabla f(0)\|^{2p}\|x\|^{2p} + (4^{2p-1}\mathtt{L}/2)\|x\|^{4p} + 4^{2p-1}|g(y)|^{2p}\ , \\
&\|\nabla f(x)\|^2 \leq 2\left\|\nabla f(0)\right\|^2 + 2\mathtt{L}^2\left\|x\right\|^2\ .
\end{aligned}
$$

Therefore, using this result Lemma 17, Lemma 47, Hölder's inequality and that $\mathbb{E}[g(Y)^2] < +\infty$, we obtain that for any $t \geq 0$ there exists $\bar{\mathtt{A}} \geq 0$ such that

$$
\begin{aligned}
&\sup_{s\in[0,t]} \mathbb{E}\left[|f(\mathbf{X}_s) - g(Y)|^{2p}\right] \leq \bar{\mathtt{A}}\ , \qquad \sup_{s\in[0,t]} \mathbb{E}\left[\|\nabla f(\mathbf{X}_s)\|^2 |f(\mathbf{X}_s) - g(Y)|^{2(p-1)}\right] \leq \bar{\mathtt{A}}\ , \\
&\sup_{s\in[0,t]} \mathbb{E}\left[\left|\langle \nabla f(\mathbf{X}_s)\nabla f(\mathbf{X}_s)^\top, \Sigma(\mathbf{X}_s)\rangle(f(\mathbf{X}_s) - g(Y))^{2(p-2)}\right|\right] \leq \bar{\mathtt{A}}\ .
\end{aligned}
$$

Combining this result, Lemma 47, that $(\int_0^t \varphi(s)\langle \nabla f(\mathbf{X}_s), \Sigma(\mathbf{X}_s)^{1/2}(f(\mathbf{X}_s) - g(Y))^{p-1}\mathrm{d}\mathbf{B}_s\rangle)_{t\geq 0}$ is a martingale and Fubini-Lebesgue's theorem we obtain that for any $t \geq 0$

$$\begin{aligned}
&\mathbb{E}\left[F_y(t, \mathbf{X}_t)\right] = \mathbb{E}\left[\mathbb{E}\left[F_Y(t, \mathbf{X}_t) | \mathcal{F}_0\right]\right] \\
&= \mathbb{E}\left[\varphi(0)(f(\mathbf{X}_0) - g(Y))^{2p}\right] + \int_0^t \varphi'(s)\mathbb{E}\left[(f(\mathbf{X}_s) - g(Y))^{2p}\right]\mathrm{d}s \\
&\quad - 2p\int_0^t (\gamma_\alpha + s)^{-\alpha}\varphi(s)\mathbb{E}\left[\|\nabla f(\mathbf{X}_s)\|^2 (f(\mathbf{X}_s) - g(y))^{2(p-1)}\right]\mathrm{d}s \\
&\quad + \gamma_\alpha p\int_0^t (\gamma_\alpha + s)^{-2\alpha}\varphi(s)\mathbb{E}\left[\langle \nabla^2 f(\mathbf{X}_s), \Sigma(\mathbf{X}_s)\rangle (f(\mathbf{X}_s) - g(Y))^{2(p-1)}\right]\mathrm{d}s \\
&\quad + 2\gamma_\alpha p(p-1)\int_0^t (\gamma_\alpha + s)^{-2\alpha}\varphi(s)\mathbb{E}\left[\langle \nabla f(\mathbf{X}_s)\nabla f(\mathbf{X}_s)^\top, \Sigma(\mathbf{X}_s)\rangle (f(\mathbf{X}_s) - g(Y))^{2(p-2)}\right]\mathrm{d}s\ .
\end{aligned}$$

■

The following lemma is a useful tool that converts results on C^2 functions to C^1 functions.

Lemma 49 *Assume* **A**1, **F**2-(a), **A**3 *and that* $\arg\min_{x\in\mathbb{R}^d} f$ *is bounded. Then there exists* $(f_\varepsilon)_{\varepsilon>0}$ *such that for any* $\varepsilon > 0$, f_ε *is convex,* C^2 *with* $\mathtt{L}$*-Lipschitz continuous gradient. In addition, there exists* $\mathtt{C} \geq 0$ *such that the following properties are satisfied.*

(a) For all $\varepsilon > 0$, f_ε *admits a minimizer* $x_\varepsilon^\star$ *and* $\limsup_{\varepsilon\to 0} f_\varepsilon(x_\varepsilon^\star) \leq f(x^\star)$.

(b) $\liminf_{\varepsilon\to 0} \|x_\varepsilon^\star\| \leq \mathtt{C}$.

(c) for any $T \geq 0$, $\lim_{\varepsilon\to 0} \mathbb{E}\left[|f_\varepsilon(\mathbf{X}_{T,\varepsilon}) - f(\mathbf{X}_T)|\right] = 0$ *, where* $(\mathbf{X}_{t,\varepsilon})_{t\geq 0}$ *is the solution of* (2) *replacing* f *by* f_ε.

Proof Let $\varphi \in \mathrm{C}_c^\infty(\mathbb{R}^d, \mathbb{R}_+)$ be an even compactly-supported function such that $\int_{\mathbb{R}^d}\varphi(z)\mathrm{d}z = 1$. For any $\varepsilon > 0$ and $x \in \mathbb{R}^d$, let $\varphi_\varepsilon(x) = \varepsilon^{-d}\varphi(x/\varepsilon)$ and $f_\varepsilon = \varphi_\varepsilon * f$. Since $\varphi \in \mathrm{C}_c^\infty(\mathbb{R}^d, \mathbb{R}_+)$ and is compactly-supported, we have $f_\varepsilon \in \mathrm{C}^\infty(\mathbb{R}^d, \mathbb{R})$. In addition, we have for any $\varepsilon > 0$, $(\nabla f)_\varepsilon = \nabla f_\varepsilon$.

First, we show that for any ε, f_ε is convex and ∇f_ε is $\mathtt{L}$-Lipschitz continuous. Let $\varepsilon > 0$, $x, y \in \mathbb{R}^d$ and $t \in [0, 1]$. Using **F**2-(a) we have

$$\begin{aligned}
f_\varepsilon(tx + (1-t)y) &= \int_{\mathbb{R}^d} f(tx + (1-t)y - z)\varphi_\varepsilon(z)\mathrm{d}z \\
&\leq \int_{\mathbb{R}^d}\left\{tf(x-z) + (1-t)f(y-z)\right\}\varphi_\varepsilon(z)\mathrm{d}z \\
&\leq tf_\varepsilon(x) + (1-t)f_\varepsilon(y)\ .
\end{aligned}$$

Hence, f_ε is convex. In addition, using **A**1 and that $\int_{\mathbb{R}^d}\varphi_\varepsilon(z)\mathrm{d}z = 1$ we have

$$\|\nabla f_\varepsilon(x) - \nabla f_\varepsilon(y)\| \leq \int_{\mathbb{R}^d} \|\nabla f(x-z) - \nabla f(y-z)\|\,\varphi_\varepsilon(z)\mathrm{d}z \leq \mathtt{L}\,\|x - y\|\ ,$$

which proves that ∇f_ε is $\mathtt{L}$-Lipschitz continuous.

Second we show that f_ε and ∇f_ε converge uniformly towards f and ∇f. Let $\varepsilon > 0$, $x \in \mathbb{R}^d$. Using the convexity of f and that φ_ε is even, we get

$$\begin{aligned} f_\varepsilon(x) - f(x) &= \int_{\mathbb{R}^d} (f(x-z) - f(x))\varphi_\varepsilon(z)\mathrm{d}z \\ &\geq -\int_{\mathbb{R}^d} \langle \nabla f(x), z\rangle \varphi_\varepsilon(z)\mathrm{d}z \\ &\geq -\langle \nabla f(x), \int_{\mathbb{R}^d} z\varphi_\varepsilon(z)\mathrm{d}z\rangle \geq 0 \;, \end{aligned} \tag{45}$$

Conversely, using the descent lemma (Nesterov, 2004, Lemma 1.2.3) and that φ_ε is even, we have

$$\begin{aligned} f_\varepsilon(x) - f(x) &= \int_{\mathbb{R}^d} (f(x-z) - f(x))\varphi_\varepsilon(z)\mathrm{d}z \\ &\leq \int_{\mathbb{R}^d} \left(-\langle \nabla f(x), z\rangle + (\mathtt{L}/2)\left\|z\right\|^2\right)\varphi_\varepsilon(z)\mathrm{d}z \\ &\leq (\mathtt{L}/2)\int_{\mathbb{R}^d} \varepsilon^2 \left\|z/\varepsilon\right\|^2 \varepsilon^{-d}\varphi(z/\varepsilon)\mathrm{d}z \leq (\mathtt{L}/2)\varepsilon^2 \int_{\mathbb{R}^d} \left\|u\right\|^2 \varphi(u)\mathrm{d}u \;. \end{aligned} \tag{46}$$

Combining (45) and (46) we get that $\lim_{\varepsilon\to 0} \|f - f_\varepsilon\|_\infty = 0$. Using **A**1 we have for any $x \in \mathbb{R}^d$

$$\begin{aligned} \|\nabla f_\varepsilon(x) - \nabla f(x)\| &\leq \|(\nabla f)_\varepsilon(x) - \nabla f(x)\| \\ &\leq \int_{\mathbb{R}^d} \|\nabla f(x-z) - \nabla f(x)\|\varphi_\varepsilon(z)\mathrm{d}z \leq \mathtt{L}\varepsilon \int_{\mathbb{R}^d} \|z\|\varphi(z)\mathrm{d}z \;, \end{aligned}$$

Hence, we obtain that $\lim_{\varepsilon\to 0} \|\nabla f_\varepsilon - \nabla f\|_\infty = 0$. Finally, since f is coercive (Bertsekas, 1997, Proposition B.9) and $(f_\varepsilon)_{\varepsilon>0}$ converges uniformly towards f we have that for any $\varepsilon > 0$, f_ε is coercive.

We divide the rest of the proof into three parts.

(a) Let $\varepsilon > 0$. Since f_ε is coercive and continuous it admits a minimizer $x_\varepsilon^\star$. In addition, we have

$$f_\varepsilon(x_\varepsilon^\star) \leq f_\varepsilon(x^\star) \leq f(x^\star) + \|f_\varepsilon - f\|_\infty \;. \tag{47}$$

Therefore, $\limsup_{\varepsilon\to 0} f_\varepsilon(x_\varepsilon^\star) \leq f(x^\star)$.

(b) Let $\varepsilon \in (0, 1]$. Using (47), we obtain that $|f_\varepsilon(x^\star)| \leq |f(x^\star)| + \sup_{\varepsilon\in(0,1]} \|f_\varepsilon - f\|_\infty$. Since f is coercive, we obtain that $(x_\varepsilon^\star)_{\varepsilon\in(0,1]}$ is bounded and therefore there exists $\mathtt{C} \geq 0$ such that $\liminf_{\varepsilon\to 0} \|x_\varepsilon^\star\| \leq \mathtt{C}$.

(c) Let $\varepsilon > 0$, $T \geq 0$ and $(\mathbf{X}_{t,\varepsilon})_{t\geq 0}$ be the solution of (2) replacing f by f_ε. Using (2), the fact that $\lim_{\varepsilon\to 0} \|\nabla f - \nabla f_\varepsilon\|_\infty = 0$, **A**1 and Grönwall's inequality (Pachpatte, 1998, Theorem 1.2.2) we have

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{X}_{T,\varepsilon} - \mathbf{X}_T\|^2\right] &\leq \mathbb{E}\left[\left\|\int_0^T (\gamma_\alpha + s)^{-\alpha}\left\{-\nabla f_\varepsilon(\mathbf{X}_{t,\varepsilon}) + \nabla f(\mathbf{X}_t)\right\}\mathrm{d}t\right\|^2\right] \\ &\leq 2\gamma_\alpha^{-2\alpha}T\int_0^T \mathbb{E}\left[\|\nabla f(\mathbf{X}_{t,\varepsilon}) - \nabla f(\mathbf{X}_t)\|^2\right]\mathrm{d}t + 2\gamma_\alpha^{-2\alpha}T^2\|\nabla f - \nabla f_\varepsilon\|_\infty^2 \\ &\leq 2\mathtt{L}\gamma_\alpha^{-2\alpha}T\int_0^T \mathbb{E}\left[\|\mathbf{X}_{t,\varepsilon} - \mathbf{X}_t\|^2\right]\mathrm{d}t + 2\gamma_\alpha^{-2\alpha}T^2\|\nabla f - \nabla f_\varepsilon\|_\infty^2 \\ &\leq 2\gamma_\alpha^{-2\alpha}T^2\|\nabla f - \nabla f_\varepsilon\|_\infty^2 \exp\left[2\mathtt{L}\gamma_\alpha^{-2\alpha}T^2\right] \;. \end{aligned} \tag{48}$$

Therefore $\lim_{\varepsilon \to 0} \mathbb{E}\left[\|\mathbf{X}_{T,\varepsilon} - \mathbf{X}_T\|^2\right] = 0$. In addition, using the Cauchy-Schwarz inequality, **A**1 and Lemma 17 we have

$$\begin{aligned}
\mathbb{E}\left[|f(\mathbf{X}_{T,\varepsilon}) - f(\mathbf{X}_T)|\right] &\leq \mathbb{E}\left[\int_0^1 \|\nabla f(\mathbf{X}_T + t(\mathbf{X}_{T,\varepsilon} - \mathbf{X}_T))\| \|\mathbf{X}_{T,\varepsilon} - \mathbf{X}_T\| \mathrm{d}t\right] \\
&\leq \mathbb{E}\left[(\|\mathbf{X}_{T,\varepsilon}\| + \|\mathbf{X}_T\| + \|x^\star\|)\|\mathbf{X}_{T,\varepsilon} - \mathbf{X}_T\|\right] \\
&\leq 3^{1/2}\left(\|x^\star\|^2 + \mathbb{E}\left[\|\mathbf{X}_T\|^2\right] + \mathbb{E}\left[\|\mathbf{X}_{T,\varepsilon}\|^2\right]\right)^{1/2} \mathbb{E}\left[\|\mathbf{X}_{T,\varepsilon} - \mathbf{X}_T\|^2\right]^{1/2} \\
&\leq 3^{1/2}(\|x^\star\| + 2\mathtt{A}_{T,1})^{1/2}(1 + \|x_0\|^2)^{1/2}\mathbb{E}\left[\|\mathbf{X}_{T,\varepsilon} - \mathbf{X}_T\|^2\right]^{1/2} . \qquad (49)
\end{aligned}$$

Therefore, using (48), (49) and the fact that $\lim_{\varepsilon \to 0} \|f - f_\varepsilon\|_\infty = 0$ we obtain that

$$\lim_{\varepsilon \to 0} \mathbb{E}\left[|f_\varepsilon(\mathbf{X}_{T,\varepsilon}) - f(\mathbf{X}_T)|\right] \leq \lim_{\varepsilon \to 0} \mathbb{E}\left[|f(\mathbf{X}_{T,\varepsilon}) - f(\mathbf{X}_T)|\right] + \lim_{\varepsilon \to 0} \|f - f_\varepsilon\|_\infty = 0 \, ,$$

which concludes the proof.

■

Lemma 50 *Let $x, y \geq 1$. Let $\alpha \in (0, 1/2]$. If $y < x$ then $x^\alpha - y^\alpha \leq x^{1-\alpha} - y^{1-\alpha}$.*

Proof Let $\lambda \in (0,1)$ such that $y = \lambda x$. Then $x^\alpha - y^\alpha = x^\alpha(1 - \lambda^\alpha) \leq x^{1-\alpha}(1 - \lambda^{1-\alpha}) = x^{1-\alpha} - y^{1-\alpha}$ because $x > 1$, $\lambda < 1$ and $\alpha \leq 1 - \alpha$. ■

E.2. Proof of Theorem 6

In this section we prove Theorem 6. We begin with Lemma 51 which is a useful result to bound $\mathbb{E}[\|\mathbf{X}_t - x^\star\|^2]$. Then, we introduce the averaging process in (50). The study of this process is central in our proof. First we establish Lemma 52 which allows to control the time-derivative of the process S. We show that the difference $\mathbb{E}[f(\mathbf{X}_T)] - f^\star$ can be rewritten as the sum of three terms involving S. We bound each one of these three terms in Lemma 53, Lemma 54 and Lemma 55, concluding the proof of Theorem 6. We finish this section with a proof of Corollary 7 which extends our result to the case where $f \in \mathrm{C}^1(\mathbb{R}^d, \mathbb{R})$.

Lemma 51 *Assume* **F**2-(a). *Let $(\mathbf{X}_t)_{t \geq 0}$ be given by (2). Then, for any $\alpha, \gamma \in (0,1)$, there exists $\mathtt{C}_{1,\alpha}^{(c)} \geq 0$ and $\mathtt{C}_{2,\alpha}^{(c)} \geq 0$ and a function $\mathbf{\Phi}_\alpha^{(c)} : \mathbb{R}_+ \to \mathbb{R}_+$ such that, for any $t \geq 0$,*

$$\mathbb{E}[\|\mathbf{X}_t - x^\star\|^2] \leq \mathtt{C}_{1,\alpha}^{(c)} \mathbf{\Phi}_\alpha^{(c)}(t + \gamma_\alpha) + \mathtt{C}_{2,\alpha}^{(c)} \, .$$

And we have

$$\mathbf{\Phi}_\alpha^{(c)}(t) = \begin{cases} t^{1-2\alpha} & \text{if } \alpha < 1/2 \, , \\ \log(t) & \text{if } \alpha = 1/2 \, , \\ 0 & \text{if } \alpha > 1/2 \, . \end{cases}$$

The values of the constants are given by

$$\mathtt{C}_{1,\alpha}^{(c)} = \begin{cases} \gamma_\alpha \eta (1-2\alpha)^{-1} & \text{if } \alpha < 1/2 \, , \\ \gamma_\alpha \eta & \text{if } \alpha = 1/2 \, , \\ 0 & \text{if } \alpha > 1/2 \, . \end{cases}$$

$$\mathtt{C}_{2,\alpha}^{(c)} = \begin{cases} \|X_0 - x^\star\|^2 & \text{if } \alpha < 1/2 \, , \\ \|X_0 - x^\star\|^2 - \gamma_\alpha \eta \log(\gamma_\alpha) & \text{if } \alpha = 1/2 \, , \\ \|X_0 - x^\star\|^2 + (2\alpha-1)^{-1}\gamma_\alpha^{2-2\alpha}\eta & \text{if } \alpha > 1/2 \, , \end{cases}$$

Proof Let $\alpha, \gamma \in (0,1)$ and $t \geq 0$. Let $(\mathbf{X}_t)_{t\geq 0}$ be given by (2). We consider the function $F : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}_+$ defined as follows

$$\forall (t,x) \in \mathbb{R} \times \mathbb{R}^d, \; F(t,x) = \|x - x^\star\|^2 \; .$$

Applying Lemma 48 to the stochastic process $(F(t, \mathbf{X}_t))_{t\geq 0}$ and using **F**2-(a) and **A**2-(a) gives that for all $t \geq 0$,

$$\begin{aligned} &\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^2\right] - \mathbb{E}\left[\|\mathbf{X}_0 - x^\star\|^2\right] \\ &= -2\int_0^T (t+\gamma_\alpha)^{-\alpha}\langle \mathbf{X}_t - x^\star, \nabla f(\mathbf{X}_t)\rangle \mathrm{d}t + \int_0^T \gamma_\alpha (t+\gamma_\alpha)^{-2\alpha}\operatorname{Tr}(\Sigma(\mathbf{X}_t))\mathrm{d}t \\ &\leq \gamma_\alpha \eta \int_0^T (t+\gamma_\alpha)^{-2\alpha}\mathrm{d}t \; . \end{aligned}$$

We now distinguish three cases:

(a) If $\alpha < 1/2$, then we have

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^2\right] &\leq \|X_0 - x^\star\|^2 + \gamma_\alpha \eta (1-2\alpha)^{-1}((T+\gamma_\alpha)^{1-2\alpha} - \gamma_\alpha^{1-2\alpha}) \\ &\leq \|X_0 - x^\star\|^2 + \gamma_\alpha \eta (1-2\alpha)^{-1}(T+\gamma_\alpha)^{1-2\alpha} \; . \end{aligned}$$

(b) If $\alpha = 1/2$, then we have

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^2\right] &\leq \|X_0 - x^\star\|^2 + \gamma_\alpha \eta (\log(T+\gamma_\alpha) - \log(\gamma_\alpha)) \\ &\leq \gamma_\alpha \eta \log(T+\gamma_\alpha) + \|X_0 - x^\star\|^2 - \gamma_\alpha \eta \log(\gamma_\alpha) \; . \end{aligned}$$

(c) If $\alpha > 1/2$, then we have

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^2\right] &\leq \|X_0 - x^\star\|^2 + \gamma_\alpha \eta (1-2\alpha)^{-1}((T+\gamma_\alpha)^{1-2\alpha} - \gamma_\alpha^{1-2\alpha}) \\ &\leq \|X_0 - x^\star\|^2 + (2\alpha-1)^{-1}\gamma_\alpha^{2-2\alpha}\eta \; . \end{aligned}$$

■

We now turn to the proof of Theorem 6. Let $f \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R})$. Let $\gamma \in (0,1)$ and $\alpha \in (0, 1/2]$ and $T \geq 1$. Let $(\mathbf{X}_t)_{t\geq 0}$ be given by (2).

Let $S: [0,T] \to [0,+\infty)$ defined by

$$\begin{cases} S(t) = t^{-1}\int_{T-t}^{T} \{\mathbb{E}[f(\mathbf{X}_s)] - f^\star\}\,\mathrm{d}s \ , & \text{if } t>0 \ , \\ S(0) = \mathbb{E}[f(\mathbf{X}_T)] \ , & \text{otherwise.} \end{cases} \tag{50}$$

With this notation we have

$$\mathbb{E}[f(\mathbf{X}_T)] - f^\star = S(0) - S(1) + S(1) - S(T) + S(T) - f^\star \ .$$

We are now going to control each one of the three terms $(S(0)-S(1))$, $(S(1)-S(T))$, $(S(T)-f^\star)$ as follows:

(a) Case $S(1) - S(T)$ (Lemma 53): we adapt the idea of suffix averaging of Shamir and Zhang (2013) to the continuous-time setting. In particular, we control the time-derivative of S in Lemma 52.

(b) Case $S(T) - f^\star$ (Lemma 54): this result is known and corresponds to the optimal convergence rate of the averaged sequence towards the minimum of f. We provide its proof for completeness.

(c) Case $S(0) - S(1)$ (Lemma 55): this last term is specific to the continuous-time setting and is a necessary modification to the classic averaging control of $S(\varepsilon) - S(T)$, established in Lemma 53 for $\varepsilon = 1$, which diverges for ε close to 0.

Before controlling each one of these terms we state the following useful lemma, which will allow us to control the derivative of S.

Lemma 52 *Assume* **A**1*,* **A**2-(a)*,* **A**3*, and* **F**2-(a)*. Then, for any $\alpha, \gamma \in (0,1)$, $T \geq 0$, $u \in [0,T]$ and Y any $\mathbb{R}^d$-valued random variable such that $\mathbb{E}[\|Y - x^\star\|^2] \leq \mathtt{C}_{1,\alpha}^{(c)}\mathbf{\Phi}_\alpha^{(c)}(T+\gamma_\alpha) + \mathtt{C}_{2,\alpha}^{(c)}$ with $\mathtt{C}_{1,\alpha}^{(c)}$ and $\mathtt{C}_{2,\alpha}^{(c)}$ given in Lemma 51, we have*

$$\begin{aligned}\int_{T-u}^{T} \mathbb{E}[f(\mathbf{X}_t) - f(Y)]\,\mathrm{d}t \leq (\mathtt{C}_1/2)\left((T+\gamma_\alpha)^\alpha - (T-u+\gamma_\alpha)^\alpha\right) \\ + (1/2)(T-u+\gamma_\alpha)^\alpha \mathbb{E}[\|\mathbf{X}_{T-u} - Y\|^2] \\ + (\mathtt{C}_1/2)\left((T+\gamma_\alpha)^{1-\alpha} - (T-u+\gamma_\alpha)^{1-\alpha}\right)\log(T+\gamma_\alpha) \ ,\end{aligned}$$

with $\mathtt{C}_1 = \max(4\mathtt{C}_{2,\alpha}^{(c)}, (\gamma_\alpha\eta + 4\alpha\mathtt{C}_{1,\alpha}^{(c)})(1-\alpha)^{-1})$, with $\mathtt{C}_{1,\alpha}^{(c)}$ and $\mathtt{C}_{2,\alpha}^{(c)}$ given in Lemma 51.

Proof For any $x_0 \in \mathbb{R}^d$ we define the function $F_{x_0}: \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ by

$$F_{x_0}(t,x) = (t+\gamma_\alpha)^\alpha \|x - x_0\|^2 \ . \tag{51}$$

Using Lemma 51 and that for any $a,b \geq 0$, $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{X}_t - Y\|^2\right] &= \mathbb{E}\left[\|(\mathbf{X}_t - x^\star) + (x^\star - Y)\|^2\right] \\ &\leq 2\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^2\right] + 2\mathbb{E}\left[\|Y - x^\star\|^2\right] \\ &\leq 2\mathtt{C}_{1,\alpha}^{(c)}\mathbf{\Phi}_\alpha^{(c)}(t+\gamma_\alpha) + 4\mathtt{C}_{2,\alpha}^{(c)} + 2\mathtt{C}_{1,\alpha}^{(c)}\mathbf{\Phi}_\alpha^{(c)}(T+\gamma_\alpha) \\ &\leq 2\mathtt{C}_{1,\alpha}^{(c)}\mathbf{\Phi}_\alpha^{(c)}(t+\gamma_\alpha) + 2\mathtt{C}_{1,\alpha}^{(c)}\mathbf{\Phi}_\alpha^{(c)}(T+\gamma_\alpha) + \mathtt{C}_{3,\alpha}^{(c)} \ .\end{aligned}$$

with $\mathtt{C}_{3,\alpha}^{(c)} = 4\mathtt{C}_{2,\alpha}^{(c)}$. This gives in particular, for every $t \in [0, T]$,

$$(t+\gamma_\alpha)^{\alpha-1}\mathbb{E}\left[\|\mathbf{X}_t - Y\|^2\right] \leq \left\{\mathtt{C}_{3,\alpha}^{(c)} + 2\mathtt{C}_{1,\alpha}^{(c)}(T+\gamma_\alpha)^{1-2\alpha}\log(T+\gamma_\alpha)\right\}(t+\gamma_\alpha)^{\alpha-1} \quad (52)$$
$$+ 2\mathtt{C}_{1,\alpha}^{(c)}\log(T+\gamma_\alpha)(t+\gamma_\alpha)^{-\alpha} \ ,$$

with $\mathtt{C}_{1,\alpha}^{(c)} = 0$ if $\alpha > 1/2$. Notice that the additional $\log(T+\gamma_\alpha)$ term is only needed in the case where $\alpha = 1/2$. For any $(t,x) \in \mathbb{R}_+ \times \mathbb{R}^d$, we have

$$\partial_t F_{x_0}(t,x) = \alpha(t+\gamma_\alpha)^{\alpha-1}\|x - x_0\|^2 \ ,$$
$$\partial_x F_{x_0}(t,x) = 2(t+\gamma_\alpha)^{\alpha}(x - x_0) \ , \quad \partial_{xx} F_{x_0}(t,x) = 2(t+\gamma_\alpha)^{\alpha} \ .$$

Using Lemma 48 on the stochastic process $(F_Y(t,\mathbf{X}_t))_{t\geq 0}$, we have that for any $u \in [0,T]$

$$\begin{aligned}&\mathbb{E}[F_Y(T,\mathbf{X}_T)] - \mathbb{E}[F_Y(T-u,\mathbf{X}_{T-u})]\\ &\quad= \int_{T-u}^{T}\alpha(t+\gamma_\alpha)^{\alpha-1}\mathbb{E}\left[\|\mathbf{X}_t - Y\|^2\right]\mathrm{d}t - 2\int_{T-u}^{T}\mathbb{E}[\langle \mathbf{X}_t - Y, \nabla f(\mathbf{X}_t)\rangle]\,\mathrm{d}t\\ &\qquad+ \int_{T-u}^{T}\gamma_\alpha(t+\gamma_\alpha)^{-\alpha}\mathbb{E}[\mathrm{Tr}(\Sigma(\mathbf{X}_t))]\,\mathrm{d}t \ .\end{aligned} \quad (53)$$

Combining this result, **F**2-(a), **A**2-(a), (51), (52) and (53) we obtain for any $u \in [0,T]$

$$\begin{aligned}&-(T-u+\gamma_\alpha)^{\alpha}\mathbb{E}\left[\|\mathbf{X}_{T-u} - Y\|^2\right]\\ &\quad\leq \mathtt{C}_{3,\alpha}^{(c)}\int_{T-u}^{T}\alpha(t+\gamma_\alpha)^{\alpha-1}\mathrm{d}t + \eta\gamma_\alpha\int_{T-u}^{T}(t+\gamma_\alpha)^{-\alpha}\mathrm{d}t\\ &\qquad+ 2\alpha\mathtt{C}_{1,\alpha}^{(c)}\log(T+\gamma_\alpha)\left\{\int_{T-u}^{T}(t+\gamma_\alpha)^{-\alpha}\mathrm{d}t + (T+\gamma_\alpha)^{1-2\alpha}\int_{T-u}^{T}(t+\gamma_\alpha)^{\alpha-1}\mathrm{d}t\right\}\\ &\qquad- 2\int_{T-u}^{T}\mathbb{E}[f(\mathbf{X}_t) - f(Y)]\,\mathrm{d}t\\ &\quad\leq \mathtt{C}_{3,\alpha}^{(c)}\left((T+\gamma_\alpha)^{\alpha} - (T-u+\gamma_\alpha)^{\alpha}\right) - 2\int_{T-u}^{T}\mathbb{E}[f(\mathbf{X}_t) - f(Y)]\,\mathrm{d}t\\ &\qquad+ (\gamma_\alpha\eta + 2\alpha\mathtt{C}_{1,\alpha}^{(c)})(1-\alpha)^{-1}\left((T+\gamma_\alpha)^{1-\alpha} - (T-u+\gamma_\alpha)^{1-\alpha}\right)\log(T+\gamma_\alpha)\\ &\qquad+ 2\mathtt{C}_{1,\alpha}^{(c)}\log(T+\gamma_\alpha)\left\{(T+\gamma_\alpha)^{\alpha} - (T-u+\gamma_\alpha)^{\alpha}\right\}(T+\gamma_\alpha)^{1-2\alpha} \ .\end{aligned}$$

Therefore, we get for any $u \in [0,T]$

$$\begin{aligned}\int_{T-u}^{T}\mathbb{E}[f(\mathbf{X}_t) - f(Y)]\,\mathrm{d}t &\leq (\mathtt{C}_1/2)\left((T+\gamma_\alpha)^{\alpha} - (T-u+\gamma_\alpha)^{\alpha}\right)\\ &\quad+ (1/2)(T-u+\gamma_\alpha)^{\alpha}\mathbb{E}\left[\|\mathbf{X}_{T-u} - Y\|^2\right]\\ &\quad+ (\mathtt{C}_1/2)\left((T+\gamma_\alpha)^{1-\alpha} - (T-u+\gamma_\alpha)^{1-\alpha}\right)\log(T+\gamma_\alpha) \ ,\end{aligned}$$

with $\mathtt{C}_1 = \max(\mathtt{C}_{3,\alpha}^{(c)}, (\gamma_\alpha\eta + 4\alpha\mathtt{C}_{1,\alpha}^{(c)})(1-\alpha)^{-1})$. ■

Lemma 53 *Assume* **A**1*,* **A**2-(a)*,* **A**3*, and* **F**2-(a)*. Then, for any $\alpha, \gamma \in (0,1)$ and $T \geq 0$ we have*

$$S(1) - S(T) \leq 2\mathtt{C}_1 \log(T+\gamma_\alpha) \log(1+T)(T+\gamma_\alpha)^{-\min(\alpha,1-\alpha)} \; ,$$

with S given in (50)*.*

Proof In the case where $\alpha \leq 1/2$, Lemma 50 gives that for all $u \in [0,T]$:

$$((T+\gamma_\alpha)^\alpha - (T-u+\gamma_\alpha)^\alpha) \leq \left((T+\gamma_\alpha)^{1-\alpha} - (T-u+\gamma_\alpha)^{1-\alpha}\right) \; ,$$

and we also have, for all $u \in [0,T]$:

$$\begin{aligned}
&(T+\gamma_\alpha)^{1-\alpha} - (T+\gamma_\alpha-u)^{1-\alpha} \\
&\quad = \left(((T+\gamma_\alpha)^{1-\alpha} - (T+\gamma_\alpha-u)^{1-\alpha})((T+\gamma_\alpha)^\alpha + (T+\gamma_\alpha-u)^\alpha)\right) \\
&\qquad \times \left((T+\gamma_\alpha)^\alpha + (T+\gamma_\alpha-u)^{-\alpha}\right)^{-1} \\
&\quad \leq \left((T+\gamma_\alpha) - (T+\gamma_\alpha-u) + (T+\gamma_\alpha)^{1-\alpha}(T+\gamma_\alpha-u)^\alpha\right. \\
&\qquad \left. -(T+\gamma_\alpha)^\alpha(T+\gamma_\alpha-u)^{1-\alpha}\right) \times (T+\gamma_\alpha)^{-\alpha} \leq 2u/(T+\gamma_\alpha)^\alpha \; .
\end{aligned} \tag{54}$$

And in the case where $\alpha > 1/2$, for all $u \in [0,T]$:

$$\left((T+\gamma_\alpha)^{1-\alpha} - (T-u+\gamma_\alpha)^{1-\alpha}\right) \leq ((T+\gamma_\alpha)^\alpha - (T-u+\gamma_\alpha)^\alpha) \; ,$$

and we also have, for all $u \in [0,T]$:

$$\begin{aligned}
&(T+\gamma_\alpha)^\alpha - (T+\gamma_\alpha-u)^\alpha \\
&\quad = \left(((T+\gamma_\alpha)^\alpha - (T+\gamma_\alpha-u)^\alpha)((T+\gamma_\alpha)^{1-\alpha} + (T+\gamma_\alpha-u)^{1-\alpha})\right) \\
&\qquad \left((T+\gamma_\alpha)^{1-\alpha} + (T+\gamma_\alpha-u)^{1-\alpha}\right)^{-1} \\
&\quad \leq \left((T+\gamma_\alpha) - (T+\gamma_\alpha-u) + (T+\gamma_\alpha)^\alpha(T+\gamma_\alpha-u)^{1-\alpha}\right. \\
&\qquad \left. -(T+\gamma_\alpha)^{1-\alpha}(T+\gamma_\alpha-u)^\alpha\right) \times (T+\gamma_\alpha)^{-1+\alpha} \leq 2u/(T+\gamma_\alpha)^{1-\alpha} \; .
\end{aligned}$$

Now, using Lemma 52 with $Y = \mathbf{X}_{T-u}$ we obtain, for all $u \in [0,T]$:

$$\mathbb{E}\left[\int_{T-u}^T f(\mathbf{X}_t) - f(\mathbf{X}_{T-u})\mathrm{d}t\right] \leq 2\mathtt{C}_1 \log(T+\gamma_\alpha)(T+\gamma_\alpha)^{-\min(\alpha,1-\alpha)}u \; . \tag{55}$$

Since S is a differentiable function and using (55), we have for all $u \in (0,T)$,

$$S'(u) = -u^{-2}\int_{T-u}^T \mathbb{E}\left[f(\mathbf{X}_t)\right]\mathrm{d}t + u^{-1}\mathbb{E}\left[f(\mathbf{X}_{T-u})\right] = -u^{-1}(S(u) - \mathbb{E}\left[f(\mathbf{X}_{T-u})\right]) \; .$$

This last result implies $-S'(u) \leq 2\mathtt{C}_1 \log(T+\gamma_\alpha)/(T+\gamma_\alpha)^{-\min(\alpha,1-\alpha)}u^{-1}$ and integrating we get

$$S(1) - S(T) \leq 2\mathtt{C}_1 \log(T+\gamma_\alpha)\log(T)(T+\gamma_\alpha)^{-\min(\alpha,1-\alpha)} \; .$$

■

Lemma 54 *Assume* **A**1*,* **A**2-(a)*,* **A**3*, and* **F**2-(a)*. Then, for any $\alpha, \gamma \in (0,1)$ and $T \geq 0$ we have*

$$S(T) - f^\star \leq 2\mathtt{C}_1 T^{-\min(\alpha, 1-\alpha)} \log(T + \gamma_\alpha) \ ,$$

with S given in (50)*.*

Proof Using Lemma 52, with $u = T$ and $Y = x^\star$, and $\|\mathbf{X}_0 - x^\star\| \leq \mathtt{C}_1$ we obtain

$$\int_0^T \mathbb{E}\left[f(X_s)\right] \mathrm{d}s - Tf^\star \leq (\mathtt{C}_1/2)\left((T+\gamma_\alpha)^\alpha - \gamma_\alpha^\alpha + \left\{(T+\gamma_\alpha)^{1-\alpha} - \gamma\right\}\log(T+\gamma_\alpha)\right) \\ + (1/2)\gamma_\alpha \mathbb{E}\left[\|\mathbf{X}_0 - x^\star\|^2\right] \ .$$

Using this result we have

$$S(T) - f^\star \leq T^{-1}\mathtt{C}_1(T+\gamma_\alpha)^{\max(1-\alpha,\alpha)}\log(T+\gamma_\alpha) \\ + \mathtt{C}_1\gamma_\alpha T^{-1}/2 \leq 2\mathtt{C}_1 T^{-\min(\alpha,1-\alpha)}\log(T+\gamma_\alpha) \ .$$

■

Lemma 55 *Assume* **A**1*,* **A**2-(a)*,* **A**3*, and* **F**2-(a)*. Then, for any $\alpha, \gamma \in (0,1)$ and $T \geq 0$ we have*

$$S(0) - S(1) \leq \mathtt{C}_1\mathtt{L}(T-1)^{-2\alpha} \ ,$$

with S given in (50)*.*

Proof We have

$$S(0) - S(1) = \mathbb{E}\left[f(\mathbf{X}_T)\right] - S(1) = \int_{T-1}^T \left(\mathbb{E}\left[f(\mathbf{X}_T)\right] - \mathbb{E}\left[f(\mathbf{X}_s)\right]\right) \mathrm{d}s \ . \tag{56}$$

Using Lemma 48 on the stochastic process $f(\mathbf{X}_t)_{t\geq 0}$ and **A**1, we have for all $s \in [T-1, T]$

$$\mathbb{E}\left[f(\mathbf{X}_T)\right] - \mathbb{E}\left[f(\mathbf{X}_s)\right] \\ = -\int_s^T (\gamma_\alpha + t)^{-\alpha}\mathbb{E}[\|\nabla f(\mathbf{X}_t)\|^2]\mathrm{d}t + (\mathtt{L}/2)\gamma_\alpha \int_s^T (t+\gamma_\alpha)^{-2\alpha}\mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_t))\right]\mathrm{d}t \\ \leq (\eta\mathtt{L}/2)\gamma_\alpha \int_s^T (t+\gamma_\alpha)^{-2\alpha}\mathrm{d}t \leq (\mathtt{C}_1\mathtt{L}/2)(s+\gamma_\alpha)^{-2\alpha}(T-s) \ .$$

Plugging this result into (56) yields

$$S(0) - S(1) \leq (\mathtt{C}_1\mathtt{L}/2)\int_{T-1}^T (T-s)(s+\gamma_\alpha)^{-2\alpha}\mathrm{d}s \leq \mathtt{C}_1\mathtt{L}(T-1+\gamma_\alpha)^{-2\alpha} \leq \mathtt{C}_1\mathtt{L}(T-1)^{-2\alpha} \ .$$

■

We now give the extension of Theorem 6 to the case where the function f is only continuously differentiable and such that $\arg\min_{\mathbb{R}^d} f$ is bounded, see Corollary 7.

Proof Let $\alpha, \gamma \in (0,1]$ and $T \geq 0$. $(f_\varepsilon)_{\varepsilon>0}$ be given by Lemma 49. Let $\delta = \min(\alpha, 1-\alpha)$. We can apply, Theorem 6 to f_ε for each $\varepsilon > 0$. Therefore there exists $\mathtt{C}_\varepsilon^{(c)}$ such that

$$\mathbb{E}\left[f(\mathbf{X}_{T,\varepsilon})\right] - f(x_\varepsilon^\star) \leq \mathtt{C}_\varepsilon^{(c)} \left[\log(T)^2 T^{-\delta} + \log(T) T^{-\delta} + T^{-\delta} + (T-1)^{-2\alpha}\right] , \tag{57}$$

where $(\mathbf{X}_{t,\varepsilon})_{t\geq 0}$ is given by (2) with $\mathbf{X}_t = x_0$ (upon replacing f by f_ε) and

$$\mathtt{C}_\varepsilon^{(c)} = 4\max(2\mathtt{C}_{2,\alpha}^{(c)} + 2\left\|x_0 - x_\varepsilon^\star\right\|^2, (\gamma_\alpha \eta + 2\alpha \mathtt{C}_{1,\alpha}^{(c)})(1-\alpha)^{-1}) .$$

Using (57) and Lemma 49 we have

$$\begin{aligned}
\mathbb{E}\left[f(\mathbf{X}_T)\right] - f^\star &\leq \liminf_{\varepsilon\to 0} \mathbb{E}\left[f_\varepsilon(\mathbf{X}_{t,\varepsilon})\right] - \limsup_{\varepsilon\to 0} f_\varepsilon(x_\varepsilon^\star) \\
&\leq \liminf_{\varepsilon\to 0} \left\{\mathbb{E}\left[f_\varepsilon(\mathbf{X}_{t,\varepsilon})\right] - f_\varepsilon(x_\varepsilon^\star)\right\} \\
&\leq \liminf_{\varepsilon\to 0} \mathtt{C}_\varepsilon^{(c)} \left[\log(T)^2 T^{-\delta} + \log(T) T^{-\delta} + T^{-\delta} + (T-1)^{-2\alpha}\right] \\
&\leq \mathtt{C}_1^{(c)} \left[\log(T)^2 T^{-\delta} + \log(T) T^{-\delta} + T^{-\delta} + (T-1)^{-2\alpha}\right] ,
\end{aligned}$$

with $\mathtt{C}_1^{(c)} = 3\max(2\mathtt{C}_{2,\alpha}^{(c)} + 4\left\|x_0\right\|^2 + 4C^2, (\gamma_\alpha \eta + 2\mathtt{C}_{1,\alpha}^{(c)})(1-\alpha)^{-1})$, where $C = \max_{y \in \arg\min_{\mathbb{R}^d} f} \|y\|$.
■

E.3. Proof of Theorem 8

In this section we prove Theorem 8. The proof is clearly more involved than the one of Theorem 6. We will follow a similar way as in the proof of Theorem 6, with more technicalities. Again, one of the main argument of the proof is the suffix averaging technique that was introduced in (Shamir and Zhang, 2013). We begin by the discrete counterpart of Lemma 51 in Lemma 56. Proposition 57 is a first step towards proving Theorem 8. It provides suboptimal bounds for $\mathbb{E}[f(X_n)] - f^\star$. In order to prove this proposition, as in the continuous-time case, we introduce the averaged process in (60). First, we control its derivative in Lemma 58 (which is the discrete-time counterpart of Lemma 52). Then, we rewrite $\mathbb{E}[f(X_n)] - f^\star$ as a sum of two terms involving S, which we bound in Lemma 59 (discrete counterpart of Lemma 53 and Lemma 55) and Lemma 60 (discrete counterpart of Lemma 54). This concludes the proof of Theorem 8 using our original bootstrapping technique. Finally, we conclude this section with an extension of our result to the case where ∇f is bounded and no longer Lipschitz continuous in Corollary 61.

Lemma 56 *Assume* **A**1, **F**2-(a), **A**2-(a). *Then for any* $\alpha, \gamma \in (0,1)$, *there exists* $\mathtt{C}_{1,\alpha}^{(d)} \geq 0$, $\mathtt{C}_{2,\alpha}^{(d)} \geq 0$ *and a function* $\mathbf{\Phi}_\alpha^{(d)} : \mathbb{R}_+ \to \mathbb{R}_+$ *such that, for any* $n \geq 0$,

$$\mathbb{E}\left[\left\|X_n - x^\star\right\|^2\right] \leq \mathtt{C}_{1,\alpha}^{(d)} \mathbf{\Phi}_\alpha^{(d)}(n+1) + \mathtt{C}_{2,\alpha}^{(d)} .$$

And we have

$$\mathbf{\Phi}_\alpha^{(d)}(t) = \begin{cases} t^{1-2\alpha} & \textit{if } \alpha < 1/2 , \\ \log(t) & \textit{if } \alpha = 1/2 , \\ 0 & \textit{if } \alpha > 1/2 . \end{cases}$$

The values of the constants are given by

$$\mathtt{C}_{1,\alpha}^{(d)} = \begin{cases} 2\gamma^2\eta(1-2\alpha)^{-1} & \text{if } \alpha < 1/2 \ , \\ \gamma^2\eta & \text{if } \alpha = 1/2 \ , \\ 0 & \text{if } \alpha > 1/2 \ . \end{cases}$$

$$\mathtt{C}_{2,\alpha}^{(d)} = \begin{cases} 2\max_{k \leq (\gamma \mathtt{L}/2)^{1/\alpha}} \mathbb{E}\left[\|X_k - x^\star\|^2\right] & \text{if } \alpha < 1/2 \ , \\ 2\max_{k \leq (\gamma \mathtt{L}/2)^{1/\alpha}} \mathbb{E}\left[\|X_k - x^\star\|^2\right] + 2\gamma^2\eta & \text{if } \alpha = 1/2 \ , \\ 2\max_{k \leq (\gamma \mathtt{L}/2)^{1/\alpha}} \mathbb{E}\left[\|X_k - x^\star\|^2\right] + \gamma^2\eta(2\alpha-1)^{-1} & \text{if } \alpha > 1/2 \ , \end{cases}$$

Proof Let $f : \mathbb{R}^d \to \mathbb{R}$ verifying assumptions **A**1 and **F**2-(a). We consider $(X_n)_{n\geq 0}$ satisfying (1). Let $x^\star \in \mathbb{R}^d$ be given by **F**2-(a). We have, using (1) and **A**2-(a) that for all $n \geq (\gamma \mathtt{L}/2)^{1/\alpha}$,

$$\begin{aligned} \mathbb{E}\left[\|X_{n+1} - x^\star\|^2 \middle| \mathcal{F}_n\right] &= \mathbb{E}\left[\|X_n - x^\star - \gamma(n+1)^{-\alpha} H(X_n, Z_{n+1})\|^2 \middle| \mathcal{F}_n\right] \\ &= \|X_n - x^\star\|^2 - 2\gamma/(n+1)^\alpha \langle X_n - x^\star, \mathbb{E}\left[H(X_n, Z_{n+1})|\mathcal{F}_n\right]\rangle \\ &\quad + \gamma^2(n+1)^{-2\alpha}\mathbb{E}\left[\|H(X_n, Z_{n+1})\|^2 \middle| \mathcal{F}_n\right] \\ &= \|X_n - x^\star\|^2 - 2\gamma/(n+1)^\alpha \langle X_n - x^\star, \nabla f(X_n)\rangle \\ &\quad + \gamma^2(n+1)^{-2\alpha}\mathbb{E}\left[\|H(X_n, Z_{n+1}) - \nabla f(X_n) + \nabla f(X_n)\|^2 \middle| \mathcal{F}_n\right] \\ &= \|X_n - x^\star\|^2 - 2\gamma/(n+1)^\alpha \langle X_n - x^\star, \nabla f(X_n)\rangle \\ &\quad + \gamma^2(n+1)^{-2\alpha}\mathbb{E}\left[\|H(X_n, Z_{n+1}) - \nabla f(X_n)\|^2 \middle| \mathcal{F}_n\right] \\ &\quad + \gamma^2(n+1)^{-2\alpha}\Big(\mathbb{E}\left[\|\nabla f(X_n)\|^2 \middle| \mathcal{F}_n\right] \\ &\qquad + 2\mathbb{E}\left[\langle H(X_n, Z_{n+1}) - \nabla f(X_n), \nabla f(X_n)\rangle|\mathcal{F}_n\right]\Big) \\ &= \|X_n - x^\star\|^2 - 2\gamma/(n+1)^\alpha \langle X_n - x^\star, \nabla f(X_n)\rangle + \gamma^2\eta(n+1)^{-2\alpha} \\ &\quad + \gamma^2(n+1)^{-2\alpha}\left\|\nabla f(X_n)\right\|^2 \\ &\leq \|X_n - x^\star\|^2 - 2\gamma/\mathtt{L}(n+1)^{-\alpha}\left\|\nabla f(X_n)\right\|^2 + \gamma^2\eta(n+1)^{-2\alpha} \\ &\quad + \gamma^2(n+1)^{-2\alpha}\left\|\nabla f(X_n)\right\|^2 \\ &\leq \|X_n - x^\star\|^2 + \gamma/(n+1)^\alpha \left\|\nabla f(X_n)\right\|^2 \left[\gamma/(n+1)^\alpha - 2/\mathtt{L}\right] + \gamma^2\eta(n+1)^{-2\alpha} \\ &\leq \|X_n - x^\star\|^2 + \gamma^2\eta(n+1)^{-2\alpha} \\ &\leq \mathbb{E}\left[\|X_n - x^\star\|^2\right] + \gamma^2\eta(n+1)^{-2\alpha} \ , \end{aligned} \tag{58}$$

where we used the co-coercivity of f. Summing the previous inequality leads to

$$\mathbb{E}\left[\|X_n - x^\star\|^2\right] - \mathbb{E}\left[\|X_0 - x^\star\|^2\right] \leq \gamma^2\eta\sum_{k=1}^{n} k^{-2\alpha} \ .$$

As in the previous proof we now distinguish three cases:

(a) If $\alpha < 1/2$, we have

$$\mathbb{E}\left[\|X_n - x^\star\|^2\right] \le \|X_0 - x^\star\|^2 + \gamma^2\eta(1-2\alpha)^{-1}(n+1)^{1-2\alpha}$$
$$\le \|X_0 - x^\star\|^2 + 2\gamma^2\eta(1-2\alpha)^{-1}n^{1-2\alpha} \; .$$

(b) If $\alpha = 1/2$, we have

$$\mathbb{E}\left[\|X_n - x^\star\|^2\right] \le \|X_0 - x^\star\|^2 + \gamma^2\eta(\log(n) + 2) \; .$$

(c) If $\alpha > 1/2$, we have

$$\mathbb{E}\left[\|X_n - x^\star\|^2\right] \le \|X_0 - x^\star\|^2 + \gamma^2\eta(2\alpha - 1)^{-1} \; .$$

■

We now turn to the proof of Theorem 8 by stating an intermediate result where we assume a condition bounding $\mathbb{E}[\|\nabla f(X_n)\|^2]$. This proposition provides non-optimal convergence rates for SGD but will be used as a central tool to improve them via a bootstrapping technique and obtain optimal convergence rates.

Proposition 57 *Let $\gamma, \alpha \in (0,1)$ and $x_0 \in \mathbb{R}^d$ and $(X_n)_{n\ge 0}$ be given by (1). Assume* **A**1, **F**2-(a), **A**2-(a). *Suppose additionally that there exists $\alpha^\star \in [0, 1/2]$, $\beta > 0$ and $\mathtt{C}_0 \ge 0$ such that for all $n \in \mathbb{N}$*

$$\mathbb{E}[\|\nabla f(X_n)\|^2] \le \begin{cases} \mathtt{C}_0 (n+1)^\beta \log(n+1) & \text{if } \alpha \le \alpha^\star \; , \\ \mathtt{C}_0 & \text{if } \alpha > \alpha^\star \; . \end{cases} \tag{59}$$

Then there exists $\tilde{\mathtt{C}}_\alpha \ge 0$ such that, for all $N \ge 1$,

$$\mathbb{E}\left[f(X_N)\right] - f^\star \le \tilde{\mathtt{C}}_\alpha \left\{(1 + \log(N+1))^2/(N+1)^{\min(\alpha, 1-\alpha)}\boldsymbol{\Psi}_\alpha(N+1) + 1/(N+1)\right\} \; ,$$

where for any $n \in \mathbb{N}$

$$\boldsymbol{\Psi}_\alpha(n) = \begin{cases} n^\beta(1 + \log(n)) & \text{if } \alpha \le \alpha^\star \; , \\ 1 & \text{if } \alpha > \alpha^\star \; . \end{cases}$$

Proof Let $\alpha, \gamma \in (0,1)$ and $N \ge 1$. Let $(X_n)_{n\ge 0}$ be given by (1). The proof is a straightforward application of Lemma 59 and Lemma 60 below with $\tilde{\mathtt{C}}_\alpha = 2\max((2\gamma)^{-1}\|X_0 - x^\star\|^2, 2\mathtt{C}^{(d)})$. ■

Let $(S_k)_{k\in\{0,\cdots,N\}}$ be given for any $k \in \mathbb{N}$ by

$$S_k = (k+1)^{-1}\sum_{t=N-k}^{N}\mathbb{E}\left[f(X_t)\right] \; . \tag{60}$$

Note that $\mathbb{E}[f(X_N)] - f^\star = (S_N - S_0) + (S_0 - f^\star)$. We are now going to control each one of the two terms $(S_0 - S_N)$ and $(S_N - f^\star)$ as follows:

(a) Case $S_n - S_0$ (Lemma 59): this is an adaption of the idea of suffix averaging of Shamir and Zhang (2013) to our setting (one of crucial difference lies into the control of the sequence $(\mathbb{E}[\nabla f(X_n)]^2)_{n \in \mathbb{N}}$ which is assumed to be uniformly bounded in Shamir and Zhang (2013)). In particular, we control the (discrete) time-derivative of S in Lemma 58.

(b) Case $S(T) - f^\star$ (Lemma 60): this result is known and corresponds to the optimal convergence rate of the averaged sequence towards the minimum of f. We provide its proof for completeness.

Before controlling each one of these terms we state the following useful lemma, which will allow us to control the derivative of S.

Lemma 58 *Assume* **A**1*,* **A**2-(a)*, and* **F**2-(a)*. In addition, assume that* (59) *holds. Then, for any* $\alpha, \gamma \in (0,1)$*,* $N \in \mathbb{N}$*,* $u \in \{0, \dots, N\}$ *and* Y *any* $\mathbb{R}^d$*-valued random variable such that* $\mathbb{E}[\|Y - x^\star\|^2] \leq \mathtt{C}_{1,\alpha}^{(d)}(N+1)^{1-2\alpha}\log(N+1) + 4\mathtt{C}_{2,\alpha}^{(d)}$ *with* $\mathtt{C}_{1,\alpha}^{(d)}$ *and* $\mathtt{C}_{2,\alpha}^{(d)}$ *given in Lemma 56, we have*

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=N-u}^{N} f(X_k) - f(X_{N-u})\right] \\
&\qquad \leq 2\mathtt{C}^{(d)}(u+1)/(N+1)^{\min(\alpha,1-\alpha)}(1+\log(N+1))\boldsymbol{\Psi}_\alpha(N+1) \\
&\qquad\quad + (2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}[\|X_{N-u} - Y\|^2] \, ,
\end{aligned}$$

with $\mathtt{C}^{(d)} = 4(\gamma/2) + (2\gamma)^{-1}(4\mathtt{C}_{2,\alpha}^{(d)} + 4\mathtt{C}_{1,\alpha}^{(d)})$ *with* $\mathtt{C}_{1,\alpha}^{(d)}$ *and* $\mathtt{C}_{1,\alpha}^{(d)}$ *given in Lemma 56 and*

$$\boldsymbol{\Psi}_\alpha(n) = \begin{cases} n^\beta(1+\log(n)) & \text{if } \alpha \leq \alpha^\star \, , \\ 1 & \text{if } \alpha > \alpha^\star \, . \end{cases}$$

Proof Let $\ell \in \{0, \cdots, N\}$, let $k \geq \ell$, let $Y \in \mathcal{F}_\ell$. Using **F**2-(a) we have

$$\begin{aligned}
\mathbb{E}\left[\|X_{k+1} - Y\|^2 \middle| \mathcal{F}_k\right] &= \mathbb{E}\left[\left\|X_k - Y - \gamma(k+1)^{-\alpha}H(X_k, Z_{k+1})\right\|^2 \middle| \mathcal{F}_k\right] \\
&= \|X_k - Y\|^2 + \gamma^2(k+1)^{-2\alpha}\mathbb{E}\left[\|H(X_k, Z_{k+1})\|^2 \middle| \mathcal{F}_k\right] \\
&\quad - 2\gamma(k+1)^{-\alpha}\langle X_k - Y, \nabla f(X_k)\rangle \\
\mathbb{E}\left[f(X_k) - f(Y)\right] &\leq (2\gamma)^{-1}(k+1)^{\alpha}\left(\mathbb{E}\left[\|X_k - Y\|^2\right] - \mathbb{E}\left[\|X_{k+1} - Y\|^2\right]\right) \\
&\quad + (\gamma/2)(k+1)^{-\alpha}\mathbb{E}\left[\mathbb{E}\left[\|H(X_k, Z_{k+1})\|^2 \middle| \mathcal{F}_k\right]\right] \\
\mathbb{E}\left[f(X_k) - f(Y)\right] &\leq (2\gamma)^{-1}(k+1)^{\alpha}\left(\mathbb{E}\left[\|X_k - Y\|^2\right] - \mathbb{E}\left[\|X_{k+1} - Y\|^2\right]\right) \\
&\quad + (\gamma/2)(k+1)^{-\alpha}\left(\eta + \mathbb{E}\left[\|\nabla f(X_k)\|^2\right]\right) \, . \qquad (61)
\end{aligned}$$

Let $u \in \{0, \cdots, N\}$. Summing now (61) between $k = N - u$ and $k = N$ gives

$$\begin{aligned} \mathbb{E}\left[\sum_{k=N-u}^{N} f(X_k) - f(Y)\right] \leq (\gamma\eta/2) \sum_{k=N-u}^{N} (k+1)^{-\alpha} \\ + (2\gamma)^{-1} \sum_{k=N-u+1}^{N} \mathbb{E}\left[\|X_k - Y\|^2\right] ((k+1)^{\alpha} - k^{\alpha}) \\ + (\gamma/2) \sum_{k=N-u}^{N} \mathbb{E}\left[\|\nabla f(X_k)\|^2\right] (k+1)^{-\alpha} \\ + (2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] . \end{aligned} \tag{62}$$

We now have to conduct separate analyses depending on the value of α.

(a) First assume that $\alpha \leq \alpha^\star$. In that case (59) gives that

$$\mathbb{E}\left[\|\nabla f(X_k)\|^2\right] \leq \mathtt{C}_0(N+1)^{\beta}\log(N+1),$$

and Lemma 56 gives that for all $k \in \{0, \ldots, N\}$,

$$\begin{aligned} \mathbb{E}\left[\|X_k - Y\|^2\right] &\leq 2\mathbb{E}\left[\|X_k - x^\star\|^2\right] + 2\mathbb{E}\left[\|Y - x^\star\|^2\right] \\ &\leq 2\mathtt{C}_{1,\alpha}^{(d)}(k+1)^{1-2\alpha}\log(k+1) + 2\mathtt{C}_{1,\alpha}^{(d)}(N+1)^{1-2\alpha}\log(N+1) + 4\mathtt{C}_{2,\alpha}^{(d)} \\ &\leq 4\mathtt{C}_{1,\alpha}^{(d)}(N+1)^{1-2\alpha}\log(N+1) + 4\mathtt{C}_{2,\alpha}^{(d)} . \end{aligned}$$

We note $\mathtt{C}_{3,\alpha}^{(d)} = 4\mathtt{C}_{2,\alpha}^{(d)}$. Combining (62) and $\mathtt{C}^{(b)} = ((\gamma\eta/2) + (\gamma/2)\mathtt{C}_0)(1-\alpha)^{-1}$ we get that

$$\begin{aligned} &\mathbb{E}\left[\sum_{k=N-u}^{N} f(X_k) - f(Y)\right] \leq (\gamma\eta/2)(1-\alpha)^{-1}\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right) \\ &+ (2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] \\ &+ (2\gamma)^{-1}\left(\mathtt{C}_{3,\alpha}^{(d)} + 4\mathtt{C}_{1,\alpha}^{(d)}(N+1)^{1-2\alpha}\log(N+1)\right)((N+1)^{\alpha} - (N-u+1)^{\alpha}) \\ &+ (\gamma/2)\mathtt{C}_0(N+1)^{\beta}\log(N+1)(1-\alpha)^{-1}\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right) \\ &\leq \mathtt{C}^{(b)}(N+1)^{\beta}(1+\log(N+1))^2\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right) \\ &+ (2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] \\ &+ (2\gamma)^{-1}\mathtt{C}_{3,\alpha}^{(d)}((N+1)^{\alpha} - (N-u)^{\alpha}) \\ &+ (2\gamma)^{-1}4\mathtt{C}_{1,\alpha}^{(d)}\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right) \\ &\leq \mathtt{C}^{(d)}(N+1)^{\beta}(1+\log(N+1))^2\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right) \\ &+ (2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] , \end{aligned}$$

where we used Lemma 50.

Notice now that, similarly to (54) we have

$$\begin{aligned}&(N+1)^{1-\alpha}-(N-u)^{1-\alpha}\\&\qquad=\left\{\left((N+1)^{1-\alpha}-(N-u)^{1-\alpha}\right)\left((N+1)^{\alpha}+(N-u)^{\alpha}\right)\right\}\left((N+1)^{\alpha}+(N-u)^{\alpha}\right)^{-1}\\&\qquad\leq 2(u+1)/(N+1)^{\alpha}\ .\end{aligned}$$

(b) Second, assume that $\alpha\in(\alpha^\star,1/2]$. Using Lemma 56, we have for all $k\in\{0,\ldots,N\}$,

$$\begin{aligned}\mathbb{E}\left[\|X_k-Y\|^2\right]&\leq 2\mathbb{E}\left[\|X_k-x^\star\|^2\right]+2\mathbb{E}\left[\|Y-x^\star\|^2\right]\\&\leq 2\mathtt{C}^{(d)}_{1,\alpha}(k+1)^{1-2\alpha}\log(k+1)+2\mathtt{C}^{(d)}_{1,\alpha}(N+1)^{1-2\alpha}\log(N+1)+4\mathtt{C}^{(d)}_{2,\alpha}\\&\leq 4\mathtt{C}^{(d)}_{1,\alpha}(N+1)^{1-2\alpha}\log(N+1)+4\mathtt{C}^{(d)}_{2,\alpha}\ .\end{aligned}$$

Using (59), (62) rewrites

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=N-u}^{N}f(X_k)-f(Y)\right]\leq(\gamma\eta/2)(1-\alpha)^{-1}\left((N+1)^{1-\alpha}-(N-u)^{1-\alpha}\right)\\&\quad+(2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u}-Y\|^2\right]\\&\quad+(2\gamma)^{-1}\left(\mathtt{C}^{(d)}_{3,\alpha}+4\mathtt{C}^{(d)}_{1,\alpha}\log(N+1)(N+1)^{1-2\alpha}\right)\left((N+1)^{\alpha}-(N-u+1)^{\alpha}\right)\\&\quad+(\gamma/2)\mathtt{C}_0(1-\alpha)^{-1}\left((N+1)^{1-\alpha}-(N-u)^{1-\alpha}\right)\\&\leq\mathtt{C}^{(b)}\left((N+1)^{1-\alpha}-(N-u)^{1-\alpha}\right)+(2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u}-Y\|^2\right]\\&\quad+(2\gamma)^{-1}\left(\mathtt{C}^{(d)}_{3,\alpha}+4\mathtt{C}^{(d)}_{1,\alpha}\right)(1+\log(N+1))\left((N+1)^{\alpha}-(N-u)^{\alpha}\right)\\&\leq\mathtt{C}^{(d)}(1+\log(N+1))\left((N+1)^{1-\alpha}-(N-u)^{1-\alpha}\right)\\&\quad+(2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u}-Y\|^2\right]\ .\end{aligned}$$

(c) Finally, assume that $\alpha>1/2$. In that case, $\alpha>\alpha^\star$ and Lemma 56 gives

$$\forall k\in\{0,\ldots,N\}\,,\ \mathbb{E}\left[\|X_k-Y\|^2\right]\leq 2\mathbb{E}\left[\|X_k-x^\star\|^2\right]+2\mathbb{E}\left[\|Y-x^\star\|^2\right]\leq 4\mathtt{C}^{(d)}_{2,\alpha}=\mathtt{C}^{(d)}_{3,\alpha}\ .$$

Using Lemma 50 and (59) we rewrite (62) as

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=N-u}^{N}f(X_k)-f(Y)\right]\leq((\gamma\eta/2)+\gamma\mathtt{C}_0/2)(1-\alpha)^{-1}\left((N+1)^{1-\alpha}-(N-u)^{1-\alpha}\right)\\&\quad+(2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u}-Y\|^2\right]\\&\quad+(2\gamma)^{-1}\mathtt{C}^{(d)}_{3,\alpha}\left((N+1)^{\alpha}-(N-u+1)^{\alpha}\right)\\&\leq\mathtt{C}^{(b)}\left((N+1)^{1-\alpha}-(N-u)^{1-\alpha}\right)+(2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u}-Y\|^2\right]\\&\quad+(2\gamma)^{-1}\mathtt{C}^{(d)}_{3,\alpha}\left((N+1)^{\alpha}-(N-u)^{\alpha}\right)\\&\leq\mathtt{C}^{(d)}\left((N+1)^{\alpha}-(N-u)^{\alpha}\right)+(2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u}-Y\|^2\right]\ .\end{aligned}$$

Notice now that, similarly to (54) we have

$$\begin{aligned}&(N+1)^{\alpha}-(N-u)^{\alpha}\\&\quad=\left\{\left((N+1)^{\alpha}-(N-u)^{\alpha}\right)\left((N+1)^{1-\alpha}+(N-u)^{1-\alpha}\right)\right\}\\&\qquad\times\left((N+1)^{1-\alpha}+(N-u)^{1-\alpha}\right)^{-1}\leq 2(u+1)/(N+1)^{1-\alpha}\ .\end{aligned}$$

Finally, putting the three cases above together we obtain

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=N-u}^{N} f(X_k)-f(X_{N-u})\right]\\&\quad\leq 2\mathtt{C}^{(d)}(u+1)/(N+1)^{\min(\alpha,1-\alpha)}(1+\log(N+1))\boldsymbol{\Psi}_{\alpha}(N+1)\\&\qquad+(2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u}-Y\|^2\right]\ ,\end{aligned}$$

with

$$\boldsymbol{\Psi}_{\alpha}(n)=\begin{cases}n^{\beta}(1+\log(n)) & \text{if } \alpha\leq\alpha^{\star}\ ,\\ 1 & \text{if } \alpha>\alpha^{\star}\ .\end{cases}$$

Note that the additional $\log(N+1)$ factor can be removed if $\alpha\neq 1/2$. ■

Lemma 59 *Assume* **A**1, **A**2-(a) *and* **F**2-(a). *In addition, assume that* (59) *holds. Then, for any* $\alpha,\gamma\in(0,1)$ *and* $N\in\mathbb{N}$ *we have*

$$S_0-S_N\leq 2\mathtt{C}^{(d)}(N+1)^{-\min(\alpha,1-\alpha)}(1+\log(N+1))^2\boldsymbol{\Psi}_{\alpha}(N+1)\ .$$

with S *given in* (60).

Proof Let $u\in\{0,\ldots,N\}$. Using Lemma 58 with the choice $Y=X_{N-u}$ gives

$$\mathbb{E}\left[\sum_{k=N-u}^{N} f(X_k)-f(X_{N-u})\right]\leq 2\mathtt{C}^{(d)}(u+1)/(N+1)^{\min(\alpha,1-\alpha)}(1+\log(N+1))\boldsymbol{\Psi}_{\alpha}(N+1)\ .$$

And then,

$$\begin{aligned}S_u&=(u+1)^{-1}\sum_{k=N-u}^{N}\mathbb{E}\left[f(X_k)\right]\\&\leq 2\mathtt{C}^{(d)}(N+1)^{-\min(\alpha,1-\alpha)}(1+\log(N+1))\boldsymbol{\Psi}_{\alpha}(N+1)+\mathbb{E}\left[f(X_{N-u})\right]\ .\end{aligned}\tag{63}$$

We have now, using (63),

$$\begin{aligned}uS_{u-1}&=(u+1)S_u-\mathbb{E}\left[f(X_{N-u})\right]\\&=uS_u+S_u-\mathbb{E}\left[f(X_{N-u})\right]\\&\leq uS_u+2\mathtt{C}^{(d)}(N+1)^{-\min(\alpha,1-\alpha)}(1+\log(N+1))\boldsymbol{\Psi}_{\alpha}(N+1)\\S_{u-1}-S_u&\leq 2\mathtt{C}^{(d)}u^{-1}(N+1)^{-\min(\alpha,1-\alpha)}\log(N+1)\\S_0-S_N&\leq 2\mathtt{C}^{(d)}(N+1)^{-\min(\alpha,1-\alpha)}(1+\log(N+1))\boldsymbol{\Psi}_{\alpha}(N+1)\sum_{u=1}^{N}(1/u)\\S_0-S_N&\leq 2\mathtt{C}^{(d)}(N+1)^{-\min(\alpha,1-\alpha)}(1+\log(N+1))^2\boldsymbol{\Psi}_{\alpha}(N+1)\ .\end{aligned}$$

■

Lemma 60 *Assume* **A**1*,* **A**2-(a) *and* **F**2-(a)*. In addition, assume that* (59) *holds. Then, for any* $\alpha, \gamma \in (0,1)$ *and* $N \in \mathbb{N}$ *we have*

$$S_N - f^\star \leq 2\mathtt{C}^{(d)}(1+\log(N+1))^2(N+1)^{-\min(\alpha,1-\alpha)}\boldsymbol{\Psi}_\alpha(N+1) \\ + (2\gamma)^{-1}(N+1)^{-1}\left\|X_0 - x^\star\right\|^2 \ .$$

with S *given in* (60).

Proof Using Lemma 58 with the choice $Y = x^\star$ and $u = N$ gives

$$(N+1)^{-1}\mathbb{E}\left[\sum_{k=0}^{N} f(X_k) - f(x^\star)\right] \leq 2\mathtt{C}^{(d)}(1+\log(N+1))(N+1)^{-\min(\alpha,1-\alpha)}\boldsymbol{\Psi}_\alpha(N+1) \\ + (2\gamma)^{-1}(N+1)^{-1}\left\|X_0 - x^\star\right\|^2$$

Therefore,

$$S_N - f^\star \leq 2\mathtt{C}^{(d)}(1+\log(N+1))^2(N+1)^{-\min(\alpha,1-\alpha)}\boldsymbol{\Psi}_\alpha(N+1) \\ + (2\gamma)^{-1}(N+1)^{-1}\left\|X_0 - x^\star\right\|^2 \ .$$

■

We can finally conclude the proof of Theorem 8.
Proof We begin by proving by induction over $m \in \mathbb{N}^*$ that the following assertion **H**2(m) is true.

H2 (m) *For any* $\alpha > 1/(m+1)$*, there exists* $\mathtt{C}_\alpha^+ > 0$ *such that for all* $n \in \mathbb{N}$*,* $\mathbb{E}[\|\nabla f(X_n)\|^2] \leq \mathtt{C}_\alpha^+$*. In addition, for any* $\alpha \leq 1/(m+1)$*, there exists* $\mathtt{C}_\alpha^- > 0$ *such that for all* $n \in \mathbb{N}$*,* $\mathbb{E}[\|\nabla f(X_n)\|^2] \leq \mathtt{C}_\alpha^- n^{1-(m+1)\alpha}(1+\log(n))^3$*.*

For $m = 1$, **H**2(1) is an immediate consequence of **A**1 and Lemma 56, with $\mathtt{C}_\alpha^+ = \mathtt{L}^2\mathtt{C}_{2,\alpha}^{(d)}$ and $\mathtt{C}_\alpha^- = \mathtt{L}^2\max(\mathtt{C}_{1,\alpha}^{(d)}, \mathtt{C}_{2,\alpha}^{(d)})$. Now, let $m \in \mathbb{N}^*$ and suppose that **H**2(m) holds. Let $\alpha \in (0,1)$. Setting $\alpha^\star = 1/(m+1)$ we have that (59) is verified with $\beta = 1-(m+1)\alpha$. Consequently, using **A**1, **F**2-(a) and **A**2-(a) we can apply Proposition 57 and for any $\alpha \leq 1/(m+1)$ we have

$$\begin{aligned} \mathbb{E}\left[f(X_N)\right] - f^\star &\leq \tilde{\mathtt{C}}_\alpha\left\{(1+\log(N+1))^2/(N+1)^{\min(\alpha,1-\alpha)}\boldsymbol{\Psi}_\alpha(N+1) + 1/(N+1)\right\} \\ &\leq \tilde{\mathtt{C}}_\alpha\left\{(1+\log(N+1))^3(N+1)^{-\alpha}(N+1)^{1-(m+1)\alpha} + 1/(N+1)\right\} \\ &\leq \tilde{\mathtt{C}}_\alpha\left\{(1+\log(N+1))^3(N+1)^{1-(m+2)\alpha} + 1/(N+1)\right\} \ . \end{aligned} \tag{64}$$

In particular, if $\alpha > 1/(m+2)$ we have the existence of $\bar{\mathtt{C}}_\alpha > 0$ such that for all $n \in \mathbb{N}$, $\mathbb{E}[f(X_n)] - f^\star \leq \bar{\mathtt{C}}_\alpha$. And using **A**1 and Lemma 40 we get that, for all $n \in \mathbb{N}$

$$\mathbb{E}[\|\nabla f(X_n)\|^2] \leq 2\mathtt{L}\mathbb{E}\left[f(X_n) - f^\star\right] \leq 2\mathtt{L}\bar{\mathtt{C}}_\alpha \ ,$$

Combining this result with (64), we get that **H**2$(m+1)$ holds with $\mathtt{C}_\alpha^+ = 2\mathtt{L}\bar{\mathtt{C}}_\alpha$ and $\mathtt{C}_\alpha^- = 2\tilde{\mathtt{C}}_\alpha$. We conclude by recursion.

Now, let $\alpha \in (0,1)$. Since $\mathbb{R}$ is archimedean, there exists $m \in \mathbb{N}^*$ such that $\alpha > 1/(m+1)$ and therefore **H**2(m) shows the existence of $\mathtt{C}_0 > 0$ such that $\mathbb{E}[\|\nabla f(X_n)\|^2] \leq \mathtt{C}_0$ for all $n \in \mathbb{N}^*$. Applying Proposition 57 gives the existence of $\mathtt{C}^{(d)} > 0$ such that for all $N \geq 1$

$$\mathbb{E}\left[f(X_N)\right] - f^\star \leq \mathtt{C}^{(d)}(1+\log(N+1))^2/(N+1)^{\min(\alpha,1-\alpha)} ,$$

with $\mathtt{C}^{(d)} = 2\tilde{\mathtt{C}}_\alpha$, concluding the proof. ■

We present now a corollary of the previous theorem under a different setting. Let us assume, as in (Shamir and Zhang, 2013), that ∇f is not Lipschitz-continuous but bounded instead.

Corollary 61 *Let $\gamma, \alpha \in (0,1)$ and $x_0 \in \mathbb{R}^d$ and $(X_n)_{n\geq 0}$ be given by (1). Assume* **F**2-(a)*,* **A**2-(a) *and ∇f bounded. Then there exists $\mathtt{C}_b^{(d)} \geq 0$ such that, for all $N \geq 1$,*

$$\mathbb{E}\left[f(X_N)\right] - f^\star \leq \mathtt{C}_b^{(d)}(1+\log(N+1))^2/(N+1)^{\min(\alpha,1-\alpha)} .$$

Proof The proof follows the same lines as the ones of Lemma 56 and Proposition 57. We show that both conclusions hold under the assumption that ∇f is bounded instead of being Lipschitz-continuous.

In order to prove that Lemma 56 still holds, let us do the following computation. We consider $(X_n)_{n\geq 0}$ satisfying (1). We have, using (1), **F**2-(a) and **A**2-(a) that for all $n \geq 0$,

$$\begin{aligned}
\mathbb{E}[\|X_{n+1} - x^\star\|^2 \,|\mathcal{F}_n] &= \mathbb{E}\left[\left\|X_n - x^\star - \gamma(n+1)^{-\alpha} H(X_n, Z_{n+1})\right\|^2 \middle| \mathcal{F}_n\right] \\
&= \|X_n - x^\star\|^2 - 2\gamma/(n+1)^\alpha \langle X_n - x^\star, \mathbb{E}\left[H(X_n, Z_{n+1})|\mathcal{F}_n\right]\rangle \\
&\quad + \gamma^2(n+1)^{-2\alpha}\mathbb{E}\left[\|H(X_n, Z_{n+1})\|^2 \middle| \mathcal{F}_n\right] \\
&= \|X_n - x^\star\|^2 - 2\gamma/(n+1)^\alpha \langle X_n - x^\star, \nabla f(X_n)\rangle \\
&\quad + \gamma^2\eta(n+1)^{-2\alpha} + \gamma^2(n+1)^{-2\alpha}\|\nabla f(X_n)\|^2 \\
\mathbb{E}[\|X_{n+1} - x^\star\|^2] &\leq \mathbb{E}[\|X_n - x^\star\|^2] + \gamma^2(\eta + \|\nabla f\|_\infty)(n+1)^{-2\alpha} .
\end{aligned}$$

And we obtain the same equation as in (58), with a different constant before the asymptotic term $(n+1)^{-2\alpha}$. Hence the conclusions of Lemma 56 still hold, because **A**1 is never used in the remaining of the proof. We can now safely apply Proposition 57 (since **A**1 is only used to use Lemma 56) with $\alpha^\star = 0$. This concludes the proof. ■

Appendix F. Convex case (under A2-(b))

In this section, we prove similar results to the ones of Appendix E under **A**2-(b). In Appendix F.1 we prove the equivalent to Appendix E.2 in this setting (in particular we recover the optimal rate in the convex setting under **A**2-(b) for continuous SGD). Similarly, in Appendix F.2 we prove the equivalent to Appendix E.3 in this setting (in particular we recover the optimal rate in the convex setting under **A**2-(b) for SGD).

F.1. Equivalent to Appendix E.2

First, we start with Lemma 62 which is an equivalent of Lemma 51. The discussion conducted at the begin of Appendix E.2 is still valid here. However, similarly to the discrete-case under **A**2-(a) we have to rely on some bootstrapping technique to conclude. The equivalent to Lemma 52 is given in Lemma 52. The intermediate result needed to apply our bootstrapping procedure is stated in Proposition 63. Lemma 65, Lemma 66 and Lemma 67 are the counterparts to Lemma 53, Lemma 54 and Lemma 55 respectively. We state and prove our main result in Theorem 68.

Lemma 62 *Assume* **A**1*,* **A**2-(b)*,* **A**3 *and* **F**2-(b)*. Let* $(\mathbf{X}_t)_{t\geq 0}$ *be given by* (2)*. Then, for any* $\alpha, \gamma \in (0,1)$*, there exists* $\mathtt{C}^{(c)}_{1,\alpha} \geq 0$ *and* $\mathtt{C}^{(c)}_{2,\alpha} \geq 0$ *and a function* $\mathbf{\Phi}^{(c)}_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ *such that, for any* $t \geq 0$*,*

$$\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^2\right] \leq \mathtt{C}^{(c)}_{1,\alpha}\mathbf{\Phi}^{(c)}_\alpha(t+\gamma_\alpha) + \mathtt{C}^{(c)}_{2,\alpha} \ .$$

And we have

$$\mathbf{\Phi}^{(c)}_\alpha(t) = \begin{cases} t^{1-2\alpha} & \textit{if } \alpha < 1/2 \ , \\ \log(t) & \textit{if } \alpha = 1/2 \ , \\ 0 & \textit{if } \alpha > 1/2 \ . \end{cases}$$

The values of the constants are given by

$$\mathtt{C}^{(c)}_{1,\alpha} = \begin{cases} \gamma_\alpha \mathtt{L}_\mathtt{T}(1-2\alpha)^{-1} & \textit{if } \alpha < 1/2 \ , \\ \gamma_\alpha \mathtt{L}_\mathtt{T} & \textit{if } \alpha = 1/2 \ , \\ 0 & \textit{if } \alpha > 1/2 \ . \end{cases}$$

$$\mathtt{C}^{(c)}_{2,\alpha} = \begin{cases} 2\max_{t\leq(\gamma_\alpha \mathtt{L}_\mathtt{T})^{1/\alpha}} \mathbb{E}\left[\|X_t - x^\star\|^2\right] & \textit{if } \alpha < 1/2 \ , \\ 2\max_{t\leq(\gamma_\alpha \mathtt{L}_\mathtt{T})^{1/\alpha}} \mathbb{E}\left[\|X_t - x^\star\|^2\right] - \gamma_\alpha \mathtt{L}_\mathtt{T}\log(\gamma_\alpha) & \textit{if } \alpha = 1/2 \ , \\ 2\max_{t\leq(\gamma_\alpha \mathtt{L}_\mathtt{T})^{1/\alpha}} \mathbb{E}\left[\|X_t - x^\star\|^2\right] + (2\alpha-1)^{-1}\gamma_\alpha^{2-2\alpha}\mathtt{L}_\mathtt{T} & \textit{if } \alpha > 1/2 \ , \end{cases}$$

Proof Let $\alpha, \gamma \in (0,1)$ and $t \geq 0$. Let $(\mathbf{X}_t)_{t\geq 0}$ be given by (2). We consider the function $F : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}_+$ given for any $(t,x) \in \mathbb{R}\times\mathbb{R}^d$ by $F(t,x) = \|x - x^\star\|^2$. Applying Lemma 48 to the stochastic process $(F(t, \mathbf{X}_t))_{t\geq 0}$ and using Lemma 42 and **F**2-(b) gives that for all $t \geq (\gamma_\alpha \mathtt{L_T})^{1/\alpha}$,

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^2\right] - \mathbb{E}\left[\|\mathbf{X}_0 - x^\star\|^2\right] \\
&\quad = -2\int_0^T (t+\gamma_\alpha)^{-\alpha}\mathbb{E}\left[\langle \mathbf{X}_t - x^\star, \nabla f(\mathbf{X}_t)\rangle\right] \mathrm{d}t + \int_0^T \gamma_\alpha (t+\gamma_\alpha)^{-2\alpha}\mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_t))\right] \mathrm{d}t \\
&\quad \leq -2\int_0^T (t+\gamma_\alpha)^{-\alpha}\mathbb{E}\left[\langle \mathbf{X}_t - x^\star, \nabla f(\mathbf{X}_t)\rangle\right] \mathrm{d}t \\
&\qquad + \int_0^T \gamma_\alpha (t+\gamma_\alpha)^{-2\alpha}\mathtt{L_T}\left[1 + \mathbb{E}\left[f(\mathbf{X}_t) - f(x^\star)\right]\right] \mathrm{d}t \\
&\quad \leq -2\int_0^T (t+\gamma_\alpha)^{-\alpha}\mathbb{E}\left[\langle \mathbf{X}_t - x^\star, \nabla f(\mathbf{X}_t)\rangle\right] \mathrm{d}t \\
&\qquad + \int_0^T \gamma_\alpha (t+\gamma_\alpha)^{-2\alpha}\mathtt{L_T}\left[1 + \mathbb{E}\left[\langle \nabla f(\mathbf{X}_t), \mathbf{X}_t - x^\star\rangle\right]\right] \mathrm{d}t \\
&\quad \leq \gamma_\alpha \mathtt{L_T}\int_0^T (t+\gamma_\alpha)^{-2\alpha}\mathrm{d}t + \int_0^T (t+\gamma_\alpha)^{-\alpha}\mathbb{E}\left[\langle \nabla f(\mathbf{X}_t), \mathbf{X}_t - x^\star\rangle\right]\left\{-2 + \gamma_\alpha\mathtt{L_T}(t+\gamma_\alpha)^{-\alpha}\right\} \mathrm{d}t \\
&\quad \leq \gamma_\alpha \mathtt{L_T}\int_0^T (t+\gamma_\alpha)^{-2\alpha}\mathrm{d}t \; .
\end{aligned}$$

We now distinguish three cases:

(a) If $\alpha < 1/2$, then we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^2\right] &\leq \|X_0 - x^\star\|^2 + \gamma_\alpha\mathtt{L_T}(1-2\alpha)^{-1}((T+\gamma_\alpha)^{1-2\alpha} - \gamma_\alpha^{1-2\alpha}) \\
&\leq \|X_0 - x^\star\|^2 + \gamma_\alpha\mathtt{L_T}(1-2\alpha)^{-1}(T+\gamma_\alpha)^{1-2\alpha} \; .
\end{aligned}$$

(b) If $\alpha = 1/2$, then we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^2\right] &\leq \|X_0 - x^\star\|^2 + \gamma_\alpha\mathtt{L_T}(\log(T+\gamma_\alpha) - \log(\gamma_\alpha)) \\
&\leq \gamma_\alpha\mathtt{L_T}\log(T+\gamma_\alpha) + \|X_0 - x^\star\|^2 - \gamma_\alpha\mathtt{L_T}\log(\gamma_\alpha) \; .
\end{aligned}$$

(c) If $\alpha > 1/2$, then we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^2\right] &\leq \|X_0 - x^\star\|^2 + \gamma_\alpha\mathtt{L_T}(1-2\alpha)^{-1}((T+\gamma_\alpha)^{1-2\alpha} - \gamma_\alpha^{1-2\alpha}) \\
&\leq \|X_0 - x^\star\|^2 + (2\alpha-1)^{-1}\gamma_\alpha^{2-2\alpha}\mathtt{L_T} \; .
\end{aligned}$$

■

Proposition 63 *Let $\gamma, \alpha \in (0,1)$ and $x_0 \in \mathbb{R}^d$ $(\mathbf{X}_t)_{t\geq 0}$ be given by (2). Assume* **A**1*,* **A**2*-(b),* **F**2*-(b) and* **A**3*. In addition, assume that there exists $\alpha^\star \in [0, 1/2]$, $\beta > 0$ and $\mathtt{C}_0 \geq 0$ such that for all $t \in [0, T]$*

$$\mathbb{E}\left[\mathrm{Tr}\,\Sigma(\mathbf{X}_t)\right] \leq \begin{cases} \mathtt{C}_0 (t+\gamma_\alpha)^\beta \log(t+\gamma_\alpha) & \textit{if } \alpha \leq \alpha^\star \; , \\ \mathtt{C}_0 & \textit{if } \alpha > \alpha^\star \; . \end{cases} \tag{65}$$

Then there exists $\tilde{\mathtt{C}}_\alpha \geq 0$ *such that, for all* $T \geq 1$*,*

$$\mathbb{E}\left[f(\mathbf{X}_T)\right] - f^\star \leq \tilde{\mathtt{C}}_\alpha \left[\log(T+\gamma_\alpha)^2 (T+\gamma_\alpha)^{-\min(\alpha, 1-\alpha)}\right] \left(1 + \mathbf{\Psi}_\alpha(T+\gamma_\alpha)\right) ,$$

with

$$\mathbf{\Psi}_\alpha(t) = \begin{cases} t^\beta & \text{if } \alpha \leq \alpha^\star \ , \\ 0 & \text{if } \alpha > \alpha^\star \ . \end{cases}$$

Proof Let $f \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R})$. Let $\gamma \in (0,1)$ and $\alpha \in (0, 1/2]$ and $T \geq 1$. Let $(\mathbf{X}_t)_{t \geq 0}$ be given by (2). Combining Lemma 65, Lemma 66 and Lemma 67 gives the desired result

$$\mathbb{E}\left[f(\mathbf{X}_T)\right] - f^\star \leq \mathtt{C}^{\widetilde{(c)}} \left[\log(T+\gamma_\alpha)^2 (T+\gamma_\alpha)^{-\min(\alpha, 1-\alpha)}\right] \left(1 + \mathbf{\Psi}_\alpha(T+\gamma_\alpha)\right) ,$$

with $\mathtt{C}^{\widetilde{(c)}} = (4/\gamma_\alpha^2)(4\mathtt{C}_1 + 2\tilde{\mathtt{C}}_1 + 2\mathtt{C}_1\mathtt{L})$. ■

Let $S : [0, T] \to [0, +\infty)$ defined by

$$\begin{cases} S(t) = t^{-1} \int_{T-t}^{T} \left\{\mathbb{E}\left[f(\mathbf{X}_s)\right] - f^\star\right\} \mathrm{d}s \ , & \text{if } t > 0 \ , \\ S(0) = \mathbb{E}\left[f(\mathbf{X}_T)\right] \ . \end{cases} \tag{66}$$

With this notation we have

$$\mathbb{E}\left[f(\mathbf{X}_T)\right] - f^\star = S(0) - S(1) + S(1) - S(T) + S(T) - f^\star \ .$$

We are now going to control each one of the three terms $(S(0) - S(1))$, $(S(1) - S(T))$, $(S(T) - f^\star)$ as follows:

(a) Case $S(1) - S(T)$ (Lemma 65): we adapt the idea of suffix averaging of Shamir and Zhang (2013) to the continuous-time setting. In particular, we control the time-derivative of S in Lemma 64 (counterpart of Lemma 52). Note that Lemma 65 is the counterpart to Lemma 53.

(b) Case $S(T) - f^\star$ (Lemma 66): this result is known and corresponds to the optimal convergence rate of the averaged sequence towards the minimum of f. We provide its proof for completeness. Note that Lemma 66 is the counterpart to Lemma 54.

(c) Case $S(0) - S(1)$ (Lemma 67): this last term is specific to the continuous-time setting and is a necessary modification to the classic averaging control of $S(\varepsilon) - S(T)$, established in Lemma 65 for $\varepsilon = 1$, which diverges for ε close to 0. Note that Lemma 67 is the counterpart to Lemma 55.

Before controlling each one of these terms we state the following useful lemma, which will allow us to control the derivative of S.

Lemma 64 *Assume* **A**1*,* **A**2-(b)*,* **A**3*, and* **F**2-(b)*. In addition, assume that* (65) *holds. Then, for any* $\alpha, \gamma \in (0,1)$*,* $T \geq 0$*,* $u \in [0, T]$ *and* Y *any* $\mathbb{R}^d$*-valued random variable such that* $\mathbb{E}[\|Y - x^\star\|^2] \leq$

$\mathtt{C}_{1,\alpha}^{(c)}\mathbf{\Phi}_{\alpha}^{(c)}(T+\gamma_\alpha)+\mathtt{C}_{2,\alpha}^{(c)}$ with $\mathtt{C}_{1,\alpha}^{(c)}$ and $\mathtt{C}_{2,\alpha}^{(c)}$ given in Lemma 62, we have

$$\begin{aligned}\int_{T-u}^{T}&\mathbb{E}\left[f(\mathbf{X}_t)-f(Y)\right]\mathrm{d}t\\ &\leq \mathtt{C}_1\left((T+\gamma_\alpha)^{\alpha}-(T-u+\gamma_\alpha)^{\alpha}\right)\\ &\quad+(1/2)(T-u+\gamma_\alpha)^{\alpha}\mathbb{E}\left[\|\mathbf{X}_{T-u}-Y\|^2\right]\\ &\quad+\mathtt{C}_1\log(T+\gamma_\alpha)\left\{(T+\gamma_\alpha)^{\alpha}-(T-u+\gamma_\alpha)^{\alpha}\right\}(T+\gamma_\alpha)^{1-2\alpha}\\ &\quad+\mathtt{C}_1\left((T+\gamma_\alpha)^{1-\alpha}-(T-u+\gamma_\alpha)^{1-\alpha}\right)\log(T+\gamma_\alpha)(1+\mathbf{\Psi}_\alpha(T+\gamma_\alpha))\ ,\end{aligned}$$

with $\mathtt{C}_1=\max(4\mathtt{C}_{2,\alpha}^{(c)}/2,(\gamma_\alpha\mathtt{C}_0/2+\mathtt{C}_{1,\alpha}^{(c)})(1-\alpha)^{-1})$ and $\mathbf{\Psi}_\alpha(t)=0$ if $\alpha>\alpha^\star$ and t^β if $\alpha\leq\alpha^\star$, with $\mathtt{C}_{1,\alpha}^{(c)}$ and $\mathtt{C}_{2,\alpha}^{(c)}$ given in Lemma 62.

Proof For any $Y\in\mathbb{R}^d$ we define the function $F_{y_0}:\mathbb{R}_+\times\mathbb{R}^d\to\mathbb{R}$ by

$$F_{y_0}(t,x)=(t+\gamma_\alpha)^{\alpha}\|x-y_0\|^2\ .$$

Using Lemma 51, that $\mathbf{\Phi}_{\alpha}^{(c)}$ is non-decreasing and that for any $a,b\geq 0$, $(a+b)^2\leq 2a^2+2b^2$, we have

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{X}_t-Y\|^2\right]&=\mathbb{E}\left[\|(\mathbf{X}_t-x^\star)+(x^\star-Y)\|^2\right]\\ &\leq 2\mathbb{E}\left[\|\mathbf{X}_t-x^\star\|^2\right]+2\mathbb{E}\left[\|Y-x^\star\|^2\right]\\ &\leq 2\mathtt{C}_{1,\alpha}^{(c)}\mathbf{\Phi}_{\alpha}^{(c)}(t+\gamma_\alpha)+4\mathtt{C}_{2,\alpha}^{(c)}+2\mathtt{C}_{1,\alpha}^{(c)}\mathbf{\Phi}_{\alpha}^{(c)}(T+\gamma_\alpha)\\ &\leq 2\mathtt{C}_{1,\alpha}^{(c)}\mathbf{\Phi}_{\alpha}^{(c)}(t+\gamma_\alpha)+2\mathtt{C}_{1,\alpha}^{(c)}\mathbf{\Phi}_{\alpha}^{(c)}(T+\gamma_\alpha)+\mathtt{C}_{3,\alpha}^{(c)}\ .\end{aligned}$$

with $\mathtt{C}_{3,\alpha}^{(c)}=4\mathtt{C}_{2,\alpha}^{(c)}$. This gives in particular, for every $t\in[0,T]$,

$$\begin{aligned}(t+\gamma_\alpha)^{\alpha-1}\mathbb{E}\left[\|\mathbf{X}_t-Y\|^2\right]\leq&\left[\mathtt{C}_{3,\alpha}^{(c)}+2\mathtt{C}_{1,\alpha}^{(c)}(T+\gamma_\alpha)^{1-2\alpha}\log(T+\gamma_\alpha)\right](t+\gamma_\alpha)^{\alpha-1}\\ &+2\mathtt{C}_{1,\alpha}^{(c)}\log(T+\gamma_\alpha)(t+\gamma_\alpha)^{-\alpha}\ ,\end{aligned}\tag{67}$$

with $\mathtt{C}_{1,\alpha}^{(c)}=0$ if $\alpha>1/2$. Notice that the additional $\log(T+\gamma_\alpha)$ term is only needed in the case where $\alpha=1/2$. For any $(t,x)\in\mathbb{R}_+\times\mathbb{R}^d$, we have

$$\begin{aligned}&\partial_t F_{y_0}(t,x)=\alpha(t+\gamma_\alpha)^{\alpha-1}\|x-y_0\|^2\ ,\\ &\partial_x F_{y_0}(t,x)=2(t+\gamma_\alpha)^{\alpha}(x-y_0)\ ,\quad \partial_{xx}F_{y_0}(t,x)=2(t+\gamma_\alpha)^{\alpha}\ .\end{aligned}$$

Using Lemma 48 on the stochastic process $(F_Y(t,\mathbf{X}_t))_{t\geq 0}$, we have that for any $u\in[0,T]$

$$\begin{aligned}\mathbb{E}\left[F_Y(T,\mathbf{X}_T)\right]-\mathbb{E}\left[F_Y(T-u,\mathbf{X}_{T-u})\right]=&\int_{T-u}^{T}\alpha(t+\gamma_\alpha)^{\alpha-1}\mathbb{E}\left[\|\mathbf{X}_t-Y\|^2\right]\mathrm{d}t\\ &-2\int_{T-u}^{T}\mathbb{E}\left[\langle\mathbf{X}_t-Y,\nabla f(\mathbf{X}_t)\rangle\right]\mathrm{d}t\\ &+\int_{T-u}^{T}\gamma_\alpha(t+\gamma_\alpha)^{-\alpha}\mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_t))\right]\mathrm{d}t\ .\end{aligned}\tag{68}$$

We distinguish now several cases depending on the value of α.

(a) If $\alpha \leq \alpha^\star$, then combining **F**2-(b), **A**2-(b), (65), (67) and (68) we obtain for any $u \in [0, T]$

$$\begin{aligned}
&- (T-u+\gamma_\alpha)^\alpha \mathbb{E}\left[\|\mathbf{X}_{T-u} - Y\|^2\right] \\
&\quad \leq \mathtt{C}_{3,\alpha}^{(c)} \int_{T-u}^T \alpha(t+\gamma_\alpha)^{\alpha-1}\mathrm{d}t + \mathtt{C}_0\gamma_\alpha \int_{T-u}^T (t+\gamma_\alpha)^{-\alpha}(t+\gamma_\alpha)^\beta \log(t+\gamma_\alpha)\mathrm{d}t \\
&\qquad + 2\alpha\mathtt{C}_{1,\alpha}^{(c)} \log(T+\gamma_\alpha)\left\{\int_{T-u}^T (t+\gamma_\alpha)^{-\alpha}\mathrm{d}t + (T+\gamma_\alpha)^{1-2\alpha}\int_{T-u}^T (t+\gamma_\alpha)^{\alpha-1}\mathrm{d}t\right\} \\
&\qquad - 2\int_{T-u}^T \mathbb{E}\left[f(\mathbf{X}_t) - f(Y)\right]\mathrm{d}t \\
&\quad \leq \mathtt{C}_{3,\alpha}^{(c)}\left((T+\gamma_\alpha)^\alpha - (T-u+\gamma_\alpha)^\alpha\right) - 2\int_{T-u}^T \mathbb{E}\left[f(\mathbf{X}_t) - f(Y)\right]\mathrm{d}t \\
&\qquad + (\gamma_\alpha\mathtt{C}_0(T+\gamma_\alpha)^\beta + 2\alpha\mathtt{C}_{1,\alpha}^{(c)})(1-\alpha)^{-1}\left((T+\gamma_\alpha)^{1-\alpha} - (T-u+\gamma_\alpha)^{1-\alpha}\right)\log(T+\gamma_\alpha) \\
&\qquad + 2\mathtt{C}_{1,\alpha}^{(c)}\log(T+\gamma_\alpha)\left\{(T+\gamma_\alpha)^\alpha - (T-u+\gamma_\alpha)^\alpha\right\}(T+\gamma_\alpha)^{1-2\alpha} .
\end{aligned}$$

(b) If $\alpha > \alpha^\star$, then combining **F**2-(b), **A**2-(b), (65) (67) and (68) we obtain for any $u \in [0, T]$

$$\begin{aligned}
&- (T-u+\gamma_\alpha)^\alpha \mathbb{E}\left[\|\mathbf{X}_{T-u} - Y\|^2\right] \\
&\quad \leq \mathtt{C}_{3,\alpha}^{(c)} \int_{T-u}^T \alpha(t+\gamma_\alpha)^{\alpha-1}\mathrm{d}t + \mathtt{C}_0\gamma_\alpha \int_{T-u}^T (t+\gamma_\alpha)^{-\alpha}\mathrm{d}t \\
&\qquad + 2\alpha\mathtt{C}_{1,\alpha}^{(c)} \log(T+\gamma_\alpha)\left\{\int_{T-u}^T (t+\gamma_\alpha)^{-\alpha}\mathrm{d}t + (T+\gamma_\alpha)^{1-2\alpha}\int_{T-u}^T (t+\gamma_\alpha)^{\alpha-1}\mathrm{d}t\right\} \\
&\qquad - 2\int_{T-u}^T \mathbb{E}\left[f(\mathbf{X}_t) - f(Y)\right]\mathrm{d}t \\
&\quad \leq \mathtt{C}_{3,\alpha}^{(c)}\left((T+\gamma_\alpha)^\alpha - (T-u+\gamma_\alpha)^\alpha\right) - 2\int_{T-u}^T \mathbb{E}\left[f(\mathbf{X}_t) - f(Y)\right]\mathrm{d}t \\
&\qquad + (\gamma_\alpha\mathtt{C}_0 + 2\alpha\mathtt{C}_{1,\alpha}^{(c)})(1-\alpha)^{-1}\left((T+\gamma_\alpha)^{1-\alpha} - (T-u+\gamma_\alpha)^{1-\alpha}\right)\log(T+\gamma_\alpha) \\
&\qquad + 2\mathtt{C}_{1,\alpha}^{(c)}\log(T+\gamma_\alpha)\left\{(T+\gamma_\alpha)^\alpha - (T-u+\gamma_\alpha)^\alpha\right\}(T+\gamma_\alpha)^{1-2\alpha} .
\end{aligned}$$

Putting this together we get for any $u \in [0, T]$

$$\begin{aligned}
&\int_{T-u}^T \mathbb{E}\left[f(\mathbf{X}_t) - f(Y)\right]\mathrm{d}t \\
&\quad \leq \mathtt{C}_1\left((T+\gamma_\alpha)^\alpha - (T-u+\gamma_\alpha)^\alpha\right) \\
&\qquad + (1/2)(T-u+\gamma_\alpha)^\alpha\mathbb{E}\left[\|\mathbf{X}_{T-u} - Y\|^2\right] \\
&\qquad + \mathtt{C}_1\log(T+\gamma_\alpha)\left\{(T+\gamma_\alpha)^\alpha - (T-u+\gamma_\alpha)^\alpha\right\}(T+\gamma_\alpha)^{1-2\alpha} \\
&\qquad + \mathtt{C}_1\left((T+\gamma_\alpha)^{1-\alpha} - (T-u+\gamma_\alpha)^{1-\alpha}\right)\log(T+\gamma_\alpha)(1 + \mathbf{\Psi}_\alpha(T+\gamma_\alpha)) ,
\end{aligned}$$

with $\mathtt{C}_1 = \max(\mathtt{C}_{3,\alpha}^{(c)}/2, (\gamma_\alpha\mathtt{C}_0/2 + \mathtt{C}_{1,\alpha}^{(c)})(1-\alpha)^{-1})$ and $\mathbf{\Psi}_\alpha(t) = 0$ if $\alpha > \alpha^\star$ and t^β if $\alpha \leq \alpha^\star$. ■

We divide the rest of the proof into three parts, to bound the quantities $S(1)-S(T)$, $S(T)-f^\star$ and $S(0)-S(1)$.

Lemma 65 *Assume* **A**1*,* **A**2-(b)*,* **A**3*, and* **F**2-(b)*. In addition, assume that* (65) *holds. Then, for any* $\alpha, \gamma \in (0,1)$ *and* $T \geq 0$ *we have*

$$S(1) - S(T) \leq 2\tilde{\mathtt{C}_1} \log(T+\gamma_\alpha) \log(1+T)(T+\gamma_\alpha)^{-\min(\alpha,1-\alpha)}(1+\boldsymbol{\Psi}_\alpha(T+\gamma_\alpha)) \ ,$$

with S *given in* (66).

Proof In the case where $\alpha \leq 1/2$, Lemma 50 gives that for all $u \in [0,T]$:

$$((T+\gamma_\alpha)^\alpha - (T-u+\gamma_\alpha)^\alpha) \leq \left((T+\gamma_\alpha)^{1-\alpha} - (T-u+\gamma_\alpha)^{1-\alpha}\right) \ ,$$

and we also have, for all $u \in [0,T]$:

$$\begin{aligned}
&(T+\gamma_\alpha)^{1-\alpha} - (T+\gamma_\alpha-u)^{1-\alpha} \\
&\quad = \left(((T+\gamma_\alpha)^{1-\alpha} - (T+\gamma_\alpha-u)^{1-\alpha})((T+\gamma_\alpha)^\alpha + (T+\gamma_\alpha-u)^\alpha)\right) \\
&\qquad \times \left((T+\gamma_\alpha)^\alpha + (T+\gamma_\alpha-u)^{-\alpha}\right)^{-1} \\
&\quad \leq \left((T+\gamma_\alpha) - (T+\gamma_\alpha-u) + (T+\gamma_\alpha)^{1-\alpha}(T+\gamma_\alpha-u)^\alpha - (T+\gamma_\alpha)^\alpha(T+\gamma_\alpha-u)^{1-\alpha}\right) \\
&\qquad \times (T+\gamma_\alpha)^{-\alpha} \leq 2u/(T+\gamma_\alpha)^\alpha \ . \qquad (69)
\end{aligned}$$

And in the case where $\alpha > 1/2$, for all $u \in [0,T]$:

$$\left((T+\gamma_\alpha)^{1-\alpha} - (T-u+\gamma_\alpha)^{1-\alpha}\right) \leq ((T+\gamma_\alpha)^\alpha - (T-u+\gamma_\alpha)^\alpha) \ ,$$

and we also have, for all $u \in [0,T]$:

$$\begin{aligned}
&(T+\gamma_\alpha)^\alpha - (T+\gamma_\alpha-u)^\alpha \\
&\quad = \left(((T+\gamma_\alpha)^\alpha - (T+\gamma_\alpha-u)^\alpha)((T+\gamma_\alpha)^{1-\alpha} + (T+\gamma_\alpha-u)^{1-\alpha})\right) \\
&\qquad \left((T+\gamma_\alpha)^{1-\alpha} + (T+\gamma_\alpha-u)^{1-\alpha}\right)^{-1} \\
&\quad \leq \left((T+\gamma_\alpha) - (T+\gamma_\alpha-u) + (T+\gamma_\alpha)^\alpha(T+\gamma_\alpha-u)^{1-\alpha} - (T+\gamma_\alpha)^{1-\alpha}(T+\gamma_\alpha-u)^\alpha\right) \\
&\qquad \times (T+\gamma_\alpha)^{-1+\alpha} \leq 2u/(T+\gamma_\alpha)^{1-\alpha} \ .
\end{aligned}$$

Now, plugging $Y = \mathbf{X}_{T-u}$ in Lemma 64 we obtain, for all $u \in [0,T]$:

$$\mathbb{E}\left[\int_{T-u}^T f(\mathbf{X}_t) - f(\mathbf{X}_{T-u}) \mathrm{d}t\right] \leq 2\tilde{\mathtt{C}_1} \log(T+\gamma_\alpha)(T+\gamma_\alpha)^{-\min(\alpha,1-\alpha)}(1+\boldsymbol{\Psi}_\alpha(T+\gamma_\alpha))u \ , \quad (70)$$

with $\tilde{\mathtt{C}_1} = (2\mathtt{C}_1 + \mathtt{C}_1(1+\gamma_\alpha^{1-2\alpha}))$.

Since S is a differentiable function and using (70), we have for all $u \in (0,T)$,

$$S'(u) = -u^{-2}\int_{T-u}^T \mathbb{E}\left[f(\mathbf{X}_t)\right] \mathrm{d}t + u^{-1}\mathbb{E}\left[f(\mathbf{X}_{T-u})\right] = -u^{-1}(S(u) - \mathbb{E}\left[f(\mathbf{X}_{T-u})\right]) \ .$$

This last result implies $-S'(u) \leq 2\tilde{\mathtt{C}_1} \log(T+\gamma_\alpha)/(T+\gamma_\alpha)^{\min(\alpha,1-\alpha)}(1+\boldsymbol{\Psi}_\alpha(T+\gamma_\alpha))u^{-1}$ and integrating we get

$$S(1) - S(T) \leq 2\tilde{\mathtt{C}_1} \log(T+\gamma_\alpha) \log(T)(T+\gamma_\alpha)^{-\min(\alpha,1-\alpha)}(1+\boldsymbol{\Psi}_\alpha(T+\gamma_\alpha)) \ .$$

■

Lemma 66 *Assume* **A**1*,* **A**2-(b)*,* **A**3*, and* **F**2-(b)*. In addition, assume that* (65) *holds. Then, for any $\alpha, \gamma \in (0,1)$ and $T \geq 0$ we have*

$$S(T) - f^\star \leq 4\mathtt{C}_1 T^{-\min(\alpha,1-\alpha)}(1 + \log(T + \gamma_\alpha))(1 + \mathbf{\Psi}_\alpha(T + \gamma_\alpha)) ,$$

with S given in (66)*.*

Proof Using Lemma 64 with $u = T$ and $Y = x^\star$, and $\|\mathbf{X}_0 - x^\star\| \leq \mathtt{C}_1$ we obtain

$$\begin{aligned}\int_0^T \mathbb{E}\left[f(X_s)\right] \mathrm{d}s - Tf^\star &\leq (\mathtt{C}_1/2)\left((T+\gamma_\alpha)^\alpha - \gamma_\alpha^\alpha\right) + (1/2)\gamma_\alpha^\alpha \mathbb{E}\left[\|\mathbf{X}_0 - x^\star\|^2\right] \\ &\quad + (\mathtt{C}_1/2)\left[(T+\gamma_\alpha)^\alpha - \gamma_\alpha^\alpha\right](T+\gamma_\alpha)^{1-2\alpha}\log(T+\gamma_\alpha) \\ &\quad + (\mathtt{C}_1/2)\log(T+\gamma_\alpha)\left[(T+\gamma_\alpha)^{1-\alpha} - \gamma_\alpha^{1-\alpha}\right](1 + \mathbf{\Psi}_\alpha(T+\gamma_\alpha)) \\ &\leq (\mathtt{C}_1/2)(T+\gamma_\alpha)^\alpha + (\mathtt{C}_1/2)\gamma_\alpha^\alpha \\ &\quad + (\mathtt{C}_1/2)(T+\gamma_\alpha)^{1-\alpha}\log(T+\gamma_\alpha) \\ &\quad + (\mathtt{C}_1/2)\log(T+\gamma_\alpha)(T+\gamma_\alpha)^{1-\alpha}(1 + \mathbf{\Psi}_\alpha(T+\gamma_\alpha)) .\end{aligned}$$

Using this result we have

$$\begin{aligned}S(T) - f^\star &\leq T^{-1}2\mathtt{C}_1(T+\gamma_\alpha)^{\max(1-\alpha,\alpha)}(1+\log(T+\gamma_\alpha))(1+\mathbf{\Psi}_\alpha(T+\gamma_\alpha)) + 2\mathtt{C}_1\gamma_\alpha^\alpha T^{-1}/2 \\ &\leq 4\mathtt{C}_1 T^{-\min(\alpha,1-\alpha)}(1+\log(T+\gamma_\alpha))(1+\mathbf{\Psi}_\alpha(T+\gamma_\alpha)) .\end{aligned}$$

■

Lemma 67 *Assume* **A**1*,* **A**2-(b)*,* **A**3*, and* **F**2-(b)*. In addition, assume that* (65) *holds. Then, for any $\alpha, \gamma \in (0,1)$ and $T \geq 0$ we have*

$$S(0) - S(1) \leq \mathtt{C}_1\mathtt{L}\log(T+\gamma_\alpha)(1+\mathbf{\Psi}_\alpha(T+\gamma_\alpha))(T-1)^{-2\alpha} ,$$

with S given in (66)*.*

Proof We have

$$S(0) - S(1) = \mathbb{E}\left[f(\mathbf{X}_T)\right] - S(1) = \int_{T-1}^T \left(\mathbb{E}\left[f(\mathbf{X}_T)\right] - \mathbb{E}\left[f(\mathbf{X}_s)\right]\right) \mathrm{d}s . \tag{71}$$

Using Lemma 48 on the stochastic process $f(\mathbf{X}_t)_{t\geq 0}$, **A**1 and (65), we have for all $s \in [T-1, T]$

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{X}_T)\right] - \mathbb{E}\left[f(\mathbf{X}_s)\right] &= -\int_s^T (\gamma_\alpha + t)^{-\alpha}\mathbb{E}[\|\nabla f(\mathbf{X}_t)\|^2]\mathrm{d}t \\ &\quad + (\mathtt{L}/2)\gamma_\alpha \int_s^T (t+\gamma_\alpha)^{-2\alpha}\mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_t))\right]\mathrm{d}t \\ &\leq (\mathtt{L}/2)\gamma_\alpha\mathtt{C}_0\log(T+\gamma_\alpha)(1+\mathbf{\Psi}_\alpha(T+\gamma_\alpha))\int_s^T (t+\gamma_\alpha)^{-2\alpha}\mathrm{d}t \\ &\leq \mathtt{C}_1\mathtt{L}\log(T+\gamma_\alpha)(1+\mathbf{\Psi}_\alpha(T+\gamma_\alpha))(s+\gamma_\alpha)^{-2\alpha}(T-s) .\end{aligned}$$

Plugging this result into (71) yields

$$
\begin{aligned}
S(0) - S(1) &\le \mathtt{C}_1 \mathtt{L} \log(T+\gamma_\alpha)(1+\boldsymbol{\Psi}_\alpha(T+\gamma_\alpha)) \int_{T-1}^{T} (T-s)(s+\gamma_\alpha)^{-2\alpha} \mathrm{d}s \\
&\le \mathtt{C}_1 \mathtt{L} \log(T+\gamma_\alpha)(1+\boldsymbol{\Psi}_\alpha(T+\gamma_\alpha))(T-1+\gamma_\alpha)^{-2\alpha} \\
&\le \mathtt{C}_1 \mathtt{L} \log(T+\gamma_\alpha)(1+\boldsymbol{\Psi}_\alpha(T+\gamma_\alpha))(T-1)^{-2\alpha} \; .
\end{aligned}
$$

■

Theorem 68 *Let $\alpha, \gamma \in (0,1)$ and $(\mathbf{X}_t)_{t\ge 0}$ be given by (2). Assume $f \in \mathrm{C}^2(\mathbb{R}^d, \mathbb{R})$,* **A**1*,* **A**2*-(b),* **A**3 *and* **F**2*-(b). Then, there exists $C \ge 0$ (explicit and given in the proof) such that for any $T \ge 1$*

$$
\mathbb{E}\left[f(\mathbf{X}_T)\right] - \min\nolimits_{\mathbb{R}^d} f \le C(1+\log(T))^2/T^{\alpha\wedge(1-\alpha)} \; .
$$

Proof We begin by proving by induction over $m \in \mathbb{N}^*$ that the following assertion **H**3(m) is true.

H3 (m) *For any $\alpha > 1/(m+1)$, there exists $\mathtt{C}_\alpha^+ > 0$ such that for all $t \ge 0$, $\mathbb{E}\left[\mathrm{Tr}(\Sigma(X_t))\right] \le \mathtt{C}_\alpha^+$. In addition, for any $\alpha \le 1/(m+1)$, there exists $\mathtt{C}_\alpha^- > 0$ such that for all $t \ge 0$, $\mathbb{E}\left[\mathrm{Tr}(\Sigma(X_t))\right] \le \mathtt{C}_\alpha^-(t+\gamma_\alpha)^{1-(m+1)\alpha}\log(t+\gamma_\alpha)^2$.*

For $m = 1$, **H**3(1) is an immediate consequence of **A**1 and Lemma 62. Now, let $m \in \mathbb{N}^*$ and suppose that **H**3(m) holds. Let $\alpha \in (0,1)$. Setting $\alpha^\star = 1/(m+1)$ we see that (65) is verified with $\beta = 1-(m+1)\alpha$. Consequently, using **A**1, **F**2-(b), **A**2-(b) we can apply Proposition 63 which shows that, for $\alpha \le 1/(m+1)$, there exists $\mathtt{C}^{\widetilde{(c)}} > 0$ such that for all $T \ge 1$,

$$
\begin{aligned}
\mathbb{E}\left[f(\mathbf{X}_T)\right] - f^\star &\le \mathtt{C}^{\widetilde{(c)}} \left\{ \log(T+\gamma_\alpha))^2/(T+\gamma_\alpha)^{\min(\alpha,1-\alpha)} \boldsymbol{\Psi}_\alpha(T+\gamma_\alpha) \right\} \\
&\le \mathtt{C}^{\widetilde{(c)}} \left\{ \log(T+\gamma_\alpha))^2 (T+\gamma_\alpha)^{-\alpha} (T+\gamma_\alpha)^{1-(m+1)\alpha} \right\} \\
&\le \mathtt{C}^{\widetilde{(c)}} \left\{ \log(T+\gamma_\alpha))^2 (T+\gamma_\alpha)^{1-(m+2)\alpha} \right\} \; . \qquad (72)
\end{aligned}
$$

In particular, if $\alpha > 1/(m+2)$ we have the existence of $\bar{\mathtt{C}}_\alpha > 0$ such that for all $n \in \{0, \cdots, N\}$, $\mathbb{E}\left[f(X_n)\right] - f^\star \le \bar{\mathtt{C}}_\alpha$. And using **A**1 and Lemma 42 we get that, for all $t \in [0,T]$,

$$
\mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_t))\right] \le \mathtt{L}_\mathtt{T}(1+\mathbb{E}\left[f(\mathbf{X}_t) - f^\star\right]) \le \mathtt{L}_\mathtt{T}(1+\bar{\mathtt{C}}_\alpha),
$$

Combining this result with (72), we get that **H**3$(m+1)$ holds with $\mathtt{C}_\alpha^+ = \mathtt{L}_\mathtt{T}(1+\bar{\mathtt{C}}_\alpha)$ and $\mathtt{C}_\alpha^- = \mathtt{C}^{\widetilde{(c)}}$. We conclude by recursion,

Now, let $\alpha \in (0,1)$. Since $\mathbb{R}$ is archimedean, there exists $m \in \mathbb{N}^*$ such that $\alpha > 1/(m+1)$ and therefore **H**3(m) shows the existence of $\mathtt{C}_0 > 0$ such that $\mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_t))\right] \le \mathtt{C}_0$ for all $t \in [0,T]$. Applying Proposition 63 gives the existence of $\mathtt{C}^{(c)} > 0$ such that for all $T \ge 1$

$$
\mathbb{E}\left[f(\mathbf{X}_T)\right] - f^\star \le \mathtt{C}^{(c)}(1+\log(T))^2/T^{\min(\alpha,1-\alpha)} \; ,
$$

concluding the proof. ■

F.2. Equivalent to Appendix E.3

First, we start with Lemma 69 which is an equivalent to Lemma 56. The discussion conducted at the begin of Appendix E.3 is still valid here (with changes in the bootstrapping used). Proposition 70, Lemma 71, Lemma 72 and Lemma 73 are the counterparts of Proposition 57, Lemma 58, Lemma 59 and Lemma 60 respectively. Finally, our main result is stated and proven in Theorem 74.

Lemma 69 *Assume* **A**1*,* **F**2-(b)*,* **A**2-(b)*. Then for any $\alpha, \gamma \in (0,1)$, there exists $\mathtt{C}_{1,\alpha}^{(d)} \geq 0$, $\mathtt{C}_{2,\alpha}^{(d)} \geq 0$ and a function $\mathbf{\Phi}_{\alpha}^{(d)} : \mathbb{R}_+ \to \mathbb{R}_+$ such that, for any $n \geq 0$,*

$$\mathbb{E}\left[\|X_n - x^\star\|^2\right] \leq \mathtt{C}_{1,\alpha}^{(d)} \mathbf{\Phi}_{\alpha}^{(d)}(n+1) + \mathtt{C}_{2,\alpha}^{(d)} \; .$$

And we have

$$\mathbf{\Phi}_{\alpha}^{(d)}(t) = \begin{cases} t^{1-2\alpha} & \text{if } \alpha < 1/2 \; , \\ \log(t) & \text{if } \alpha = 1/2 \; , \\ 0 & \text{if } \alpha > 1/2 \; . \end{cases}$$

The values of the constants are given by

$$\mathtt{C}_{1,\alpha}^{(d)} = \begin{cases} 2\gamma^2\eta(1-2\alpha)^{-1} & \text{if } \alpha < 1/2 \; , \\ \gamma^2\eta & \text{if } \alpha = 1/2 \; , \\ 0 & \text{if } \alpha > 1/2 \; . \end{cases}$$

$$\mathtt{C}_{2,\alpha}^{(d)} = \begin{cases} 2\max_{k \leq (\gamma\mathtt{L}/2)^{1/\alpha}} \mathbb{E}\left[\|X_k - x^\star\|^2\right] & \text{if } \alpha < 1/2 \; , \\ 2\max_{k \leq (\gamma\mathtt{L}/2)^{1/\alpha}} \mathbb{E}\left[\|X_k - x^\star\|^2\right] + 2\gamma^2\eta & \text{if } \alpha = 1/2 \; , \\ 2\max_{k \leq (\gamma\mathtt{L}/2)^{1/\alpha}} \mathbb{E}\left[\|X_k - x^\star\|^2\right] + \gamma^2\eta(2\alpha-1)^{-1} & \text{if } \alpha > 1/2 \; , \end{cases}$$

Proof Let $f : \mathbb{R}^d \to \mathbb{R}$ verifying assumptions **A**1 and **F**2-(b). We consider $(X_n)_{n\geq 0}$ satisfying (1). Let $x^\star \in \mathbb{R}^d$. We have, using (1), Lemma 41 and **F**2-(b) that for all $n \geq (\gamma\mathtt{L}_\mathtt{T})^{1/\alpha}$,

$$\begin{aligned}
&\mathbb{E}[\|X_{n+1} - x^\star\|^2 | \mathcal{F}_n] = \mathbb{E}[\|X_n - x^\star - \gamma(n+1)^{-\alpha}\nabla\tilde{f}(X_n, Z_{n+1})\|^2 | \mathcal{F}_n] \\
&\quad = \|X_n - x^\star\|^2 - 2\gamma/(n+1)^\alpha \langle X_n - x^\star, \mathbb{E}[\nabla\tilde{f}(X_n, Z_{n+1})|\mathcal{F}_n]\rangle \\
&\quad\quad + \gamma^2(n+1)^{-2\alpha}\mathbb{E}[\|\nabla\tilde{f}(X_n, Z_{n+1})\|^2|\mathcal{F}_n] \\
&\quad \leq \|X_n - x^\star\|^2 - 2\gamma/(n+1)^\alpha \langle X_n - x^\star, \nabla f(X_n)\rangle \\
&\quad\quad + \mathtt{L}_\mathtt{T}\gamma^2(n+1)^{-2\alpha}\left(f(X_n) - f(x^\star) + 1\right) \\
&\quad \leq \|X_n - x^\star\|^2 - 2\gamma/(n+1)^\alpha \langle X_n - x^\star, \nabla f(X_n)\rangle \\
&\quad\quad + \mathtt{L}_\mathtt{T}\gamma^2(n+1)^{-2\alpha}\langle\nabla f(X_n), X_n - x^\star\rangle + \mathtt{L}_\mathtt{T}\gamma^2(n+1)^{-2\alpha} \\
&\quad \leq \|X_n - x^\star\|^2 + \gamma/(n+1)^\alpha\langle\nabla f(X_n), X_n - x^\star\rangle\left[\mathtt{L}_\mathtt{T}\gamma/(n+1)^\alpha - 2\right] + \gamma^2\mathtt{L}_\mathtt{T}(n+1)^{-2\alpha} \\
&\quad \leq \|X_n - x^\star\|^2 + \gamma^2\mathtt{L}_\mathtt{T}(n+1)^{-2\alpha} \\
&\mathbb{E}[\|X_{n+1} - x^\star\|^2] \leq \mathbb{E}[\|X_n - x^\star\|^2] + \gamma^2\mathtt{L}_\mathtt{T}(n+1)^{-2\alpha} \; .
\end{aligned}$$

Summing the previous inequality leads to

$$\mathbb{E}\left[\|X_n - x^\star\|^2\right] - \mathbb{E}\left[\|X_0 - x^\star\|^2\right] \leq \gamma^2\mathtt{L}_\mathtt{T}\sum_{k=1}^n k^{-2\alpha} \; .$$

As in the previous proof we now distinguish three cases:

(a) If $\alpha < 1/2$, we have

$$\begin{aligned}\mathbb{E}\left[\|X_n - x^\star\|^2\right] &\leq \|X_0 - x^\star\|^2 + \gamma^2 \mathtt{L}_\mathtt{T}(1-2\alpha)^{-1}(n+1)^{1-2\alpha} \\ &\leq \|X_0 - x^\star\|^2 + 2\gamma^2 \mathtt{L}_\mathtt{T}(1-2\alpha)^{-1}n^{1-2\alpha} \ .\end{aligned}$$

(b) If $\alpha = 1/2$, we have $\mathbb{E}[\|X_n - x^\star\|^2] \leq \|X_0 - x^\star\|^2 + \gamma^2 \mathtt{L}_\mathtt{T}(\log(n) + 2)$.

(c) If $\alpha > 1/2$, we have $\mathbb{E}[\|X_n - x^\star\|^2] \leq \|X_0 - x^\star\|^2 + \gamma^2 \mathtt{L}_\mathtt{T}(2\alpha - 1)^{-1}$.

■

In order to prove the theorem we will need an intermediate proposition.

Proposition 70 *Let $\gamma, \alpha \in (0,1)$ and $x_0 \in \mathbb{R}^d$ and $(X_n)_{n\geq 0}$ be given by (1). Assume* **A**1*,* **F**2*-(b),* **A**2*-(b). In addition, assume that there exists $\alpha^\star \in [0,1/2]$, $\beta > 0$ and $\mathtt{C}_0 \geq 0$ such that for all $n \in \{0, \cdots, N\}$*

$$\mathbb{E}[\|\nabla \tilde{f}(X_n, Z)\|^2] \leq \begin{cases} \mathtt{C}_0(n+1)^\beta \log(n+1) & \text{if } \alpha \leq \alpha^\star \ , \\ \mathtt{C}_0 & \text{if } \alpha > \alpha^\star \ . \end{cases} \tag{73}$$

Then there exists $\tilde{\mathtt{C}}_\alpha \geq 0$ such that, for all $N \geq 1$,

$$\mathbb{E}[f(X_N)] - f^\star \leq \tilde{\mathtt{C}}_\alpha \left\{(1 + \log(N+1))^2/(N+1)^{\min(\alpha, 1-\alpha)} \mathbf{\Psi}_\alpha(N+1) + 1/(N+1)\right\} \ ,$$

with

$$\mathbf{\Psi}_\alpha(n) = \begin{cases} n^\beta & \text{if } \alpha \leq \alpha^\star \ , \\ 1 & \text{if } \alpha > \alpha^\star \ . \end{cases}$$

Proof Let $\alpha, \gamma \in (0,1)$ and $N \geq 1$. Let $(X_n)_{n\geq 0}$ be given by (1). And finally, combining (78) and (80) together gives, for $\tilde{\mathtt{C}}_\alpha = 2\max((2\gamma)^{-1}\|X_0 - x^\star\|^2, 2\mathtt{C}^{(d)})$,

■

Let $(S_k)_{k\in\{0,\cdots,N\}}$ defined for any $k \in \{0, \ldots, N\}$ by

$$S_k = (k+1)^{-1} \sum_{t=N-k}^{N} \mathbb{E}[f(X_t)] \ . \tag{74}$$

Note that $\mathbb{E}[f(X_N)] - f^\star = (S_N - S_0) + (S_0 - f^\star)$. We are now going to control each one of the two terms $(S_0 - S_N)$ and $(S_N - f^\star)$ as follows:

(a) Case $S_n - S_0$ (Lemma 72): this is an adaption of the idea of suffix averaging of Shamir and Zhang (2013) to our setting (one of crucial difference lies into the control of the sequence $(\mathbb{E}[\nabla f(X_n)]^2)_{n\in\mathbb{N}}$ which is assumed to be uniformly bounded in Shamir and Zhang (2013)). In particular, we control the (discrete) time-derivative of S in Lemma 71 (counterpart to Lemma 58). Note that Lemma 72 is the counterpart to Lemma 59.

(b) Case $S(T) - f^\star$ (Lemma 73): this result is known and corresponds to the optimal convergence rate of the averaged sequence towards the minimum of f. We provide its proof for completeness. Note that Lemma 73 is the counterpart to Lemma 60.

Before controlling each one of these terms we state the following useful lemma, which will allow us to control the derivative of S.

Lemma 71 *Assume* **A**1*,* **A**2-(b)*, and* **F**2-(b)*. In addition, assume that* (73) *holds. Then, for any* $\alpha, \gamma \in (0,1)$*,* $N \in \mathbb{N}$*,* $u \in \{0, \ldots, N\}$ *and* Y *any* $\mathbb{R}^d$*-valued random variable such that* $\mathbb{E}[\|Y - x^\star\|^2] \leq \mathtt{C}_{1,\alpha}^{(d)}(N+1)^{1-2\alpha}\log(N+1) + 4\mathtt{C}_{2,\alpha}^{(d)}$ *with* $\mathtt{C}_{1,\alpha}^{(d)}$ *and* $\mathtt{C}_{2,\alpha}^{(d)}$ *given in Lemma* 69*, we have*

$$\mathbb{E}\left[\sum_{k=N-u}^{N} f(X_k) - f(Y)\right] \leq 2\mathtt{C}^{(d)}(u+1)/(N+1)^{\min(\alpha,1-\alpha)}(1+\log(N+1))\boldsymbol{\Psi}_\alpha(N+1) \\ + (2\gamma)^{-1}(N-u+1)^\alpha \mathbb{E}[\|X_{N-u} - Y\|^2] \;,$$

with

$$\boldsymbol{\Psi}_\alpha(n) = \begin{cases} n^\beta & \text{if } \alpha \leq \alpha^\star \;, \\ 1 & \text{if } \alpha > \alpha^\star \;, \end{cases}$$

and $\mathtt{C}^{(d)} = 4((\gamma/2)\mathtt{C}_0)(1-\alpha)^{-1} + (2\gamma)^{-1}(\mathtt{C}_{2,\alpha}^{(d)} + 4\mathtt{C}_{1,\alpha}^{(d)})$.

Proof Let $\ell \in \{0, \cdots, N\}$, let $k \geq \ell$, let $Y \in \mathcal{F}_\ell$. Using **F**2-(b) we have

$$\begin{aligned}
\mathbb{E}\left[\|X_{k+1} - Y\|^2 \middle| \mathcal{F}_k\right] &= \mathbb{E}\left[\left\|X_k - Y - \gamma(k+1)^{-\alpha}\nabla\tilde{f}(X_k, Z_{k+1})\right\|^2 \middle| \mathcal{F}_k\right] \\
&= \|X_k - Y\|^2 + \gamma^2(k+1)^{-2\alpha}\mathbb{E}\left[\left\|\nabla\tilde{f}(X_k, Z_{k+1})\right\|^2 \middle| \mathcal{F}_k\right] \\
&\quad - 2\gamma(k+1)^{-\alpha}\langle X_k - Y, \nabla f(X_k)\rangle \\
\mathbb{E}\left[f(X_k) - f(Y)\right] &\leq (2\gamma)^{-1}(k+1)^\alpha\left(\mathbb{E}\left[\|X_k - Y\|^2\right] - \mathbb{E}\left[\|X_{k+1} - Y\|^2\right]\right) \\
&\quad + (\gamma/2)(k+1)^{-\alpha}\mathbb{E}\left[\mathbb{E}\left[\left\|\nabla\tilde{f}(X_k, Z_{k+1})\right\|^2 \middle| \mathcal{F}_k\right]\right] \\
\mathbb{E}\left[f(X_k) - f(Y)\right] &\leq (2\gamma)^{-1}(k+1)^\alpha\left(\mathbb{E}\left[\|X_k - Y\|^2\right] - \mathbb{E}\left[\|X_{k+1} - Y\|^2\right]\right) \\
&\quad + (\gamma/2)(k+1)^{-\alpha}\mathbb{E}\left[\left\|\nabla\tilde{f}(X_k, Z_{k+1})\right\|^2\right] \;. \qquad (75)
\end{aligned}$$

Let $u \in \{0, \cdots, N\}$. Summing now (75) between $k = N-u$ and $k = N$ gives

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=N-u}^{N} f(X_k) - f(Y)\right] &\leq (2\gamma)^{-1}\sum_{k=N-u+1}^{N}\mathbb{E}\left[\|X_k - Y\|^2\right]((k+1)^\alpha - k^\alpha) \\
&\quad + (\gamma/2)\sum_{k=N-u}^{N}\mathbb{E}\left[\left\|\nabla\tilde{f}(X_k, Z_{k+1})\right\|^2\right](k+1)^{-\alpha} \\
&\quad + (2\gamma)^{-1}(N-u+1)^\alpha\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] \;. \qquad (76)
\end{aligned}$$

In the following we will take for Y either $x^\star$ or X_m for $m \in [0, N]$. We now have to run separate analyses depending on the value of α.

(a) If $\alpha \leq \alpha^\star$, then (73) gives that

$$\mathbb{E}\left[\|\nabla f(X_k, Z_{k+1})\|^2\right] \leq \mathtt{C}_0 (N+1)^\beta \log(N+1),$$

and Lemma 69 gives that for all $k \in \{0, \ldots, N\}$,

$$\begin{aligned}\mathbb{E}\left[\|X_k - Y\|^2\right] &\leq 2\mathbb{E}\left[\|X_k - x^\star\|^2\right] + 2\mathbb{E}\left[\|Y - x^\star\|^2\right] \\ &\leq 2\mathtt{C}_{1,\alpha}^{(d)}(k+1)^{1-2\alpha}\log(k+1) + 2\mathtt{C}_{1,\alpha}^{(d)}(N+1)^{1-2\alpha}\log(N+1) + 4\mathtt{C}_{2,\alpha}^{(d)} \\ &\leq 4\mathtt{C}_{1,\alpha}^{(d)}(N+1)^{1-2\alpha}\log(N+1) + 4\mathtt{C}_{2,\alpha}^{(d)} \,.\end{aligned}$$

We define $\mathtt{C}_{3,\alpha}^{(d)} = 4\mathtt{C}_{2,\alpha}^{(d)}$. Using (76) with $\mathtt{C}^{(b)} = ((\gamma/2)\mathtt{C}_0)(1-\alpha)^{-1}$, we get

$$\begin{aligned}\mathbb{E}\left[\sum_{k=N-u}^{N} f(X_k) - f(Y)\right] &\leq (2\gamma)^{-1}(N-u+1)^\alpha \mathbb{E}\left[\|X_{N-u} - Y\|^2\right] \\ &\quad + (2\gamma)^{-1}\left(\mathtt{C}_{3,\alpha}^{(d)} + 4\mathtt{C}_{1,\alpha}^{(d)}(N+1)^{1-2\alpha}\log(N+1)\right)\left((N+1)^\alpha - (N-u+1)^\alpha\right) \\ &\quad + (\gamma/2)\mathtt{C}_0(N+1)^\beta\log(N+1)(1-\alpha)^{-1}\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right) \\ &\leq \mathtt{C}^{(b)}(N+1)^\beta\log(N+1)\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right) \\ &\quad + (2\gamma)^{-1}(N-u+1)^\alpha\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] \\ &\quad + (2\gamma)^{-1}\mathtt{C}_{3,\alpha}^{(d)}\left((N+1)^\alpha - (N-u)^\alpha\right) \\ &\quad + (2\gamma)^{-1}4\mathtt{C}_{1,\alpha}^{(d)}\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right)\log(N+1) \\ &\leq \mathtt{C}^{(d)}(N+1)^\beta(1+\log(N+1))\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right) \\ &\quad + (2\gamma)^{-1}(N-u+1)^\alpha\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] \,,\end{aligned}$$

where we used Lemma 50. Similarly to (69) we have

$$\begin{aligned}&(N+1)^{1-\alpha} - (N-u)^{1-\alpha} \\ &\qquad = \left\{\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right)\left((N+1)^\alpha + (N-u)^\alpha\right)\right\}\left((N+1)^\alpha + (N-u)^\alpha\right)^{-1} \\ &\qquad \leq 2(u+1)/(N+1)^\alpha \,.\end{aligned}$$

(b) If $\alpha \in (\alpha^\star, 1/2]$, then Lemma 69 gives that for all $k \in \{0, \ldots, N\}$,

$$\begin{aligned}\mathbb{E}\left[\|X_k - Y\|^2\right] &\leq 2\mathbb{E}\left[\|X_k - x^\star\|^2\right] + 2\mathbb{E}\left[\|Y - x^\star\|^2\right] \\ &\leq 2\mathtt{C}_{1,\alpha}^{(d)}(k+1)^{1-2\alpha}\log(k+1) + 2\mathtt{C}_{1,\alpha}^{(d)}(N+1)^{1-2\alpha}\log(N+1) + 4\mathtt{C}_{2,\alpha}^{(d)} \\ &\leq 4\mathtt{C}_{1,\alpha}^{(d)}(N+1)^{1-2\alpha}\log(N+1) + 4\mathtt{C}_{2,\alpha}^{(d)} \,.\end{aligned}$$

Combining (73) and (76) we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=N-u}^{N} f(X_k) - f(Y)\right] \leq (2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] \\
&\quad + (2\gamma)^{-1}\left(\mathtt{C}_{3,\alpha}^{(d)} + 4\mathtt{C}_{1,\alpha}^{(d)}\log(N+1)(N+1)^{1-2\alpha}\right)\left((N+1)^{\alpha} - (N-u+1)^{\alpha}\right) \\
&\quad + (\gamma/2)\mathtt{C}_0(1-\alpha)^{-1}\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right) \\
&\leq \mathtt{C}^{(b)}\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right) + (2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] \\
&\quad + (2\gamma)^{-1}\left(\mathtt{C}_{3,\alpha}^{(d)} + 4\mathtt{C}_{1,\alpha}^{(d)}\right)(1+\log(N+1))\left((N+1)^{\alpha} - (N-u)^{\alpha}\right) \\
&\leq \mathtt{C}^{(d)}(1+\log(N+1))\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right) \\
&\quad + (2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] .
\end{aligned}
$$

(c) If $\alpha > 1/2$, then $\alpha > \alpha^\star$ and Lemma 69 gives

$$
\forall k \in \{0, \ldots, N\}\,,\ \mathbb{E}\left[\|X_k - Y\|^2\right] \leq 2\mathbb{E}\left[\|X_k - x^\star\|^2\right] + 2\mathbb{E}\left[\|Y - x^\star\|^2\right] \leq 4\mathtt{C}_{2,\alpha}^{(d)} = \mathtt{C}_{3,\alpha}^{(d)} .
$$

Using Lemma 50, (73) and (76) we have

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=N-u}^{N} f(X_k) - f(Y)\right] \leq (\gamma\mathtt{C}_0/2)(1-\alpha)^{-1}\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right) \\
&\quad + (2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] \\
&\quad + (2\gamma)^{-1}\mathtt{C}_{3,\alpha}^{(d)}\left((N+1)^{\alpha} - (N-u+1)^{\alpha}\right) \\
&\leq \mathtt{C}^{(b)}\left((N+1)^{1-\alpha} - (N-u)^{1-\alpha}\right) + (2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] \\
&\quad + (2\gamma)^{-1}\mathtt{C}_{3,\alpha}^{(d)}\left((N+1)^{\alpha} - (N-u)^{\alpha}\right) \\
&\leq \mathtt{C}^{(d)}\left((N+1)^{\alpha} - (N-u)^{\alpha}\right) + (2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] .
\end{aligned}
$$

Similarly to (69) we have

$$
\begin{aligned}
(N+1)^{\alpha} - (N-u)^{\alpha} &= \left\{\left((N+1)^{\alpha} - (N-u)^{\alpha}\right)\left((N+1)^{1-\alpha} + (N-u)^{1-\alpha}\right)\right\} \\
&\qquad \times \left((N+1)^{1-\alpha} + (N-u)^{1-\alpha}\right)^{-1} \\
&\leq 2(u+1)/(N+1)^{1-\alpha} .
\end{aligned}
$$

Finally, putting the three cases above together we obtain

$$
\begin{aligned}
\mathbb{E}\left[\sum_{k=N-u}^{N} f(X_k) - f(Y)\right] &\leq 2\mathtt{C}^{(d)}(u+1)/(N+1)^{\min(\alpha,1-\alpha)}(1+\log(N+1))\boldsymbol{\Psi}_{\alpha}(N+1) \\
&\qquad + (2\gamma)^{-1}(N-u+1)^{\alpha}\mathbb{E}\left[\|X_{N-u} - Y\|^2\right] ,
\end{aligned}
$$

with

$$\mathbf{\Psi}_\alpha(n) = \begin{cases} n^\beta & \text{if } \alpha \leq \alpha^\star \ , \\ 1 & \text{if } \alpha > \alpha^\star \ . \end{cases}$$

Note that the additional $\log(N+1)$ factor can be removed if $\alpha \neq 1/2$. ■

Lemma 72 *Assume* **A**1*,* **A**2-(b) *and* **F**2-(b)*. In addition, assume that* (73) *holds. Then, for any* $\alpha, \gamma \in (0,1)$ *and* $N \in \mathbb{N}$ *we have*

$$S_0 - S_N \leq 2\mathtt{C}^{(d)}(N+1)^{-\min(\alpha,1-\alpha)}(1+\log(N+1))^2\mathbf{\Psi}_\alpha(N+1) \ .$$

with S *given in* (74)*.*

Proof Let $u \in \{0,\dots,N\}$. Using Lemma 71 with the choice $Y = X_{N-u}$ gives

$$\mathbb{E}\left[\sum_{k=N-u}^{N} f(X_k) - f(X_{N-u})\right] \leq 2\mathtt{C}^{(d)}(u+1)/(N+1)^{\min(\alpha,1-\alpha)}(1+\log(N+1))\mathbf{\Psi}_\alpha(N+1) \ .$$

And then,

$$\begin{aligned} S_u &= (u+1)^{-1}\sum_{k=N-u}^{N}\mathbb{E}\left[f(X_k)\right] \\ &\leq 2\mathtt{C}^{(d)}(N+1)^{-\min(\alpha,1-\alpha)}(1+\log(N+1))\mathbf{\Psi}_\alpha(N+1) + \mathbb{E}\left[f(X_{N-u})\right] \ . \end{aligned} \tag{77}$$

We have now, using (77),

$$\begin{aligned} uS_{u-1} &= (u+1)S_u - \mathbb{E}\left[f(X_{N-u})\right] \\ &= uS_u + S_u - \mathbb{E}\left[f(X_{N-u})\right] \\ &\leq uS_u + 2\mathtt{C}^{(d)}(N+1)^{-\min(\alpha,1-\alpha)}(1+\log(N+1))\mathbf{\Psi}_\alpha(N+1) \\ S_{u-1} - S_u &\leq 2\mathtt{C}^{(d)}u^{-1}(N+1)^{-\min(\alpha,1-\alpha)}\log(N+1) \\ S_0 - S_N &\leq 2\mathtt{C}^{(d)}(N+1)^{-\min(\alpha,1-\alpha)}(1+\log(N+1))\mathbf{\Psi}_\alpha(N+1)\sum_{u=1}^{N}(1/u) \\ S_0 - S_N &\leq 2\mathtt{C}^{(d)}(N+1)^{-\min(\alpha,1-\alpha)}(1+\log(N+1))^2\mathbf{\Psi}_\alpha(N+1) \ . \end{aligned} \tag{78}$$

■

Lemma 73 *Assume* **A**1*,* **A**2-(b) *and* **F**2-(b)*. In addition, assume that* (73) *holds. Then, for any* $\alpha, \gamma \in (0,1)$ *and* $N \in \mathbb{N}$ *we have*

$$\begin{aligned} S_N - f^\star \leq 2\mathtt{C}^{(d)}&(1+\log(N+1))^2(N+1)^{-\min(\alpha,1-\alpha)}\mathbf{\Psi}_\alpha(N+1) \\ &+ (2\gamma)^{-1}(N+1)^{-1}\left\|X_0 - x^\star\right\|^2 \ . \end{aligned} \tag{79}$$

with S *given in* (74)*.*

Proof Using Lemma 71 with the choice $Y = x^\star$ and $u = N$ gives

$$(N+1)^{-1}\mathbb{E}\left[\sum_{k=0}^{N} f(X_k) - f(x^\star)\right] \leq 2\mathtt{C}^{(d)}(1+\log(N+1))(N+1)^{-\min(\alpha,1-\alpha)}\boldsymbol{\Psi}_\alpha(N+1) \\ + (2\gamma)^{-1}(N+1)^{-1}\left\|X_0 - x^\star\right\|^2$$

Therefore,

$$S_N - f^\star \leq 2\mathtt{C}^{(d)}(1+\log(N+1))^2(N+1)^{-\min(\alpha,1-\alpha)}\boldsymbol{\Psi}_\alpha(N+1) \\ + (2\gamma)^{-1}(N+1)^{-1}\left\|X_0 - x^\star\right\|^2 \ . \tag{80}$$

■

Theorem 74 *Let $\gamma, \alpha \in (0,1)$ and $(X_n)_{n\geq 0}$ be given by (1). Assume* **A**1, **A**2-(b) *and* **F**2-(b). *Then, there exists $C \geq 0$ (explicit and given in the proof) such that for any $N \geq 1$,*

$$\mathbb{E}\left[f(X_N)\right] - \min_{\mathbb{R}^d} f \leq C(1+\log(N+1))^2/(N+1)^{\alpha\wedge(1-\alpha)} \ .$$

Proof We begin by proving by induction over $m \in \mathbb{N}^*$ that the following assertion **H**4(m) is true.

H4 (m) *For any $\alpha > 1/(m+1)$, there exists $\mathtt{C}_\alpha^+ > 0$ such that for all $n \in \mathbb{N}$, $\mathbb{E}[\|\nabla\tilde{f}(X_n, Z)\|^2] \leq \mathtt{C}_\alpha^+$. In addition, for any $\alpha \leq 1/(m+1)$, there exists $\mathtt{C}_\alpha^- > 0$ such that for all $n \in \mathbb{N}$, $\mathbb{E}[\|\nabla\tilde{f}(X_n, Z)\|^2] \leq \mathtt{C}_\alpha^- n^{1-(m+1)\alpha}(1+\log(n))^2$.*

For $m = 1$, **H**4(1) is an immediate consequence of **A**1 and Lemma 69, with $\mathtt{C}_\alpha^+ = \mathtt{L}^2\mathtt{C}_{2,\alpha}^{(d)}$ and $\mathtt{C}_\alpha^- = \mathtt{L}^2\max(\mathtt{C}_{1,\alpha}^{(d)}, \mathtt{C}_{2,\alpha}^{(d)})$. Now, let $m \in \mathbb{N}^*$ and assume that **H**4(m) holds. Let $\alpha \in (0,1)$. Setting $\alpha^\star = 1/m+1$ we see that (73) is verified with $\beta = 1-(m+1)\alpha$. Consequently, using **A**1, **F**2-(b), **A**2-(b) we can apply Proposition 70 which shows that, for $\alpha \leq 1/(m+1)$

$$\begin{aligned}\mathbb{E}\left[f(X_N)\right] - f^\star &\leq \tilde{\mathtt{C}}_\alpha\left\{(1+\log(N+1))^2/(N+1)^{\min(\alpha,1-\alpha)}\boldsymbol{\Psi}_\alpha(N+1) + 1/(N+1)\right\} \\ &\leq \tilde{\mathtt{C}}_\alpha\left\{(1+\log(N+1))^2(N+1)^{-\alpha}(N+1)^{1-(m+1)\alpha} + 1/(N+1)\right\} \\ &\leq \tilde{\mathtt{C}}_\alpha\left\{(1+\log(N+1))^2(N+1)^{1-(m+2)\alpha} + 1/(N+1)\right\} \ . \end{aligned} \tag{81}$$

In particular, if $\alpha > 1/(m+2)$ we have the existence of $\bar{\mathtt{C}}_\alpha > 0$ such that for all $n \in \{0, \cdots, N\}$, $\mathbb{E}\left[f(X_n)\right] - f^\star \leq \bar{\mathtt{C}}_\alpha$. And using **A**1 and Lemma 41 we get that, for all $n \in \{0, \cdots, N\}$

$$\mathbb{E}[\|\nabla\tilde{f}(X_n, Z)\|^2] \leq \mathtt{L}_\mathtt{T}(1 + \mathbb{E}\left[f(X_n) - f^\star\right]) \leq \mathtt{L}_\mathtt{T}(1 + \bar{\mathtt{C}}_\alpha) \ ,$$

Combining this result with (81), we get that **H**4$(m+1)$ holds with $\mathtt{C}_\alpha^+ = \mathtt{L}_\mathtt{T}(1+\bar{\mathtt{C}}_\alpha)$ and $\mathtt{C}_\alpha^- = 2\tilde{\mathtt{C}}_\alpha$. Finally this proves that **H**4(m) is true for any $n \geq 1$ by induction

Now, let $\alpha \in (0,1)$. Since $\mathbb{R}$ is archimedean, there exists $m \in \mathbb{N}^*$ such that $\alpha > 1/(m+1)$ and therefore **H**4(m) shows the existence of $\mathtt{C}_0 > 0$ such that $\mathbb{E}[\|\nabla\tilde{f}(X_n, Z)\|^2] \leq \mathtt{C}_0$ for all $n \in \mathbb{N}^*$. Applying Proposition 70 gives the existence of $\mathtt{C}^{(d)} > 0$ such that for all $N \geq 1$

$$\mathbb{E}\left[f(X_N)\right] - f^\star \leq \mathtt{C}^{(d)}(1+\log(N+1))^2/(N+1)^{\min(\alpha,1-\alpha)} \ ,$$

with $\mathtt{C}^{(d)} = 2\tilde{\mathtt{C}}_\alpha$, concluding the proof. ■

Appendix G. Weakly Quasi-Convex Case

In this section we give the proofs of the results presented in Section 5. We prove Theorem 10 in Appendix G.1. Technical lemmas are gathered in Appendix G.2. We control the norm of $\mathbb{E}[\|\mathbf{X}_t - x^\star\|^{2p}]^{1/p}$ in the convex framework in Appendix G.3. The proof of Appendix G.4 is presented in Appendix G.4. Its discrete counterpart is given Appendix G.5. Finally, we conclude this section with the proof of Theorem 12 in Appendix G.6.

G.1. Proof of Theorem 10

Without loss of generality, we assume that $f^\star = 0$. Let $\alpha, \gamma \in (0,1)$, $x_0 \in \mathbb{R}^d$, $a_t = \gamma_\alpha + t$, $\ell_t = 1 + \log(1 + \gamma_\alpha^{-1} t)$ for any $t \geq 0$ and $\delta = \min(\delta_1, \delta_2)$ with δ_1 and δ_2 given in Theorem 10. Using Lemma 48, we have for any $t \geq 0$

$$\begin{aligned}
&\mathbb{E}\left[f(\mathbf{X}_t) a_t^\delta \ell_t^{-\varepsilon}\right] - f(x_0)\gamma_\alpha^\delta \\
&\quad = \int_0^t \Big\{ -\ell_s^{-\varepsilon} a_s^{\delta-\alpha} \mathbb{E}[\|\nabla f(\mathbf{X}_s)\|^2] + (\gamma_\alpha/2)\ell_s^{-\varepsilon} a_s^{\delta-2\alpha} \mathbb{E}\left[\langle \nabla^2 f(\mathbf{X}_s), \Sigma(\mathbf{X}_s)\rangle\right] \\
&\qquad + \delta \ell_s^{-\varepsilon} a_s^{\delta-1} \mathbb{E}\left[f(\mathbf{X}_s)\right] - \varepsilon \ell_s^{-\varepsilon-1} a_s^{\delta} \mathbb{E}\left[f(\mathbf{X}_s)\right] \Big\} \mathrm{d}s \; .
\end{aligned}$$

Define for any $t \geq 0$, $\mathcal{E}(t) = \mathbb{E}[f(\mathbf{X}_t)] a_t^\delta \ell_t^{-\varepsilon}$. $(t \mapsto \mathcal{E}(t))$ is differentiable and using **A**1 and **A**2 we have for any $t > 0$,

$$\mathrm{d}\mathcal{E}(t)/\mathrm{d}t \leq -\ell_t^{-\varepsilon} a_t^{\delta-\alpha} \mathbb{E}\left[\|\nabla f(\mathbf{X}_t)\|^2\right] + (\gamma_\alpha/2)\ell_t^{-\varepsilon} a_t^{\delta-2\alpha} \mathtt{L}\eta + \delta a_t^{-1} \mathcal{E}(t) \; .$$

Using, **F**3 and Hölder's inequality we have for any $t \geq 0$

$$\tau \mathbb{E}\left[f(\mathbf{X}_t)\right] \leq \mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^{r_2 r_3}\right]^{r_3^{-1}} \mathbb{E}[\|\nabla f(\mathbf{X}_t)\|^2]^{r_1/2} \; .$$

Noting that $(r_3 r_1)^{-1} = r_1^{-1} - 1/2$, we get for any $t \geq 0$

$$\begin{aligned}
\mathbb{E}[\|\nabla f(\mathbf{X}_t)\|^2] &\geq \tau^{2r_1^{-1}} \mathbb{E}\left[f(\mathbf{X}_t)\right]^{2r_1^{-1}} \mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^{r_2 r_3}\right]^{1-2r_1^{-1}} \\
&\geq \tau^{2r_1^{-1}} C_{\beta,\varepsilon}^{1-2r_1^{-1}} a_t^{\beta(1-2r_1^{-1})} \ell_t^{\varepsilon(1-2r_1^{-1})} \mathbb{E}\left[f(\mathbf{X}_t)\right]^{2r_1^{-1}} \\
&\geq \tau^{2r_1^{-1}} C_{\beta,\varepsilon}^{1-2r_1^{-1}} a_t^{\beta(1-2r_1^{-1})-2r_1^{-1}\delta} \ell_t^{\varepsilon(1-2r_1^{-1})-2r_1^{-1}-\varepsilon} \mathcal{E}(t)^{2r_1^{-1}} \; .
\end{aligned}$$

Therefore, we have for any $t \geq 0$

$$\mathrm{d}\mathcal{E}(t)/\mathrm{d}t \leq -\tau^{2r_1^{-1}} C_{\beta,\varepsilon}^{1-2r_1^{-1}} a_t^{(1-2r_1^{-1})(\delta+\beta)-\alpha} \mathcal{E}(t)^{2r_1^{-1}} + \gamma_\alpha \ell_t^{-\varepsilon} a_t^{\delta-2\alpha} \mathtt{L}\eta + \delta a_t^{-1} \mathcal{E}(t) \; .$$

Let $\mathtt{D}_3 = \max(\mathtt{D}_1, \mathtt{D}_2)$ with

$$\begin{aligned}
\mathtt{D}_1 &= (|\delta| C_{\beta,\varepsilon}^{2r_1^{-1}-1} \tau^{-2r_1^{-1}} \gamma_\alpha^{(2r_1^{-1}-1)(\delta+\beta)+\alpha-1})^{(2r_1^{-1}-1)^{-1}} \; , \\
\mathtt{D}_2 &= ((\mathtt{L}\eta/2) C_{\beta,\varepsilon}^{2r_1^{-1}-1} \tau^{-2r_1^{-1}} \gamma_\alpha^{(2r_1^{-1}-1)(\delta+\beta)+\delta-\alpha+1})^{r_1/2} \; .
\end{aligned}$$

If $\mathcal{E}(t) \geq \mathtt{D}_3$ then $\mathrm{d}\mathcal{E}(t)/\mathrm{d}t \leq 0$. Let $\mathtt{D} = \max(\mathtt{D}_3, \mathcal{E}(0))$, then for any $t \geq 0$, $\mathcal{E}(t) \leq \mathtt{D}$, which concludes the proof.

G.2. Technical lemmas

Lemma 75 *Assume that f is continuous, that $x^\star \in \arg\min_{x\in\mathbb{R}^d} f(x)$ and that there exist $c, R \geq 0$ such that for any $x \in \mathbb{R}^d$ with $\|x - x^\star\| \geq R$ we have $f(x) - f(x^\star) \geq c\|x - x^\star\|$. Let $p \in \mathbb{N}$, X a d-dimensional random variable and $\mathtt{D}_4 \geq 1$ such that $\mathbb{E}[(f(X) - f(x^\star))^{2p}] \leq \mathtt{D}_4$. Then there exists $\mathtt{D}_5 \geq 0$ such that*

$$\mathbb{E}\left[\|X - x^\star\|^{2p}\right] \leq \mathtt{D}_5\mathtt{D}_4 \ .$$

Proof Since f is continuous there exists $\mathtt{a} \geq 0$ such that for any $x \in \mathbb{R}^d$, $f(x) - f(x^\star) \geq c\|x - x^\star\| - \mathtt{a}$. Therefore, using Jensen's inequality and that $\mathtt{D}_4 \geq 1$ we have

$$\begin{aligned}
\mathbb{E}\left[\|X - x^\star\|^{2p}\right] &\leq c^{-2p}\sum_{k=0}^{2p}\binom{k}{2p}\mathbb{E}\left[(f(X) - f(x^\star))^k\right]\mathtt{a}^{2p-k} \\
&\leq c^{-2p}\sum_{k=0}^{2p}\binom{k}{2p}\mathbb{E}\left[(f(X) - f(x^\star))^{2p}\right]^{k/(2p)}\mathtt{a}^{2p-k} \\
&\leq c^{-2p}\sum_{k=0}^{2p}\binom{k}{2p}\mathtt{D}_4^{k/(2p)}\mathtt{a}^{2p-k} \leq \mathtt{D}_5\mathtt{D}_4 \ ,
\end{aligned}$$

with $\mathtt{D}_5 = c^{-2p}\sum_{k=0}^{2p}\binom{k}{2p}\mathtt{a}^{2p-k}$. ■

Lemma 76 *Assume* **F3** *with $r_1 = r_2 = 1$. Then for any $p \in \mathbb{N}$ with $p \geq 2$ and d-dimensional random variable X we have*

$$\mathbb{E}\left[\|\nabla f(X)\|^2 (f(X) - f(x^\star))^{p-1}\right] \geq \mathbb{E}\left[(f(X) - f(x^\star))^p\right]^{1+1/p}\mathbb{E}\left[\|X - x^\star\|^{2p}\right]^{-1/p} \ ,$$

Proof Let $p \in \mathbb{N}$ with $p \geq 2$ and let $\varpi = 2p/(p+1)$. Using **F3** we have for any $x \in \mathbb{R}^d$

$$\|x - x^\star\|^{\varpi}\|\nabla f(x)\|^{\varpi}(f(x) - f(x^\star))^{\varpi(p-1)/2} \geq (f(x) - f(x^\star))^{\varpi(p+1)/2} \geq (f(x) - f(x^\star))^p \ .$$

Let $\varsigma = 2\varpi^{-1} = 1 + p^{-1}$ and $\varkappa$ such that $\varsigma^{-1} + \varkappa^{-1} = 1$. Using Hölder's inequality the fact that $\varkappa\varpi = 2p$ we have

$$\begin{aligned}
&\mathbb{E}\left[\|X - x^\star\|^{\varpi}\|\nabla f(X)\|^{\varpi}(f(X) - f(x^\star))^{\varpi(p-1)/2}\right] \\
&\qquad\leq \mathbb{E}\left[\|\nabla f(X)\|^2 (f(X) - f(x^\star))^{p-1}\right]^{1/\varsigma}\mathbb{E}\left[\|X - x^\star\|^{2p}\right]^{1/\varkappa} \ .
\end{aligned}$$

Since, $\varkappa^{-1} = (1+p)^{-1}$ we have

$$\mathbb{E}\left[\|\nabla f(X)\|^2 (f(X) - f(x^\star))^{p-1}\right] \geq \mathbb{E}\left[(f(X) - f(x^\star))^p\right]^{1+1/p}\mathbb{E}\left[\|X - x^\star\|^{2p}\right]^{-1/p} \ ,$$

which concludes the proof. ■

Lemma 77 *Let $\alpha, \gamma \in (0,1)$. Assume that* **F**3b *holds then for any $p \in \mathbb{N}$, there exists $\mathtt{D}_{p,4} \geq 0$ such that for any $t \geq 0$*

$$\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^{2p}\right]^{1/p} \leq \mathtt{D}_{p,4}\left\{1 + (\gamma_\alpha + t)^{1-2\alpha}\right\} \; .$$

Proof Let $\alpha, \gamma \in (0,1)$ and $p \in \mathbb{N}$. Let $\mathcal{E}_{t,p} = \mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^{2p}\right]$. Using Lemma 47 and Lemma 48 we have for any $t > 0$

$$\begin{aligned}
\mathrm{d}\mathcal{E}_{t,p}/\mathrm{dt} &= -2p(\gamma_\alpha + t)^{-\alpha}\mathbb{E}\left[\langle \nabla f(\mathbf{X}_t), \mathbf{X}_t - x^\star\rangle \|\mathbf{X}_t - x^\star\|^{2(p-1)}\right] \\
&\quad + p\gamma_\alpha(\gamma_\alpha + t)^{-2\alpha}\left\{\mathbb{E}\left[\mathrm{Tr}(\Sigma(\mathbf{X}_t))\,\|\mathbf{X}_t - x^\star\|^{2(p-1)}\right]\right. \\
&\quad \left. +2(p-1)\mathbb{E}\left[\langle (\mathbf{X}_t - x^\star)^\top(\mathbf{X}_t - x^\star), \Sigma(\mathbf{X}_t)\rangle\,\|\mathbf{X}_t - x^\star\|^{2(p-2))}\right]\right\} \\
&\leq 2p\gamma_\alpha\eta(2p-1)(\gamma_\alpha + t)^{-2\alpha}\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^{2(p-1)}\right] \\
&\leq p\gamma_\alpha\eta(2p-1)(\gamma_\alpha + t)^{-2\alpha}\mathcal{E}_{t,(p-1)} \; . \qquad (82)
\end{aligned}$$

If $p = 1$, the proposition holds and by recursion and using (82) we obtain the result for $p \in \mathbb{N}$. ■

G.3. Control of the norm in the convex case

Proposition 78 *Let $\alpha, \gamma \in (0,1)$. Let $m \in [0,2]$ and $\varphi > 0$ such that for any $p \in \mathbb{N}$ there exists $\mathtt{D}_{p,2} \geq 0$ such that for any $t \geq 0$, $\mathbb{E}[\|\mathbf{X}_t - x^\star\|^{2p}]^{1/p} \leq \mathtt{D}_{p,1}\{1 + (\gamma_\alpha + t)^{m-\varphi\alpha}\}$. Assume* **A**1 *and* **F**3b *and that there exist $R \geq 0$ and $c > 0$ such that for any $x \in \mathbb{R}^d$, with $\|x\| \geq R$, $f(x) - f(x^\star) \geq c\,\|x - x^\star\|$. Then, for any $p \in \mathbb{N}$, there exists $\mathtt{D}_{p,2} \geq 0$ such that for any $t \geq 0$,*

$$\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^{2p}\right]^{1/p} \leq \mathtt{D}_{p,2}\{1 + (\gamma_\alpha + t)^{m-(1+\varphi)\alpha}\} \; .$$

Proof If $\alpha \geq m/\varphi$ the proof is immediate since $\sup_{t\geq 0}\{\mathbb{E}[\|\mathbf{X}_t - x^\star\|^{2p}]^{1/p}\} < +\infty$. Now assume that $\alpha < m/\varphi$. Let $p \in \mathbb{N}$, $\delta_p = p(1+\varphi)\alpha - pm$ and $(t \mapsto \mathcal{E}_{t,p})$ such that for any $t \geq 0$, $\mathcal{E}_{t,p} = (f(\mathbf{X}_t) - f(x^\star))^{2p}(\gamma_\alpha + t)^{\delta_p}$. Using Lemma 48 we have for any $t > 0$

$$\begin{aligned}
\mathrm{d}\mathcal{E}_{t,p}/\mathrm{dt} &= -2p(\gamma_\alpha + t)^{-\alpha+\delta_p}\mathbb{E}\left[\|\nabla f(\mathbf{X}_t)\|^2\,(f(\mathbf{X}_t) - f(x^\star))^{2p-1}\right] \qquad (83) \\
&\quad + p\gamma_\alpha(\gamma_\alpha + t)^{-2\alpha+\delta_p}\left\{\mathbb{E}\left[\langle \nabla^2 f(\mathbf{X}_t), \Sigma(\mathbf{X}_t)\rangle(f(\mathbf{X}_t) - f(x^\star))^{2p-1}\right]\right. \\
&\quad \left. + (2p-1)\mathbb{E}\left[\langle \nabla f(\mathbf{X}_t)\nabla f(\mathbf{X}_t)^\top, \Sigma(\mathbf{X}_t)\rangle(f(\mathbf{X}_t) - f(x^\star)^{2p-2})\right]\right\} + \delta_p(\gamma_\alpha + t)^{-1}\mathcal{E}_{t,p} \; .
\end{aligned}$$

Combining (83), Lemma 47, Lemma 40, Lemma 76 and the fact that for any $t \geq 0$, $\mathbb{E}[\|\mathbf{X}_t - x^\star\|^{4p}]^{1/(2p)} \leq \mathtt{D}_{p,1}\{1 + (\gamma_\alpha + t)^{m - \varphi\alpha}\}$ we get

$$\begin{aligned}
\mathrm{d}\mathcal{E}_{t,p}/\mathrm{d}t \leq & -2p(\gamma_\alpha + t)^{-\alpha + \delta_p} \mathbb{E}\left[(f(\mathbf{X}_t) - f(x^\star))^{2p}\right]^{1+1/(2p)} \mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^{4p}\right]^{-1/(2p)} \\
& + p\gamma_\alpha(\gamma_\alpha + t)^{-2\alpha + \delta_p} \left\{\mathtt{L}\eta\mathbb{E}\left[(f(\mathbf{X}_t) - f(x^\star))^{2p-1}\right]\right. \\
& \left. + \mathtt{L}(2p-1)\eta^2\mathbb{E}\left[(f(\mathbf{X}_t) - f(x^\star))^{2p-1}\right]\right\} + \delta_p(\gamma_\alpha + t)^{-1}\mathcal{E}_{t,p} \\
\leq & -2p(\gamma_\alpha + t)^{-\alpha - \delta_p/(2p)} \mathcal{E}_{t,p}^{1+1/(2p)} \mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^{2p}\right]^{-1/(2p)} \\
& + p\gamma_\alpha(d + 2p - 1)\mathtt{L}\eta(1+\eta)(\gamma_\alpha + t)^{-2\alpha + \delta_p/(2p)} \mathcal{E}_{t,p}^{1-1/(2p)} + \delta_p(\gamma_\alpha + t)^{-1}\mathcal{E}_{t,p} \\
\leq & -2p(\gamma_\alpha + t)^{-\alpha - \delta_p/(2p)} \mathcal{E}_{t,p}^{1+1/(2p)} \mathtt{D}_{p,1}^{-1}\{1 + (\gamma_\alpha + t)^{m - \varphi\alpha}\}^{-1} \\
& + p\gamma_\alpha(d + 2p - 1)\mathtt{L}\eta(1+\eta)(\gamma_\alpha + t)^{-2\alpha + \delta_p/p} \mathcal{E}_{t,p}^{1-1/p} + \delta_p(\gamma_\alpha + t)^{-1}\mathcal{E}_{t,p} \\
\leq & -2p(\gamma_\alpha + t)^{(\varphi-1)\alpha - \delta_p/(2p) - m} \mathcal{E}_{t,p}^{1+1/(2p)} \mathtt{D}_{p,1}^{-1}\{1 + (\gamma_\alpha + t)^{-m + \varphi\alpha}\}^{-1} \\
& + 2p\gamma_\alpha(d + 2p - 1)\mathtt{L}\eta(1+\eta)(\gamma_\alpha + t)^{-2\alpha + \delta_p/(2p)} \mathcal{E}_{t,p}^{1-1/(2p)} + \delta_p(\gamma_\alpha + t)^{-1}\mathcal{E}_{t,p} \\
\leq & -p\mathtt{D}_{p,1}^{-1}\{1 + \gamma_\alpha^{-m + \varphi\alpha}\}^{-1}(\gamma_\alpha + t)^{(\varphi-1)\alpha - \delta_p/(2p) - m} \mathcal{E}_{t,p}^{1+1/(2p)} \\
& + 2p\gamma_\alpha(d + 2p - 1)\mathtt{L}\eta(1+\eta)(\gamma_\alpha + t)^{-2\alpha + \delta_p/(2p)} \mathcal{E}_{t,p}^{1-1/(2p)} + \delta_p(\gamma_\alpha + t)^{-1}\mathcal{E}_{t,p} \; .
\end{aligned}$$

Since $m \in [0, 2]$, we have that $1 - m + (\varphi - 1)\alpha \geq (1 + \varphi)\alpha/2 - m/2$. Hence,

$$(1 - \varphi)\alpha - \delta_p/(2p) - m \leq 2\alpha + \delta_p/(2p) \; , \qquad (1 - \varphi)\alpha - \delta_p/(2p) - m \leq 1 \; .$$

Therefore, using Lemma 3, there exists $\mathtt{D}_p^{(a)} \geq 1$ such that for any $t \geq 0$, $\mathcal{E}_{t,p} \leq \mathtt{D}_p^{(a)}$. Hence, for any $t \geq 0$,

$$\mathbb{E}\left[(f(\mathbf{X}_t) - f(x^\star))^{2p}\right] \leq \mathtt{D}_p^{(a)}(1 + (\gamma_\alpha + t)^{pm - p(1+\varphi)\alpha}) \; .$$

Using Lemma 75, there exists $\mathtt{D}_5 \geq 0$ such that

$$\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^{2p}\right] \leq \mathtt{D}_5(1 + (\gamma_\alpha + t)^{pm - p(1+\varphi)\alpha}) \; ,$$

which concludes the proof upon using that for any $a, b \geq 0$, $(a + b)^{1/2} \leq a^{1/2} + b^{1/2}$. ■

The following corollary is of independent interest.

Corollary 79 *Let $\alpha, \gamma \in (0, 1)$. Assume* **F**2 *and that* $\arg\min_{\mathbb{R}^d} f$ *is bounded. Then, for any $p \geq 0$ and $t \geq 0$,*

$$\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^p\right] < +\infty \; .$$

Proof Without loss of generality we assume that $x^\star = 0$ and $f(x^\star) = 0$. First, since $\arg\min_{\mathbb{R}^d} f$ is bounded, there exists $\tilde{R} \geq 0$ such that for any $x \in \mathbb{R}^d$ with $\|x\| \geq \tilde{R}$, $f(x) > 0$. Let $\mathsf{S} = \{x \in \mathbb{R}^d, \|x\| = 1\}$ and consider $m : \mathsf{S} \to (0, +\infty)$ such that for any $\theta \in \mathsf{S}$, $m(\theta) = f(\tilde{R}\theta)$. m is continuous since f is convex and therefore it attains its minimum and there exists $m^\star > 0$ such that for any $\theta \in \mathsf{S}$, $m(\theta) \geq m^\star$. Let $x \in \mathbb{R}^d$ with $\|x\| \geq 2\tilde{R}$. Since $f_x : [0, +\infty) \to \mathbb{R}$ such that $f_x(t) = f(tx)$ is convex we have

$$(f(x) - f(\tilde{R}x/\|x\|))(\|x\| - \tilde{R})^{-1} \geq (f(\tilde{R}x/\|x\|))\tilde{R}^{-1} \geq m^\star\tilde{R}^{-1} \; .$$

Therefore, there exists $c > 0$ and $R \geq 0$ such that for any $x \in \mathbb{R}^d$ with $\|x\| \geq R$, $f(x) \geq c\|x\|$. Let $p \in \mathbb{N}$. Noticing that **F**2 implies that **F**3b holds we can apply Lemma 77 and Proposition 78 with $m = 1$ and $\varphi = 2$. Applying repeatedly Proposition 78 we obtain that there exists $\mathtt{D}_p \geq 0$ such that

$$\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^{2p}\right]^{1/p} \leq \mathtt{D}_p\{1 + (\gamma_\alpha + t)^{m - \lceil \alpha^{-1} \rceil \alpha}\} \\ \leq \mathtt{D}_p\{1 + (\gamma_\alpha + t)^{m - \lceil m/\alpha \rceil \alpha}\} \leq \mathtt{D}_p\{1 + \gamma_\alpha^{m - \lceil m/\alpha \rceil \alpha}\} ,$$

which concludes the proof. ■

G.4. Proof of Corollary 11

Let $\alpha, \gamma \in (0, 1)$ and $X_0 \in \mathbb{R}^d$. Using Lemma 48, we have for any $t \geq 0$

$$\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^2\right] = \|X_0 - x^\star\|^2 - \int (\gamma_\alpha + s)^{-\alpha} \langle f(\mathbf{X}_s), \mathbf{X}_s - x^\star \rangle \mathrm{d}s \\ + (\gamma_\alpha/2) \int (\gamma_\alpha + s)^{-2\alpha} \langle \Sigma(\mathbf{X}_s), \nabla^2 f(\mathbf{X}_s) \rangle \mathrm{d}s . \tag{84}$$

Let $\mathcal{E}_t = \mathbb{E}[\|\mathbf{X}_t - x^\star\|^2]$. Using, (84) we have for any $t \geq 0$,

$$\mathcal{E}_t' \leq -(\gamma_\alpha + t)^{-\alpha} \mathbb{E}\left[\langle \nabla f(\mathbf{X}_s), \mathbf{X}_s - x^\star \rangle\right] + (\gamma_\alpha \mathtt{L}\eta/2)(\gamma_\alpha + t)^{-2\alpha} . \tag{85}$$

We divide the proof into three parts.

(a) First, assume that **F**3b holds. Combining this result and (85), we get that for any $t \geq 0$, $\mathcal{E}_t' \leq \gamma_\alpha \mathtt{L}\eta^2 d(\gamma_\alpha + t)^{-2\alpha}$. Therefore, there exist $\beta, \varepsilon \geq 0$ and $C_{\beta,\varepsilon} \geq 0$ such that $\mathbb{E}[\|\mathbf{X}_t - x^\star\|^2] < C_{\beta,\varepsilon}(\gamma_\alpha + t)^{-\beta}(1 + \log(1 + \gamma_\alpha^{-1} t))^\varepsilon$ with $\beta = 0$ and $\varepsilon = 0$ if $\alpha > 1/2$, $\beta = 1 - 2\alpha$ and $\varepsilon = 0$ if $\alpha < 1/2$ and $\beta = 0$ and $\varepsilon = 1$ if $\alpha = 1/2$. Combining this result and Theorem 10 concludes the proof.

(b) We can apply Lemma 77 and Proposition 78 with $m = 1$ and $\varphi = 2$. Applying repeatedly Proposition 78 we obtain that there exists $\mathtt{D}_p \geq 0$ such that

$$\mathbb{E}\left[\|\mathbf{X}_t - x^\star\|^{2p}\right]^{1/p} \\ \leq \mathtt{D}_p\{1 + (\gamma_\alpha + t)^{m - \lceil \alpha^{-1} \rceil \alpha}\} \leq \mathtt{D}_p\{1 + (\gamma_\alpha + t)^{m - \lceil m/\alpha \rceil \alpha}\} \leq \mathtt{D}_p\{1 + \gamma_\alpha^{m - \lceil m/\alpha \rceil \alpha}\} ,$$

which concludes the proof.

(c) Finally, assume that there exists $R \geq 0$ such that for any $x \in \mathbb{R}^d$ with $\|x\| \geq R$, $\langle \nabla f(x), x - x^\star \rangle \geq \mathtt{m}\, \|x - x^\star\|^2$. Therefore, since $(x \mapsto \nabla f(x))$ is continuous, there exists $\mathtt{a} \geq 0$ such that for any $x \in \mathbb{R}^d$, $\langle \nabla f(x), x - x^\star \rangle \geq \mathtt{m}\, \|x - x^\star\|^2 - \mathtt{a}$. Combining this result and (85), we get that for any $t \geq 0$,

$$\mathcal{E}_t' \leq -\mathtt{m}(\gamma_\alpha + t)^{-\alpha} \mathcal{E}_t + (\gamma_\alpha + t)^{-\alpha} \mathtt{a} + \gamma_\alpha \mathtt{L}\eta(\gamma_\alpha + t)^{-2\alpha}$$

Hence, if $\mathcal{E}_t \geq \max(\mathtt{a}/\mathtt{m}, \mathtt{L}\eta)$ we have that $\mathcal{E}_t' \leq 0$ and for any $t \geq 0$, $\mathcal{E}_t \leq \max(\mathtt{a}/\mathtt{m}, \mathtt{L}\eta, \mathcal{E}_0)$ and is bounded. Therefore, there exist $\beta, \varepsilon \geq 0$ and $C_{\beta,\varepsilon} \geq 0$ such that $\mathbb{E}[\|\mathbf{X}_t - x^\star\|^2] < C_{\beta,\varepsilon}(\gamma_\alpha + t)^{-\beta}(1 + \log(1 + \gamma_\alpha^{-1} t))^\varepsilon$ with $\beta = \varepsilon = 0$, which concludes the proof.

G.5. Discrete counterpart of Corollary 11

Corollary 80 *Let $\alpha, \gamma \in (0,1)$ and $x_0 \in \mathbb{R}^d$. Assume* **A**1*,* **A**2*. Then we have:*

(a) if **F**3b *holds then, there exists $\mathtt{D} \geq 0$ such that for any $N \in \mathbb{N}^\star$*

$$\mathbb{E}\left[f(X_N)\right] - f^\star \leq \mathtt{D}\left[N^{(1-3\alpha)/2} + N^{-\alpha/2} + N^{\alpha-1}\right] ,$$

(b) if **F**3 *holds and if there exists $R \geq 0$ such that for any $x \in \mathbb{R}^d$ with $\|x\| \geq R$, $\langle \nabla f(x), x - x^\star \rangle \geq \mathtt{m}\|x - x^\star\|^2$, then there exists $\mathtt{D} \geq 0$ such that for any $N \in \mathbb{N}^\star$*

$$\mathbb{E}\left[f(X_N)\right] - f^\star \leq \mathtt{D}\left[N^{-\alpha/2} + N^{\alpha-1}\right] .$$

Proof Let $\alpha, \gamma \in (0,1)$ and $x_0 \in \mathbb{R}^d$. We have for any $n \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\|X_{n+1} - x^\star\|^2\right] &= \mathbb{E}\left[\|X_n - x^\star\|^2\right] + 2\mathbb{E}\left[\langle X_n - x^\star, X_{n+1} - X_n\rangle\right] + \mathbb{E}\left[\|X_{n+1} - X_n\|^2\right] \\ &\leq \mathbb{E}\left[\|X_n - x^\star\|^2\right] - 2\gamma(n+1)^{-\alpha}\mathbb{E}\left[\langle X_n - x^\star, \nabla f(X_n)\rangle\right] \\ &\quad + 2\gamma^2(n+1)^{-2\alpha}\mathbb{E}\left[\|\nabla f(X_n)\|^2\right] + 2\gamma(n+1)^{-2\alpha}\eta . \end{aligned} \tag{86}$$

We now divide the proof into two parts.

(a) Using **F**3b and Lemma 40 we have for any $x \in \mathbb{R}^d$,

$$\langle \nabla f(x), x - x^\star \rangle \geq \tau(f(x) - f(x^\star)) \geq \tau \|\nabla f(x)\|^2 /(2\mathtt{L}) . \tag{87}$$

Using **A**1, (86) and (87) we have for any $n \geq (4\gamma\mathtt{L}/\tau)^{1/\alpha}$

$$\begin{aligned}\mathbb{E}\left[\|X_{n+1} - x^\star\|^2\right] &\leq \mathbb{E}\left[\|X_n - x^\star\|^2\right] + 2\gamma(n+1)^{-\alpha}(-\tau/(2\mathtt{L}) + \gamma(n+1)^{-\alpha})\mathbb{E}\left[\|\nabla f(X_n)\|^2\right] \\ &\quad + 2\gamma(n+1)^{-2\alpha}\eta \\ &\leq \mathbb{E}\left[\|X_n - x^\star\|^2\right] + 2\gamma(n+1)^{-2\alpha}\eta . \end{aligned}$$

Therefore, there exist $\beta, \varepsilon \geq 0$ and $C_{\beta,\varepsilon} \geq 0$ such that $\mathbb{E}[\|\mathbf{X}_n - x^\star\|^2] < C_{\beta,\varepsilon}(n+1)^{-\beta}(1 + \log(1+n))^\varepsilon$ with $\beta = 0$ and $\varepsilon = 0$ if $\alpha > 1/2$, $\beta = 1 - 2\alpha$ and $\varepsilon = 0$ if $\alpha < 1/2$ and $\beta = 0$ and $\varepsilon = 1$ if $\alpha = 1/2$. Combining this result and Theorem 12 concludes the proof.

(b) Finally, assume that there exists $R \geq 0$ such that for any $x \in \mathbb{R}^d$ with $\|x\| \geq R$, $\langle \nabla f(x), x - x^\star \rangle \geq \mathtt{m}\|x - x^\star\|^2$. Therefore, since $(x \mapsto \nabla f(x))$ is continuous, there exists $\mathtt{a} \geq 0$ such that for any $x \in \mathbb{R}^d$, $\langle \nabla f(x), x - x^\star \rangle \geq \mathtt{m}\|x - x^\star\|^2 - \mathtt{a}$. Combining this result and (86) we get that for any $n \in \mathbb{N}$ such that $n \geq (2/\gamma)^{\alpha^{-1}}$

$$\mathbb{E}\left[\|X_{n+1} - x^\star\|^2\right] \leq (1 - \gamma(n+1)^{-\alpha})\mathbb{E}\left[\|X_n - x^\star\|^2\right] + 2\gamma(n+1)^{-\alpha}\mathtt{a} + 2\gamma^2(n+1)^{-2\alpha}\eta .$$

Hence, if $n \geq (2/\gamma)^{-\alpha^{-1}}$ and $\mathbb{E}[\|X_n - x^\star\|^2] \geq \max(2\mathtt{a}, 2\gamma\eta)$ then $\mathbb{E}[\|X_{n+1} - x^\star\|^2] \leq \mathbb{E}[\|X_n - x^\star\|^2]$. Therefore, we obtain by recursion that for any $n \in \mathbb{N}$, that $(\mathbb{E}[\|X_n - x^\star\|^2])_{n\in\mathbb{N}}$ is bounded which concludes the proof by applying Theorem 12.

■

G.6. Proof of Theorem 12

Without loss of generality, we assume that $f^\star = 0$. Let $\alpha, \gamma \in (0,1)$, $x_0 \in \mathbb{R}^d$. Let $\delta = \min(\delta_1, \delta_2)$, with δ_1, δ_2 given in Theorem 12 and let $(E_k)_{k \in \mathbb{N}}$ such that for any $k \in \mathbb{N}$, $E_k = (k+1)^\delta \mathbb{E}\left[f(X_k)\right](1+\log(k+1))^{-\varepsilon}$. There exists $c_\delta \in \mathbb{R}$ such that for any $x \in [0,1]$, $(1+x)^\delta \leq 1 + c_\delta x$. Hence, for any $n \in \mathbb{N}$ we have

$$(n+2)^\delta - (n+1)^\delta \leq (n+1)^\delta \left\{(1+(n+1)^{-1})^\delta - 1\right\} \leq c_\delta (n+1)^{\delta-1} . \tag{88}$$

Using (Nesterov, 2004, Lemma 1.2.3) and **A**2 we have for any $n \in \mathbb{N}$ such that $n \geq (2\mathtt{L}\gamma)^{1/\alpha}$

$$\begin{aligned}
&\mathbb{E}\left[f(X_{n+1}) | \mathcal{F}_n\right] \leq f(X_n) - \gamma(n+1)^{-\alpha}\mathbb{E}\left[\langle \nabla f(X_n), H(X_n, Z_{n+1})\rangle | \mathcal{F}_n\right] \\
&\qquad + (\mathtt{L}/2)\gamma^2(n+1)^{-2\alpha}\mathbb{E}\left[\left\|H(X_n, Z_{n+1})\right\|^2 \middle| \mathcal{F}_n\right] \\
\mathbb{E}\left[f(X_{n+1})\right] &\leq \mathbb{E}\left[f(X_n)\right] - \gamma(n+1)^{-\alpha}\mathbb{E}\left[\left\|\nabla f(X_n)\right\|^2\right] \\
&\qquad + \mathtt{L}\gamma^2(n+1)^{-2\alpha}\mathbb{E}\left[\left\|\nabla f(X_n)\right\|^2\right] + \mathtt{L}\gamma^2(n+1)^{-2\alpha}\eta \\
&\leq \mathbb{E}\left[f(X_n)\right] - \gamma(n+1)^{-\alpha}\left\{1 - \mathtt{L}\gamma(n+1)^{-\alpha}\right\}\mathbb{E}\left[\left\|\nabla f(X_n)\right\|^2\right] + \mathtt{L}\gamma^2(n+1)^{-2\alpha}\eta \\
&\leq \mathbb{E}\left[f(X_n)\right] - \gamma(n+1)^{-\alpha}\mathbb{E}\left[\left\|\nabla f(X_n)\right\|^2\right]/2 + \mathtt{L}\gamma^2(n+1)^{-2\alpha}\eta .
\end{aligned} \tag{89}$$

Combining (88) and (89) we get for any $n \in \mathbb{N}$ such that $n \geq (2\mathtt{L}\gamma)^{1/2}$

$$\begin{aligned}
E_{n+1} - E_n &= (n+2)^\delta \mathbb{E}\left[f(X_{n+1})\right](1+\log(n+2))^{-\varepsilon} - (n+1)^\delta \mathbb{E}\left[f(X_n)\right](1+\log(n+1))^{-\varepsilon} \\
&\leq (1+\log(n+1))^{-\varepsilon}\left[\left\{(n+2)^\delta - (n+1)^\delta\right\}(\mathbb{E}\left[f(X_{n+1})\right]) \right. \\
&\qquad \left. + (n+1)^\delta\left\{\mathbb{E}\left[f(X_{n+1})\right] - \mathbb{E}\left[f(X_n)\right]\right\}\right] \\
&\leq (1+\log(n+1))^{-\varepsilon}\left[\left\{(n+2)^\delta - (n+1)^\delta\right\}(\mathbb{E}\left[f(X_n)\right] + \mathtt{L}\gamma^2(n+1)^{-2\alpha}\eta) \right. \\
&\qquad \left. + (n+1)^\delta\left\{-\gamma(n+1)^{-\alpha}\mathbb{E}\left[\left\|\nabla f(X_n)\right\|^2\right]/2 + \mathtt{L}\gamma^2(n+1)^{-2\alpha}\eta\right\}\right] \\
&\leq (1+\log(n+1))^{-\varepsilon}\left[c_\delta(n+1)^{\delta-1}(\mathbb{E}\left[f(X_n)\right] + 2\gamma^2(n+1)^{-2\alpha}\eta) \right. \\
&\qquad \left. + (n+1)^\delta\left\{-\gamma(n+1)^{-\alpha}\mathbb{E}\left[\left\|\nabla f(X_n)\right\|^2\right]/2 + \mathtt{L}\gamma^2(n+1)^{-2\alpha}\eta\right\}\right] \\
&\leq c_\delta E_n + 2\mathtt{L}\gamma^2(1+c_\delta)(n+1)^{\delta-2\alpha}(1+\log(n+1))^{-\varepsilon}\eta \\
&\quad - \gamma(n+1)^{\delta-\alpha}(1+\log(n+1))^{-\varepsilon}\mathbb{E}\left[\left\|\nabla f(X_n)\right\|^2\right]/2 .
\end{aligned} \tag{90}$$

Using (3) and the fact that for any $k \in \mathbb{N}$, $\mathbb{E}\left[\left\|X_k - x^\star\right\|^{r_2 r_3}\right] \leq C_{\beta,\varepsilon}(k+1)^\beta(1+\log(1+k))^\varepsilon$ and Hölder's inequality and that $r_1 r_3 = 2(2r_1^{-1} - 1)^{-1}$, we have for any $k \in \mathbb{N}$

$$\mathbb{E}\left[\left\|\nabla f(X_k)\right\|^2\right] \geq \mathbb{E}\left[f(X_k)\right]^{2r_1^{-1}} C_{\beta,\varepsilon}^{-(2r_1^{-1}-1)^{-1}} \tau^{2r_1^{-1}} (k+1)^{-\beta(2r_1^{-1}-1)} (1+\log(k+1))^{-\varepsilon(2r_1^{-1}-1)} . \tag{91}$$

Combining (90) and (91) we get that for any $n \in \mathbb{N}$ with $n \geq (4\gamma)^{1/\alpha}$

$$\begin{aligned} E_{n+1} - E_n \leq c_\delta E_n + 2\mathtt{L}\gamma^2(1+c_\delta)(n+1)^{\delta-2\alpha}(1+\log(n+1))^{-\varepsilon}\eta \\ - \gamma(n+1)^{\delta-\alpha-\beta(2r_1^{-1}-1)}\mathbb{E}\left[f(X_n)\right]^{2r_1^{-1}} C_{\beta,\varepsilon}^{-(2r_1^{-1}-1)^{-1}} \tau^{2r_1^{-1}}(1+\log(n+1))^{-\varepsilon 2r_1^{-1}}/2 \\ \leq c_\delta E_n + 2\mathtt{L}\gamma^2(1+c_\delta)(n+1)^{\delta-2\alpha}(1+\log(n+1))^{-\varepsilon}\eta \\ - \gamma(n+1)^{\alpha-(\delta+\beta)(2r_1^{-1}-1)} E_n^{2r_1^{-1}} C_{\beta,\varepsilon}^{-(2r_1^{-1}-1)^{-1}} \tau^{2r_1^{-1}}/2 \ . \end{aligned}$$

Let $\mathtt{D}_3 = \max(\mathtt{D}_1, \mathtt{D}_2)$ with

$$\begin{cases} \mathtt{D}_1 = (2|c_\delta| C_{\beta,\varepsilon}^{2r_1^{-1}-1} \tau^{-2r_1^{-1}})^{2r_1^{-1}-1} \ , \\ \mathtt{D}_2 = (4\mathtt{L}\gamma^2(1+c_\delta) C_{\beta,\varepsilon}^{2r_1^{-1}-1} \tau^{-2r_1^{-1}})^{r_1/2} \ . \end{cases}$$

If $E_n \geq \mathtt{D}_3$ and $n \geq (4\gamma)^{1/\alpha}$ then $E_{n+1} \leq E_n$. Therefore, we obtain by recursion that $E_n \leq \mathtt{D}$ with $\mathtt{D} = \max(E_0, \dots, E_{\lceil (2\mathtt{L}\gamma)^{1/\alpha} \rceil}, \mathtt{D}_3)$.