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

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

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

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

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 $dx(t)/dt = -\nabla f(x(t))$. 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 \((X_t)_{t \geq 0}\) 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 \( t \mapsto \mathbb{E}[f(X_t)] - \min_{\mathbb{R}^d} f \) 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
Convergence and approximation for SGD

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 $O\left(\frac{n^{-1/2}}{2}\right)$, with stepsize $\gamma_n = O\left(\frac{n^{-1/2}}{2}\right)$, 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 $f \in C^1(\mathbb{R}^d, \mathbb{R})$, an objective function satisfying the following regularity condition.

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

We consider the general case where we do not have access to $\nabla f$ 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 $(\mathcal{Z}, \mathcal{F}, \pi_Z)$ and $\eta \geq 0$ such that one of the following conditions holds:

(a) There exists a function $H : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$ such that for any $x \in \mathbb{R}^d$,
$$\int_{\mathcal{Z}} H(x, z) d\pi_Z(z) = \nabla f(x), \quad \int_{\mathcal{Z}} \|H(x, z) - \nabla f(x)\|^2 d\pi_Z(z) \leq \eta.$$ 

(b) There exists a function $\tilde{f} : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ such that for all $z \in \mathcal{Z}$, $\tilde{f}(\cdot, z) \in C^1(\mathbb{R}^d, \mathbb{R})$ is $L$-smooth. In addition, there exists $x^* \in \mathbb{R}^d$ such that for any $x \in \mathbb{R}^d$
$$\int_{\mathcal{Z}} \tilde{f}(x, z) d\pi_Z(z) = f(x), \quad \int_{\mathcal{Z}} \nabla \tilde{f}(x, z) d\pi_Z(z) = \nabla f(x), \quad \int_{\mathcal{Z}} \|\nabla \tilde{f}(x^*, z)\|^2 d\pi_Z(z) \leq \eta.$$  

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

The first setting A2-(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 A2-(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 $z \in \mathcal{Z}$. 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 A1 and A2, 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

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

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 $(Z, \mathcal{P})$ such that for any $n \in \mathbb{N}$, $Z_n$ is distributed according to $\pi^Z$. We now turn to the continuous counterpart of (1). Define for any $x \in \mathbb{R}^d$, the semi-definite positive matrix $\Sigma(x) = \pi^Z(\{H(x, \cdot) - \nabla f(x)\} H(x, \cdot) - \nabla f(x))$ and, for $\alpha \in [0, 1)$, consider the time inhomogeneous SDE,

$$dX_t = -(\gamma_\alpha + t)^{-\alpha} \{\nabla f(X_t)dt + \gamma_\alpha^{1/2} \Sigma(X_t)^{1/2}dB_t\},$$

where $\gamma_\alpha = \gamma^{1/(1-\alpha)}$ and $(B_t)_{t \geq 0}$ 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 $M \geq 0$ such that for any $x, y \in \mathbb{R}^d$, $\|\Sigma(x)^{1/2} - \Sigma(y)^{1/2}\| \leq M \|x - y\|$.

Indeed, using (Karatzas and Shreve, 1991, Chapter 5, Theorem 2.5), strong solutions $(X_t)_{t \geq 0}$ exist if A1 and A3 hold. Condition A3 can be hard to check in practice and can be replaced by the following stronger (but easier to verify) assumption: $\Sigma \in C^2(\mathbb{R}^d, \mathbb{R}^{d \times d})$ with bounded Hessian, see (Stroock and Varadhan, 2007, Theorem 5.2.3). In the sequel, $(X_t)_{t \geq 0}$ 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 $(X_t)_{t \geq 0}$ 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 $(\tilde{X}_t)_{t \geq 0}$ be the linear interpolation of $(X_n)_{n \in \mathbb{N}}$, i.e., for any $t \in [n\gamma_\alpha, (n + 1)\gamma_\alpha]$, $n \in \mathbb{N}$, $\tilde{X}_t = ((t - n\gamma_\alpha)X_{n+1} + ((n + 1)\gamma_\alpha - t)X_n)/\gamma_\alpha$, 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 $\Sigma(X_{n\gamma_\alpha})$, we have the following approximation,

$$\tilde{X}_{(n+1)\gamma_\alpha} - \tilde{X}_{n\gamma_\alpha} = X_{n+1} - X_n \approx -\gamma(n + 1)^{-\alpha}H(\tilde{X}_{n\gamma_\alpha}, Z_{n+1})$$

$$\approx -\gamma_\alpha(n\gamma_\alpha + \gamma_\alpha)^{-\alpha} \{\nabla f(\tilde{X}_{n\gamma_\alpha}) + \Sigma(\tilde{X}_{n\gamma_\alpha})^{1/2}G_{n+1}\}$$

$$\approx -\int_{n\gamma_\alpha}^{(n+1)\gamma_\alpha} (s + \gamma_\alpha)^{-\alpha}\nabla f(\tilde{X}_s)ds - \gamma_\alpha^{1/2}\int_{n\gamma_\alpha}^{(n+1)\gamma_\alpha} (s + \gamma_\alpha)^{-\alpha}\Sigma(\tilde{X}_s)^{1/2}dB_s,$$

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 $W_2 : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to [0, +\infty]$ (where $\mathcal{P}(\mathbb{R}^d)$ is the set of probability measures over $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and given for any $\mu_1, \mu_2 \in \mathcal{P}(\mathbb{R}^d)$ by $W_2^2(\mu_1, \mu_2) = \inf_{\Lambda \in \Gamma(\mu_1, \mu_2)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|v_1 - v_2\|^2 d\Lambda(v_1, v_2)$, where $\Gamma(\mu_1, \mu_2) \subset \mathcal{P}(\mathbb{R}^{2d})$ is the set of transference plans between $\mu_1$ and $\mu_2$, i.e., $\Lambda \in \Gamma(\mu_1, \mu_2)$ if for any $A \in \mathcal{B}(\mathbb{R}^d)$, $\Lambda(A \times \mathbb{R}^d) = \mu_1(A)$ and $\Lambda(\mathbb{R}^d \times A) = \mu_2(A)$.

4
Theorem 1  Let \( \bar{\gamma} > 0 \) and \( \alpha \in [0, 1) \). Assume A1, A2-(b) and A3. 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 \( \pi^\varepsilon \), there exists \((B_t)_{t \geq 0}\) such that the following hold:

(a) \((B_t)_{t \geq 0}\) 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 \( n \leq n_T = [T/\gamma_\alpha] \) and \( \gamma_\alpha = \gamma^{1/(1-\alpha)} \) we have

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

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

\[
\varepsilon^2 = \sup_{n \in \{0, \ldots, n_T\}} \mathbb{E}[W_2^2(\nu^\delta(X_{n\gamma_\alpha}), \nu^\varepsilon(X_{n\gamma_\alpha}))] ,
\]

where for any \( \bar{x} \in \mathbb{R}^d \), \( \nu^\delta(\bar{x}) \) is the distribution of \( H(\bar{x}, Z_n) \) and \( \nu^\varepsilon(\bar{x}) \) is the distribution of \( \nabla f(\bar{x}) + \Sigma^{1/2}(\bar{x})G \), 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 \((Y_t)_{t \geq 0}\) and its EM discretization \((Y_n)_{n \in \mathbb{N}}\) given by \( Y_0 = Y_0 \in \mathbb{R}^d \) and for any \( n \in \mathbb{N} \)

\[
dY_t = b(Y_t)dt + \sigma(Y_t)dB_t , \quad Y_{n+1} = Y_n + \gamma b(Y_n) + \sqrt{\gamma} \sigma(Y_n)G_{n+1} , \tag{5}
\]

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 \( \mathbb{E}[G_n G_n^T] = \text{Id} \). 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 \( \sigma \). If \( \sigma \) 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 \( b = 0 \), \( \sigma = \text{Id} \), (but it holds under mild conditions on \( b \) and \( \sigma \)), there exists \( C \geq 0 \) such that for any \( T \geq 0 \), \( \gamma > 0 \), \( n \in \mathbb{N} \), \( n\gamma \leq T \), \( \mathbb{E}^{1/2}[\|Y_{n\gamma} - Y_n\|^2] \geq C\sqrt{T} \), i.e., no strong approximation holds. The behavior is different for SGD for which we obtain a strong approximation of order \( O(\gamma^{1/2} \varepsilon) \), regardless the structure of the noise.

(b) We remark that a strong approximation of order \( O(\gamma^{1/2}) \) 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 \( O(\gamma^{1/2} \varepsilon) \), where \( \varepsilon \) is a measure of the distance between the noise and some Gaussian distribution in \( W_2 \). 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 \( \mathbb{E}[\|X_{n\gamma} - Y_n\|^2] = O(\gamma^p) \)
(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 A2-(b) which implies that for any \( z \in \mathbb{Z} \), \( x \mapsto H(x, z) \) is Lipschitz continuous. It is not clear if our results can be extended to A2-(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 \((Z, \mathcal{Z}) = (Y^M, \mathcal{Y}^\otimes M)\), \( \pi^Z = \pi^{\otimes M} \) with \( M \in \mathbb{N} \), \( \pi \) a probability measure on the Polish space \((\mathcal{Y}, \mathcal{Y})\) and for any \( x \in \mathbb{R}^d \), \( z = \{y_i\}_{i=1}^M \), let

\[
H(x, z) = (1/M) \sum_{i=1}^M \nabla \tilde{f}(x, y_i).
\]

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 A1, A2-(b) and A3 (with respect to \((Y, \mathcal{Y}, \pi)\)). Let \( H \) be given by (6). Assume that there exists \( x^* \in \mathbb{R}^d \), \( C, p \geq 0 \) such that for any \( x \in \mathbb{R}^d \) and \( y \in \mathcal{Y} \)

\[
\int_{\mathcal{Y}} \|\nabla \tilde{f}(x^*, y)\|^4 \pi(y) < +\infty, \quad \|\Sigma_f(x)^{-1/2}\| \leq C(1 + \|x\|^p).
\]

Then, there exists a random variable \(( (B_t)_{t \geq 0}, (Z_n)_{n \in \mathbb{N}} ) \) 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 \( n \gamma \alpha \leq T \gamma_0 = \gamma^{1/(1-\alpha)} \) we have

\[
\mathbb{E}^{1/2}[\|X_{n \gamma_0} - X_n\|^2] \leq C(\gamma^{\alpha}M^{-1} + \gamma)(1 + \log(\gamma^{-1})), \quad \text{with} \quad \delta = \min(1, (2 - 2\alpha)^{-1}).
\]

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 \( \mathcal{V} : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}_+ \) which depend on the conditions imposed on the function \( f \). Then, we show that \( t \mapsto v(t) = \mathbb{E}[\mathcal{V}(t, X_t)] \) satisfies an ODE and prove that it is bounded using the following simple lemma.

**Lemma 3** Let \( F \in C^1(\mathbb{R}_+ \times \mathbb{R}, \mathbb{R}) \) and \( v \in C^1(\mathbb{R}_+, \mathbb{R}_+) \) such that for all \( t \geq 0 \), \( dv(t)/dt \leq F(t, v(t)). \) 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 $d v(t_1)/dt < 0$ 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 $n \in \mathbb{N}$, $F(n, \cdot) \in C^1(\mathbb{R}, \mathbb{R})$. 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 A2, we consider two versions of the strong convexity hypothesis, either directly on $f$ if A2-(a) holds or on $\vec{f}$ if A2-(b) holds.

**F1** Either one of the following conditions holds:

(a) Case A2-(a): $f$ is $\mu$-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 A2-(b): for all $z \in \mathbb{Z}$, $\vec{f}(\cdot, z)$ is $\mu$-strongly convex.

Note that F1-(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 $x^*$ the unique minimizer of $f$ (which exists under F1).

**Theorem 5** Let $\alpha, \gamma \in (0, 1)$ and $(X_t)_{t \geq 0}$ be given by (2). Assume A1, A2, A3 and F1. Then there exists $C \geq 0$ (explicit in the proof) such that for any $T \geq 1$, $\mathbb{E}[\|X_T - x^*\|^2] \leq CT^{-\alpha}$.

This result holds for both versions of F1 and we present below a proof under F1-(a). The result under F1-(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 $\mathcal{E}(t) = \mathbb{E}[\langle t + \gamma \alpha \rangle^\alpha \|X_t - x^*\|^2]$, with $\gamma_0 = \gamma^{1/(1-\alpha)}$. Using Dynkin’s formula, see Lemma 48, we have for any $t \geq 0$,

$$
\mathcal{E}(t) = \mathcal{E}(0) + \alpha \int_0^t \mathcal{E}(s) ds + \int_0^t \gamma_0 \frac{\mathbb{E}[\langle t + \gamma \alpha \rangle^\alpha \|X_t - x^*\|^2]}{(s + \gamma_0)^\alpha} ds - 2 \int_0^t \mathbb{E}[\langle \nabla f(X_s), X_s - x^* \rangle] ds.
$$

We now differentiate this expression with respect to $t$ and using F1-(a) and A2-(a), we get for any $t > 0$,

$$
d\mathcal{E}(t)/dt = \alpha \mathcal{E}(t)(t + \gamma_0)^{-1} - 2\mathbb{E}[\langle \nabla f(X_t), X_t - x^* \rangle] + \gamma_0 \mathbb{E}[\langle \nabla f(X_t) \rangle](t + \gamma_0)^{-\alpha}
$$

$$
\leq \alpha \mathcal{E}(t)/(t + \gamma_0) - 2\mu \mathbb{E}[\|X_t - x^*\|^2] + \gamma_0 \eta/(t + \gamma_0)^\alpha
$$

$$
\leq F(t, \mathcal{E}(t)) = \alpha \mathcal{E}(t)(t + \gamma_0)^{-1} - 2\mu \mathcal{E}(t)(t + \gamma_0)^{-\alpha} + \gamma_0 \eta(t + \gamma_0)^{-\alpha},
$$
CONVERGENCE AND APPROXIMATION FOR SGD

where we have used that \( \text{Tr}(\Sigma(x)) \leq \eta \) for any \( x \in \mathbb{R}^d \) by A2-(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 \( \mu, \eta \) and the initial condition. However, in Theorem 34 we obtain that (i) the constant in front of the asymptotic term \( T^{-\alpha} \) scales as \( \eta/\mu \) 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 \( \mu \)-strongly convex, i.e., for any \( x, y \in \mathbb{R}^d \), \( \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2 \). 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 \( \mu \)-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 \( (\mathbb{E}[f(X_t)] - \min_{\mathbb{R}^d} f)_{t \geq 0} \) 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 A2 we consider.

F2 Either one of the following conditions holds:
(a) Case A2-(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 \( x^* \in \arg\min_{\mathbb{R}^d} f \).
(b) Case A2-(b): for all \( z \in \mathbb{Z} \), \( \tilde{f}(\cdot, z) \) is convex and there exists a minimizer \( x^* \in \arg\min_{\mathbb{R}^d} f \).

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 $(X_t)_{t \geq 0}$ be given by (2). Assume $f \in C^2(\mathbb{R}^d, \mathbb{R})$, A1, A2, A3 and F2. Then, there exists $C \geq 0$ (explicit and given in the proof) such that for any $T \geq 1$

$$\mathbb{E}[f(X_T)] - \min_{\mathbb{R}^d} f \leq C(1 + \log(T))^2/T^{\alpha \wedge (1-\alpha)}.$$ 

To the best of our knowledge, these non-asymptotic results are new for the continuous process $(X_t)_{t \geq 0}$ defined by (2). Note that for $\alpha = 1/2$ the convergence rate is of order $O(T^{-1/2} \log^2(T))$ 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

$$(\text{Agarwal et al., 2012}) \quad \text{and is in accordance with the tight bounds derived in the discrete case under}$$

$$(\text{additional assumptions (Shamir and Zhang, 2013)}).$$

The general proof is postponed to Appendix E.2. Additional assumptions (Shamir and Zhang, 2013). The general proof is postponed to Appendix E.2. 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.

Corollary 7  Let $\alpha, \gamma \in (0, 1)$ and $(X_t)_{t \geq 0}$ be given by (2). Assume that arg min$_{\mathbb{R}^d} f$ is bounded, A1, A2, A3 and F2. Then, there exists $C \geq 0$ (explicit and given in the proof) such that for any $T \geq 1$, 

$$\mathbb{E}[f(X_T)] - \min_{\mathbb{R}^d} f \leq C(1 + \log(T))^2/T^{\alpha \wedge (1-\alpha)}.$$ 

The proof is postponed to Appendix E.2 and 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 $O(N^{-1/2} \log(N + 1))$ 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 A1 (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 $\alpha (\alpha = 1/2$ 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 $(\varphi_p)_{p \in \mathbb{N}^*}$, where for any $x \in \mathbb{R}$, $p \in \mathbb{N}^*$,

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

Figure 3: Convergence rates for $\varphi_p$ match the theoretical results of Theorem 8 asymptotically.

For any $p \in \mathbb{N}$, $\varphi_p$ satisfies and A1 and F2. Denoting $\alpha_p^*$ the non-increasing rate $\alpha$ for which the convergence rate $\tau_p^*$ is maximum, we experimentally check that $\lim_{p \to +\infty} \tau_p^* = 1/2$ and $\lim_{p \to +\infty} \alpha_p^* = 1/2$. Note also that $\alpha_p^*$ 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 A1, A2 and F2. 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 $x \in \mathbb{R}^d$

$$\|\nabla f(x)\|^{r_1} \|x - x^*\|^{r_2} \geq \tau (f(x) - f(x^*)),$$

where $x^* \in \text{arg min}_{\mathbb{R}^d} f \neq \emptyset$.

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 $x \in \mathbb{R}^d$

$$\langle \nabla f(x), x - x^* \rangle \geq \tau (f(x) - f(x^*)),$$

where $x^* \in \text{arg min}_{\mathbb{R}^d} f \neq \emptyset$.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Thm.8 (L)</th>
<th>(BM'11) (B, L)</th>
<th>(BM'11) (L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha \in (0, 1/3)$</td>
<td>$\alpha$</td>
<td>$\times$</td>
<td>$\times$</td>
</tr>
<tr>
<td>$(1/3, 1/2)$</td>
<td>$\alpha$</td>
<td>$(3\alpha - 1)/2$</td>
<td>$\times$</td>
</tr>
<tr>
<td>$(1/2, 2/3)$</td>
<td>$1 - \alpha$</td>
<td>$\alpha/2$</td>
<td>$\alpha/2$</td>
</tr>
<tr>
<td>$(2/3, 1)$</td>
<td>$1 - \alpha$</td>
<td>$1 - \alpha$</td>
<td>$1 - \alpha$</td>
</tr>
</tbody>
</table>

Table 1: Convergence rates for convex SGD (B: Bounded gradients, L: Lipschitz gradient).
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 \textbf{F3} 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 \( \hat{\tau} > 0 \) such that for any \( x \in \mathbb{R}^d \),

\[
\|\nabla f(x)\|^r \geq \hat{\tau}(f(x) - f(x^*)) , \quad \text{where } x^* \in \arg\min_{\mathbb{R}^d} f \neq \emptyset .
\] (7)

In this case, \textbf{F3} is satisfied with \( r_1 = r, r_2 = 0 \) and \( \tau = \hat{\tau} \). 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 \textbf{F3}, one can still derive convergence of order \( \alpha \) 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 \((X_t)_{t \geq 0}\) be given by (2). Assume \( f \in C^2(\mathbb{R}^d, \mathbb{R}) \), \textbf{A1}, \textbf{A2-(a)}, \textbf{A3} and \textbf{F3}. In addition, assume that there exist \( \beta, \varepsilon \geq 0 \) and \( C_{\beta, \varepsilon} \geq 0 \) such that for any \( t \geq 0 \),

\[
\mathbb{E}[\|X_t - x^*\|^{r_3}] \leq C_{\beta, \varepsilon}(\gamma_{\alpha} + t)^{\beta}(1 + \log(1 + \gamma_{\alpha}^{-1}t))^\varepsilon ,
\]

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 \( T \geq 1 \)

\[
\mathbb{E}[f(X_T)] - \min_{\mathbb{R}^d} f \leq CT^{-\delta}(1 + \log(1 + \alpha^{-1}T))\varepsilon ,
\] (8)

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

The proof is postponed to Appendix G. First, note that if \( f \) satisfies a Kurdyka-Łojasiewicz condition of type (7) then \textbf{F3} 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 \textbf{F1}.

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

**Corollary 11** Let \( \alpha, \gamma \in (0, 1) \) and \((X_t)_{t \geq 0}\) given by (2). Assume \( f \in C^2(\mathbb{R}^d, \mathbb{R}) \), \textbf{A1}, \textbf{A2-(a)}, \textbf{A3}.

(a) If \textbf{F3b} holds, then there exists \( C \geq 0 \) such that for any \( T \geq 1 \)

\[
\mathbb{E}[f(X_T)] - \min_{\mathbb{R}^d} f \leq C[T^{(1-3\alpha)/2} + T^{-\alpha/2} + T^{\alpha-1}] .
\]

(b) If \textbf{F3b} holds and there exist \( R \geq 0 \) and \( c > 0 \) such that for any \( x \in \mathbb{R}^d \) with \( \|x - x^*\| \geq R \),

\[
f(x) - f(x^*) \geq c\|x - x^*\| \quad \text{then there exists } C \geq 0 \text{ such that for any } T \geq 1
\]

\[
\mathbb{E}[f(X_T)] - \min_{\mathbb{R}^d} f \leq C[T^{-\alpha/2} + T^{\alpha-1}] .
\] (9)
Corollary 11-(a) \(\alpha \in (0,1/3)\) \(\times\) \(\alpha/2\) \(\times\)
\(\alpha \in (1/3,1/2)\) \((3\alpha - 1)/2\) \(\alpha/2\) \(\times\)
\(\alpha = 1/2\) \(1/4 + \log\) \(1/4 + \log\) \(\times\)
\(\alpha \in (1/2,2/3)\) \(\alpha/2\) \(1 - \alpha\) \(2\alpha - 1\)
\(\alpha \in (2/3,1)\) \(1 - \alpha\) \(1 - \alpha\) \(1 - \alpha\)

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

(c) If \(F_3\) holds and if there exist \(R \geq 0\) and \(m > 0\) such that for any \(x \in \mathbb{R}^d\) with \(\|x - x^*\| \geq R\), \((\nabla f(x), x - x^*) \geq m \|x - x^*\|^2\), 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 \(t \mapsto \mathbb{E}[\|X_t - x^*\|^2]\) using either the SDE satisfied by \((\|X_t - x^*\|^2)_{t \geq 0}\) in the case of (a) and (c), or the SDE satisfied by \((f(X_t) - \min_{x \in \mathbb{R}^d} f)_{t \geq 0}\) in the case of (b).

Under \(F_3\)b, 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 \(n \in \mathbb{N}\), \(\mathbb{E}[\|X_n - x^*\|^{r_2}] \leq C_{\beta,\varepsilon}(n + 1)^{\beta}(1 + \log(1 + n))^\varepsilon\), where \(r_2 = (1 - r_1/2)^{-1}\). Then, there exists \(C \geq 0\) (explicit and given in the proof) such that for any \(N \geq 1\)

\[
\mathbb{E}[f(X_N)] - \min_{x \in \mathbb{R}^d} f \leq CN^{-\delta_1 \land \delta_2} (1 + \log(1 + N))^\varepsilon,
\]

where \(\delta_1, \delta_2\) 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 inhomogeneous 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.
Convergence and approximation for SGD

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


Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-


Convergence and approximation for SGD


### 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 A2-(a). The counterpart of these results in the case where A2-(b) holds is presented in Appendix D. In Appendix E, we analyze the convex setting under A2-(a). Again, the counterpart of these results in the case where A2-(b) holds is given in Appendix F. We conclude with the proofs of the weakly quasi-convex setting in Appendix G.

### Contents

A Organization of the appendix

---

18
## 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 $A$ 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 \(A_{\alpha,r} \geq 0\) such that for any \(N \in \mathbb{N}\) with \(N \gamma_\alpha \leq T\) we have

\[
g_r \sum_{k=0}^{N-1} (k+1)^{-\alpha r} \leq \begin{cases} A_{\alpha,r} \gamma_r^{1 + \log(\gamma^{-1})} (1 + \log(T)), & \text{if } \alpha \geq 1/r, \\ A_{\alpha,r} \gamma_r^{\gamma_\alpha \alpha r - 1} T^{1-\alpha r}, & \text{otherwise}. \end{cases}
\]

**Proof** Let \(r > 0\), \(\gamma > 0\) and \(\alpha \in [0, 1)\). If \(\alpha > 1/r\) then there exists \(A_{\alpha,r} \geq 0\) such that

\[
g_r \sum_{k=0}^{N-1} (k+1)^{-\alpha r} \leq A_{\alpha,r} \gamma_r.
\]

If \(\alpha < 1/r\) then there exists \(A_{\alpha,r} \geq 0\) such that

\[
g_r \sum_{k=0}^{N-1} (k+1)^{-\alpha r} \leq A_{\alpha,r} \gamma_r N^{-\alpha r + 1} \leq A_{\alpha,r} \gamma_r \gamma_\alpha^{\alpha r - 1} T^{1-\alpha r}.
\]

If \(\alpha = 1/r\) then there exists \(A_{\alpha,r} \geq 0\) such that

\[
g_r \sum_{k=0}^{N-1} (k+1)^{-\alpha r} \leq \gamma_r (1 + \log(N)) \leq A_{\alpha,r} \gamma_r (1 + \log(T))(1 + \log(\gamma^{-1})).
\]

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

The following lemma derives upper-bound from the regularity assumption A1 and A3.

**Lemma 15** Assume A1 and A3. Then there exists \(C \geq 0\) such that for any \(x \in \mathbb{R}^d\),

\[
\|\nabla f(x)\| \leq C(1 + \|x\|), \quad \|
abla^{1/2} f(x)\| \leq C(1 + \|x\|), \quad \|\Sigma(x)\| \leq C(1 + \|x\|^2).
\]
Proof First, we have for any \( x \in \mathbb{R}^d \) using \( A1 \)

\[
\| \nabla f(x) \| \leq \| \nabla f(0) \| + L \| x \| \leq (\| \nabla f(0) \| + L)(1 + \| x \|) .
\]

Similarly, we have for any \( x \in \mathbb{R}^d \) using \( A3 \),

\[
\| \Sigma^{1/2}(x) \| \leq (\| \nabla f(0) \| + M)(1 + \| x \|) .
\]

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 \((B_t)_{t \geq 0}\) 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 \( A1 \) and \( A3 \). Let \( \alpha \in [0, 1) \), \( \bar{\gamma} > 0 \) and \( \gamma \in (0, \bar{\gamma}] \). Then, there exists \((B_t)_{t \geq 0}\) 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 \( \pi^Z \) and is independent from \( \mathcal{K}_k \), where for any \( k \in \mathbb{N} \)

\[
\mathcal{K}_k = \sigma(\{B_t, Z_j : t \in [0, k\gamma_\alpha] \} , j \in \{1, \ldots, k\} ) ,
\]

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

(b) For any \( k \in \mathbb{N} \), there exists \( \mathcal{A}_k \in \mathcal{B}(\mathbb{R}^d) \) such that \( \mathbb{P}(X_{k\gamma_\alpha} \in \mathcal{A}_k) = 1 \) and for any \( \bar{x} \in \mathcal{A}_k \)

\[
W^2_2(\nu^d_k(\bar{x}), \nu^\mathcal{F}_k(\bar{x})) = \mathbb{E} \left[ \left\| H(\bar{x}, Z_{k+1}) - \nabla f(\bar{x}) - \gamma_\alpha^{1/2} \Sigma^{1/2}(\bar{x}) \int_{k\gamma_\alpha}^{(k+1)\gamma_\alpha} dB_s \right\| ^2 \right] ,
\]

where \((X_t)_{t \geq 0}\) is a unique strong solution of (2) starting from \( X_0 = X_0 \in \mathbb{R}^d \), \( \nu^d(\bar{x}) \) is the distribution of \( H(\bar{x}, Z_0) \) and \( \nu^\mathcal{F}(\bar{x}) \) is the distribution of \( \nabla f(\bar{x}) + \Sigma^{1/2}(\bar{x})G \) 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 \((B^\alpha_{t\gamma_\alpha})_{t \in [0,\gamma_\alpha]} \), \((Z_k)_{k \in \mathbb{N}}\) such that for any \( k \in \mathbb{N} \), the following assertion \( H1(k) \) is true.
\textbf{H1 (k)} We have that $\{B^k_{t+1}\}_{t \in [0, \gamma_α]}$ and $Z_{k+1}$ are independent from $\mathcal{H}_k = \sigma(\{B^j_t, Z_j : t \in [0, \gamma_α], j \in \{1, \ldots, k-1\}\}$ (with $\mathcal{H}_0 = \{\emptyset, \Omega\}$) and there exists $A_k \in \mathcal{B}(\mathbb{R}^d)$ such that $\mathbb{P}(Y_{k\gamma_α} \in A_k) = 1$ and for any $\bar{x} \in A_k$,

$$W_2(\nu^d(\bar{x}), \nu^c(\bar{x})) = \mathbb{E} \left[ \left\| H(\bar{x}, Z_{k+1}) - \nabla f(\bar{x}) - \gamma_α^{-1/2} \Sigma^{1/2}(\bar{x}) \int_0^{\gamma_α} dB^{k+1}_s \right\|^2 \right],$$

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

$$Y_t = X_0 + \sum_{j=0}^{n_t-1} \left( - \int_0^{\gamma_α} ((j + 1)\gamma_α + s)^{-\alpha} \nabla f(Y_{j\gamma_α+s}) ds \right.$$

$$- \gamma_α^{1/2} \int_0^{\gamma_α} ((j + 1)\gamma_α + s)^{-\alpha} \Sigma^{1/2}(Y_{j\gamma_α+s}) dB^{j+1}_s \bigg)$$

$$- \int_0^{t-n_t\gamma_α} ((n_t + 1)\gamma_α + s)^{-\alpha} \nabla f(Y_{n_t\gamma_α+s}) ds$$

$$- \gamma_α^{1/2} \int_0^{t-n_t\gamma_α} ((n_t + 1)\gamma_α + s)^{-\alpha} \Sigma^{1/2}(Y_{n_t\gamma_α+s}) dB^{n_t+1}_s,$$

where $n_t = \lfloor t/\gamma_α \rfloor$. Denote $\mu_k$ the distribution of $\{(B^j_t)_{t \in [0, \gamma_α]}, Z_j\}_{j \in \{0, \ldots, k\}}$.

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

$$F_k(x, w) = \nabla f(x) + \gamma_α^{-1/2} \Sigma^{1/2}(x) \int_0^{\gamma_α} dw_s.$$

Since $x \mapsto F(x, w)$ is continuous for $\pi^B$-almost every $w \in C_{\gamma_α}$ by $A1$ and $A3$, and for any $x \in \mathbb{R}^d$, $w \mapsto F(x, w)$ is measurable, we get that $F$ is measurable using (Aliprantis and Border, 2006, Lemma 4.51). In addition, using $A1$, $A3$ and (Rogers and Williams, 2000, Theorem 10.4), for any $k \in \mathbb{N}$, there exists a measurable mapping $\tilde{S}_{k+1} : \mathbb{R}^d \times C_{\gamma_α} \to \mathbb{R}^d$ such that for any Brownian motion $(B_t)_{t \in [0, \gamma_α]}$, $\tilde{S}_{k+1}(\tilde{x}, (B_t)_{t \in [0, \gamma_α]}) = Y^k_{\gamma_α}$, where $(Y^k_t)_{t \in [0, \gamma_α]}$ is the unique strong solution to the following SDE: for any $t \in [0, \gamma_α]$

$$Y^k_t = \tilde{x} - \int_0^t ((k+1)\gamma_α + s)^{-\alpha} \nabla f(Y^k_s) ds - \gamma_α^{1/2} \int_0^t ((k+1)\gamma_α + s)^{-\alpha} \Sigma^{1/2}(Y^k_s) dB_s.$$

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

$$S_k(\{(w^j_t)_{t \in [0, \gamma_α]}, Z_j\}_{j=0}^k)$$

$$= \tilde{S}_k(\tilde{S}_{k-1}(\ldots (\tilde{S}_1(\tilde{S}_0(X_0, (w^0_t)_{t \in [0, \gamma_α]}), (w^1_t)_{t \in [0, \gamma_α]})) \ldots, (w^{k-1}_t)_{t \in [0, \gamma_α]}), (w^k_t)_{t \in [0, \gamma_α]})) \ldots.$$
Finally, for any \( k \in \mathbb{N} \), let \( \mathcal{F}_k : \mathbb{R}^{k+1} \times C_{\gamma_0} \to \mathbb{R}^d \) and \( \mathcal{H}_k : \mathbb{R}^{k+1} \times Z \to \mathbb{R}^d \) such that for any \( k \in \mathbb{N} \), \( \{(w_t^j)_{t \in [0, \gamma_0]}, Z^j_{j=0} \in \mathbb{R}^{k+1}, (w_t)_{t \in [0, \gamma_0]} \in C_{\gamma_0} \) and \( z \in Z \)

\[
\mathcal{F}_k(\{(w_t^j)_{t \in [0, \gamma_0]}, Z^j_{j=0} \}, (w_t)_{t \in [0, \gamma_0]}, (w_t)_{t \in [0, \gamma_0]}, (w_t)_{t \in [0, \gamma_0]}),
\mathcal{H}_k(\{(w_t^j)_{t \in [0, \gamma_0]}, Z^j_{j=0} \}, (w_t)_{t \in [0, \gamma_0]}, (w_t)_{t \in [0, \gamma_0]}), z).
\]

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

\[
Q_k(u, A) = \operatorname{opt}(\mathcal{F}_k(u, \cdot)_{\#} \pi^B, \mathcal{H}_k(u, \cdot)_{\#} \pi^Z)(A),
\]

where for all \( u \in \mathbb{R}^{k+1} \), \( \operatorname{opt}(\mathcal{F}_k(u, \cdot)_{\#} \pi^B, \mathcal{H}_k(u, \cdot)_{\#} \pi^Z) \) is the optimal transference plan between \( \mathcal{F}_k(u, \cdot)_{\#} \pi^B \) and \( \mathcal{H}_k(u, \cdot)_{\#} \pi^Z \) w.r.t. the \( W_2 \), which exists by (Villani, 2009, Theorem 4.1).

Note that for any \( k \in \mathbb{N} \), \( Q_k \) is well-defined since \( u \mapsto \operatorname{opt}(\mathcal{F}_k(u, \cdot)_{\#} \pi^B, \mathcal{H}_k(u, \cdot)_{\#} \pi^Z) \) 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 \( H1(k) \) is true. Second, we show that we can construct a Brownian motion from the random variables introduced in \( H1(k) \) for any \( k \in \mathbb{N} \) such that the proposition holds.

(a) We start by proving that \( H1(0) \) holds. Let \( \mu_0 \in \mathcal{P}(C_{\gamma_0} \times Z) \) be any coupling between \( \pi^B \) and \( \pi^Z \). Let \( \eta^0 \in \mathcal{P}(E \times C_{\gamma_0} \times \mathbb{R}^d \times E) \), \( \eta^1 \in \mathcal{P}(\mathbb{R}^d \times E \times \mathbb{R}^d) \) and \( \eta^2 \in \mathcal{P}(E \times \mathbb{R}^d \times E \times Z) \) such that

\[
\eta^0 = (\operatorname{Id}(\mathcal{F}_0, \Pi_1))_{\#}(\mu_0 \otimes \pi^B), \quad \eta^2 = (\Pi_1, \mathcal{H}_0, \operatorname{Id})_{\#}(\mu_0 \otimes \pi^Z),
\]

where \( \Pi_1 \) is the projection on the first variable. In addition, let \( \eta^1 = \mu_0 \otimes \mathbb{Q}_0 \), i.e., for any \( A \in B(E) \) and \( B_1, B_2 \in B(\mathbb{R}^d) \) we have

\[
\eta^1(B_1 \times A \times B_2) = \int_A Q_0(x, B_1 \times B_2) d\mu_0(x).
\]

Note that \( \eta^1 = (\mathcal{F}_0, \operatorname{Id})_{\#}(\mu_0 \otimes \pi^B) = \eta^1_{12} \). Therefore, using the gluing lemma (Ambrosio et al., 2008, Lemma 5.3.2) (which is valid since \( E \), \( C_{\gamma_0} \) and \( Z \) are Polish spaces), there exists a probability measure \( \tilde{\eta} \in \mathcal{P}(E \times C_{\gamma_0} \times \mathbb{R}^d \times E \times \mathbb{R}^d) \) such that \( \tilde{\eta}_{1234} = \eta^0 \) and \( \tilde{\eta}_{1234} = \eta^1 \). In addition note that \( \tilde{\eta}_{45} = (\operatorname{Id}, \mathcal{H}_0)_{\#}(\mu_0 \otimes \pi^Z) = \eta^2_{12} \). Therefore, using the gluing lemma, there exists a probability measure \( \eta \in \mathcal{P}(E \times C_{\gamma_0} \times \mathbb{R}^d \times E \times \mathbb{R}^d \times Z \times E) \) such that \( \eta_{12345} = \tilde{\eta} \) and \( \eta_{4567} = \eta^1 \). In particular, \( \eta_{1234} = \eta^0 \) and \( \eta_{4567} = \eta^1 \). Let \( (U_i)_{i \in \{1,...,7\}} \) be a random variable with distribution \( \eta \). Then, using that \( \eta_{1234} = \eta^0 \) and \( \eta_{4567} = \eta^1 \), we have almost surely

\[
U_3 = \mathcal{F}_0(U_1, U_2), \quad U_4 = U_1, \quad U_4 = U_7, \quad U_5 = \mathcal{H}_0(U_7, U_6).
\]

Therefore, we get that

\[
(\eta_1, \ldots, \eta_7) = (U_1, U_2, \mathcal{F}_0(U_1, U_2), U_1, \mathcal{H}_0(U_1, U_6), U_6, U_1).
\]

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 \( \mu_1 \in \mathcal{P}(E \times C_{\gamma_0} \times Z) \) such that

\[
\eta = (\Pi_1, \Pi_2, \mathcal{F}_0(\Pi_1, \Pi_2), \Pi_1, \mathcal{H}_0(\Pi_1, \Pi_2), \Pi_3, \Pi_1)_{\#} \mu_1.
\]
Let \(((B^0_t)_{t \in [0, \gamma_\alpha]}, Z_0, (B^1_t)_{t \in [0, \gamma_\alpha]}, Z_1)\) be a random variable with distribution \(\mu_1\). Then \((B^1_t)_{t \in [0, \gamma_\alpha]}\) and \(Z_1\) are independent from \((B^0_t)_{t \in [0, \gamma_\alpha]}, Z_0\), \((B^1_t)_{t \in [0, \gamma_\alpha]}\) has distribution \(\pi^B\) and \(Z_1\) has distribution \(\pi^Z\). Hence, \((B^1_t)_{t \in [0, \gamma_\alpha]}\) 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 \(u \in E\) and \(A \in B(\mathbb{R}^d \times \mathbb{R}^d)\) by

\[
R_0(u, A) = \int_{\mathbb{R}^d} \mathbb{P}_0(u, \Pi_2(\cdot)), \mathbb{H}_0(u, \Pi_3(\cdot))) \# \mu_1 d\mu_0(u) ,
\]

Note that \(\mu_1 = \mu_0 \otimes R_0\). But by definition of \(\mu_1\) we also have that \(\mu_1 = \mu_0 \otimes Q_0\). Therefore, \(\mu_0\) almost surely we have \(R_0(x, \cdot) = Q_0(x, \cdot)\), which concludes the proof of (11). Therefore \(H1(0)\) holds. Assume that \(H1(k)\) is true with \(k \in \mathbb{N}\). Then \(H1(k + 1)\) holds. The proof is similar to the one for \(H1(0)\) upon replacing \(\mu_0\) by \(\mu_k\), \(\mathcal{F}_0\) by \(\mathcal{F}_k\), \(\mathcal{H}_0\) by \(\mathcal{H}_k\) and \(Q_0\) by \(Q_k\). We conclude by recursion.

(b) Finally, it remains to define a Brownian motion \((B_t)_{t \geq 0}\) such that for any \(k \in \mathbb{N}\), \(\mathcal{K}_k = \mathcal{H}_k\) and \((X_t)_{t \geq 0} = (Y_t)_{t \geq 0}\). For any \(t \geq 0\), let \(n_t = \lfloor t/\gamma_\alpha \rfloor\) and \((B^1_t)_{t \geq 0}\) such that for any \(t \geq 0\)

\[
B_t = B^0_{n_t \gamma_\alpha} + \sum_{k=0}^{n_t-1} B^k_{\gamma_\alpha} .
\]

Since \((B^k_{t \in [0, \gamma_\alpha]}\)\)\(k \in \mathbb{N}\) is a sequence of independent Brownian motion, we get that \((B_t)_{t \geq 0}\) is a Brownian motion. In addition, there exists a measurable bijection mapping \((B_t)_{t \geq 0}\) to \((B^k_{t \in [0, \gamma_\alpha]}\)\)\(k \in \mathbb{N}\) and therefore for any \(k \in \mathbb{N}\), \(\mathcal{K}_k = \mathcal{H}_k\). Finally, we have that \((Y_t)_{t \geq 0}\) solution to (12) is a solution to (2) with initial condition \(Y_0 = X_0\), i.e., \((Y_t)_{t \geq 0} = (X_t)_{t \geq 0}\) which concludes the proof.

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]

\[
\]
Proof We prove the result under A2-(b). The proof under A2-(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 \( g_p \in C^2(\mathbb{R}^d, [0, +\infty)) \) 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 \((X_t)_{t \geq 0}\) be a solution to (2). We have for any \( x \in \mathbb{R}^d \)

\[
\nabla g_p(x) = 2p \|x\|^{2(p-1)} x, \quad \nabla^2 g_p(x) = 4p(p-1) \|x\|^{2(p-2)} xx^\top + 2p \|x\|^{2(p-1)} \text{Id}.
\] (13)

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

\[
\mathbb{E} \left[ g_p(X_{t \wedge \tau_n}) | \mathcal{F}_s \right] - \mathbb{E} \left[ g_p(X_s) \right] = \mathbb{E} \left[ \int_s^{t \wedge \tau_n} (-\gamma_p + u)^{-\alpha} \langle \nabla f(X_u), \nabla g_p(X_u) \rangle | \mathcal{F}_s \right] \, du \\
\quad + \left(\gamma_p/2\right) \mathbb{E} \left[ \int_s^{t \wedge \tau_n} (\gamma_p + u)^{-2\alpha} \langle \Sigma(X_u), \nabla^2 g_p(X_u) \rangle \, du \right] | \mathcal{F}_s \right].
\] (14)

Using A1, (13) and the Cauchy-Schwarz inequality we get that for any \( u \in [s, s + T] \)

\[
|\langle \nabla f(X_u), \nabla g_p(X_u) \rangle| \leq 2p\|X_u\|^{2(p-1)} \left\{ |\langle \nabla f(X_u) - \nabla f(0), X_u \rangle| + \|\nabla f(0)\|\|X_u\| \right\} \\
\leq 2p(L + \|\nabla f(0)\|)g_p(X_u).
\] (15)

In addition, using A1, Lemma 15, (13) and the Cauchy-Schwarz inequality we get that for any \( u \in [s, s + T] \)

\[
|\langle \Sigma(X_u), \nabla^2 g_p(X_u) \rangle| \leq C(1 + \|X_u\|^2) \|\nabla^2 g_p(X_u)\| \\
\leq C(1 + \|X_u\|^2)(8p(p-1)d + 2pd) \|X_u\|^{2(p-1)} \\
\leq 4Cd(p-1) + 1)g_p(X_u).
\] (16)

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

\[
\mathbb{E} \left[ g_p(X_{t \wedge \tau_n}) | \mathcal{F}_s \right] - g_p(X_s) \\
\leq 2p(L + \|\nabla f(0)\|) \mathbb{E} \left[ \int_s^{t \wedge \tau_n} g_p(X_u) \, du \right] | \mathcal{F}_s \right] + \bar{\gamma}_p(2p-1) \mathbb{E} \left[ \int_s^{t \wedge \tau_n} g_p(X_u) \, du \right] | \mathcal{F}_s \right] \\
\leq \left\{ 2p(L + \|\nabla f(0)\|) + 4\gamma_p Cd(p-1) + 1 \right\} \int_s^t \mathbb{E} \left[ g_p(X_{u \wedge \tau_n}) | \mathcal{F}_u \right] \, du.
\]

Using Grönwall’s lemma we obtain

\[
\mathbb{E} \left[ g_p(X_{t \wedge \tau_n}) | \mathcal{F}_s \right] \leq g_p(X_s) \exp \left[ T \left\{ 2p(L + \|\nabla f(0)\|) + 4\gamma_p Cd(p-1) + 1 \right\} \right]
\]

We conclude upon using Fatou’s lemma and remarking that \( \lim_{n} \tau_n = +\infty \), since \( X_t \) 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,

\[
\|X_{k+1}\|^{2p} = \|A_k + B_k\|^{2p} = \{\|A_k\|^2 + 2\langle A_k, B_k \rangle + \|B_k\|^2\}^p \\
\leq \sum_{i=0}^{p} \sum_{j=0}^{i} \binom{p}{i} \binom{i}{j} \|A_k\|^{2(p-i)+j} \|B_k\|^{2i-j} \times 2^j \\
\leq \|A_k\|^{2p} + 2^p \sum_{i=1}^{p} \sum_{j=0}^{i} \binom{p}{i} \binom{i}{j} \|A_k\|^{2(p-i)+j} \|B_k\|^{2i-j}.
\]

Using \(A_1\), there exists \(\tilde{\alpha}_{T,1}^{(a)}, \tilde{\alpha}_{T,1}^{(b)}, \tilde{\alpha}_{T,1}^{(c)} \geq 0\) such that for any \(\ell \in \{0, \ldots, 2p\}\)

\[
\|A_k\|^\ell \leq \sum_{m=0}^{\ell} \binom{\ell}{m} (1 + \gamma(k + 1)^{-\alpha})^m \|X_k\|^m (\gamma(k + 1)^{-\alpha} \|\nabla f(0)\|)^{\ell-m} \\
\leq (1 + \gamma(k + 1)^{-\alpha} \tilde{\alpha}_{T,1}^{(a)} \|X_k\|^\ell + \gamma(k + 1)^{-\alpha} \tilde{\alpha}_{T,1}^{(b)} (1 + \|X_k\|^{2p})) \\
\leq (1 + \gamma(k + 1)^{-\alpha} \tilde{\alpha}_{T,1}^{(c)}(1 + \|X_k\|^{2p}).
\]

In addition, we have that for any \(\ell \in \{1, \ldots, p\}, x \in \mathbb{R}^d\) and \(z \in \mathbb{Z}\),

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

Therefore, there exists \(\eta_\ell > 0\) such that for any \(\ell \in \mathbb{N}\),

\[
\mathbb{E} \|B_k\|^{2\ell} |G_k| \leq \gamma^{2\ell}(k + 1)^{-2\alpha\ell} \eta_\ell(1 + \|X\|^{2\ell}).
\]

Combining this result, (17), Jensen’s inequality and that \(f\) we have

\[
\mathbb{E} \|X_{k+1}\|^{2p} |G_k| \leq \|X_k\|^{2p} + \gamma(k + 1)^{-\alpha} \tilde{\alpha}_{T,1}^{(a)} + \tilde{\alpha}_{T,1}^{(b)} (1 + \|X_k\|^{2p}) \\
+ 2^{p+1} (1 + \gamma(k + 1)^{-\alpha} \tilde{\alpha}_{T,1}^{(c)})(1 + \|X_k\|^{2p}) \sum_{i=1}^{p} \sum_{j=0}^{i} \binom{p}{i} \binom{i}{j} \eta_1^{1/2} \gamma^{2i-j}(k + 1)^{-\alpha(2i-j)}.
\]

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

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

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

\[
\|
\]

**Lemma 18** Let \(p \in \mathbb{N}, \gamma > 0\) and \(\alpha \in [0, 1)\). Assume \(A_1, A_2\) \(\text{(b)}, A_3\) and that there exists \(x^*\) such that \(\int_Z |H(x^*, z)|^{2p} d\pi_Z(z) < +\infty\). Then for any \(T \geq 0\), there exists \(\kappa_{T,2} \geq 0\) such that for any \(\gamma \in (0, \gamma]\), \(k \in \mathbb{N}\) with \((k + 1)\gamma_0 \leq T, t \in [k\gamma_0, (k + 1)\gamma_0]\), we have

\[
\mathbb{E} \|X_{k+1} - X_k\|^{2p} |G_k| \leq \kappa_{T,2}(k + 1)^{-2\alpha p}\gamma^{2p}(1 + \|X_k\|^{2p}) \\
\mathbb{E} \|X_t - X_{k\gamma_0}\|^{2p} |F_{k\gamma_0}| \leq \kappa_{T,2}(k + 1)^{-2\alpha p}\gamma^{2p}(1 + \|X_{k\gamma_0}\|^{2p})
\]

where \((X_k)_{k \in \mathbb{N}}\) satisfies the recursion and \((X_t)_{t \geq 0}\) 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 $(X_s)_{s \geq 0}$ be a solution to (2). Using A1, A2, 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

\[
\mathbb{E} \left[ \|X_t - X_{k\gamma_{\alpha}}\|^{2p} \right. \left| \mathcal{F}_{k\gamma_{\alpha}} \right] \\
\leq 2^{2p-1} \mathbb{E} \left[ \left\| \frac{1}{k\gamma_{\alpha}} \int_{k\gamma_{\alpha}}^t (\gamma_{\alpha} + s)^{-\alpha} \nabla f(X_s) ds \right\|^{2p} \right. \left| \mathcal{F}_{k\gamma_{\alpha}} \right] \\
+ 2^{2p-1}\gamma_{\alpha}^{-p} \mathbb{E} \left[ \left\| \frac{1}{k\gamma_{\alpha}} \int_{k\gamma_{\alpha}}^t (\gamma_{\alpha} + s)^{-\alpha} \Sigma(X_s)_{1/2} d\mathbb{B}_s \right\|^{2p} \right. \left| \mathcal{F}_{k\gamma_{\alpha}} \right] \\
\leq 2^{2p-1}\gamma_{\alpha}^{-p} \int_{k\gamma_{\alpha}}^t (\gamma_{\alpha} + s)^{-2\alpha} \mathbb{E} \left[ \left\| \nabla f(X_s) \right\|^{2p} \left| \mathcal{F}_{k\gamma_{\alpha}} \right] ds \\
+ B_p 2^{p-1}\gamma_{\alpha}^{-p} \left( \int_{k\gamma_{\alpha}}^t (\gamma_{\alpha} + s)^{-2\alpha} \mathbb{E} \left[ \text{Tr}(\Sigma(X_s)) \left| \mathcal{F}_{k\gamma_{\alpha}} \right] ds \right) \right. \left[ \mathcal{F}_{k\gamma_{\alpha}} \right] \\
\leq 2^{2p-1}\gamma_{\alpha}^{-p} \int_{k\gamma_{\alpha}}^t (\gamma_{\alpha} + s)^{-2\alpha} \mathbb{E} \left[ \left\| \nabla f(X_s) \right\|^{2p} \left| \mathcal{F}_{k\gamma_{\alpha}} \right] ds \\
+ \int_{k\gamma_{\alpha}}^t \mathbb{E} \left[ \text{Tr}(\Sigma(X_s)) \left| \mathcal{F}_{k\gamma_{\alpha}} \right] ds \right] \\
\leq 2^{4p}(1 + L^2\gamma_{\alpha}^{-1}(k + 1)^{-2\alpha} B_p + 1) \int_{k\gamma_{\alpha}}^t C^{2p} \left( 1 + \mathbb{E} \left[ \left\| X_s \right\|^{2p} \left| \mathcal{F}_{k\gamma_{\alpha}} \right] \right) \right] ds \\
\leq 2^{4p}(1 + L^2\gamma_{\alpha}^{-1}(k + 1)^{-2\alpha} B_p + 1) \left( 1 + \sup_{s \in [k\gamma_{\alpha}, t]} \mathbb{E} \left[ \left\| X_s \right\|^{2p} \left| \mathcal{F}_{k\gamma_{\alpha}} \right] \right) \right] \\
\leq 2^{4p}(1 + L^2\gamma_{\alpha}^{-1}(k + 1)^{-2\alpha} B_p + 1) (1 + A_{T,1} g_p(X_{k\gamma_{\alpha}})),
\]

(b) Let $(X_n)_{n \in \mathbb{N}}$ which satisfies the recursion (1). Using A1 and A2-(b) we get that

\[
\mathbb{E} \left[ \|X_{k+1} - X_k\|^{2p} \mid \mathcal{G}_k \right] \\
= \mathbb{E} \left[ \| -\gamma (k + 1)^{-\alpha} (H(X_k, Z_{k+1}) - H(x^*, Z_{k+1}) + H(x^*, Z_{k+1})) \|^{2p} \mid \mathcal{G}_k \right] \\
\leq 2^{2p}\gamma_{\alpha}^{-2p}(k + 1)^{-2\alpha} \left( L^{2p} \| X_k - x^* \|^{2p} + \mathbb{E} \left[ \| H(x^*, Z_{k+1}) \|^{2p} \mid \mathcal{G}_k \right] \right) \\
\leq 2^{4p}\gamma_{\alpha}^{-2p}(k + 1)^{-2\alpha} \left( L^{2p} \| X_k \|^{2p} + L^{2p} \| x^* \|^{2p} + \mathbb{E} \left[ \| H(x^*, Z_{k+1}) \|^{2p} \mid \mathcal{G}_k \right] \right),
\]

which concludes the proof.

\[\blacksquare\]

**B.4. Mean-square approximation**

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

\[
\bar{X}_{k+1} = \bar{X}_k - \gamma (1 + k)^{-\alpha} \left\{ \nabla f(\bar{X}_k) + \frac{1}{2} \alpha_{\alpha} \Sigma^{1/2}(\bar{X}_k)G_{k+1} \right\},
\]
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 \((\overline{X}_t)_{t \geq 0}\). Consider the stochastic process \((\overline{X}_t)_{t \geq 0}\) defined by \(\overline{X}_0 = X_0\) and solution of the following SDE

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

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

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

with \(G_k = \gamma_{\alpha}^{-1/2} \int_{k \gamma_{\alpha}}^{(k+1) \gamma_{\alpha}} dB_s\). Hence, for any \(k \in \mathbb{N}\), \(\overline{X}_{k \gamma_{\alpha}}\) has the same distribution as \(X_k\) given by (1) with \(H(x, z) = \nabla f(x) + \Sigma(x)^{1/2} z\), \((Z, \mathcal{Z}) = (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))\) and \(\pi^Z\) 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 \((\overline{X}_t)_{t \geq 0}\) 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 \((\overline{X}_t)_{t \geq 0}\) 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 \((\overline{X}_t)_{t \geq 0}\) and the continuous-time process \((X_t)_{t \geq 0}\). We combine these estimates in Proposition 21. We conclude by proving Proposition 22 which is a restatement of Theorem 1.

Recall that \(n_T = \lceil T/\gamma_{\alpha} \rceil\). In what follows, we denote

\[
\varepsilon^2 = \sup_{k \in \{0, \ldots, n_T\}} \mathbb{E} \left[ \mathbf{W}_2^2(\nu^d_k(X_{k \gamma_{\alpha}}), \nu^d_k(X_{k \gamma_{\alpha}})) \right], \quad (18)
\]

where for any \(\tilde{x} \in \mathbb{R}^d\), \(\nu^d_k(\tilde{x})\) is the distribution of \(H(\tilde{x}, Z_{n+1})\), \(\nu^d_k(\tilde{x})\) is the distribution of \(\nabla f(\tilde{x}) + \gamma_{\alpha}^{-1/2} \Sigma^{1/2}(\tilde{x}) \int_{k \gamma_{\alpha}}^{(k+1) \gamma_{\alpha}} (\gamma_{\alpha} + s)^{-\alpha} dB_s\).

**Lemma 19** Assume \(A_1\) and \(A_3\). Let \(\gamma > 0\) and \(\alpha \in [0, 1)\). Then for any \(T \geq 0\), there exists \(A_{T, 3} \geq 0\) 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

\[
\mathbb{E} \left[ \|\overline{X}_{(k+1) \gamma_{\alpha}} - \hat{X}_{k+1}\|^2 \right] \leq A_{T, 3} \gamma^2 (k+1)^{-2\alpha} \varepsilon^2,
\]

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

\[
\hat{X}_{k+1} = X_{k \gamma_{\alpha}} - (k+1)^{-\alpha} H(X_{k \gamma_{\alpha}}, Z_{k+1}),
\]

\[
\hat{X}_t = X_{k \gamma_{\alpha}} - \gamma_{\alpha}^{-1} \gamma (1 + k)^{-\alpha} \left\{ (t - k \gamma_{\alpha}) \nabla f(X_{k \gamma_{\alpha}}) + \gamma_{\alpha}^{1/2} \Sigma^{1/2}(X_{k \gamma_{\alpha}}) \int_{k \gamma_{\alpha}}^t dB_s \right\},
\]

We recall that for any \(\tilde{x} \in \mathbb{R}^d\), \(\nu^d_k(\tilde{x})\) is the distribution of \(H(\tilde{x}, Z_{n+1})\), \(\nu^d_k(\tilde{x})\) is the distribution of \(\nabla f(\tilde{x}) + \gamma_{\alpha}^{-1/2} \Sigma^{1/2}(\tilde{x}) \int_{k \gamma_{\alpha}}^{(k+1) \gamma_{\alpha}} (\gamma_{\alpha} + s)^{-\alpha} dB_s\).
Lemma 20

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

\[
\mathbb{E} \left[ \left\| X_{(k+1)\gamma} - X_{k+1} \right\|^2 \right] = \gamma^2 (k+1)^{-2\alpha} \mathbb{E} \left[ \left\| \nabla f(X_{k\gamma}) + \Sigma^{1/2}(X_0) G_k - H(X_{k\gamma}, Z_k) \right\|^2 \right] \\
\leq 2\gamma^2 (k+1)^{-2\alpha} \mathbb{E} \left[ \frac{W_2^2(v_k^d(X_{k\gamma}), v_k^c(X_{k\gamma}))}{k^{2\alpha}} \right].
\]

which concludes the proof upon using (33).

\[\blacksquare\]

Lemma 20

Let \( \bar{\gamma} > 0 \) and \( \alpha \in [0, 1) \). Assume A1 and A3. Then for any \( T \geq 0 \), there exists \( k_{T,4} \geq 0 \) such that for any \( \gamma \in (0, \bar{\gamma}] \), \( k \in \mathbb{N} \) with \( (k+1)\gamma \leq T \) and \( X_0 \in \mathbb{R}^d \) we have

\[
\mathbb{E} \left[ \left\| X_{(k+1)\gamma} - \tilde{X}_{(k+1)\gamma} \right\|^2 \right] \leq \mathbb{E} \left[ F_{k\gamma} \right] \leq A_{T,4} \left\{ \gamma^4(k+1)^{-4\alpha} + \gamma^2 (k+1)^{-2(1+\alpha)} \right\} \left( 1 + \left\| X_{k\gamma} \right\|^2 \right) ,
\]

where \( (X_t)_{t \geq 0} \) is the solution of (2) and for any \( t \in [k\gamma, (k+1)\gamma] \) we have

\[
\tilde{X}_t = X_{k\gamma} - \gamma_\alpha^{-1}(1+k)^{-\alpha} \left\{ (t-k\gamma) \nabla f(X_{k\gamma}) + \gamma_\alpha^{1/2} \Sigma^{1/2}(X_{k\gamma}) \int_{k\gamma}^t dB_s \right\} .
\]  

\[\blacksquare\]

Proof

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

\[
\mathbb{E} \left[ \left\| X_{(k+1)\gamma} - \tilde{X}_{(k+1)\gamma} \right\|^2 \right] \leq \mathbb{E} \left[ \left\| \frac{1}{k\gamma} \int_{k\gamma}^{(k+1)\gamma} \gamma_\alpha s^{-\alpha} \nabla f(X_s) ds + \gamma_\alpha^{1/2} \Sigma(X_s)^{1/2} dB_s \right\|^2 \right] \\
\leq 2\mathbb{E} \left[ \left\| \frac{1}{\gamma_\alpha^{1/2}} \int_{k\gamma}^{(k+1)\gamma} (1+\gamma_\alpha^{-1}s)^{-\alpha} \nabla f(X_s) ds + \gamma_\alpha^{-1/2}(k+1)^{-\alpha} \Sigma(X_{k\gamma})^{1/2} dB_s \right\|^2 \right].
\]
We now treat each term separately.

Using Jensen’s inequality, Itô isometry, Fubini-Tonelli’s theorem, A1, A3 and Lemma 18 we have

\[
E \left[ \left\| \int_{k \gamma_{\alpha}}^{(k+1) \gamma_{\alpha}} \left\{ (k + 1)^{-\alpha} \nabla f(X_{k \gamma_{\alpha}}) - (1 + \gamma_{\alpha}^{-1} s)^{-\alpha} \nabla f(X_s) \right\} \, ds \right\|^2 \bigg| \mathcal{F}_{k \gamma_{\alpha}} \right] \tag{21}
\]

\[
\leq 2 \left( k + 1 \right)^{-2 \alpha} \left[ \int_{k \gamma_{\alpha}}^{(k+1) \gamma_{\alpha}} \mathbb{E} \left[ \left\| \nabla f(X_{k \gamma_{\alpha}}) - (1 + \gamma_{\alpha}^{-1} s)^{-\alpha} \nabla f(X_s) \right\|^2 \bigg| \mathcal{F}_{k \gamma_{\alpha}} \right] \, ds \right] + \eta \left( k + 1 \right)^{-\alpha} \int_{k \gamma_{\alpha}}^{(k+1) \gamma_{\alpha}} \left[ (1 + \gamma_{\alpha}^{-1} s)^{-2 \alpha} \right] \, ds
\]

\[
\leq 2 \gamma_{\alpha} \left( k + 1 \right)^{-2 \alpha} \mathbb{E} \sup_{s \in [k \gamma_{\alpha}, (k+1) \gamma_{\alpha}]} \left\{ \left\| \nabla f(X_{k \gamma_{\alpha}}) - (1 + \gamma_{\alpha}^{-1} s)^{-\alpha} \nabla f(X_s) \right\|^2 \right\} \mathbb{E} \left[ \left\| \nabla f(X_s) - \nabla f(X_{k \gamma_{\alpha}}) \right\|^2 \bigg| \mathcal{F}_{k \gamma_{\alpha}} \right]
\]

\[
\leq 2 \gamma_{\alpha} \left( k + 1 \right)^{-2 \alpha} \mathbb{E} \left[ \left\| \nabla f(X_{k \gamma_{\alpha}}) \right\|^2 \left( 1 + \gamma_{\alpha}^{-1} s \right)^{-2 \alpha} \right] \mathbb{E} \left[ \left\| \nabla f(X_s) \right\|^2 \bigg| \mathcal{F}_{k \gamma_{\alpha}} \right]
\]

Using Jensen’s inequality, Fubini-Tonelli’s theorem, the fact that for any \( u > 0 \), \( u^{-\alpha} \leq (u + 1)^{-\alpha} \leq \alpha u^{-(\alpha+1)} \), A1 and Lemma 18 we get that

\[
E \left[ \left\| \int_{k \gamma_{\alpha}}^{(k+1) \gamma_{\alpha}} \left\{ (k + 1)^{-\alpha} \nabla f(X_{k \gamma_{\alpha}}) - (1 + \gamma_{\alpha}^{-1} s)^{-\alpha} \nabla f(X_s) \right\} \, ds \right\|^2 \bigg| \mathcal{F}_{k \gamma_{\alpha}} \right] \tag{22}
\]

\[
\leq 2 \gamma_{\alpha} \mathbb{E} \sup_{s \in [k \gamma_{\alpha}, (k+1) \gamma_{\alpha}]} \left\{ \left\| \nabla f(X_{k \gamma_{\alpha}}) - (1 + \gamma_{\alpha}^{-1} s)^{-\alpha} \nabla f(X_s) \right\|^2 \mathbb{E} \left[ \left\| \nabla f(X_s) - \nabla f(X_{k \gamma_{\alpha}}) \right\|^2 \bigg| \mathcal{F}_{k \gamma_{\alpha}} \right] \right\}
\]

\[
\leq 2 \gamma_{\alpha} \left( k + 1 \right)^{-2 \alpha} \mathbb{E} \left[ \left\| \nabla f(X_{k \gamma_{\alpha}}) \right\|^2 \left( 1 + \gamma_{\alpha}^{-1} s \right)^{-2 \alpha} \right] \mathbb{E} \left[ \left\| \nabla f(X_s) \right\|^2 \bigg| \mathcal{F}_{k \gamma_{\alpha}} \right]
\]

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

\[
A_{T,4} = 4 \left[ \beta^2 A_{T,2} + \eta \alpha^2 + \alpha^2 \left( \left\| \nabla f(0) \right\|^2 + L^2 \right) \right] + L^2 A_{T,2} \gamma^2
\]

\[
\text{CONVERGENCE AND APPROXIMATION FOR SGD}
\]

30
Proposition 21 Let $\tilde{\gamma} > 0$ and $\alpha \in [0, 1)$. Assume A1, A2-(b) and A3. Then for any $T \geq 0$, there exists $k_{T,5} \geq 0$ such that for any $\gamma \in (0, \tilde{\gamma}]$, $k \in \mathbb{N}$ with $(k + 1)\gamma \leq T$ and $X_0 \in \mathbb{R}^d$ we have
\[
\mathbb{E} \left[ \|X_{(k+1)\gamma} - \tilde{X}_{k+1}\|^2 \right] \leq k_{T,5} \left\{ \gamma^4 (k + 1)^{-4\alpha} + \gamma^2 (k + 1)^{-2\alpha} \epsilon^2 \right\} \left( 1 + \|X_{k\gamma}\|^2 \right),
\]
where $(X_t)_{t \geq 0}$ is the solution of (2) and for any $k \in \mathbb{N}$
\[
\tilde{X}_{k+1} = X_{k\gamma} - \gamma (k + 1)^{-\alpha} H(X_{k\gamma}, Z_{k+1}).
\]
Proof The proof is straightforward upon combining Lemma 19 and Lemma 20.

Proposition 22 Let $\tilde{\gamma} > 0$, $\alpha \in [0, 1)$ and $\gamma \in (0, \tilde{\gamma}]$. Assume A1, A2-(b) and A3. Then there exists a coupling $((B_t)_{t \geq 0}, (Z_n)_{n \in \mathbb{N}})$ 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 $\pi^Z$.

(b) For any $T \geq 0$, there exists $C \geq 0$ such that for any $\gamma \in (0, \tilde{\gamma}]$, $n \in \mathbb{N}$ with $\gamma \alpha = \gamma^1/(1-\alpha)$, $n\gamma \leq T$ we have
\[
\mathbb{E}^{1/2} \left[ \|X_{n\gamma} - X_n\|^2 \right] \leq C (\gamma^\delta \epsilon + \gamma) (1 + \log(\gamma^{-1})) \leq \gamma^{1/(1-\alpha)} + 1 + \gamma^{-1}, \quad \text{with} \ \delta = \min(1, (2 - 2\alpha)^{-1}),
\]
where $(X_t)_{t \geq 0}$ is solution of (2), $(X_n)_{n \in \mathbb{N}}$ is defined by (1) with $X_0 = X_0 \in \mathbb{R}^d$ and
\[
\epsilon^2 = \sup_{k \in \{0, \ldots, n_T\}} \mathbb{E} \left[ W_2^2(\nu_k^d(X_{k\gamma}), \nu_k^e(X_{k\gamma})) \right], \quad (23)
\]
where for any $\tilde{x} \in \mathbb{R}^d$, $\nu_k^d(\tilde{x})$ is the distribution of $H(\tilde{x}, Z_{n+1})$, $\nu_k^e(\tilde{x})$ is the distribution of $\nabla f(\tilde{x}) + \gamma^{-1/2} \epsilon^{1/2} (\tilde{x}) \int (k+1)\gamma_0 (\gamma_0 + s)^{-\alpha} d\mathbb{B}_s$.

Proof Let $\alpha \in [0, 1)$, $\gamma > 0$, $\gamma \in (0, \tilde{\gamma}]$, $k \in \mathbb{N}$, 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 $k \in \mathbb{N}$, $E_k = \mathbb{E}[\|X_{k\gamma} - X_k\|^2]$. Note that $E_0 = 0$. Let $\tilde{X}_{k+1} = X_{k\gamma} - \gamma (k + 1)^{-\alpha} H(X_{k\gamma}, Z_{k+1})$. We have
\[
E_{k+1} = \mathbb{E} \left[ \|X_{(k+1)\gamma} - X_{k+1}\|^2 \right] \leq \mathbb{E} \left[ \|X_{(k+1)\gamma} - \tilde{X}_{k+1} + \tilde{X}_{k+1} - X_{k+1}\|^2 \right]
\]
\[
= \mathbb{E} \left[ \|X_{(k+1)\gamma} - \tilde{X}_{k+1}\|^2 \right] + 2 \mathbb{E} \left[ \langle X_{(k+1)\gamma} - \tilde{X}_{k+1}, \tilde{X}_{k+1} - X_{k+1} \rangle \right] + \mathbb{E} \left[ \|\tilde{X}_{k+1} - X_{k+1}\|^2 \right]
\]
\[
= \mathbb{E} \left[ \|X_{(k+1)\gamma} - \tilde{X}_{k+1}\|^2 \right] + \mathbb{E} \left[ \|\tilde{X}_{k+1} - X_{k+1}\|^2 \right]
\]
\[
+ 2 \mathbb{E} \left[ \langle X_{(k+1)\gamma} - \tilde{X}_{k+1}, X_{k\gamma} - X_{k} \rangle \right] + 2 \gamma (k + 1)^{-\alpha} \mathbb{E} \left[ \langle X_{(k+1)\gamma} - \tilde{X}_{k+1}, H(X_{k}, Z_{k+1}) - H(X_{k\gamma}, Z_{k+1}) \rangle \right].
\]
Let $a_k = \gamma^4(k + 1)^{-4\alpha} + \gamma^2(k + 1)^{-2\alpha}$ and $a_k^2 = \gamma^4(k + 1)^{-4\alpha} + \epsilon^2\gamma^2(k + 1)^{-2\alpha}$. We now bound each of the four terms appearing in (24)

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

$$
\mathbb{E} \left[ \left\| X_{(k+1)\gamma_a} - \bar{X}_{k+1} \right\|^2 \right] = \mathbb{E} \left[ \mathbb{E} \left[ \left\| X_{(k+1)\gamma_a} - \bar{X}_{k+1} \right\|^2 \middle| F_{k\gamma_a} \right] \right] 
\leq \mathbb{E} \left[ A_{T,5}(\gamma^4(k + 1)^{-4\alpha} + \epsilon^2\gamma^2(k + 1)^{-2\alpha}) \left( 1 + \left\| X_{k\gamma_a} \right\|^2 \right) \right]
\leq k_{T,1} A_{T,5}(\gamma^4(k + 1)^{-4\alpha} + \epsilon^2\gamma^2(k + 1)^{-2\alpha}) \left( 1 + \left\| X_0 \right\|^2 \right) \leq A_{T,6}^{(a)} a_k^2 ,
$$
with $A_{T,6}^{(a)} \geq 0$ which does not depend on $\gamma$ 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

$$
\mathbb{E} \left[ \left\| \bar{X}_{k+1} - X_{k+1} \right\|^2 \right] = \mathbb{E} \left[ \left\| X_{k\gamma_a} - X_{k} - \gamma(k + 1)^{-\alpha} (H(X_{k\gamma_a}, Z_{k+1}) - H(X_{k}, Z_{k+1})) \right\|^2 \right]
\leq (1 + \gamma L(k + 1)^{-\alpha} \epsilon^2 \mathbb{E}[\left\| X_{k\gamma_a} - X_{k} \right\|^2]
\leq (1 + 2\gamma L(k + 1)^{-\alpha} + \epsilon^2 L^2(k + 1)^{-2\alpha}) E_k \leq (1 + A_{T,6}^{(b)} a_k^{1/2}) E_k ,
$$
with $A_{T,6}^{(b)} \geq 0$ which does not depend on $\gamma$ and $k$.

(c) In what follows, let $\bar{X}_{(k+1)\gamma_a} = X_{k\gamma_a} - \gamma(k + 1)^{-\alpha} \{ \nabla f(X_{k\gamma_a}) + \Sigma(X_{k\gamma_a})^{1/2} G_k \}$, with $G_k = \gamma^{-1/2} \int_{k\gamma_a}^{(k+1)\gamma_a} dB_k$. Let $b_k = \gamma^3(k + 1)^{-3\alpha} + \gamma(k + 1)^{-2(1+\alpha/2)}$. Using $A_2$ we have $\mathbb{E}[\bar{X}_{(k+1)\gamma_a} | K_k] = \mathbb{E}[\bar{X}_{k+1} | K_k]$. 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

$$
\mathbb{E} \left[ \left\langle X_{(k+1)\gamma_a} - \bar{X}_{k+1}, X_{k\gamma_a} - X_{k} \right\rangle \right]
= \mathbb{E} \left[ \mathbb{E} \left[ \left\langle X_{(k+1)\gamma_a} - \bar{X}_{k+1}, X_{k\gamma_a} - X_{k} \right\rangle \middle| K_k \right] \right]
= \mathbb{E} \left[ \mathbb{E} \left[ \left\langle X_{(k+1)\gamma_a} - \bar{X}_{(k+1)\gamma_a}, X_{k\gamma_a} - X_{k} \right\rangle \middle| K_k \right] \right]
\leq \mathbb{E} \left[ \mathbb{E}^{1/2} \left[ \left\| X_{(k+1)\gamma_a} - \bar{X}_{(k+1)\gamma_a} \right\| \left\| K_k \right\| \left\| X_{k\gamma_a} - X_{k} \right\| \right] \right]
\leq \mathbb{E}^{1/2} \left[ \left\| X_{(k+1)\gamma_a} - \bar{X}_{(k+1)\gamma_a} \right\|^2 \right] \mathbb{E}^{1/2} \left[ \left\| X_{k\gamma_a} - X_{k} \right\|^2 \right]
\leq A_{T,4}^{(c)} A_{T,4}^{1/2} \left\{ \gamma^4(k + 1)^{-4\alpha} + \gamma^2(k + 1)^{-2(1+\alpha)} \right\}^{1/2} (1 + \left\| X_0 \right\|^2) E_k^{1/2}
\leq A_{T,4}^{1/2} A_{T,4}^{1/2} \left\{ \gamma^3(k + 1)^{-3\alpha} \right\} \left\{ 1 + \left\| X_0 \right\|^2 \right\}^{1/2} (1 + \left\| X_0 \right\|^2) E_k^{1/2}
\leq A_{T,6}^{(c)} \left\{ \gamma^3(k + 1)^{-3\alpha} + (k + 1)^{-2(1+\alpha)} \right\} / 2 + a_k^2 E_k / 2 \leq A_{T,6}^{(c)} b_k + a_k^2 E_k .
$$
with $A_{T,6}^{(c)} \geq 0$ which does not depend on $\gamma$ and $k$. 

32
(d) Finally, using the Cauchy-Schwarz inequality, (25), A2 and A1 and that for any \(a, b \geq 0\), 
\((a + b)^{1/2} \leq a^{1/2} + b^{1/2}\), we have
\[
\gamma(k + 1) - \alpha E \left[ (X_{(k+1)} - \bar{X}_{k+1}, H(X_k, Z_k) - H(X_{k\gamma}^k, Z_{k+1})) \right] 
\leq \gamma(k + 1) - \alpha E^{1/2} \left[ \|X_{(k+1)} - \bar{X}_{k+1}\|^2 \right] E^{1/2} \left[ \|H(X_k, Z_k) - H(X_{k\gamma}^k, Z_{k+1})\|^2 \right] 
\leq (\bar{\alpha})^{1/2} \gamma(k + 1) - \alpha k^{1/2} E_k \leq \bar{\alpha}^{(d)} \gamma E_k .
\]

with \(\bar{\alpha}^{(d)} \geq 0\) which does not depend on \(\gamma\) and \(k\).

Let \(\bar{\alpha}\) = \(\bar{\alpha}^{(a)} + \bar{\alpha}^{(b)} + \bar{\alpha}^{(c)} + \bar{\alpha}^{(d)}\). Finally, we have using (25), (26), (27) and (28) in (24)
\[
E_{k+1} \leq \bar{\alpha}^{(a)} k^{1/2} E_k + \bar{\alpha}^{(d)} (a_k^2 + b_k) .
\] (29)

We denote \(v_k = \bar{\alpha}^{(a)} (a_k^2 + a_k)\) and \(w_k = \bar{\alpha}^{(d)} (a_k^2 + b_k)\). Using Lemma 14 and that \(a_k^{1/2} \leq \gamma(k + 1)^{-\alpha} + \gamma^2 (k + 1)^{-2\alpha}\), there exists \(\bar{\alpha}^{(e)} \geq 0\) which does not depend on \(\gamma\) and \(k\) such that
\[
\sum_{k=0}^{N-1} v_k \leq \bar{\alpha}^{(e)} .
\] (30)

In addition, we have that for any \(k \in \mathbb{N}\),
\[
v_k \leq \bar{\alpha}^{(a)} (\gamma^2 (k + 1)^{-2\alpha} + \gamma^4 (k + 1)^{-3\alpha} + \gamma^6 (k + 1)^{-4\alpha} + \gamma^8 (k + 1)^{-6\alpha} + \gamma (k + 1)^{-2(1+\alpha/2)}) .
\]

Using that \(\gamma \gamma^\alpha = \gamma^\alpha\) and Lemma 14 there exists \(\bar{\alpha}^{(f)} \geq 0\) which does not depend on \(\gamma\) and \(k\) such that
\[
\sum_{k=0}^{N-1} v_k \leq \begin{cases} 
\bar{\alpha}^{(f)} (1 + \log(\gamma^{-1})) & \text{if } \alpha \geq 1/2 , \\
\bar{\alpha}^{(f)} \gamma^2 & \text{if } \alpha < 1/2 .
\end{cases}
\] (31)

Using (29) and Lemma 13 we obtain that
\[
E_k \leq \sum_{k=0}^{N-1} w_k + \exp \left[ \sum_{k=0}^{N-1} v_k \right] \sum_{k=0}^{N-1} v_k w_k 
\leq \sum_{k=0}^{N-1} w_k + \exp \left[ \sum_{k=0}^{N-1} v_k \right] \left( \sum_{k=0}^{N-1} v_k \right)^2 .
\] (32)

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 \(W_2\), see Bonis (2020). In Proposition 24, we show that contrary to the SDE setting the gradient flow has an error of order at least \(O(M^{-1})\).
Corollary 23  Let $\bar{\gamma} > 0$ and $\alpha \in [0, 1)$. Assume A1, A2-(b) and A3 (with respect to $(\mathcal{Y}, \mathcal{Y}^\prime, \pi)$). Let $H$ be given by (6). Assume that there exists $x^* \in \mathbb{R}^d$, $C, p \geq 0$ such that for any $x \in \mathbb{R}^d$ and $y \in \mathcal{Y}$

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


Then, there exists a random variable $((B_t)_{t \geq 0}, (Z_n)_{n \in \mathbb{N}})$ 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 $n\gamma \leq T \gamma = \gamma^{1/(1-\alpha)}$ we have

$$\mathbb{E}^{1/2} [||X_{n\gamma} - X_n||^2] \leq C(\gamma^\delta M^{-1} + \gamma)(1 + \log(\gamma^{-1})), \quad \text{with } \delta = \min(1, (2 - 2\alpha)^{-1}).$$

Proof  Let $\bar{\gamma} > 0$, $\alpha \in [0, 1)$ and $M \in \mathbb{N}$. Applying Theorem 1, there exists a random variable $((B_t)_{t \geq 0}, (Z_n)_{n \in \mathbb{N}})$ 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 $n\gamma \leq T \gamma = \gamma^{1/(1-\alpha)}$ we have

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

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

$$\varepsilon^2 = \sup_{k \in \{0, \ldots, nT\}} \mathbb{E} \left[ W_2^2(\nu^{\alpha}(X_{k\gamma^\delta}), \nu^{\alpha}(X_{k\gamma^\delta})) \right], \quad (33)$$

where for any $\bar{x} \in \mathbb{R}^d$, $\nu^{\alpha}(\bar{x})$ is the distribution of $H(\bar{x}, Z_{n+1})$ and $\nu^{\alpha}(\bar{x})$ is the distribution of $\nabla f(\bar{x}) + \gamma^\alpha \Sigma^{1/2}(\bar{x}) \int_{k\gamma}^{(k+1)\gamma} (\alpha + s)^{-\alpha} dB_s$. Our goal is now to control $\varepsilon$ in this specific setting. Let $x \in \mathbb{R}^d$, $k \in \mathbb{N}$ and $(X_k^T, X_k^\tau)$ be an optimal coupling between $\nu_k^\alpha$ and $\nu_k^\varepsilon$. Note that $X_k^\tau$ is a Gaussian random variable with mean $\nabla f(x)$ and covariance matrix $\Sigma(x)$, where $\Sigma(x) = (1/M)\Sigma_f(x)$ with $\Sigma_f = \pi[(\nabla f(x, \cdot) - \nabla f(x))(\nabla f(x, \cdot) - \nabla f(x))^T]$. In particular, we get that

$$W_2^2(\nu^{\alpha}(x), \nu^{\varepsilon}(x)) \leq M^{-1}||\Sigma_f^{1/2}(x)||^2 W_2^2(\tilde{\nu}(x), \nu(x)), \quad (34)$$

where $\tilde{\nu}(x)$ is the distribution of $M^{-1} \sum_{k=1}^{M} \Sigma_f(x)^{1/2} \{ \nabla f(x, Y_i) - \nabla f(x) \}$ with $\{ Y_i \}_{i=1}^{M}$ distributed according to $\pi \otimes M$ and $\nu$ the distribution of a Gaussian random variable with zero mean and identity covariance matrix. Denote $\{ Y_i \}_{i=1}^{M} = (\Sigma_f(x)^{-1/2}(\nabla f(x, Y_i) - \nabla f(x))\}_{i=1}^{M}$. The random variables $\{ Y_i \}_{i=1}^{M}$ are i.i.d., $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}[Y_i Y_i^\top] = \text{Id}$. In addition, using A1 and A2-(b) we have that

$$\|Y_1\| \leq 8\|\Sigma_f(x)^{-1/2}\|^4(\|\nabla f(x, U_1)\|^4 + \|\nabla f(x)\|^4)$$
$$\leq 216C^4(1 + \|x\|^p)^4(\|\nabla f(x^*, U_1)\|^4 + \|\nabla f(0)\|^4 + \|\nabla f(0)\|^4 + \|\nabla f(0)\|^4 + \|\nabla f(0)\|^4 + \|\nabla f(0)\|^4).$$

Combining this result and the fact $\int_\mathcal{X} ||\nabla \tilde{f}(x^*, y)||^4 d\pi(y) < +\infty$, 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 $x \in \mathbb{R}^d$

$$W_2^2(\nu_k^{\alpha}(x), \nu_k^{\varepsilon}(x)) \leq C M^{-1/2}(1 + \|x\|^{2q}).$$

Using Lemma 17, there exists $C \geq 0$ such that $\varepsilon \leq C M^{-1}$, which concludes the proof. \(\blacksquare\)
Proposition 24  Let α ∈ [0, 1/2), T ≥ 0 and γ > 0 such that for any γ ∈ (0, γ], (T − γα)1−2α − γα1−2α ≥ (T/2)1−2α. Let f = 0 and f : Rd → R such that for any x ∈ Rd and z ∈ Rd, f(x, z) = ⟨x, z⟩, (Z, ζ) = ((Rd)M, B(Rd)⊗M), πZ = π⊗M with M ∈ N, π a Gaussian distribution with zero mean and identity covariance matrix. In this case, for any γ ∈ (0, γ] we have for n = ⌊T/γα⌋

\[ \mathbb{E}^{1/2} [\|X_{n\gamma} - X_n\|^2] \geq M^{-1/2} \gamma^2 (1 - 2\alpha)^{-1/2} (T/2)^{1/2 - \alpha}, \]  

with δ = min(1, (2 - 2α)^{-1}),

where (X_t)_{t≥0} is the solution of dX_t = −∇f(X_t)dt and (X_n)_{n∈\mathbb{N}} is a solution of (1).

Proof  Note that for any t ≥ 0, X_t = 0. In addition, for any n ∈ N, we have X_n = (1/M)γ \sum_{k=0}^{n−1}(k+1)^{−α} \sum_{m=1}^{M} Z_{k,m}, where \{Z_{k,m} : k, m ∈ \mathbb{N}\} is a collection of independent Gaussian random variables with zero mean and identity covariance matrix. Therefore we get that

\[ \mathbb{E} \left[ \|X_n\|^2 \right] = (1/M)\gamma^2 \sum_{k=0}^{n−1}(k+1)^{−2\alpha} \geq (1/M)\gamma^2 \int_1^n t^{−2\alpha} dt \geq M^{−1/2}\gamma^2\gamma^{−1}(1−2\alpha)^{−1/2}(T−\gamma\alpha)^{1−2\alpha} \gamma^{1−2\alpha}, \]

which concludes the proof. ■

B.6. Weak approximation

We also derive weak approximation estimates of order 1. Note that in the case where α ≥ 1/2, these weak results are a direct consequence of Theorem 1. Denote by G_{p,k} the set of k-times continuously differentiable functions g such that there exists K ≥ 0 such that for any x ∈ Rd, max(∥∇g(x)∥, . . . , ∥∇^k g(x)∥) ≤ K(1 + ∥x∥^p). We state our main result in Proposition 25.

Proposition 25  Let γ > 0, α ∈ [0, 1) and p ∈ N. Assume that f ∈ G_{p,4}, Σ^{1/2} ∈ G_{p,3}. A1, A2-(b) and A3. Let g ∈ G_{p,2}. In addition, assume that for any m ∈ N there exists x* ∈ Rd such that \( \int_\mathbb{R}^d \|H(x,z)\|^{2m+1} \|d\pi\| < +\infty \). Then for any T ≥ 0, there exists C ≥ 0 such that for any γ ∈ (0, γ], n ∈ N with γα = γ/(1−α), nγα ≤ T we have

\[ |\mathbb{E} [g(X_{n\gamma} − (X_n))]| \leq C\gamma(1 + \log(\gamma^{-1})) . \]

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 \( Y_{n+1} = Y_n + \gamma b(Y_n) + \sqrt{\gamma} \sigma(Y_n) \mathcal{G}_{n+1} \) 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 α ∈ [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 G_p is the set of twice continuously differentiable functions from Rd to R such that for any g ∈ G_p, there exists K ≥ 0 such that for any x ∈ Rd

\[ \max \{ \|\nabla g(x)\|, \|\nabla^2 g(x)\| \} \leq K(1 + \|x\|^p) , \]

with p ∈ N.
Lemma 26  Let \( p \in \mathbb{N}, g \in \mathbb{G}_p \) and let \( K \geq 0 \) as in (35). Then, for any \( x, y \in \mathbb{R}^d \)
\[
|g(y) - g(x) - \langle \nabla g(x), y - x \rangle| \leq K(1 + \|x\|^p + \|y\|^p) \|x - y\|^2.
\]

**Proof** Using that for any \( x \mapsto \|x\|^p \) is convex, and Cauchy-Schwarz inequality we get for any \( x, y \in \mathbb{R}^d \)
\[
|g(x) - g(y) - \langle \nabla g(x), y - x \rangle| \leq \int_0^1 |\nabla^2 g(x + t(y - x))(y - x)\| \, dt \\
\leq \|x - y\|^2 \int_0^1 |\nabla^2 g(x + t(y - x))(y - x)\| \, dt \\
\leq K(1 + \|x\|^p + \|y\|^p) \|x - y\|^2.
\]

\[
\blacksquare
\]

Proposition 27  Let \( \tilde{\gamma} > 0 \) and \( \alpha \in [1/2, 1) \) and \( p \in \mathbb{N} \). Assume A1, A2-(b) and A3. In addition, assume that for any \( m \in \mathbb{N} \) there exists \( x^* \in \mathbb{R}^d \) such that \( \int_{\mathbb{Z}} \|H(x^*, z)\|^{2m} \, d\pi^Z(z) < +\infty \). Then for any \( T \geq 0 \) and \( g \in \mathbb{G}_p \), there exists \( K_{T,\gamma} \geq 0 \) such that for any \( \gamma \in (0, \tilde{\gamma}], k \in \mathbb{N} \) with \( k\gamma \leq T \) and \( X_0 \in \mathbb{R}^d \) we have
\[
\mathbb{E} [\|g(X_{k\gamma}) - g(X_k)\|_{\mathcal{F}_k}] \leq K_{T,\gamma} (1 + \log(\gamma^{-1}))
\]
where \( (X_k)_{k \in \mathbb{N}} \) satisfies the recursion (1) and \( (X_t)_{t \geq 0} \) is the solution of (2) with \( X_0 = X_0 \)

**Proof** Let \( p \in \mathbb{N}, g \in \mathbb{G}_p, \alpha \in [1/2, 1), \tilde{\gamma} > 0, \gamma \in (0, \tilde{\gamma}], k \in \mathbb{N} \), 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
\[
|g(x) - g(y)| \leq \int_0^1 |\langle \nabla g(x + t(y - x)), y - x \rangle| \, dt \leq \|y - x\| \int_0^1 \|\nabla g(x + t(y - x))\| \, dt \\
\leq \|y - x\| K(1 + \|x\|^p + \|y\|^p).
\]
Combining this result, Proposition 22, Lemma 17 and the Cauchy-Schwarz inequality we get that
\[
\mathbb{E} [\|g(X_{k\gamma}) - g(X_k)\|] \leq K_{T,\gamma} (1 + \log(\gamma^{-1}))(A_{T,\gamma} + \bar{A}_{T,\gamma})^{1/2}(1 + \|X_0\|^{2p})^{1/2},
\]
which concludes the proof.

\[
\blacksquare
\]

Proposition 28  Let \( p \in \mathbb{N} \) and \( g \in \mathbb{G}_p \). Let \( \tilde{\gamma} > 0 \) and \( \alpha \in [0, 1) \). Assume A1, A2-(b), A3 and that for any \( m \in \mathbb{N} \) there exists \( x^* \in \mathbb{R}^d \) such that \( \int_{\mathbb{Z}} \|H(x^*, z)\|^{2m} \, d\pi^Z(z) < +\infty \). Then for any \( T \geq 0 \), there exists \( K_{T,\gamma} \geq 0 \) such that for any \( \gamma \in (0, \tilde{\gamma}], k \in \mathbb{N} \) with \( (k + 1)\gamma \leq T \) and \( X_0 \in \mathbb{R}^d \) we have
\[
|\mathbb{E} [g(X_{(k+1)\gamma}) - g(X_{k+1})|\mathcal{G}_k]| \leq K_{T,\gamma} \left\{ \gamma^2(k + 1)^{-2\alpha} + \gamma(k + 1)^{-(1+\alpha)} \right\} (1 + \|X_{k\gamma}\|^{p+2}),
\]
where \( (X_k) \) is the solution of (1) and \( X_t \) is the solution of
\[
X_t = X_k - \int_{k\gamma}^t (s + \gamma\alpha)^{-\alpha} \nabla f(X_s) \, ds + \gamma_{\alpha}^{1/2} \int_{k\gamma}^t (s + \gamma\alpha)^{-\alpha} \Sigma(X_s)^{1/2} \, dB_s.
\]

36
Proof Let $X_{(k+1)\gamma} = X_k - \gamma(k+1)^{-\alpha} \{ \nabla f(X_{k+1}) + \Sigma(X_k)^{1/2}G_k \}$, with $G_k = \gamma^{1/2} f^{(k+1)\gamma} dB_s$.
Using A2 we have $E [X_{(k+1)\gamma}] = E [X_{k+1} | G_k]$. Using Lemma 17, Lemma 18, Lemma 20, Lemma 26 and the Cauchy-Schwarz inequality we have
\[
\left| \mathbb{E} \left[ g(X_{(k+1)\gamma}) - g(X_{k+1}) \right| G_k \right| \\
\leq \mathbb{E} \left[ |\nabla g(X_k)|, X_{(k+1)\gamma} - X_{k+1} | G_k \right] \\
+ K \mathbb{E} \left[ \left\| X_{(k+1)\gamma} - X_k \right\|^2 (1 + \|X_k\|^p + \|X_{(k+1)\gamma}\|^p) \right| G_k \\
+ K \mathbb{E} \left[ \left\| X_{k+1} - X_k \right\|^2 (1 + \|X_k\|^p + \|X_{k+1}\|^p) \right| G_k \\
\leq K(1 + \|X_k\|^p) \mathbb{E} \left[ \left\| X_{(k+1)\gamma} - X_{k+1} \right\|^2 \right| G_k \\
+ 3^{1/2} K \mathbb{E} \left[ \left\| X_{(k+1)\gamma} - X_k \right\|^4 \right]^{1/2} (1 + A_{T,1})^{1/2} (1 + \|X_k\|^p) \\
+ 3^{1/2} K \mathbb{E} \left[ \left\| X_{k+1} - X_k \right\|^4 \right]^{1/2} (1 + A_{k+1})^{1/2} (1 + \|X_k\|^p) \\
\leq K(1 + \|X_k\|^p) A_{T,4}^{1/2} \left\{ \gamma^2 (k+1)^{-2\alpha} + \gamma (k+1)^{-1/2} \right\} (1 + \|X_k\|) \\
+ 3^{1/2} K (k+1)^{-2\alpha} (1 + \|X_k\|^2) (1 + A_T,1)^{1/2} (1 + \|X_k\|^p) \\
+ 3^{1/2} K (k+1)^{-2\alpha} (1 + \|X_k\|^2) (1 + A_T,1)^{1/2} (1 + \|X_k\|^p) \\
+ 3^{1/2} K (k+1)^{-2\alpha} (1 + \|X_k\|^2) (1 + A_T,1)^{1/2} (1 + \|X_k\|^p) \\
+ 3^{1/2} K (k+1)^{-2\alpha} (1 + \|X_k\|^2) (1 + A_T,1)^{1/2} (1 + \|X_k\|^p) ,
\]
which concludes the proof. \(\blacksquare\)

**Proposition 29** Let $\gamma > 0$ and $\alpha \in [0, 1]$. Assume that $f \in \mathbb{G}_{p,4}$, $\Sigma^{1/2} \in \mathbb{G}_{p,3} A_1$, A2-(b) and A3. Let $p \in \mathbb{N}$ and $g \in \mathbb{G}_{p,2}$. In addition, assume that for any $m \in \mathbb{N}$ there exists $x^* \in \mathbb{R}^d$ such that $\int_{\mathbb{R}^d} \|H(x^*, z)\|^{2m} d\mathbb{P} Z(z) < +\infty$. Then for any $T \geq 0$, there exists $A_{T,9} \geq 0$ such that for any $\gamma \in (0, \bar{\gamma})$, $k \in \mathbb{N}$ with $k \gamma \leq T$ and $X_0 \in \mathbb{R}^d$ we have
\[
\left| \mathbb{E} \left[ g(X_{k\gamma}) - g(X_k) \right] \right| \leq A_{T,9} \gamma (1 + \log(\gamma^{-1})) ,
\]
where $(X_k)_{k \in \mathbb{N}}$ satisfies the recursion (1) and $(X_t)_{t \geq 0}$ is the solution of (2) with $X_0 = X_0$.

**Proof** For any $k \in \mathbb{N}$ with $k \gamma \leq T$, let $g_k(x) = \mathbb{E} [g(X_{k\gamma})]$ with $X_0 = x$. 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 (Blagoveschenskii and Freidlin, 1961) or (Kunita, 1981, Proposition 2.1), that there exists $m \in \mathbb{N}$ and $K \geq 0$ such that for any $k \in \mathbb{N}$ $g_k \in C^m(\mathbb{R}^d, \mathbb{R})$ and
\[
\max \{ \|g_k(x)\|, \ldots, \|\nabla^m g_k(x)\| \} \leq K(1 + \|x\|^p) .
\]
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 $k \gamma \leq T$, let $h_k^{(1)}(x) = \mathbb{E} [g_k(X_{k+1})]$ with $X_k = x$ and $h_k^{(2)}(x) = \mathbb{E} [g_k(X_{(k+1)\gamma})]$ with $X_{k\gamma} = x$. Using Proposition 28 we have for any $k \in \mathbb{N}$, $k \gamma \leq T$
\[
\left| h_k^{(1)}(x) - h_k^{(2)}(x) \right| \leq A_{T,8} \left\{ \gamma^2 (k+1)^{-2\alpha} + \gamma (k+1)^{-1/2} \right\} (1 + \|x\|^m) ,
\]
37
Therefore, using Lemma 17 we have for any \( k \in \mathbb{N} \) with \( k \gamma \alpha \leq T \) and \( j \leq k \),
\[
\left| \mathbb{E} \left[ h_{k-j-1}^{(1)}(X_j) - h_{k-j-1}^{(2)}(X_j) \right] \right| \leq \tilde{A}_{T,1} A_{T,8} \left\{ \gamma^2 (k+1)^{-2\alpha} + \gamma (k+1)^{-(1+\alpha)} \right\} (1+\|X_0\|^{m+2}).
\] (36)

Now, let \( k \in \mathbb{N} \) with \( k \gamma \alpha \leq T \) and consider the family \( \{ (X_j^\ell)_{\ell \in \mathbb{N}} : j = 0, \ldots, N \} \), defined by the following recursion: for any \( j \in \{ 0, \ldots, N \} \), \( X_j^0 = X_0 \), and for any \( \ell \in \mathbb{N} \):

(a) if \( \ell \geq j \),
\[
X_{\ell+1}^j = X_j^j - \gamma (k+1)^{-\alpha} H(X_j^\ell, Z_{\ell+1}),
\]
(b) if \( \ell < j \), \( X_{\ell+1}^j = X_j^{(\ell+1)\gamma} \), where \( X_j^{\ell\gamma} = X_j^\ell \) and for any \( t \in [\ell \gamma \alpha, (\ell+1) \gamma \alpha] \) we have
\[
X_j^t = X_j^\ell - \int_{\ell \gamma \alpha}^{t} (\gamma \alpha + s)^{-\alpha} \nabla f(X_s^\ell) \, ds - \gamma^{1/2} \int_{\ell \gamma \alpha}^{t} (\gamma \alpha + s)^{-\alpha} \Sigma^{1/2}(X_s^\ell) \, dB_s.
\]

We have
\[
\left| \mathbb{E} [g(X_{k\gamma}^\alpha) - g(X_k)] \right| = \left| \mathbb{E} [g(X_k^0) - g(X_k^0)] \right| = \sum_{j=0}^{k-1} \left| \mathbb{E} [g(X_{k+j}^{j+1}) - g(X_k^j)] \right|.
\]

Using (36) we get
\[
\left| \mathbb{E} [g(X_{k+j}^{j+1}) - g(X_k^j)] \right| = \left| \mathbb{E} \left[ \mathbb{E} [g(X_k^j) - g(X_k^{j+1}) | X_j^\ell] \right] \right| \\
\leq \tilde{A}_{T,1} A_{T,8} \left\{ \gamma^2 (k+1)^{-2\alpha} + \gamma (k+1)^{-(1+\alpha)} \right\} (1+\|X_0\|^{m+2}) \\
\leq A_{T,9}^{(a)} \gamma^2 (k+1)^{-2\alpha} + \gamma (k+1)^{-(1+\alpha)},
\]

with \( A_{T,9}^{(a)} \geq 0 \) which does not depend on \( k \) or \( \gamma \). In addition, using Lemma 14 there exists \( A_{T,9}^{(b)} \geq 0 \) such that
\[
\sum_{k=0}^{N-1} \left\{ \gamma^2 (k+1)^{-2\alpha} + \gamma (k+1)^{-(1+\alpha)} \right\} \leq A_{T,9}^{(b)} \gamma.
\]

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 A2-(a). We refer to Appendix D for similar results under A2-(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 \((\mathbb{E}[f(X_t)] - \min_{\mathbb{R}^d} f)\) \(_{t \geq 0}\). 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 \((X_t)_{t \geq 0}\) be given by (2). Assume \(A1, A2\)-(a), \(A3\) and \(F1\)-(a). Then there exists \(C \geq 0\) such that for any \(T > 0\), \(\mathbb{E}[f(X_T)] - \min_{\mathbb{R}^d} f \leq CT^{-\alpha}\).

**Proof** The proof is a direct consequence of \(A1\), (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 \(A2\)-(a) and \(F1\)-(a). Then there exists \(B_3 > 0\) such that for all \(N \geq 1\),

\[
\mathbb{E}[\|X_N - x^*\|^2] \leq B_3 N^{-\alpha}.
\]

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 \(F1\)-(a) we get for all \(n \geq 0\),

\[
\mathbb{E}[\|X_{n+1} - x^*\|^2 | F_n] = \mathbb{E}[\|X_n - x^* - \gamma(n + 1)^{-\alpha} H(X_n, Z_{n+1})\|^2 | F_n] = \|X_n - x^*\|^2 + \gamma^2(n + 1)^{-2\alpha} \mathbb{E}[\|H(X_n, Z_{n+1})\|^2 | F_n] - 2\gamma(n + 1)^{-\alpha} \mathbb{E}[\langle X_n - x^*, H(X_n, Z_{n+1}) \rangle | F_n] \leq \|X_n - x^*\|^2 + \gamma^2(n + 1)^{-2\alpha} [\eta + \|\nabla f(X_n)\|^2] - 2\gamma(n + 1)^{-\alpha} \langle X_n - x^*, \nabla f(X_n) \rangle.
\]

Therefore, we have

\[
\mathbb{E}[\|X_{n+1} - x^*\|^2] \leq \mathbb{E}[\|X_n - x^*\|^2] [1 - 2\gamma(n + 1)^{-\alpha} \mu + \gamma^2(n + 1)^{-2\alpha} L^2] + \eta \gamma^2(n + 1)^{-2\alpha}.
\] (37)

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

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

39
Therefore, in the case where \( \alpha < 1 \), there exists \( n_0 \geq 0 \) such that for all \( n \geq n_0 \),
\[
v_{n+1} - v_n \leq -\gamma \mu u_n + \eta \gamma^2 (n + 1)^{-\alpha} \\
\leq -\gamma \mu n^{-\alpha} v_n + \eta \gamma^2 (n + 1)^{-\alpha} \leq (n + 1)^{-\alpha} (-\gamma \mu v_n + \eta \gamma^2).
\]
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 \( n \geq n_1 \),
\[
v_{n+1} - v_n \leq \left[ (1/2 - \gamma \mu) + \gamma^2 L^2 (n + 1)^{-\alpha} + \alpha n^{\alpha-1} \right] u_n + \eta \gamma^2 (n + 1)^{-\alpha}.
\]
Using Lemma 4 this shows that, for \( \alpha \in (0, 1] \), there exists a constant \( B_3 > 0 \) such that for all \( n \geq 0 \), \( v_n \leq B_3 \). This proves the result.

Using A1 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 A1, A2-(a) and F1-(a). Then there exists \( B_4 > 0 \) such that for all \( N \geq 1 \),
\[
\mathbb{E} [f(X_N) - f^*] \leq B_4 N^{-\alpha}.
\]
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 \( F(t, x) \geq -f(x)g(t) \) 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 \( F \in C^1(\mathbb{R}_+ \times \mathbb{R}, \mathbb{R}) \) and \( v \in C^1(\mathbb{R}_+, \mathbb{R}_+) \) such that for all \( t \geq 0 \), \( dv(t)/dt \leq F(t, v(t)) \). Assume that there exist \( f : \mathbb{R} \rightarrow \mathbb{R}, g \in C(\mathbb{R}_+, \mathbb{R}_+), t_0 > 0, A > 0 \) and \( \beta > 0 \) such that the following conditions hold.

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

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

(c) For any \( u \geq A \), \( f(u) > \beta u \).

Then, for any \( t \geq 0 \),
\[
v(t) \leq \max\{A, \exp[\beta (G(t_0) - G(t))] \max_{s \in [0, t_0]} v(s)\},
\]
with \( G(t) = \int_0^t g(s)ds \).
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 $t \geq 0$, $g(t) \geq 0$, we have for any $t \in (0, T]$

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

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

$$d v_T(t)/dt \leq -f(v_T(t))g(t) + \beta v_T(t)g(t) < 0 .$$

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

\[ \text{Theorem 34} \quad \text{Let } \alpha, \gamma \in (0, 1) \text{ and } (X_t)_{t \geq 0} \text{ be given by (2). Assume A1, A2-(a), A3 and F1-(a). Then for any } T \geq 0, \]

$$\mathbb{E}[\|X_T - x^*\|^2] \leq \max \left\{ 4\gamma_\alpha \eta/\mu, C \mathbb{E}[\|X_0 - x^*\|^2] \exp[-\mu(\gamma_\alpha + T)^{1-\alpha}/(2-2\alpha)] \right\}(\gamma_\alpha + T)^{-\alpha},$$

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 $\mathcal{E}(t) = \mathbb{E}[(t + \gamma_\alpha)^\alpha \|X_t - x^*\|^2]$, with $\gamma_\alpha = \gamma^{1/(1-\alpha)}$. Using Dynkin’s formula, see Lemma 48, we have for any $t \geq 0$,

$$\mathcal{E}(t) = \mathcal{E}(0) + \alpha \int_0^t \mathcal{E}(s) \frac{ds}{s + \gamma_\alpha} + \frac{\mathbb{E}[\text{Tr}(\Sigma(X_s))] - \mathbb{E}[\text{Tr}(\Sigma(X_0))]}{(s + \gamma_\alpha)\alpha^\alpha} - 2 \int_0^t \mathbb{E}[(\nabla f(X_s), X_s - x^*)] \frac{ds}{s + \gamma_\alpha}.$$

We now differentiate this expression with respect to $t$ and using F1 and A2, we get for any $t > 0$,

$$d \mathcal{E}(t)/dt = \alpha \mathcal{E}(t)(t + \gamma_\alpha)^{-\alpha} - 2 \mathbb{E}[(\nabla f(X_t), X_t - x^*)] + \gamma_\alpha \mathbb{E}[\text{Tr}(\Sigma(X_t))] (t + \gamma_\alpha)^{-\alpha} \leq \alpha \mathcal{E}(t)/(t + \gamma_\alpha) - 2 \mu \mathbb{E}[\|X_t - x^*\|^2] + \gamma_\alpha \eta/(t + \gamma_\alpha)^\alpha \leq F(t, \mathcal{E}(t)) = \alpha \mathcal{E}(t)(t + \gamma_\alpha)^{-\alpha} - 2 \mu \mathcal{E}(t)(t + \gamma_\alpha)^{-\alpha} + \gamma_\alpha \eta(t + \gamma_\alpha)^{-\alpha},$$

where we have used in the penultimate line that $\text{Tr}(\Sigma(x)) \leq \eta$ for any $x \in \mathbb{R}^d$ by A2. Let $t_0 = \max((\alpha/\mu)^{1/(1-\alpha)} - \gamma_\alpha, \gamma_\alpha)$. We have for any $t \geq t_0$ and $u \geq 0$

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

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 $u \geq A$, $f(u) > \mu u/2$ and therefore condition (c) of Lemma 33 is satisfied. Applying Lemma 33, we obtain that for any $t \geq 0$

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

41
with $B = \exp[\mu(\gamma + t_0)^{1-\alpha}/(2 - 2\alpha)] \max_{s \in [0, t_0]} \mathcal{E}(s)$. We have that $\max_{s \in [0, t_0]} \mathcal{E}(s) \leq (t_0 + \gamma)\alpha^{2} \max_{s \in [0, t_0]} \mathbb{E}[\|X_s - x^*\|^2]$. Using Dynkin’s formula, see Lemma 48, we have for any $t \geq 0$,
\[
\mathbb{E}[\|X_t - x^*\|^2] \leq \mathbb{E}[\|X_0 - x^*\|^2] + \eta\Psi(\alpha, t_0),
\]
with
\[
\Psi(\alpha, t_0) = \begin{cases} 
\gamma^2/(2\alpha - 1) & \text{if } 2\alpha > 1, \\
\gamma \alpha \log(\gamma^{-1}(t_0 + \gamma)) & \text{if } 2\alpha = 1, \\
\gamma(\gamma + t_0)^{1-2\alpha}/(1 - 2\alpha) & \text{otherwise}.
\end{cases}
\]
We conclude the proof upon setting $C = (1 + \eta\Psi(\alpha, t_0)) \exp[\mu(\gamma + t_0)^{1-\alpha}/(2 - 2\alpha)](\gamma + t_0)^{\alpha}$.

**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 $x \in \mathbb{R}^d$,
\[
f(x) - f(x^*) \leq c\|\nabla f(x)\|^2.
\]
(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 $(X_t)_{t \geq 0}$ be given by (2). Assume A1, A2-(a), A3 and that $f$ verifies (39). Then there exists $B_5 > 0$ such that for any $T > 0$,
\[
\mathbb{E}[f(X_T) - f^*] \leq B_5T^{-\alpha}.
\]

**Proof** Let $\alpha, \gamma \in (0, 1)$ and $(X_t)_{t \geq 0}$ be given by (2). Without loss of generality we can assume that $f^* = \min_{x \in \mathbb{R}^d} f(x) = 0$. We note $\mathcal{E}(t) = (t + \gamma)^{\alpha}\mathbb{E}[f(X_t)]$ and we apply Lemma 48 to the stochastic process $(t + \gamma)^{\alpha}f(X_t)_{t \geq 0}$, and using A1, A2-(a), A3, (39) and Lemma 47 this gives, for all $t > 0$,
\[
\mathcal{E}(t) - \mathcal{E}(0) = \int_0^t \alpha(s + \gamma)^{\alpha-1}\mathbb{E}[f(X_s)] \, ds - \int_0^t \mathbb{E}\left[\|\nabla f(X_s)\|^2\right] \, ds
\]
\[
+ (\gamma/2) \int_0^t (s + \gamma)^{-\alpha}\mathbb{E}\left[\operatorname{Tr}(\nabla^2 f(X_s)\Sigma(X_s))\right] \, ds
\]
\[
d\mathcal{E}(t)/dt \leq \alpha\mathcal{E}(t)(t + \gamma)^{-1} - (1/c)\mathcal{E}(t)(t + \gamma)^{-\alpha} + (\gamma/2)L\eta(t + \gamma)^{-\alpha}.
\]
We can now apply Lemma 3 to $F(t, x) = \alpha x(t + \gamma)^{-1} - (1/c)x(t + \gamma)^{-\alpha} + (\gamma/2)L\eta(t + \gamma)^{-\alpha}$ with $t_0 = (2\gamma)^{1/(1-\alpha)}$ and $A = 2\gamma cL\eta$, which shows the existence of $B_5 > 0$ such that for all $t > 0$, $\mathcal{E}(t) \leq B_5$, 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 $B_6 > 0$ such that for all $N \geq 1$,

$$
\mathbb{E} [f(X_N) - f^*] \leq B_6 N^{-\alpha}.
$$

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

$$
\mathbb{E} [f(X_{n+1}) | \mathcal{F}_n] = \mathbb{E} [f(X_n) - \gamma/(n + 1)^\alpha H(X_n, Z_{n+1})] | \mathcal{F}_n] 
\leq f(X_n) - \gamma/(n + 1)^\alpha \mathbb{E} [\langle \nabla f(X_n), H(X_n, Z_{n+1}) \rangle] | \mathcal{F}_n] 
+ \gamma^2/(n + 1)^{2\alpha}(L/2)\mathbb{E} [\|H(X_n, Z_{n+1})\|^2] | \mathcal{F}_n] 
\leq f(X_n) - \gamma/(n + 1)^\alpha \|\nabla f(X_n)\|^2 + (L\gamma^2/2)(n + 1)^{-2\alpha}[\eta + \|\nabla f(X_n)\|^2] 
\mathbb{E} [f(X_{n+1})] - f^* \leq \mathbb{E} [f(X_n)] - f^* + \gamma(n + 1)^{-\alpha}\mathbb{E} [\|\nabla f(X_n)\|^2] [1 + (L\gamma^2/2)(n + 1)^{-\alpha}] 
+ (L\gamma^2/2)(n + 1)^{-2\alpha}\eta.
$$

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

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

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

$$
v_{n+1} - v_n = (n + 1)^\alpha u_{n+1} - n^\alpha u_n 
= (n + 1)^\alpha (u_{n+1} - u_n) + u_n[(n + 1)^\alpha - n^\alpha] 
\leq -(\gamma c^{-1}/2)u_n + (L\gamma^2\eta/2)(n + 1)^{-\alpha} + u_n n^\alpha [(1 + 1/n)^\alpha - 1] 
\leq u_n(-\gamma c^{-1}/2 + \alpha n^{\alpha - 1}) + (L\gamma^2\eta/2)(n + 1)^{-\alpha}.
$$

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

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

This proves the existence of $B_6 > 0$ such that for all $n \geq 0$, $v_n \leq B_6$, which concludes the proof. ■

**Appendix D. Strongly convex case (under $A_2$-(b))**

This section gathers 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 $\text{Tr} \Sigma$ and $E[\|\nabla \tilde{f}\|^2]$. We will note $L_T = 6L^2 + 4L + 3\eta$.

**Lemma 37** Assume $A2$-(b). Then, for all $x \in \mathbb{R}^d$, we have

$$\int_Z \|\nabla \tilde{f}(x,z)\|^2 d\pi^Z(z) \leq L_T(\|x - x^*\|^2 + 1).$$

**Proof** Let $x \in \mathbb{R}^d$. Using $A2$-(b) we have

$$\int_Z \|\nabla \tilde{f}(x,z)\|^2 d\pi^Z(z) = \int_Z \|\nabla \tilde{f}(x,z) - \nabla f(x^*, z) + \nabla \tilde{f}(x^*, z)\|^2 d\pi^Z(z)$$

$$\leq 2 \int_Z \|\nabla \tilde{f}(x,z) - \nabla f(x^*, z)\|^2 + \int_Z \|\nabla \tilde{f}(x^*, z)\|^2 d\pi^Z(z)$$

$$\leq 2L^2\|x - x^*\|^2 + 2\eta,$$

which concludes the proof. □

**Lemma 38** Assume $A2$-(b). Then, for all $x \in \mathbb{R}^d$, we have

$$\text{Tr}(\Sigma(x)) \leq L_T \left(1 + \|x - x^*\|^2\right).$$

**Proof** Let $x \in \mathbb{R}^d$. Using $A2$-(b) we have

$$\text{Tr}(\Sigma(x)) = \text{Tr}\left(\int_Z (\nabla \tilde{f}(x,z) - \nabla f(x))(\nabla \tilde{f}(x,z) - \nabla f(x))^\top d\pi^Z(z)\right)$$

$$= \int_Z \|\nabla \tilde{f}(x,z) - \nabla f(x)\|^2 d\pi^Z(z)$$

$$= \int_Z \|\nabla \tilde{f}(x,z) - \nabla \tilde{f}(x^*, z) + \nabla \tilde{f}(x^*, z) - \nabla f(x)\|^2 d\pi^Z(z)$$

$$\leq 3 \int_Z \left(\|\nabla \tilde{f}(x,z) - \nabla \tilde{f}(x^*, z)\|^2 + \|\nabla \tilde{f}(x^*, z)\|^2 + \|\nabla f(x)\|^2\right) d\pi^Z(z)$$

$$\leq 6L^2\|x - x^*\|^2 + 3\eta \leq L_T \left(1 + \|x - x^*\|^2\right),$$

which concludes the proof. □

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

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

$$\|a + b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle a, b \rangle$$

$$\geq \|a\|^2 + \|b\|^2 - 2\|a\|\|b\| \geq \|a\|^2 + \|b\|^2 - 2\|b\|^2 - \|a\|^2 / 2 \geq \|a\|^2 / 2 - \|b\|^2.$$ □
Lemma 40 Let $f \in C^1(\mathbb{R}^d, \mathbb{R})$. Assume that there exists $L \geq 0$ such that for any $x, y \in \mathbb{R}^d$, $\nabla f$ is $L$-Lipschitz. Then for any $x \in \mathbb{R}^d$

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

(40)

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

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

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 $z \in Z$. For all $x \in \mathbb{R}^d$, we have

$$\int_Z \|\nabla \tilde{f}(x, z)\|^2 d\pi^Z(z) \leq L_T (f(x) - f(x^*)) + 1) .$$

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

$$\tilde{f}(x, z) - \tilde{f}(x^*, z) \geq \langle \nabla \tilde{f}(x^*, z), x - x^* \rangle + \left(1/2L\right)\|\nabla \tilde{f}(x, z) - \nabla \tilde{f}(x^*, z)\|^2$$
$$f(x) - f(x^*) \geq \left(1/2L\right)\mathbb{E}[\|\nabla \tilde{f}(x, z) - \nabla \tilde{f}(x^*, z)\|^2]$$
$$\geq \left(1/4L\right)\mathbb{E}[\|\nabla \tilde{f}(x, z)\|^2] - \left(1/2L\right)\mathbb{E}[\|\nabla f(x)\|^2] ,$$

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 $z \in Z$. For all $x \in \mathbb{R}^d$, we have

$$\text{Tr}(\Sigma(x)) \leq L_T (f(x) - f(x^*)) + 1) .$$

Proof Let $x \in \mathbb{R}^d$. Then using $A_2$-(b) and Lemma 41 we have

$$\text{Tr}(\Sigma(x)) = \int_Z \|\nabla \tilde{f}(x, z) - \nabla f(x)\|^2 d\pi^Z(z)$$
$$= \int_Z \|\nabla \tilde{f}(x, z)\|^2 + \|\nabla f(x)\|^2 - 2\langle \nabla \tilde{f}(x, z), \nabla f(x) \rangle d\pi^Z(z)$$
$$= \int_Z \|\nabla \tilde{f}(x, z)\|^2 d\pi^Z(z) - \|\nabla f(x)\|^2$$
$$\leq \int_Z \|\nabla \tilde{f}(x, z)\|^2 d\pi^Z(z) \leq L_T (f(x) - f(x^*) + 1) ,$$

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 $(X_t)_{t \geq 0}$ be given by (2). Assume A1, A2-(b), A3 and F1-(b). Then there exists $C \geq 0$ (explicit in the proof) such that for any $T \geq 1$, $E[\|X_T - x^*\|^2] \leq CT^{-\alpha}$.

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

$$E(t) = E(0) + \alpha \int_0^t \frac{E(s)}{s + \gamma_\alpha} ds + \int_0^t \gamma_\alpha E[\text{Tr}(\Sigma(X_s))] \frac{ds}{s + \gamma_\alpha} - 2 \int_0^t E[\langle \nabla f(X_s), X_s - x^* \rangle] ds.$$

We now differentiate this expression with respect to $t$ and using F1, A2 and Lemma 38, we get for any $t > 0$,

$$dE(t)/dt = \alpha E(t)(t + \gamma_\alpha)^{-1} - 2E[(\nabla f(X_t), X_t - x^*)] + \gamma_\alpha E[\text{Tr}(\Sigma(X_t))] (t + \gamma_\alpha)^{-\alpha}$$

$$\leq \alpha E(t)(t + \gamma_\alpha)^{-1} - 2\mu E[\|X_t - x^*\|^2] + \gamma_\alpha \mu_T (t + \gamma_\alpha)^{-\alpha} + \gamma_\alpha \mu_T E[\|X_t - x^*\|^2] (t + \gamma_\alpha)^{-\alpha}$$

$$\leq \alpha E(t)(t + \gamma_\alpha)^{-1} - 2\mu E(t)(t + \gamma_\alpha)^{-\alpha} + \gamma_\alpha \mu_T (t + \gamma_\alpha)^{-\alpha} + \gamma_\alpha \mu_T E(t)(t + \gamma_\alpha)^{-2\alpha}.$$

Hence, using Lemma 3 we get, for any $t \geq 0$, $E(t) \leq B$, 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 A2-(b) and F1-(b). Then there exists $B_3 > 0$ such that for all $N \geq 1$,

$$E[\|X_N - x^*\|^2] \leq B_3 N^{-\alpha}.$$

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 F1-(b) and Lemma 37, we get for all $n \geq 0$,

$$E[\|X_{n+1} - x^*\|^2 | F_n] = E[\|X_n - x^* - \gamma(n + 1)^{-\alpha}\nabla f(X_n, Z_{n+1})\|^2 | F_n]$$

$$= \|X_n - x^*\|^2 + \gamma^2 (n + 1)^{-2\alpha} E[\|\nabla f(X_n, Z_{n+1})\|^2 | F_n]$$

$$- 2\gamma(n + 1)^{-\alpha} E[\langle X_n - x^*, \nabla f(X_n, Z_{n+1}) \rangle | F_n]$$

$$\leq \|X_n - x^*\|^2 + L_T \gamma^2 (n + 1)^{-2\alpha} \left[ 1 + \|X_n - x^*\|^2 \right]$$

$$- 2\gamma(n + 1)^{-\alpha} \langle X_n - x^*, \nabla f(X_n) \rangle.$$

Therefore, we have

$$E[\|X_{n+1} - x^*\|^2] \leq E[\|X_n - x^*\|^2] \left[ 1 - 2\gamma(n + 1)^{-\alpha} \mu + \gamma^2 (n + 1)^{-2\alpha} L_T \right] + L_T \gamma^2 (n + 1)^{-2\alpha}.$$

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

\[
v_{n+1} - v_n = (n + 1)^\alpha u_{n+1} - n^\alpha u_n
= (n + 1)^\alpha (u_{n+1} - u_n) + u_n((n + 1)^\alpha - n^\alpha)
\leq [-2\gamma + \gamma^2 L_T(n + 1)^{-\alpha}] u_n + L_T \gamma^2 (n + 1)^{-\alpha} + u_n n^\alpha [(1 + 1/n)^\alpha - 1]
\leq [-2\gamma + \gamma^2 L_T(n + 1)^{-\alpha} + \alpha n^{\alpha-1}] u_n + L_T \gamma^2 (n + 1)^{-\alpha}.
\]

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

\[
v_{n+1} - v_n \leq -\gamma mu_n + L_T \gamma^2 (n + 1)^{-\alpha}
\leq -\gamma mu^\alpha v_n + L_T \gamma^2 (n + 1)^{-\alpha}
\leq (n + 1)^{-\alpha} (\gamma mu_n + L_T \gamma^2).
\]

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 $n \geq n_1$,

\[
v_{n+1} - v_n \leq [(1/2 - \gamma\mu) + \gamma^2 L_T(n + 1)^{-\alpha} + \alpha n^{\alpha-1}] u_n + L_T \gamma^2 (n + 1)^{-\alpha}.
\]

Using Lemma 4 this shows that, for $\alpha \in (0, 1]$, there exists a constant $B_3 > 0$ such that for all $n \geq 0$, $v_n \leq B_3$. 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 $(X_t)_{t \geq 0}$ be given by (2). Assume A1, A2-(b), A3 and that $f$ verifies (39). Then there exists $B_5 > 0$ such that for any $T > 0$,

\[
\mathbb{E}\left[f(X_T) - f^*\right] \leq B_5 T^{-\alpha}.
\]

**Proof** Let $\alpha, \gamma \in (0, 1)$ and $(X_t)_{t \geq 0}$ be given by (2). Without loss of generality we can assume that $f^* = \min_{x \in \mathbb{R}^d} f(x) = 0$. We note $\mathcal{E}(t) = (t + \gamma_0) \mathbb{E}\left[f(X_t)\right]$ and we apply Lemma 48 to the stochastic process $((t + \gamma_0) f(X_t))_{t \geq 0}$, and using A1, A2-(b), (39) and Lemma 42 this gives, for all $t > 0$,

\[
\mathcal{E}(t) - \mathcal{E}(0) = \int_0^t \alpha (s + \gamma_0)^{-\alpha} \mathbb{E}\left[f(X_s)\right] ds - \int_0^t \mathbb{E}\left[\|\nabla f(X_s)\|^2\right] ds
+ (\gamma_0/2) \int_0^t (s + \gamma_0)^{-\alpha} \mathbb{E}\left[\text{Tr}(\nabla^2 f(X_s) \Sigma(X_s))\right] ds
\]
\[
d\mathcal{E}(t)/dt \leq \alpha \mathcal{E}(t)(t + \gamma_0)^{-1} - (1/c)\mathcal{E}(t)(t + \gamma_0)^{-\alpha} + (\gamma_0/2) L_T (1 + \mathbb{E}\left[f(X_t)\right]) (t + \gamma_0)^{-\alpha}
\leq \alpha \mathcal{E}(t)(t + \gamma_0)^{-1} - (1/c)\mathcal{E}(t)(t + \gamma_0)^{-\alpha} + (\gamma_0/2) L_T \mathcal{E}(t)(t + \gamma_0)^{-2\alpha}
+ (\gamma_0/2) L_T (t + \gamma_0)^{-\alpha}.
\]

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 A1, A2-(b) and that $f$ verifies (39). Then there exists $B_6 > 0$ such that for all $N \geq 1$,
\[
E[f(X_N) - f^*] \leq B_6 N^{-\alpha} .
\]
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 A1) and Lemma 41 gives
\[
E[f(X_{n+1})|F_n] = E\left[f(X_n) - \gamma/(n+1)^{\alpha} \nabla \tilde{f}(X_n, Z_{n+1}) \bigg| F_n \right] \\
\leq f(X_n) - \gamma/(n+1)^{\alpha} E\left[\nabla f(X_n), \nabla \tilde{f}(X_n, Z_{n+1}) \bigg| F_n \right] \\
+ \gamma^2/(n+1)^{2\alpha} (L/2) E\left[\left\| \nabla \tilde{f}(X_n, Z_{n+1}) \right\|^2 \bigg| F_n \right] \\
\leq f(X_n) - \gamma/(n+1)^{\alpha} \left\| \nabla f(X_n) \right\|^2 + (L\gamma^2/2)(n+1)^{-2\alpha} L_T [1 + f(X_n)] \\
E[f(X_{n+1})] - f^* \leq E[f(X_n)] - f^* - \gamma(n+1)^{-\alpha} E\left[\left\| \nabla f(X_n) \right\|^2 \bigg| F_n \right] \\
+ (LL\gamma^2/2)(n+1)^{-2\alpha} + (LLT\gamma^2/2)(n+1)^{-2\alpha} (E[f(X_n)] - f^*) .
\]
This shows the existence of $n_2 \geq 0$ such that using (39) we have for all $n \geq n_2$,
\[
E[f(X_{n+1})] - f^* \leq E[f(X_n)] - f^* [1 - (\gamma c^{-1}/2)(n+1)^{-\alpha}] + (LLT\gamma^2/2)(n+1)^{-2\alpha} .
\]
We note now for all $n \geq 0$, $u_n = E[f(X_n)] - f^*$ and $v_n = n^\alpha u_n$. We have
\[
v_{n+1} - v_n = (n+1)^{\alpha} u_{n+1} - n^\alpha u_n \\
= (n+1)^{\alpha} (u_{n+1} - u_n) + u_n ((n+1)^{\alpha} - n^\alpha) \\
\leq - (\gamma c^{-1}/2) u_n + (L\gamma^2 L_T/2)(n+1)^{-\alpha} + u_n n^\alpha [(1 + 1/n)^{\alpha} - 1] \\
\leq u_n (- (\gamma c^{-1}/2) + c n^{\alpha-1}) + (L\gamma^2 L_T/2)(n+1)^{-\alpha} .
\]
If $\alpha < 1$, or if $1 - \gamma c^{-1}/2 < 0$ we have the existence of $n_3 \geq n_2$ and $\tilde{B} > 0$ such that for all $n \geq n_3$,
\[
v_{n+1} - v_n \leq - \tilde{B} u_n + (L\gamma^2 L_T/2)(n+1)^{-\alpha} \\
\leq \{ - \tilde{B} v_n + (L\gamma^2 L_T/2) \} (n+1)^{-\alpha}
\]
This proves the existence of $B_6 > 0$ such that for all $n \geq 0$,
\[
v_n \leq B_6 ,
\]
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 A2-(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 $f \in C^2(\mathbb{R}^d, \mathbb{R})$. Assume $A1$ and $A2-(a)$. Then for any $x \in \mathbb{R}^d$ we have

$$\left| \langle \nabla^2 f(x), \Sigma(x) \rangle \right| \leq L \eta \, , \quad \left| \langle \nabla f(x) \nabla f(x)^\top, \Sigma(x) \rangle \right| \leq \eta^2 \| \nabla f(x) \|^2 \, .$$

Similarly, assume $A1$ and $A2-(b)$. Then there exists $C \geq 0$ such that for any $x \in \mathbb{R}^d$ we have

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

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 $A1$ we have $\| \nabla^2 f(x) \| \leq L$ for all $x \in \mathbb{R}^d$. In addition, denoting $(\lambda_i)_{i \in \{1, \ldots, d\}}$ the eigenvalues of $\Sigma(x)$, using that $\Sigma$ is positive semi-definite and $A2$ we have

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

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

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

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

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

Lemma 48 Let $\alpha \in [0, 1)$ and $\gamma > 0$. Assume $f, g \in C^2(\mathbb{R}^d, \mathbb{R})$, $A1$, $A2$ and $A3$ and let $(X_t)_{t \geq 0}$ solution of (2). Then for any $\varphi \in C^1([0, +\infty), \mathbb{R})$, $Y \in \mathcal{F}_0$ and $\mathbb{E} \left[ \| Y \|^2 + |g(Y)| \right] < +\infty$, we have the following results:

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

$$\mathbb{E} \left[ \| X_t - Y \|^2 \varphi(t) \right] = \mathbb{E} \left[ \| X_0 - Y \|^2 \varphi(0) \right]$$

$$- 2 \int_0^t (\gamma + s)^{-\alpha} \varphi(s) \mathbb{E} [\langle \nabla f(X_s), X_s - Y \rangle] \, ds$$

$$+ \gamma \int_0^t (\gamma + s)^{-2\alpha} \varphi(s) \mathbb{E} [\text{Tr}(\Sigma(X_s))] \, ds + \int_0^t \varphi'(s) \mathbb{E} [\| X_s - Y \|^2] \, ds \, . \quad (42)$$

(b) For any $t \geq 0$

$$\mathbb{E} \left[ (f(X_t) - g(Y)) \varphi(t) \right] = \mathbb{E} \left[ (f(X_0) - g(Y)) \varphi(0) \right]$$

$$- \int_0^t (\gamma + s)^{-\alpha} \varphi(s) \mathbb{E} [\| \nabla f(X_s) \|^2] \, ds$$

$$+ (\gamma / 2) \int_0^t (\gamma + s)^{-2\alpha} \varphi(s) \mathbb{E} [\langle \nabla^2 f(X_s), \Sigma(X_s) \rangle] \, ds$$

$$+ \int_0^t \varphi'(s) \mathbb{E} [f(X_s) - g(Y)] \, ds \, .$$
(c) If \( \mathbb{E}[\|Y\|^{2p}] < +\infty \), then for any \( t \geq 0 \)
\[
\mathbb{E} \left[ \|X_t - Y\|^{2p} \varphi(t) \right] = \mathbb{E} \left[ \|X_0 - Y\|^{2p} \varphi(0) \right] \\
- 2p \int_0^t (\gamma_0 + s)^{-\alpha} \varphi(s) \mathbb{E} \left[ \|\nabla f(X_s), X_s - Y\| \|X_t - Y\|^{2(p-1)} \right] ds \\
+ \gamma_0 p \int_0^t (\gamma_0 + s)^{-2\alpha} \varphi(s) \mathbb{E} \left[ \text{Tr}(\Sigma(X_s)) \|X_s - Y\|^{2(p-1)} \right] ds \\
+ \gamma_0 2p(p-1) \int_0^t (\gamma_0 + s)^{-2\alpha} \varphi(s) \mathbb{E} \left[ (\Sigma(X_s), (X_t - Y)(X_t - Y)^\top \|X_s - Y\|^{2(p-2)} \right] ds \\
+ \int_0^t \varphi'(s) \mathbb{E} \left[ (f(X_s) - g(Y))^{2p} \right] ds .
\]

(d) If \( \mathbb{E}[|g(Y)|^p] < +\infty \), then for any \( t \geq 0 \)
\[
\mathbb{E} \left[ (f(X_t) - g(Y))^p \varphi(t) \right] = \mathbb{E} \left[ (f(X_0) - g(Y))^p \varphi(0) \right] \\
- p \int_0^t (\gamma_0 + s)^{-\alpha} \varphi(s) \mathbb{E} \left[ \|\nabla f(X_s)\|^2 (f(X_s) - g(Y))^{p-1} \right] ds \\
+ \gamma_0 p(p/2) \int_0^t (\gamma_0 + s)^{-2\alpha} \varphi(s) \mathbb{E} \left[ (\nabla f(X_s), \Sigma(X_s))(f(X_s) - g(Y))^{p-2} \right] ds \\
+ \gamma_0 p(p-1)/2 \int_0^t (\gamma_0 + s)^{-2\alpha} \varphi(s) \mathbb{E} \left[ (\nabla f(X_s)\nabla f(X_s)^\top, \Sigma(X_s))(f(X_s) - g(Y))^{p-2} \right] ds \\
+ \int_0^t \varphi'(s) \mathbb{E} \left[ (f(X_s) - g(Y))^p \right] ds .
\]

**Proof** Let \( \alpha \in [0, 1) \), \( \gamma > 0 \) and \((X_t)_{t \geq 0}\) the solution of (2). Note that for any \( t \geq 0 \), we have
\[
\langle X \rangle_t = \gamma_0 \int_0^t (\gamma_0 + s)^{-2\alpha} \text{Tr}(\Sigma(X_s)) ds.
\]

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 \((X_t)_{t \geq 0}\) is a strong solution of (2) we have that \((X_t)_{t \geq 0}\) is a continuous semimartingale. Using this result, the fact that \( F \in C^{1,2}([0, +\infty), \mathbb{R}^d) \) and Itô’s lemma (Karatzas and
Shreve, 1991, Chapter 3, Theorem 3.6) we obtain that for any \( t \geq 0 \) almost surely

\[
F_y(t, X_t) = F_y(0, X_0) + \int_0^t \partial_1 F_y(s, X_s) \, ds + \int_0^t \langle \partial_2 F_y(s, X_s), dX_s \rangle \\
+ (1/2) \int_0^t \langle \partial_{2,2} F_y(s, X_s), d\langle X \rangle_s \rangle \\
= F_y(0, X_0) + \int_0^t \varphi'(s) \|X_s - y\|^2 \, ds + \int_0^t \langle \partial_2 F_y(s, X_s), dX_s \rangle \\
+ (1/2) \int_0^t \langle \partial_{2,2} F_y(s, X_s), d\langle X \rangle_s \rangle \\
= F_y(0, X_0) + \int_0^t \varphi'(s) \|X_s - y\|^2 \, ds - 2 \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \langle \nabla f(X_s), X_s - y \rangle \, ds \\
+ 2\gamma_\alpha^{1/2} \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \langle X_s - y, \Sigma(X_s)^{1/2} dB_s \rangle \\
+ \gamma_\alpha \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \text{Tr}(\Sigma(X_s)) \, ds .
\] (43)

Using A1 have for any \( x \in \mathbb{R}^d \),

\[
\langle \nabla f(x), x - y \rangle \leq \|\nabla f(0)\| \|x - y\| + L \|x\| \|x - y\| .
\]

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 \( \bar{\alpha} \geq 0 \) such that

\[
\sup_{s \in [0,t]} \mathbb{E}[\|X_s - Y\|^2] \leq \bar{\alpha} , \quad \sup_{s \in [0,t]} \mathbb{E}[\langle \nabla f(X_s), X_s - Y \rangle] \leq \bar{\alpha} .
\] (44)

In addition, we have using Lemma 15 that for any \( t \geq 0 \), \( \mathbb{E}[\text{Tr}(\Sigma(X_s))] = \mathbb{E}[\text{Tr}(\Sigma(X_s))] \leq C(1 + \bar{\alpha}) \) if A2-(b) holds or \( \mathbb{E}[\text{Tr}(\Sigma(X_s))] = \mathbb{E}[\text{Tr}(\Sigma(X_s))] \leq \eta \) if A2-(a) holds. Combining these results, (44), (43), that \( \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(t) \langle X_t - Y, \Sigma(X_t)^{1/2} dB_t \rangle_{t \geq 0} \) is a martingale and Fubini-Lebesgue’s theorem we obtain for any \( t \geq 0 \)

\[
\mathbb{E} \left[ \varphi(t) \|X_t - Y\|^2 \right] = \mathbb{E} \left[ \mathbb{E} \left[ F_y(t, X_t) | \mathcal{F}_0 \right] \right] \\
= \mathbb{E} \left[ \varphi(0) \|X_0 - Y\|^2 \right] + \int_0^t \varphi'(s) \mathbb{E} \left[ \|X_s - Y\|^2 \right] \, ds \\
- 2 \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \mathbb{E} [\langle \nabla f(X_s), X_s - Y \rangle] \, ds \\
+ \gamma_\alpha \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \mathbb{E} [\text{Tr}(\Sigma(X_s))] \, ds ,
\]

which concludes the proof of (42).

(b) Second, let \( y \in \mathbb{R}^d \) and \( F : [0, +\infty) \times \mathbb{R}^d \rightarrow \mathbb{R}^d \) 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 \( (X_t)_{t \geq 0} \) is a continuous semi-martingale, the fact that \( F \in C^{1,2}([0, +\infty), \mathbb{R}^d) \)
and Itô’s lemma (Karatzas and Shreve, 1991, Chapter 3, Theorem 3.6) we obtain that for any \( t \geq 0 \) almost surely

\[
F_y(t, X_t) = F_y(0, X_0) + \int_0^t \partial_1 F_y(s, X_s) \, ds + \int_0^t \langle \partial_2 F_y(s, X_s), dX_s \rangle \\
+ (1/2) \int_0^t \langle \partial_{2,2} F_y(s, X_s), d\langle X_s \rangle \rangle \\
= F_y(0, X_0) + \int_0^t \varphi'(s)(f(X_s) - g(y)) \, ds + \int_0^t \langle \partial_2 F_y(s, X_s), dX_s \rangle \\
+ (1/2) \int_0^t \langle \partial_{2,2} F_y(s, X_s), d\langle X_s \rangle \rangle \\
= F_y(0, X_0) + \int_0^t \varphi'(s)(f(X_s) - g(y)) \, ds - \int_0^t (\gamma \alpha + s)^{-\alpha} \varphi(s) \| \nabla f(X_s) \|^2 \, ds \\
+ \gamma^{1/2} \int_0^t (\gamma \alpha + s)^{-\alpha} \varphi(s) \langle \nabla f(X_s), \Sigma(X_s) \rangle^{1/2} dB_s \\
+ (\gamma \alpha / 2) \int_0^t (\gamma \alpha + s)^{-2\alpha} \varphi(s) \langle \nabla^2 f(X_s), \Sigma(X_s) \rangle \, ds.
\]

Using A1 and that for any \( a, b \geq 0, (a + b)^2 \leq 2(a^2 + b^2) \) we have for any \( x, y \in \mathbb{R}^d \),

\[
|f(x) - g(y)| \leq |f(0)| + \| \nabla f(0) || x \| + (L/2) \| x \|^2 + |g(y)|, \quad \| \nabla f(x) \|^2 \leq 2 \| \nabla f(0) \|^2 + 2L^2 \| x \|^2.
\]

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 \( \bar{\alpha} \geq 0 \) such that

\[
\sup_{s \in [0,t]} \mathbb{E}[|f(X_s) - g(Y)|] \leq \bar{\alpha}, \quad \sup_{s \in [0,t]} \mathbb{E}[|\nabla f(X_s)|^2] \leq \bar{\alpha}, \quad \sup_{s \in [0,t]} \mathbb{E}[\| \nabla^2 f(X_s), \Sigma(X_s) \|] \leq \bar{\alpha}.
\]

Combining this result, Lemma 47, the fact that \( (\int_0^t \varphi(s) \langle \nabla f(X_s), \Sigma(X_s) \rangle^{1/2} dB_s)_{t \geq 0} \) is a martingale and Fubini-Lebesgue’s theorem we obtain that for any \( t \geq 0 \)

\[
\mathbb{E} \left[ F_y(t, X_t) \right] = \mathbb{E} \left[ \mathbb{E} \left[ F_y(t, X_t) \big| F_0 \right] \right] \\
= \mathbb{E} \left[ \varphi(0)(f(X_0) - g(Y)) \right] + \int_0^t \varphi'(s) \mathbb{E} \left[ (f(X_s) - g(Y)) \right] \, ds \\
- \int_0^t (\gamma \alpha + s)^{-\alpha} \varphi(s) \mathbb{E} \left[ \| \nabla f(X_s) \|^2 \right] \, ds \\
+ (\gamma \alpha / 2) \int_0^t (\gamma \alpha + s)^{-2\alpha} \varphi(s) \mathbb{E} \left[ \| \nabla^2 f(X_s), \Sigma(X_s) \| \right] \, ds.
\]

(c) Let \( y \in \mathbb{R}^d \) and \( F_y : [0, +\infty) \times \mathbb{R}^d \) such that for any \( t \in [0, +\infty), x, y \in \mathbb{R}^d \), \( F_y(t, x) = \varphi(t) \| x - y \|^{2p} \). Using that \( (X_t)_{t \geq 0} \) is a continuous semi-martingale, that \( F_y \in C^{1,2}([0, +\infty), \mathbb{R}^d) \) and Itô’s lemma (Karatzas and Shreve, 1991, Chapter 3, Theorem 3.6) we obtain that for any \( t \geq 0 \)
Combining these results, Lemma 47, and for any using this result Lemma 17, Cauchy-Schwarz’s inequality and that
Using almost surely
\[
F_y(t, X_t) = F_y(0, X_0) + \int_0^t \partial_1 F_y(s, X_s) ds + \int_0^t \langle \partial_2 F_y(s, X_s), dX_s \rangle \\
+ (1/2) \int_0^t \langle \partial_2, F_y(s, X_s), d\langle X_s \rangle_s \rangle \\
= F_y(0, X_0) + \int_0^t \varphi'(s) \|X_s - y\|^{2p} ds + \int_0^t \langle \partial_2 F_y(s, X_s), dX_s \rangle \\
+ (1/2) \int_0^t \langle \partial_2, F_y(s, X_s), d\langle X_s \rangle_s \rangle \\
= F_y(0, X_0) + \int_0^t \varphi'(s) \|X_s - y\|^{2p} ds \\
- 2p \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \langle \nabla f(X_s), X_s - y \rangle \|X_s - y\|^{2(p-1)} ds \\
+ 2p \gamma_\alpha \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \langle X_s - y, \Sigma(X_s) \rangle \|X_s - y\|^{2(p-1)} dB_s \\
+ p \gamma_\alpha \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \langle (X_s - y)\Sigma(X_s) (X_s - y)\rangle \|X_s - y\|^{2(p-1)} ds \\
+ 2p(p-1) \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \langle X_s - y, X_s - y \rangle \Sigma(X_s) \|X_s - y\|^{2(p-2)} ds.
\]

Using A1 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 \(\tilde{A} \geq 0\) such that
\[
\sup_{s \in [0, t]} \mathbb{E} \left[\|X_s - Y\|^{2p} \right] \leq \tilde{A}, \quad \sup_{s \in [0, t]} \mathbb{E} \left[\|\nabla f(X_s), X_s - Y\|^{2(p-1)} \right] \leq \tilde{A},
\]
and
\[
\sup_{s \in [0, t]} \mathbb{E} \left[\|\Sigma^{1/2}(X_s)\| \|X_s - Y\|^{2p-1} \right] \leq \tilde{A}, \quad \sup_{s \in [0, t]} \mathbb{E} \left[\|\Sigma(X_s)\| \|X_s - y\|^{2(p-1)} \right] \leq \tilde{A}.
\]

Combining these results, Lemma 47, that \(\left(\int_0^t \varphi(s) \langle \nabla f(X_s), \Sigma(X_s)^{1/2}(f(X_s) - g(Y))^{p-1} dB_s \rangle \right)_{t \geq 0}\) is a martingale and Fubini-Lebesgue’s theorem we obtain that for any \(t \geq 0\)
\[
\mathbb{E} \left[ F_y(t, X_t) \right] = \mathbb{E} \left[ \mathbb{E} \left[ F_Y(t, X_t) | F_0 \right] \right] \\
= \mathbb{E} \left[ \varphi(0) \|X_0 - Y\|^{2p} \right] + \int_0^t \varphi'(s) \mathbb{E} \left[\|X_s - Y\|^{2p} \right] ds \\
- 2p \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \mathbb{E} \left[\langle \nabla f(X_s), X_s - y \rangle \|X_s - y\|^{2(p-1)} \right] ds \\
+ \gamma_\alpha p \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \mathbb{E} \left[\text{Tr}(\Sigma(X_s)) \|X_s - y\|^{2(p-1)} \right] ds \\
+ 2\gamma_\alpha p(p-1) \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \mathbb{E} \left[\langle (X_s - y)\nabla (X_s - y)\rangle^\top, \Sigma(X_s) \rangle \|X_s - y\|^{2(p-2)} \right] ds.
\]
(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 \((X_t)_{t \geq 0}\) is a continuous semi-martingale, the fact that \( F \in C^{1,2}([0, +\infty), \mathbb{R}^d)\) and Itô's lemma (Karatzas and Shreve, 1991, Chapter 3, Theorem 3.6) we obtain that for any \( t \geq 0 \) almost surely

\[
F_y(t, X_t) = F_y(0, X_0) + \int_0^t \partial_1 F_y(s, X_s) ds + \int_0^t \langle \partial_2 F_y(s, X_s), dX_s \rangle \\
+ (1/2) \int_0^t \langle \partial_{22} F_y(s, X_s), d(X_s)^2 \rangle \\
= F_y(0, X_0) + \int_0^t \varphi'(s)(f(X_s) - g(y))^{2p} ds \\
+ \int_0^t \langle \partial_2 F_y(s, X_s), dX_s \rangle + (1/2) \int_0^t \langle \partial_{22} F_y(s, X_s), d(X_s)^2 \rangle \\
= F_y(0, X_0) + \int_0^t \varphi'(s)(f(X_s) - g(y))^{2p} ds \\
- 2p \int_0^t (\gamma + s)^{-\alpha} \varphi(s) \|\nabla f(X_s)\|^2 (f(X_s) - g(y))^{2(p-1)} ds \\
+ 2p \int_0^t \varphi(s) \langle \nabla f(X_s), \Sigma(X_s)^{1/2} (f(X_s) - g(y))^{2(p-1)} dB_s \rangle \\
+ p \gamma \int_0^t (\gamma + s)^{-2\alpha} \varphi(s) \langle \nabla^2 f(X_s), \Sigma(X_s) \rangle (f(X_s) - g(y))^{2(p-1)} ds \\
+ 2p(p-1) \int_0^t (\gamma + s)^{-2\alpha} \varphi(s) \langle \nabla f(X_s) \nabla f(X_s)^\top, \Sigma(X_s) \rangle (f(X_s) - g(y))^{2(p-2)} ds
\]

Using A1 and that for any \( a, b \geq 0 \), \((a + b)^2 \leq 2(a^2 + b^2)\) we have for any \( x, y \in \mathbb{R}^d \),

\[
|f(x) - g(y)|^{2p} \leq 4^{2p-1} |f(0)|^{2p} + 4^{2p-1} \|\nabla f(0)\|^2 |x|^{2p} + (4^{2p-1} L/2) \|x\|^{4p} + 4^{2p-1} |g(y)|^{2p}, \\
\|\nabla f(x)\|^2 \leq 2 \|\nabla f(0)\|^2 + 2L^2 \|x\|^2.
\]

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 \( \tilde{A} \geq 0 \) such that

\[
\sup_{s \in [0, t]} \mathbb{E} \left[ |f(X_s) - g(Y)|^{2p} \right] \leq \tilde{A}, \\
\sup_{s \in [0, t]} \mathbb{E} \left[ \|\nabla f(X_s)\|^2 |f(X_s) - g(Y)|^{2(p-1)} \right] \leq \tilde{A}, \\
\sup_{s \in [0, t]} \mathbb{E} \left[ \langle \nabla f(X_s) \nabla f(X_s)^\top, \Sigma(X_s) \rangle (f(X_s) - g(Y))^{2(p-2)} \right] \leq \tilde{A}.
\]
Combining this result, Lemma 47, that \( \int_0^t \varphi(s) \langle \nabla f(X_s), \Sigma(X_s)^{1/2}(f(X_s) - g(Y))^{p-1}dB_s \rangle_{t \geq 0} \) is a martingale and Fubini-Lebesgue’s theorem we obtain that for any \( t \geq 0 \)

\[
\mathbb{E}[F_{\gamma}(t, X_t)] = \mathbb{E}[\mathbb{E}[F_{\gamma}(t, X_t)|F_0]]
\]

\[
= \mathbb{E}[\varphi(0)(f(X_0) - g(Y))^{2p}] + \int_0^t \varphi'(s) \mathbb{E}[(f(X_s) - g(Y))^{2p}] ds
- 2p \int_0^t (\gamma_\alpha + s)^{-\alpha} \varphi(s) \mathbb{E}\left[\|\nabla f(X_s)\|^{2} (f(X_s) - g(y))^2(p-1)\right] ds
+ \gamma_\alpha p \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \mathbb{E}\left[\langle \nabla^2 f(X_s), \Sigma(X_s)(f(X_s) - g(Y))^2(p-1)\right] ds
+ 2\gamma_\alpha p(1 - 1) \int_0^t (\gamma_\alpha + s)^{-2\alpha} \varphi(s) \mathbb{E}\left[\langle \nabla f(X_s)\rangle \Sigma(X_s)(f(X_s) - g(Y))^2(p-2)\right] ds.
\]

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

**Lemma 49** Assume A1, F2-(a), A3 and that arg min\( _{x \in \mathbb{R}^d} f \) is bounded. Then there exists \( f_\varepsilon \varepsilon > 0 \) such that for any \( \varepsilon > 0 \), \( f_\varepsilon \) is convex, \( C^2 \) with \( L \)-Lipschitz continuous gradient. In addition, there exists \( C \geq 0 \) such that the following properties are satisfied.

(a) For all \( \varepsilon > 0 \), \( f_\varepsilon \) admits a minimizer \( x^*_\varepsilon \) and \( \lim_{\varepsilon \to 0} f_\varepsilon(x^*_\varepsilon) \leq f(x^*) \).

(b) \( \lim \inf_{\varepsilon \to 0} \|x^*_\varepsilon\| \leq C \).

(c) for any \( T \geq 0 \), \( \lim_{\varepsilon \to 0} \mathbb{E}[|f_\varepsilon(X_{T, \varepsilon}) - f(X_T)|] = 0 \), where \( (X_{t, \varepsilon})_{t \geq 0} \) is the solution of (2) replacing \( f \) by \( f_\varepsilon \).

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

First, we show that for any \( \varepsilon \), \( f_\varepsilon \) is convex and \( \nabla f_\varepsilon \) is L-Lipschitz continuous. Let \( \varepsilon > 0 \), \( x, y \in \mathbb{R}^d \) and \( t \in [0, 1] \). Using F2-(a) we have

\[
f_\varepsilon(tx + (1-t)y) = \int_{\mathbb{R}^d} \varphi_\varepsilon(z)(tx + (1-t)y - z) dz \
\leq \int_{\mathbb{R}^d} \{tf(x-z) + (1-t)f(y-z)\} \varphi_\varepsilon(z) dz \
\leq tf_\varepsilon(x) + (1-t)f_\varepsilon(y).
\]

Hence, \( f_\varepsilon \) is convex. In addition, using A1 and that \( \int_{\mathbb{R}^d} \varphi_\varepsilon(z) dz = 1 \) we have

\[
\|\nabla f_\varepsilon(x) - \nabla f_\varepsilon(y)\| \leq \int_{\mathbb{R}^d} \|\nabla f(x-z) - \nabla f(y-z)\| \varphi_\varepsilon(z) dz \leq L \|x-y\,
\]

which proves that \( \nabla f_\varepsilon \) is L-Lipschitz continuous.
Second we show that $f_\varepsilon$ and $\nabla f_\varepsilon$ converge uniformly towards $f$ and $\nabla f$. Let $\varepsilon > 0$, $x \in \mathbb{R}^d$. Using the convexity of $f$ and that $\varphi_\varepsilon$ is even, we get

$$f_\varepsilon(x) - f(x) = \int_{\mathbb{R}^d} (f(x - z) - f(x)) \varphi_\varepsilon(z) \, dz$$

$$\geq - \int_{\mathbb{R}^d} \langle \nabla f(x), z \rangle \varphi_\varepsilon(z) \, dz$$

$$\geq - (\nabla f(x), \int_{\mathbb{R}^d} z \varphi_\varepsilon(z) \, dz) \geq 0,$$  \hspace{1cm} (45)

Conversely, using the descent lemma (Nesterov, 2004, Lemma 1.2.3) and that $\varphi_\varepsilon$ is even, we have

$$f_\varepsilon(x) - f(x) = \int_{\mathbb{R}^d} (f(x - z) - f(x)) \varphi_\varepsilon(z) \, dz$$

$$\leq \int_{\mathbb{R}^d} \left(-\langle \nabla f(x), z \rangle + (L/2) \|z\|^2\right) \varphi_\varepsilon(z) \, dz$$

$$\leq (L/2) \int_{\mathbb{R}^d} \varepsilon^2 \|z/\varepsilon\|^2 e^{-d} \varphi(z/\varepsilon) \, dz \leq (L/2)\varepsilon^2 \int_{\mathbb{R}^d} \|u\|^2 \varphi(u) \, du.$$ \hspace{1cm} (46)

Combining (45) and (46) we get that $\lim_{\varepsilon \to 0} \|f - f_\varepsilon\|_\infty = 0$. Using A1 we have for any $x \in \mathbb{R}^d$

$$\|\nabla f_\varepsilon(x) - \nabla f(x)\| \leq \|(\nabla f)_\varepsilon(x) - \nabla f(x)\|$$

$$\leq \int_{\mathbb{R}^d} \|\nabla f(x - z) - \nabla f(x)\| \varphi_\varepsilon(z) \, dz \leq L \varepsilon \int_{\mathbb{R}^d} \|z\| \varphi(z) \, dz,$$

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 $\varepsilon > 0$, $f_\varepsilon$ is coercive.

We divide the rest of the proof into three parts.

(a) Let $\varepsilon > 0$. Since $f_\varepsilon$ is coercive and continuous it admits a minimizer $x^*_\varepsilon$. In addition, we have

$$f_\varepsilon(x^*_\varepsilon) \leq f(x^*) \leq f(x^*) + \|f_\varepsilon - f\|_\infty.$$ \hspace{1cm} (47)

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

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

(c) Let $\varepsilon > 0$, $T \geq 0$ and $(X_{t,\varepsilon})_{t \geq 0}$ be the solution of (2) replacing $f$ by $f_\varepsilon$. Using (2), the fact that $\lim_{\varepsilon \to 0} \|\nabla f - \nabla f_\varepsilon\|_\infty = 0$, A1 and Grönwall’s inequality (Pachpatte, 1998, Theorem 1.2.2) we have

$$\mathbb{E} \left[\|X_{T,\varepsilon} - X_T\|^2\right] \leq \mathbb{E} \left[\left\|\int_0^T \left((\alpha_\varepsilon + s)^{-\alpha} \{ - \nabla f_\varepsilon(X_{t,\varepsilon}) + \nabla f(X_t)\}\right) dt\right|^2\right]$$

$$\leq 2\gamma_\varepsilon^{-2\alpha} T \left\|\int_0^T \left(\|\nabla f(X_{t,\varepsilon}) - \nabla f(X_t)\|^2\right) dt + 2\gamma_\varepsilon^{-2\alpha} T^2 \|\nabla f - \nabla f_\varepsilon\|_\infty^2\right\|$$

$$\leq 2L \gamma_\varepsilon^{-2\alpha} T \left\|\int_0^T \left(\|X_{t,\varepsilon} - X_t\|^2\right) dt + 2\gamma_\varepsilon^{-2\alpha} T^2 \|\nabla f - \nabla f_\varepsilon\|_\infty^2\right\|$$

$$\leq 2\gamma_\varepsilon^{-2\alpha} T^2 \|\nabla f - \nabla f_\varepsilon\|_\infty^2 \exp \left[2L \gamma_\varepsilon^{-2\alpha} T^2\right].$$ \hspace{1cm} (48)
Therefore $\lim_{\varepsilon \to 0} \mathbb{E} \left[ \|X_{T,\varepsilon} - X_T\|^2 \right] = 0$. In addition, using the Cauchy-Schwarz inequality, A1 and Lemma 17 we have

$$\mathbb{E} \left[ |f(X_{T,\varepsilon}) - f(X_T)| \right] \leq \mathbb{E} \left[ \int_0^1 \|\nabla f(X_T + t(X_{T,\varepsilon} - X_T))\| \|X_{T,\varepsilon} - X_T\| \, dt \right]$$

\[ \leq \mathbb{E} \left[ \|X_{T,\varepsilon}\| + \|X_T\| + \|x^*\| \right] \|X_{T,\varepsilon} - X_T\| \]

\[ \leq 3^{1/2} \left( \|x^*\|^2 + \mathbb{E} \left[ \|X_T\|^2 \right] + \mathbb{E} \left[ \|X_{T,\varepsilon}\|^2 \right] \right)^{1/2} \mathbb{E} \left[ \|X_{T,\varepsilon} - X_T\|^2 \right]^{1/2} \]

\[ \leq 3^{1/2} (\|x^*\|^2 + 2A_{T,1})^{1/2} (1 + \|x_0\|^2)^{1/2} \mathbb{E} \left[ \|X_{T,\varepsilon} - X_T\|^2 \right]^{1/2}. \]  \(49\)

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

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

which concludes the proof.

\[ \square \]

**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 $\alpha \leq 1 - \alpha$.

\[ \square \]

**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 $\mathbb{E}[\|X_t - x^*\|^2]$. 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 $\mathbb{E}[f(X_T)] - f^*$ 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 $f \in C^1(\mathbb{R}^d, \mathbb{R})$.

**Lemma 51** Assume F2-(a). Let $(X_t)_{t \geq 0}$ be given by (2). Then, for any $\alpha, \gamma \in (0, 1)$, there exists $c_{1,\alpha}^{(c)} \geq 0$ and $c_{2,\alpha}^{(c)} \geq 0$ and a function $\Phi_{\alpha}^{(c)} : \mathbb{R}_+ \to \mathbb{R}_+$ such that, for any $t \geq 0$,

$$\mathbb{E}[\|X_t - x^*\|^2] \leq c_{1,\alpha}^{(c)} \Phi_{\alpha}^{(c)}(t + \gamma_\alpha) + c_{2,\alpha}^{(c)}.$$

And we have

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

57
The values of the constants are given by

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

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

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

\[ \forall (t, x) \in \mathbb{R} \times \mathbb{R}^d, F(t, x) = \|x - x^*\|^2 . \]

Applying Lemma 48 to the stochastic process \((F(t, X_t))_{t \geq 0}\) and using F2-(a) and A2-(a) gives that for all \( t \geq 0 \),

\[
\begin{align*}
\mathbb{E} \left[ \|X_t - x^*\|^2 \right] - \mathbb{E} \left[ \|X_0 - x^*\|^2 \right] \\
= -2 \int_0^T (t + \gamma_\alpha)^{-\alpha} \langle X_t - x^*, \nabla f(X_t) \rangle dt + \int_0^T \gamma_\alpha (t + \gamma_\alpha)^{-2\alpha} \text{Tr}(\Sigma(X_t)) dt \\
\leq \gamma_\alpha \eta \int_0^T (t + \gamma_\alpha)^{-2\alpha} dt .
\end{align*}
\]

We now distinguish three cases:

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

\[
\begin{align*}
\mathbb{E} \left[ \|X_t - x^*\|^2 \right] &\leq \|X_0 - x^*\|^2 + \gamma_\alpha \eta (1 - 2\alpha)^{-1}((T + \gamma_\alpha)^{1-2\alpha} - \gamma_\alpha^{1-2\alpha}) \\
&\leq \|X_0 - x^*\|^2 + \gamma_\alpha \eta (1 - 2\alpha)^{-1}(T + \gamma_\alpha)^{1-2\alpha} .
\end{align*}
\]

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

\[
\begin{align*}
\mathbb{E} \left[ \|X_t - x^*\|^2 \right] &\leq \|X_0 - x^*\|^2 + \gamma_\alpha \eta (\log(T + \gamma_\alpha) - \log(\gamma_\alpha)) \\
&\leq \gamma_\alpha \eta \log(T + \gamma_\alpha) + \|X_0 - x^*\|^2 - \gamma_\alpha \eta \log(\gamma_\alpha) .
\end{align*}
\]

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

\[
\begin{align*}
\mathbb{E} \left[ \|X_t - x^*\|^2 \right] &\leq \|X_0 - x^*\|^2 + \gamma_\alpha \eta (1 - 2\alpha)^{-1}((T + \gamma_\alpha)^{1-2\alpha} - \gamma_\alpha^{1-2\alpha}) \\
&\leq \|X_0 - x^*\|^2 + (2\alpha - 1)^{-1} \gamma_\alpha^2 \eta^2 .
\end{align*}
\]

We now turn to the proof of Theorem 6. Let \( f \in C^2(\mathbb{R}^d, \mathbb{R}) \). Let \( \gamma \in (0, 1) \) and \( \alpha \in (0, 1/2] \) and \( T \geq 1 \). Let \((X_t)_{t \geq 0}\) be given by (2).
Let \( S : [0, T] \to [0, +\infty) \) defined by
\[
\begin{align*}
S(t) &= t^{-1} \int_{T-t}^T \{ \mathbb{E}[f(X_s)] - f^* \} \, ds, \quad \text{if } t > 0, \\
S(0) &= \mathbb{E}[f(X_T)],
\end{align*}
\]
(50)
With this notation we have
\[
\mathbb{E}[f(X_T)] - f^* = S(0) - S(1) + S(1) - S(T) + S(T) - f^*.
\]
We are now going to control each one of the three terms \((S(0) - S(1)), (S(1) - S(T)), (S(T) - f^*)\) 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 \( S(T) - f^* \) (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 \( \varepsilon \) 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 A1, A2-(a), A3, and F2-(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 \( \mathbb{E}[\|Y - x^*\|^2] \leq C_{1,\alpha}^{(c)} \Phi_\alpha^{(c)}(T + \gamma_\alpha) + C_{2,\alpha}^{(c)} \) with \( C_{1,\alpha}^{(c)} \) and \( C_{2,\alpha}^{(c)} \) given in Lemma 51, we have
\[
\int_{T-u}^T \mathbb{E}[f(X_t) - f(Y)] \, dt \leq (C_1/2) ((T + \gamma_\alpha)^\alpha - (T - u + \gamma_\alpha)^\alpha)
+ (1/2)(T - u + \gamma_\alpha)^\alpha \mathbb{E}[[X_{T-u} - Y]^2]
+ (C_1/2) ((T + \gamma_\alpha)^{1-\alpha} - (T - u + \gamma_\alpha)^{1-\alpha}) \log(T + \gamma_\alpha),
\]
with \( C_1 = \max(4C_{2,\alpha}^{(c)}, (\gamma_\alpha \eta + 4\alpha C_{1,\alpha}^{(c)})(1 - \alpha)^{-1}) \), with \( C_{1,\alpha}^{(c)} \) and \( C_{2,\alpha}^{(c)} \) 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
\[
F_{x_0}(t, x) = (t + \gamma_\alpha)^\alpha \|x - x_0\|^2.
\]
(51)
Using Lemma 51 and that for any \( a, b \geq 0 \), \((a + b)^2 \leq 2a^2 + 2b^2\), we have
\[
\mathbb{E} \left[ \|X_t - Y\|^2 \right] = \mathbb{E} \left[ \|(X_t - x^*) + (x^* - Y)\|^2 \right]
\leq 2 \mathbb{E} \left[ \|X_t - x^*\|^2 \right] + 2 \mathbb{E} \left[ \|Y - x^*\|^2 \right]
\leq 2C_{1,\alpha}^{(c)} \Phi_\alpha^{(c)} (t + \gamma_\alpha) + 4C_{2,\alpha}^{(c)} + 2C_{1,\alpha}^{(c)} \Phi_\alpha^{(c)} (T + \gamma_\alpha)
\leq 2C_{1,\alpha}^{(c)} \Phi_\alpha^{(c)} (t + \gamma_\alpha) + 2C_{1,\alpha}^{(c)} \Phi_\alpha^{(c)} (T + \gamma_\alpha) + C_3^{(c)}.
\]
Therefore, we get for any \( u \in C \) with \( C \) and \( \alpha \) such that

\[
(t + \gamma_\alpha)^{\alpha-1} \mathbb{E} \left[ \|X_t - Y\|^2 \right] \leq \left\{ C_{3,\alpha}^{(c)} + 2 \mathcal{C}_{1,\alpha}^{(c)} (T + \gamma_\alpha)^{1-2\alpha} \log(T + \gamma_\alpha) \right\} (t + \gamma_\alpha)^{\alpha-1} \tag{52}
\]

with \( C_{1,\alpha}^{(c)} = 0 \) 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

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

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

\[
\mathbb{E} [F_Y(T, X_T)] - \mathbb{E} [F_Y(T - u, X_{T-u})] = \int_{T-u}^T \alpha (t + \gamma_\alpha)^{\alpha-1} \mathbb{E} \left[ \|X_t - Y\|^2 \right] dt - 2 \int_{T-u}^T \mathbb{E} \langle [X_t - Y, \nabla f(X_t)] \rangle dt \tag{53}
\]

\[
+ \int_{T-u}^T \gamma_\alpha (t + \gamma_\alpha)^{-\alpha} \mathbb{E} \left[ \text{Tr}(\Sigma(X_t)) \right] dt.
\]

Combining this result, \( F2-(a), A2-(a), (51), (52) \) and \( (53) \) we obtain for any \( u \in [0, T] \)

\[
-(T - u + \gamma_\alpha)^{\alpha} \mathbb{E} \left[ \|X_{T-u} - Y\|^2 \right] \leq C_{3,\alpha}^{(c)} \int_{T-u}^T \alpha (t + \gamma_\alpha)^{\alpha-1} dt + \eta_\gamma \int_{T-u}^T (t + \gamma_\alpha)^{-\alpha} dt
\]

\[
+ 2 \alpha \mathcal{C}_{1,\alpha}^{(c)} \log(T + \gamma_\alpha) \left\{ \int_{T-u}^T (t + \gamma_\alpha)^{\alpha-1} dt + (T + \gamma_\alpha)^{1-2\alpha} \int_{T-u}^T (t + \gamma_\alpha)^{\alpha-1} dt \right\}
\]

\[
- 2 \int_{T-u}^T \mathbb{E} \left[ f(X_t) - f(Y) \right] dt
\]

\[
\leq C_{3,\alpha}^{(c)} ((T + \gamma_\alpha)^{\alpha} - (T - u + \gamma_\alpha)^{\alpha}) - 2 \int_{T-u}^T \mathbb{E} [f(X_t) - f(Y)] dt
\]

\[
+ (\gamma_\alpha \eta + 2 \alpha \mathcal{C}_{1,\alpha}^{(c)} (1 - \alpha)^{-1} ((T + \gamma_\alpha)^{1-\alpha} - (T - u + \gamma_\alpha)^{1-\alpha}) \log(T + \gamma_\alpha)
\]

\[
+ 2 \alpha \mathcal{C}_{1,\alpha}^{(c)} \log(T + \gamma_\alpha) \{ (T + \gamma_\alpha)^{\alpha} - (T - u + \gamma_\alpha)^{\alpha} \} (T + \gamma_\alpha)^{1-2\alpha}.
\]

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

\[
\int_{T-u}^T \mathbb{E} \left[ f(X_t) - f(Y) \right] dt \leq (C_1/2) ((T + \gamma_\alpha)^{\alpha} - (T - u + \gamma_\alpha)^{\alpha})
\]

\[
+ (1/2)(T - u + \gamma_\alpha)^{\alpha} \mathbb{E} \left[ \|X_{T-u} - Y\|^2 \right]
\]

\[
+ (C_1/2) ((T + \gamma_\alpha)^{1-\alpha} - (T - u + \gamma_\alpha)^{1-\alpha}) \log(T + \gamma_\alpha),
\]

with \( C_1 = \max(C_{3,\alpha}^{(c)}, (\gamma_\alpha \eta + 4 \alpha \mathcal{C}_{1,\alpha}^{(c)} (1 - \alpha)^{-1}) \).
Lemma 53 Assume A1, A2-(a), A3, and F2-(a). Then, for any \( \alpha, \gamma \in (0, 1) \) and \( T \geq 0 \) we have
\[
S(1) - S(T) \leq 2c_1 \log(T + \gamma_\alpha) \log(1 + T)(T + \gamma_\alpha)^{-\min(\alpha, 1-\alpha)} ,
\]
with \( S \) given in (50).

Proof In the case where \( \alpha \leq 1/2 \), Lemma 50 gives that for all \( u \in [0, T] \):
\[
((T + \gamma_\alpha)^{\alpha} - (T - u + \gamma_\alpha)^{\alpha}) \leq ((T + \gamma_\alpha)^{1-\alpha} - (T - u + \gamma_\alpha)^{1-\alpha}) ,
\]
and we also have, for all \( u \in [0, T] \):
\[
(T + \gamma_\alpha)^{1-\alpha} - (T + \gamma_\alpha - u)^{1-\alpha} \\
= \left( ((T + \gamma_\alpha)^{1-\alpha} - (T + \gamma_\alpha - u)^{1-\alpha})((T + \gamma_\alpha)^{\alpha} + (T + \gamma_\alpha - u)^{\alpha}) \right) \\
\times ((T + \gamma_\alpha)^{\alpha} + (T + \gamma_\alpha - u)^{\alpha})^{-1} \\
\leq ((T + \gamma_\alpha) - (T + \gamma_\alpha - u) + (T + \gamma_\alpha)^{1-\alpha}(T + \gamma_\alpha - u)^{\alpha} \\
- (T + \gamma_\alpha)^{\alpha}(T + \gamma_\alpha - u)^{1-\alpha}) \times (T + \gamma_\alpha)^{-\alpha} \leq 2u/(T + \gamma_\alpha)^{\alpha} .
\]

And in the case where \( \alpha > 1/2 \), for all \( u \in [0, T] \):
\[
((T + \gamma_\alpha)^{1-\alpha} - (T - u + \gamma_\alpha)^{1-\alpha}) \leq ((T + \gamma_\alpha)^{\alpha} - (T - u + \gamma_\alpha)^{\alpha}) ,
\]
and we also have, for all \( u \in [0, T] \):
\[
(T + \gamma_\alpha)^{\alpha} - (T + \gamma_\alpha - u)^{\alpha} \\
= \left( ((T + \gamma_\alpha)^{\alpha} - (T + \gamma_\alpha - u)^{\alpha})((T + \gamma_\alpha)^{1-\alpha} + (T + \gamma_\alpha - u)^{1-\alpha}) \right) \\
\times ((T + \gamma_\alpha)^{1-\alpha} + (T + \gamma_\alpha - u)^{1-\alpha})^{-1} \\
\leq ((T + \gamma_\alpha) - (T + \gamma_\alpha - u) + (T + \gamma_\alpha)^{\alpha}(T + \gamma_\alpha - u)^{1-\alpha} \\
- (T + \gamma_\alpha)^{1-\alpha}(T + \gamma_\alpha - u)^{\alpha}) \times (T + \gamma_\alpha)^{-1+\alpha} \leq 2u/(T + \gamma_\alpha)^{1-\alpha} .
\]
Now, using Lemma 52 with \( Y = X_{T-u} \) we obtain, for all \( u \in [0, T] \):
\[
\mathbb{E} \left[ \int_{T-u}^{T} f(X_t) - f(X_{T-u})dt \right] \leq 2c_1 \log(T + \gamma_\alpha)(T + \gamma_\alpha)^{-\min(\alpha, 1-\alpha)u} .
\]

Since \( S \) is a differentiable function and using (55), we have for all \( u \in (0, T) \),
\[
S'(u) = -u^{-2} \int_{T-u}^{T} \mathbb{E} [f(X_t)] dt + u^{-1} \mathbb{E} [f(X_{T-u})] = -u^{-1}(S(u) - \mathbb{E} [f(X_{T-u})]) .
\]
This last result implies \( -S'(u) \leq 2c_1 \log(T + \gamma_\alpha)/(T + \gamma_\alpha)^{-\min(\alpha, 1-\alpha)u^{-1}} \) and integrating we get
\[
S(1) - S(T) \leq 2c_1 \log(T + \gamma_\alpha) \log(T)(T + \gamma_\alpha)^{-\min(\alpha, 1-\alpha)} .
\]
Lemma 54 Assume A1, A2-(a), A3, and F2-(a). Then, for any \( \alpha, \gamma \in (0, 1) \) and \( T \geq 0 \) we have
\[
S(T) - f^* \leq 2C_1 T^{-\min(\alpha, 1-\alpha)} \log(T + \gamma_0) ,
\]
with \( S \) given in (50).

Proof Using Lemma 52, with \( u = T \) and \( Y = x^* \), and \( \|X_0 - x^*\| \leq C_1 \) we obtain
\[
\int_0^T \mathbb{E} [f(X_s)] ds - T f^* \leq \left( C_1/2 \right) \left( (T + \gamma_0)^\alpha - \gamma_0^\alpha + \{ (T + \gamma_0)^{1-\alpha} - \gamma \} \log(T + \gamma_0) \right)
\]
\[
+ \left( 1/2 \right) \gamma_0 \mathbb{E} \left[ \|X_0 - x^*\|^2 \right] .
\]
Using this result we have
\[
S(T) - f^* \leq T^{-1} C_1 (T + \gamma_0)^{\max(1-\alpha, \alpha)} \log(T + \gamma_0)
\]
\[
+ C_1 \gamma_0 T^{-1}/2 \leq 2C_1 T^{-\min(\alpha, 1-\alpha)} \log(T + \gamma_0) .
\]

Lemma 55 Assume A1, A2-(a), A3, and F2-(a). Then, for any \( \alpha, \gamma \in (0, 1) \) and \( T \geq 0 \) we have
\[
S(0) - S(1) \leq C_1 L (T - 1)^{-2\alpha} ,
\]
with \( S \) given in (50).

Proof We have
\[
S(0) - S(1) = \mathbb{E} [f(X_T)] - S(1) = \int_{T-1}^T (\mathbb{E} [f(X_T)] - \mathbb{E} [f(X_s)]) ds .
\]
Using Lemma 48 on the stochastic process \( f(X_t)_{t \geq 0} \) and A1, we have for all \( s \in [T - 1, T] \)
\[
\mathbb{E} [f(X_T)] - \mathbb{E} [f(X_s)]
\]
\[
= - \int_s^T (\gamma_0 + t)^{-\alpha} \mathbb{E} [\|\nabla f(X_t)\|^2] dt + (L/2) \gamma_0 \int_s^T (t + \gamma_0)^{-2\alpha} \mathbb{E} [\text{Tr}(\Sigma(X_t))] dt
\]
\[
\leq (\eta L/2) \gamma_0 \int_s^T (t + \gamma_0)^{-2\alpha} dt \leq (C_1 L/2)(s + \gamma_0)^{-2\alpha} (T - s) .
\]
Plugging this result into (56) yields
\[
S(0) - S(1) \leq (C_1 L/2) \int_{T-1}^T (T - s)(s + \gamma_0)^{-2\alpha} ds \leq C_1 L (T - 1 + \gamma_0)^{-2\alpha} \leq C_1 L (T - 1)^{-2\alpha} .
\]

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 $f_\varepsilon$ for each $\varepsilon > 0$. Therefore there exists $C^{(c)}_\varepsilon$ such that
\[
\mathbb{E}[f(X_{T, \varepsilon})] - f(x^*_\varepsilon) \leq C^{(c)}_\varepsilon \left[ \log(T)^2 T^{-\delta} + \log(T) T^{-\delta} + T^{-\delta} + (T - 1)^{-2\alpha} \right], \tag{57}
\]
where $(X_{t, \varepsilon})_{t \geq 0}$ is given by (2) with $X_t = x_0$ (upon replacing $f$ by $f_\varepsilon$) and
\[
C^{(c)}_\varepsilon = 4 \max(2C^{(c)}_{2, \alpha} + 2 \|x_0 - x^*_\varepsilon\|^2, (\gamma_\alpha \eta + 2\alpha C^{(c)}_{1, \alpha})(1 - \alpha)^{-1}).
\]
Using (57) and Lemma 49 we have
\[
\mathbb{E}[f(X_T)] - f^* \leq \liminf_{\varepsilon \to 0} \mathbb{E}[f_\varepsilon(X_{t, \varepsilon})] - \limsup_{\varepsilon \to 0} f_\varepsilon(x^*_\varepsilon)
\leq \liminf_{\varepsilon \to 0} \{ \mathbb{E}[f_\varepsilon(X_{t, \varepsilon})] - f_\varepsilon(x^*_\varepsilon) \}
\leq \liminf_{\varepsilon \to 0} C^{(c)}_\varepsilon \left[ \log(T)^2 T^{-\delta} + \log(T) T^{-\delta} + T^{-\delta} + (T - 1)^{-2\alpha} \right]
\leq C^{(c)}_1 \left[ \log(T)^2 T^{-\delta} + \log(T) T^{-\delta} + T^{-\delta} + (T - 1)^{-2\alpha} \right],
\]
with $C^{(c)}_1 = 3 \max(2C^{(c)}_{2, \alpha} + 4 \|x_0\|^2 + 4C^2, (\gamma_\alpha \eta + 2C^{(c)}_{1, \alpha})(1 - \alpha)^{-1})$, where $C = \max_{y \in \text{argmin}_y f} \|y\|$.

\section*{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 $\mathbb{E}[f(X_n)] - f^*$. 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 $\mathbb{E}[f(X_n)] - f^*$ 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 $\nabla f$ is bounded and no longer Lipschitz continuous in Corollary 61.

**Lemma 56** Assume A1, F2-(a), A2-(a). Then for any $\alpha, \gamma \in (0, 1)$, there exists $C^{(d)}_{1, \alpha} \geq 0, C^{(d)}_{2, \alpha} \geq 0$ and a function $\Phi^{(d)}_\alpha : \mathbb{R}_+ \to \mathbb{R}_+$ such that, for any $n \geq 0$,
\[
\mathbb{E}\left[ \|X_n - x^*\|^2 \right] \leq C^{(d)}_{1, \alpha} \Phi^{(d)}_\alpha(n + 1) + C^{(d)}_{2, \alpha}.
\]

And we have
\[
\Phi^{(d)}_\alpha(t) = \begin{cases} 
 t^{1 - 2\alpha} & \text{if } \alpha < 1/2, \\
 \log(t) & \text{if } \alpha = 1/2, \\
 0 & \text{if } \alpha > 1/2.
\end{cases}
\]
The values of the constants are given by

\[
C_{1,\alpha}^{(d)} = \begin{cases} 
2\gamma^2\eta (1 - 2\alpha)^{-1} & \text{if } \alpha < 1/2 , \\
\gamma^2\eta & \text{if } \alpha = 1/2 , \\
0 & \text{if } \alpha > 1/2 . 
\end{cases}
\]

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

**Proof** Let \( f : \mathbb{R}^d \to \mathbb{R} \) verifying assumptions A1 and F2-(a). We consider \( (X_n)_{n \geq 0} \) satisfying (1). Let \( x^* \in \mathbb{R}^d \) be given by F2-(a). We have, using (1) and A2-(a) that for all \( n \geq (\gamma L/2)^{1/\alpha} \),

\[
\mathbb{E} \left[ \| X_{n+1} - x^* \|^2 \left| F_n \right. \right] = \mathbb{E} \left[ \| X_n - x^* - \gamma(n + 1)^{-\alpha} H(X_n, Z_{n+1}) \|^2 \left| F_n \right. \right] 
= \| X_n - x^* \|^2 - 2\gamma/(n + 1)^\alpha (X_n - x^*, \mathbb{E} [H(X_n, Z_{n+1})\left| F_n \right.]) 
+ \gamma^2(n + 1)^{-2\alpha} \mathbb{E} \left[ \| H(X_n, Z_{n+1}) \|^2 \left| F_n \right. \right] 
= \| X_n - x^* \|^2 - 2\gamma/(n + 1)^\alpha (X_n - x^*, \nabla f(X_n)) 
+ \gamma^2(n + 1)^{-2\alpha} \mathbb{E} \left[ \| H(X_n, Z_{n+1}) - \nabla f(X_n) + \nabla f(X_n) \|^2 \left| F_n \right. \right] 
= \| X_n - x^* \|^2 - 2\gamma/(n + 1)^\alpha (X_n - x^*, \nabla f(X_n)) 
+ \gamma^2(n + 1)^{-2\alpha} \mathbb{E} \left[ \| \nabla f(X_n) \|^2 \left| F_n \right. \right] 
+ 2\mathbb{E} [(H(X_n, Z_{n+1}) - \nabla f(X_n), \nabla f(X_n)) | F_n] 
\leq \| X_n - x^* \|^2 - 2\gamma/L(n + 1)^{-\alpha} \| \nabla f(X_n) \|^2 + \gamma^2\eta(n + 1)^{-2\alpha} 
+ \gamma^2(n + 1)^{-2\alpha} \| \nabla f(X_n) \|^2 
\leq \| X_n - x^* \|^2 + \gamma/(n + 1)^\alpha \| \nabla f(X_n) \|^2 \left[ \gamma/(n + 1)^\alpha - 2/L \right] + \gamma^2\eta(n + 1)^{-2\alpha} 
\leq \| X_n - x^* \|^2 + \gamma^2\eta(n + 1)^{-2\alpha} 
\leq \mathbb{E} \left[ \| X_n - x^* \|^2 \right] + \gamma^2\eta(n + 1)^{-2\alpha} , 
\] (58)

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

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

As in the previous proof we now distinguish three cases:
(a) If $\alpha < 1/2$, we have
\[
E \left[ \|X_n - x^*\|^2 \right] \leq \|X_0 - x^*\|^2 + \gamma^2 \eta (1 - 2\alpha)^{-1} (n + 1)^{1 - 2\alpha}
\]
\[
\leq \|X_0 - x^*\|^2 + 2 \gamma^2 \eta (1 - 2\alpha)^{-1} n^{1 - 2\alpha}.
\]

(b) If $\alpha = 1/2$, we have
\[
E \left[ \|X_n - x^*\|^2 \right] \leq \|X_0 - x^*\|^2 + \gamma^2 \eta (\log(n) + 2).
\]

(c) If $\alpha > 1/2$, we have
\[
E \left[ \|X_n - x^*\|^2 \right] \leq \|X_0 - x^*\|^2 + \gamma^2 \eta (2\alpha - 1)^{-1}.
\]

We now turn to the proof of Theorem 8 by stating an intermediate result where we assume a condition bounding $E[\|\nabla f(X_n)\|^2]$. 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 $(X_n)_{n \geq 0}$ be given by (1). Assume $A_1, F_2-(a), A_2-(a)$. Suppose additionally that there exists $\alpha^* \in [0, 1/2], \beta > 0$ and $C_0 \geq 0$ such that for all $n \in \mathbb{N}
\[
E[\|\nabla f(X_n)\|^2] \leq \begin{cases} 
C_0(n + 1)^\beta \log(n + 1) & \text{if } \alpha \leq \alpha^*, \\
C_0 & \text{if } \alpha > \alpha^*.
\end{cases}
\]

Then there exists $\tilde{C}_\alpha \geq 0$ such that, for all $N \geq 1$,
\[
E[f(X_N)] - f^* \leq \tilde{C}_\alpha \left\{ (1 + \log(N + 1))^2/(N + 1)^{\min(\alpha, 1-\alpha)} \Psi_\alpha(N + 1) + 1/(N + 1) \right\},
\]
where for any $n \in \mathbb{N}$
\[
\Psi_\alpha(n) = \begin{cases} 
n^\beta (1 + \log(n)) & \text{if } \alpha \leq \alpha^*, \\
1 & \text{if } \alpha > \alpha^*.
\end{cases}
\]

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

Let $(S_k)_{k \in \{0, \ldots, N\}}$ be given for any $k \in \mathbb{N}$ by
\[
S_k = (k + 1)^{-1} \sum_{t=N-k}^{N} E[f(X_t)].
\]

Note that $E[f(X_N)] - f^* = (S_N - S_0) + (S_0 - f^*)$. We are now going to control each one of the two terms $(S_0 - S_N)$ and $(S_N - f^*)$ 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 $(\mathbb{E}[
abla f(X_n)]^2)_{n \in \mathbb{N}}$ 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 $S(T) - f^*$ (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 A1, A2-(a), and F2-(a). In addition, assume that (59) 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 $\mathbb{E}[[Y - x^*]^2] \leq C_{1,\alpha}(N+1)^{1-2\alpha}\log(N+1) + 4C_{2,\alpha}^d$ with $C_{1,\alpha}^d$ and $C_{2,\alpha}^d$ given in Lemma 56, we have

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

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

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

**Proof** Let $\ell \in \{0, \ldots, N\}$, let $k \geq \ell$, let $Y \in \mathcal{F}_\ell$. Using F2-(a) we have

$$
\mathbb{E}\left[\|X_{k+1} - Y\|^2|\mathcal{F}_\ell\right] = \mathbb{E}\left[\|X_k - Y - \gamma(k+1)^{-\alpha}H(X_k, Z_{k+1})\|^2|\mathcal{F}_\ell\right] \\
= \|X_k - Y\|^2 + \gamma^2(k+1)^{-2\alpha}\mathbb{E}\left[\|H(X_k, Z_{k+1})\|^2|\mathcal{F}_\ell\right] \\
- 2\gamma(k+1)^{-\alpha}\langle X_k - Y, \nabla f(X_k) \rangle \\
\mathbb{E}[f(X_k) - f(Y)] \leq (2\gamma)^{-1}(k+1)^\alpha \left( \mathbb{E}\left[\|X_k - Y\|^2\right] - \mathbb{E}\left[\|X_{k+1} - Y\|^2\right] \right) \\
+ (\gamma/2)(k+1)^{-\alpha}\mathbb{E}\left[\|H(X_k, Z_{k+1})\|^2|\mathcal{F}_\ell\right] \\
\mathbb{E}[f(X_k) - f(Y)] \leq (2\gamma)^{-1}(k+1)^\alpha \left( \mathbb{E}\left[\|X_k - Y\|^2\right] - \mathbb{E}\left[\|X_{k+1} - Y\|^2\right] \right) \\
+ (\gamma/2)(k+1)^{-\alpha} \left( \eta + \mathbb{E}\left[\|
abla f(X_k)\|^2\right] \right).
$$

(61)
Let $u \in \{0, \cdots, N\}$. Summing now (61) between $k = N - u$ and $k = N$ gives
\[
\mathbb{E} \left[ \sum_{k=N-u}^{N} f(X_k) - f(Y) \right] \leq (\gamma \eta / 2) \sum_{k=N-u}^{N} (k+1)^{-\alpha} + (2\gamma)^{-1} \sum_{k=N-u+1}^{N} \mathbb{E} \left[ \|X_k - Y\|^2 \right] ((k+1)^{\alpha} - k^{\alpha}) + (\gamma / 2) \sum_{k=N-u}^{N} \mathbb{E} \left[ \|\nabla f(X_k)\|^2 \right] (k+1)^{-\alpha} + (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \|X_{N-u} - Y\|^2 \right].
\]

(62)

We now have to conduct separate analyses depending on the value of $\alpha$.

(a) First assume that $\alpha \leq \alpha^*$. In that case (59) gives that
\[
\mathbb{E} \left[ \|\nabla f(X_k)\|^2 \right] \leq C_0 (N + 1)^{\beta} \log(N + 1),
\]
and Lemma 56 gives that for all $k \in \{0, \ldots, N\},$
\[
\mathbb{E} \left[ \|X_k - Y\|^2 \right] \leq 2\mathbb{E} \left[ \|X_k - x^*\|^2 \right] + 2\mathbb{E} \left[ \|Y - x^*\|^2 \right] \\
\leq 2C_{1,\alpha}^{(d)} (k+1)^{1-2\alpha} \log(k+1) + 2C_{1,\alpha}^{(d)} (N+1)^{1-2\alpha} \log(N + 1) + 4C_{2,\alpha}^{(d)} \\
\leq 4C_{1,\alpha}^{(d)} (N+1)^{1-2\alpha} \log(N + 1) + 4C_{2,\alpha}^{(d)}. 
\]

We note $C_{3,\alpha}^{(d)} = 4C_{2,\alpha}^{(d)}$. Combining (62) and $C^{(b)} = ((\gamma \eta / 2) + (\gamma / 2)C_0)(1-\alpha)^{-1}$ we get that
\[
\mathbb{E} \left[ \sum_{k=N-u}^{N} f(X_k) - f(Y) \right] \leq (\gamma \eta / 2)(1-\alpha)^{-1} ((N+1)^{1-\alpha} - (N-u)^{1-\alpha}) \\
+ (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \|X_{N-u} - Y\|^2 \right] \\
+ (2\gamma)^{-1} \left( C_{3,\alpha}^{(d)} + 4C_{1,\alpha}^{(d)} (N+1)^{1-2\alpha} \log(N + 1) \right) ((N+1)^{\alpha} - (N-u+1)^{\alpha}) \\
+ (\gamma / 2)C_0 (N+1)^{\beta} \log(N + 1)(1-\alpha)^{-1} ((N+1)^{1-\alpha} - (N-u)^{1-\alpha}) \\
\leq C^{(b)} (N+1)^{\beta} (1 + \log(N + 1))^2 ((N+1)^{1-\alpha} - (N-u)^{1-\alpha}) \\
+ (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \|X_{N-u} - Y\|^2 \right] \\
+ (2\gamma)^{-1} C_{3,\alpha}^{(d)} ((N+1)^{\alpha} - (N-u)^{\alpha}) \\
+ (2\gamma)^{-1} 4C_{1,\alpha}^{(d)} ((N+1)^{1-\alpha} - (N-u)^{1-\alpha}) \\
\leq C^{(d)} (N+1)^{\beta} (1 + \log(N + 1))^2 ((N+1)^{1-\alpha} - (N-u)^{1-\alpha}) \\
+ (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \|X_{N-u} - Y\|^2 \right],
\]

where we used Lemma 50.
(N + 1)^{1-\alpha} - (N - u)^{1-\alpha} \\
= \{ ((N + 1)^{1-\alpha} - (N - u)^{1-\alpha}) ((N + 1)^{\alpha} + (N - u)^{\alpha}) \} ((N + 1)^{\alpha} + (N - u)^{\alpha})^{-1} \\
\leq 2(u + 1)/(N + 1)^\alpha.

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

\[
E \left[ \|X_k - Y\|^2 \right] \leq 2E \left[ \|X_k - x^*\|^2 \right] + 2E \left[ \|Y - x^*\|^2 \right] \\
\leq 2C^{(d)}(k + 1)^{1-2\alpha} \log(k + 1) + 2C^{(d)}(N + 1)^{1-2\alpha} \log(N + 1) + 4C^{(d)} \\
\leq 4C^{(d)}(N + 1)^{1-2\alpha} \log(N + 1) + 4C^{(d)}.
\]

Using (59), (62) rewrites

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

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

\[
\forall k \in \{0, \ldots, N\}, E \left[ \|X_k - Y\|^2 \right] \leq 2E \left[ \|X_k - x^*\|^2 \right] + 2E \left[ \|Y - x^*\|^2 \right] \leq 4C^{(d)} = C^{(d)}.
\]

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

\[
E \left[ \sum_{k=N-u}^{N} f(X_k) - f(Y) \right] \leq ((\gamma \eta/2) + \gamma C^{(d)}/2)(1 - \alpha)^{-1} ((N + 1)^{1-\alpha} - (N - u)^{1-\alpha}) \\
+ (2\gamma)^{-1}(N - u + 1)^{\alpha}E \left[ \|X_{N-u} - Y\|^2 \right] \\
+ (2\gamma)^{-1}C^{(d)}_{3,\alpha}((N + 1)^{\alpha} - (N - u + 1)^{\alpha}) \\
\leq C^{(b)}((N + 1)^{1-\alpha} - (N - u)^{1-\alpha}) + (2\gamma)^{-1}(N - u + 1)^{\alpha}E \left[ \|X_{N-u} - Y\|^2 \right] \\
+ (2\gamma)^{-1}C^{(d)}_{3,\alpha}((N + 1)^{\alpha} - (N - u)^{\alpha}) \\
\leq C^{(d)}((N + 1)^{\alpha} - (N - u)^{\alpha}) + (2\gamma)^{-1}(N - u + 1)^{\alpha}E \left[ \|X_{N-u} - Y\|^2 \right].
\]
Notice now that, similarly to (54) we have
\[(N + 1)^{\alpha} - (N - u)^{\alpha} = \{((N + 1)^{\alpha} - (N - u)^{\alpha}) \times ((N + 1)^{1-\alpha} + (N - u)^{1-\alpha})\} \times ((N + 1)^{1-\alpha} + (N - u)^{1-\alpha})^{-1} \leq 2(u + 1)/(N + 1)^{1-\alpha}.\]

Finally, putting the three cases above together we obtain
\[
\mathbb{E} \left[ \sum_{k=N-u}^{N} f(X_k) - f(X_{N-u}) \right] \\
\leq 2c^{(d)}(u + 1)/(N + 1)^{\min(\alpha, 1-\alpha)}(1 + \log(N + 1))\Psi_{\alpha}(N + 1) \\
+ (2\gamma)^{-1}(N - u + 1)^{\alpha} \mathbb{E} \left[ \|X_{N-u} - Y\|^2 \right],
\]
with
\[
\Psi_{\alpha}(n) = \begin{cases} n^\beta(1 + \log(n)) & \text{if } \alpha \leq \alpha^*, \\ 1 & \text{if } \alpha > \alpha^*. \end{cases}
\]

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

**Lemma 59** Assume A1, A2-(a) and F2-(a). In addition, assume that (59) holds. Then, for any \(\alpha, \gamma \in (0, 1)\) and \(N \in \mathbb{N}\) we have
\[S_0 - S_N \leq 2c^{(d)}(N + 1)^{-\min(\alpha, 1-\alpha)}(1 + \log(N + 1))^2\Psi_{\alpha}(N + 1).\]

**Proof** Let \(u \in \{0, \ldots, N\}\). Using Lemma 58 with the choice \(Y = X_{N-u}\) gives
\[
\mathbb{E} \left[ \sum_{k=N-u}^{N} f(X_k) - f(X_{N-u}) \right] \leq 2c^{(d)}(u + 1)/(N + 1)^{\min(\alpha, 1-\alpha)}(1 + \log(N + 1))\Psi_{\alpha}(N + 1).
\]

And then,
\[
S_u = (u + 1)^{-1} \sum_{k=N-u}^{N} \mathbb{E} [f(X_k)] \\
\leq 2c^{(d)}(N + 1)^{-\min(\alpha, 1-\alpha)}(1 + \log(N + 1))\Psi_{\alpha}(N + 1) + \mathbb{E} [f(X_{N-u})]. \tag{63}
\]

We have now, using (63),
\[
u S_{u-1} = (u + 1)S_u - \mathbb{E} [f(X_{N-u})] \\
= uS_u + S_u - \mathbb{E} [f(X_{N-u})] \\
\leq uS_u + 2c^{(d)}(N + 1)^{-\min(\alpha, 1-\alpha)}(1 + \log(N + 1))\Psi_{\alpha}(N + 1) \\
S_{u-1} - S_u \leq 2c^{(d)}(N + 1)^{-\min(\alpha, 1-\alpha)} \log(N + 1) \\
S_0 - S_N \leq 2c^{(d)}(N + 1)^{-\min(\alpha, 1-\alpha)}(1 + \log(N + 1))^2\Psi_{\alpha}(N + 1) \\
S_0 - S_N \leq 2c^{(d)}(N + 1)^{-\min(\alpha, 1-\alpha)}(1 + \log(N + 1))^2\Psi_{\alpha}(N + 1).
Lemma 60 Assume A1, A2-(a) and F2-(a). In addition, assume that (59) holds. Then, for any \( \alpha, \gamma \in (0, 1) \) and \( N \in \mathbb{N} \) we have

\[
S_N - f^* \leq 2C(d)(1 + \log(N + 1))^2(N + 1)^{-\min(\alpha, 1-\alpha)}\Psi_{\alpha}(N + 1) + (2\gamma)^{-1}(N + 1)^{-1}\|X_0 - x^*\|^2 .
\]

with \( S \) given in (60).

Proof Using Lemma 58 with the choice \( Y = x^* \) and \( u = N \) gives

\[
(N + 1)^{-1}E \left[ \sum_{k=0}^{N} f(X_k) - f(x^*) \right] \leq 2C(d)(1 + \log(N + 1))(N + 1)^{-\min(\alpha, 1-\alpha)}\Psi_{\alpha}(N + 1) + (2\gamma)^{-1}(N + 1)^{-1}\|X_0 - x^*\|^2
\]

Therefore,

\[
S_N - f^* \leq 2C(d)(1 + \log(N + 1))^2(N + 1)^{-\min(\alpha, 1-\alpha)}\Psi_{\alpha}(N + 1) + (2\gamma)^{-1}(N + 1)^{-1}\|X_0 - x^*\|^2 .
\]

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 \( H2(m) \) is true.

\( H2 \) (m) For any \( \alpha > 1/(m+1) \), there exists \( C_\alpha^+ > 0 \) such that for all \( n \in \mathbb{N} \), \( E[\|\nabla f(X_n)\|^2] \leq C_\alpha^+ \). In addition, for any \( \alpha \leq 1/(m+1) \), there exists \( C_\alpha^- > 0 \) such that for all \( n \in \mathbb{N} \), \( E[\|\nabla f(X_n)\|^2] \leq C_\alpha^- n^{1-(m+1)\alpha}(1 + \log(n))^3 \).

For \( m = 1 \), \( H2(1) \) is an immediate consequence of A1 and Lemma 56, with \( C_\alpha^+ = L^2C_{1,\alpha}^{(d)} \) and \( C_\alpha^- = L^2 \max(C_{1,\alpha}^{(d)}, C_{2,\alpha}^{(d)}) \). Now, let \( m \in \mathbb{N}^* \) and suppose that \( H2(m) \) holds. Let \( \alpha \in (0, 1) \). Setting \( \alpha^* = 1/(m + 1) \) we have that (59) is verified with \( \beta = 1 - (m + 1)\alpha \). Consequently, using A1, F2-(a) and A2-(a) we can apply Proposition 57 and for any \( \alpha \leq 1/(m + 1) \) we have

\[
E[f(X_N)] - f^* \leq C_\alpha \left\{ (1 + \log(N + 1))^2/(N + 1)^{\min(\alpha, 1-\alpha)}\Psi_{\alpha}(N + 1) + 1/(N + 1) \right\}
\]

\[
\leq C_\alpha \left\{ (1 + \log(N + 1))^3(N + 1)^{-\alpha}(N + 1)^{1-(m+1)\alpha} + 1/(N + 1) \right\}
\]

\[
\leq C_\alpha \left\{ (1 + \log(N + 1))^3(N + 1)^{1-(m+2)\alpha} + 1/(N + 1) \right\} . \tag{64}
\]

In particular, if \( \alpha > 1/(m+2) \) we have the existence of \( \bar{C}_\alpha > 0 \) such that for all \( n \in \mathbb{N} \), \( E[f(X_n)] - f^* \leq \bar{C}_\alpha \). And using A1 and Lemma 40 we get that, for all \( n \in \mathbb{N} \)

\[
E[\|\nabla f(X_n)\|^2] \leq 2E[f(X_n) - f^*] \leq 2L\bar{C}_\alpha,
\]

70
Combining this result with (64), we get that $H_2(m + 1)$ holds with $C^+_0 = 2Lc_\alpha$ and $C^-_0 = 2\bar{c}_\alpha$. 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 $C_0 > 0$ such that $\mathbb{E}[\|\nabla f(X_n)\|^2] \leq C_0$ for all $n \in \mathbb{N}^*$. Applying Proposition 57 gives the existence of $C(d) > 0$ such that for all $N \geq 1$

$$\mathbb{E}[f(X_N)] - f^* \leq C(d)(1 + \log(N + 1))^2/(N + 1)^{\min(\alpha, 1 - \alpha)},$$

with $C(d) = 2\bar{c}_\alpha$, 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 $\nabla f$ 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 $\nabla f$ bounded. Then there exists $C_b^{(d)} > 0$ such that, for all $N \geq 1$,

$$\mathbb{E}[f(X_N)] - f^* \leq C_b^{(d)}(1 + \log(N + 1))^2/(N + 1)^{\min(\alpha, 1 - \alpha)}.$$

**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 $\nabla f$ 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 $n \geq 0$,

$$\mathbb{E}[\|X_{n+1} - x^*\|^2 | \mathcal{F}_n] = \mathbb{E} \left[ \|X_n - x^* - \gamma(n + 1)^{-\alpha}H(X_n, Z_{n+1})\|^2 | \mathcal{F}_n \right]
= \|X_n - x^*\|^2 - 2\gamma/(n + 1)^{\alpha}\langle X_n - x^*, \mathbb{E}[H(X_n, Z_{n+1}) | \mathcal{F}_n] \rangle
+ \gamma^2(n + 1)^{-2\alpha} \mathbb{E} \left[ \|H(X_n, Z_{n+1})\|^2 | \mathcal{F}_n \right]
= \|X_n - x^*\|^2 - 2\gamma/(n + 1)^{\alpha}\langle X_n - x^*, \nabla f(X_n) \rangle
+ \gamma^2\eta(n + 1)^{-2\alpha} + \gamma^2(n + 1)^{-2\alpha} \|\nabla f(X_n)\|^2
\geq \mathbb{E}[\|X_{n+1} - x^*\|^2] \leq \mathbb{E}[\|X_n - x^*\|^2] + \gamma^2(\eta + \|\nabla f\|_\infty)(n + 1)^{-2\alpha}.

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 $\alpha^* = 0$. This concludes the proof.

**Appendix F. Convex case (under $A_2$-(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 A2-(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 A1, A2-(b), A3 and F2-(b). Let \((X_t)_{t \geq 0}\) be given by (2). Then, for any \(\alpha, \gamma \in (0, 1)\), there exists \(c_1^{(c)} \geq 0\) and \(c_2^{(c)} \geq 0\) and a function \(\Phi_{\alpha}^{(c)} : \mathbb{R}_+ \rightarrow \mathbb{R}_+\) such that, for any \(t \geq 0\),

\[
\mathbb{E}\left[\|X_t - x^*\|^2\right] \leq c_1^{(c)} \Phi_{\alpha}^{(c)}(t) + c_2^{(c)}.
\]

And we have

\[
\Phi_{\alpha}^{(c)}(t) = \begin{cases} 
 t^{1-2\alpha} & \text{if } \alpha < 1/2 , \\
 \log(t) & \text{if } \alpha = 1/2 , \\
 0 & \text{if } \alpha > 1/2 .
\end{cases}
\]

The values of the constants are given by

\[
\begin{align*}
C_{1,\alpha}^{(c)} &= \begin{cases} 
 \gamma \alpha L_T (1 - 2\alpha)^{-1} & \text{if } \alpha < 1/2 , \\
 \gamma \alpha L_T & \text{if } \alpha = 1/2 , \\
 0 & \text{if } \alpha > 1/2 .
\end{cases} \\
C_{2,\alpha}^{(c)} &= \begin{cases} 
 2 \max_{t \leq (\gamma \alpha L_T)^{1/\alpha}} \mathbb{E}\left[\|X_t - x^*\|^2\right] & \text{if } \alpha < 1/2 , \\
 2 \max_{t \leq (\gamma \alpha L_T)^{1/\alpha}} \mathbb{E}\left[\|X_t - x^*\|^2\right] - \gamma \alpha L_T \log(\gamma \alpha) & \text{if } \alpha = 1/2 , \\
 2 \max_{t \leq (\gamma \alpha L_T)^{1/\alpha}} \mathbb{E}\left[\|X_t - x^*\|^2\right] + (2\alpha - 1)^{-1} \gamma \alpha^{2-2\alpha} L_T & \text{if } \alpha > 1/2 .
\end{cases}
\end{align*}
\]

**Proof** Let \(\alpha, \gamma \in (0, 1)\) and \(t \geq 0\). Let \((X_t)_{t \geq 0}\) be given by (2). We consider the function \(F : \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}_+\) given for any \((t, x) \in \mathbb{R} \times \mathbb{R}^d\) by \(F(t, x) = \|x - x^*\|^2\). Applying Lemma 48 to the
stochastic process $(F(t, X_t))_{t \geq 0}$ and using Lemma 42 and F2-(b) gives that for all $t \geq (\gamma_L + 1/\alpha)$,
\[
\mathbb{E} \left[ \left\| X_t - x^* \right\|^2 \right] - \mathbb{E} \left[ \left\| X_0 - x^* \right\|^2 \right]
\leq -2 \int_0^T (t + \gamma_L)^{-\alpha} \mathbb{E} \left[ \langle X_t - x^*, \nabla f(X_t) \rangle \right] dt + \int_0^T \gamma_L (t + \gamma_L)^{-2\alpha} \mathbb{E} \left[ \text{Tr} \left( \Sigma(X_t) \right) \right] dt
\leq -2 \int_0^T (t + \gamma_L)^{-\alpha} \mathbb{E} \left[ \langle X_t - x^*, \nabla f(X_t) \rangle \right] dt
+ \int_0^T \gamma_L (t + \gamma_L)^{-2\alpha} \left[ 1 + \mathbb{E} \left[ \langle \nabla f(X_t), X_t - x^* \rangle \right] \right] dt
\leq \gamma_L (t + \gamma_L)^{-2\alpha} dt + \int_0^T (t + \gamma_L)^{-\alpha} \mathbb{E} \left[ \langle \nabla f(X_t), X_t - x^* \rangle \right] \{ -2 + \gamma_L (t + \gamma_L)^{-\alpha} \} dt
\leq \gamma_L (t + \gamma_L)^{-2\alpha} dt.
\]

We now distinguish three cases:

(a) If $\alpha < 1/2$, then we have
\[
\mathbb{E} \left[ \left\| X_t - x^* \right\|^2 \right] \leq \left\| X_0 - x^* \right\|^2 + \gamma_L (1 - 2\alpha)^{-1} (\gamma + \gamma_L)^{1 - 2\alpha} \left( \left\| X_0 - x^* \right\|^2 + \gamma_L (1 - 2\alpha)^{-1} (\gamma + \gamma_L)^{1 - 2\alpha} \right).
\]

(b) If $\alpha = 1/2$, then we have
\[
\mathbb{E} \left[ \left\| X_t - x^* \right\|^2 \right] \leq \left\| X_0 - x^* \right\|^2 + \gamma_L (\log(\gamma_L) - \log(\gamma_L))
\leq \gamma_L \log(\gamma_L) + \left\| X_0 - x^* \right\|^2 - \gamma_L \log(\gamma_L).
\]

(c) If $\alpha > 1/2$, then we have
\[
\mathbb{E} \left[ \left\| X_t - x^* \right\|^2 \right] \leq \left\| X_0 - x^* \right\|^2 + \gamma_L (1 - 2\alpha)^{-1} (\gamma + \gamma_L)^{1 - 2\alpha} \left( \left\| X_0 - x^* \right\|^2 + (2\alpha - 1)^{-1} \gamma_L^{2 - 2\alpha} \right).
\]

\[\blacksquare\]

**Proposition 63** Let $\gamma, \alpha \in (0, 1)$ and $x_0 \in \mathbb{R}^d$ $(X_t)_{t \geq 0}$ be given by (2). Assume A1, A2-(b), F2-(b) and A3. In addition, assume that there exists $\alpha^* \in [0, 1/2]$, $\beta > 0$ and $C_0 \geq 0$ such that for all $t \in [0, T]$
\[
\mathbb{E} \left[ \text{Tr} \Sigma(X_t) \right] \leq \begin{cases} C_0 (t + \gamma_L)^{\beta \log(t + \gamma_L)} & \text{if } \alpha \leq \alpha^*, \\ C_0 & \text{if } \alpha > \alpha^*. \end{cases}
\] (65)
Then there exists $\tilde{C}_\alpha \geq 0$ such that, for all $T \geq 1$,  

$$
E[f(X_T)] - f^* \leq \tilde{C}_\alpha \left[ \log(T + \gamma_\alpha)^2(T + \gamma_\alpha)^{-\min(\alpha,1-\alpha)} \right] \left( 1 + \Psi_\alpha(T + \gamma_\alpha) \right),
$$

with  

$$
\Psi_\alpha(t) = \begin{cases} 
t^\beta & \text{if } \alpha \leq \alpha^*, \\
0 & \text{if } \alpha > \alpha^*. 
\end{cases}
$$

**Proof** Let $f \in C^2(\mathbb{R}^d, \mathbb{R})$. Let $\gamma \in (0,1)$ and $\alpha \in (0,1/2]$ and $T \geq 1$. Let $(X_t)_{t \geq 0}$ be given by (2). Combining Lemma 65, Lemma 66 and Lemma 67 gives the desired result  

$$
E[f(X_T)] - f^* \leq \tilde{c}(c) \left[ \log(T + \gamma_\alpha)^2(T + \gamma_\alpha)^{-\min(\alpha,1-\alpha)} \right] \left( 1 + \Psi_\alpha(T + \gamma_\alpha) \right),
$$

with $\tilde{c}(c) = (4/\gamma_\alpha^2)(4C_1 + 2\tilde{C}_1 + 2C_1L)$.

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

$$
\begin{cases} 
S(t) = t^{-1} \int_T^t \{ \mathbb{E}[f(X_s)] - f^* \} \, ds, & \text{if } t > 0, \\
S(0) = \mathbb{E}[f(X_T)]. 
\end{cases}
$$

With this notation we have  

$$
E[f(X_T)] - f^* = S(0) - S(1) + S(1) - S(T) + S(T) - f^*.
$$

We are now going to control each one of the three terms $(S(0) - S(1))$, $(S(1) - S(T))$, $(S(T) - f^*)$ 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 $S(T) - f^*$ (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 $\varepsilon$ 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 A1, A2-(b), A3, and F2-(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 $\mathbb{E}[\|Y - x^*\|^2] \leq 74$.
\[ C_{1,\alpha}^{(c)} \Phi_\alpha^{(c)}(T + \gamma_\alpha) + C_{2,\alpha}^{(c)} \text{ with } C_{1,\alpha}^{(c)} \text{ and } C_{2,\alpha}^{(c)} \text{ given in Lemma 62, we have} \]
\[ \int_{T-u}^{T} \mathbb{E} \left[ f(X_t) - f(Y) \right] \, dt \]
\[ \leq C_1 \left( (T + \gamma_\alpha) - (T - u + \gamma_\alpha) \right) \]
\[ + (1/2)(T - u + \gamma_\alpha) \mathbb{E} \left[ \|X_{T-u} - Y\|^2 \right] \]
\[ + C_1 \log(T + \gamma_\alpha) \left\{ (T + \gamma_\alpha) - (T - u + \gamma_\alpha) \right\} (T + \gamma_\alpha)^{1-2\alpha} \]
\[ + C_1 \left( (T + \gamma_\alpha)^{1-\alpha} - (T - u + \gamma_\alpha)^{1-\alpha} \right) \log(T + \gamma_\alpha)(1 + \Psi_\alpha(T + \gamma_\alpha)), \]

with \( C_1 = \max(4C_{2,\alpha}^{(c)}/2, (\gamma_\alpha C_0/2 + C_{1,\alpha}^{(c)})(1 - \alpha)^{-1}) \) and \( \Psi_\alpha(t) = 0 \) if \( \alpha > \alpha^* \) and \( t^\beta \) if \( \alpha \leq \alpha^* \), with \( C_{1,\alpha}^{(c)} \) and \( C_{2,\alpha}^{(c)} \) 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 \( \Phi_\alpha^{(c)} \) is non-decreasing and that for any \( a, b \geq 0 \), \((a + b)^2 \leq 2a^2 + 2b^2\), we have
\[ \mathbb{E} \left[ \|X_t - Y\|^2 \right] = \mathbb{E} \left[ \|(X_t - x^*) + (x^* - Y)\|^2 \right] \]
\[ \leq 2\mathbb{E} \left[ \|X_t - x^*\|^2 \right] + 2\mathbb{E} \left[ \|Y - x^*\|^2 \right] \]
\[ \leq 2C_{1,\alpha}^{(c)} \Phi_\alpha^{(c)}(t + \gamma_\alpha) + 4C_{2,\alpha}^{(c)} + 2C_{1,\alpha}^{(c)} \Phi_\alpha^{(c)}(T + \gamma_\alpha) \]
\[ \leq 2C_{1,\alpha}^{(c)} \Phi_\alpha^{(c)}(t + \gamma_\alpha) + 2C_{1,\alpha}^{(c)} \Phi_\alpha^{(c)}(T + \gamma_\alpha) + C_{3,\alpha}^{(c)}. \]

with \( C_{3,\alpha}^{(c)} = 4C_{2,\alpha}^{(c)}. \) This gives in particular, for every \( t \in [0, T] \),
\[ (t + \gamma_\alpha)^{\alpha-1} \mathbb{E} \left[ \|X_t - Y\|^2 \right] \leq \left[ C_{3,\alpha}^{(c)} + 2C_{1,\alpha}^{(c)}(T + \gamma_\alpha)^{1-2\alpha} \log(T + \gamma_\alpha) \right] (t + \gamma_\alpha)^{\alpha-1} \]
\[ + 2C_{1,\alpha}^{(c)} \log(T + \gamma_\alpha)(t + \gamma_\alpha)^{-\alpha}, \]

with \( C_{1,\alpha}^{(c)} = 0 \) 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
\[ \partial_t F_{y_0}(t, x) = \alpha(t + \gamma_\alpha)^{\alpha-1} \|x - y_0\|^2, \]
\[ \partial_x F_{y_0}(t, x) = 2(t + \gamma_\alpha)^\alpha (x - y_0), \quad \partial_{xx} F_{y_0}(t, x) = 2(t + \gamma_\alpha)^\alpha. \]

Using Lemma 48 on the stochastic process \((F_Y(t, X_t))_{t \geq 0}\), we have that for any \( u \in [0, T] \)
\[ \mathbb{E} \left[ F_Y(T, X_T) \right] - \mathbb{E} \left[ F_Y(T, u, X_{T-u}) \right] = \int_{T-u}^{T} \alpha(t + \gamma_\alpha)^{\alpha-1} \mathbb{E} \left[ \|X_t - Y\|^2 \right] \, dt \]
\[ - 2 \int_{T-u}^{T} \mathbb{E} \left[ \langle X_t - Y, \nabla f(X_t) \rangle \right] \, dt \]
\[ + \int_{T-u}^{T} \gamma_\alpha(t + \gamma_\alpha)^{-\alpha} \mathbb{E} \left[ \text{Tr}(\Sigma(X_t)) \right] \, dt. \]

We distinguish now several cases depending on the value of \( \alpha. \)
(a) If $\alpha \leq \alpha^*$, then combining F2-(b), A2-(b), (65), (67) and (68) we obtain for any $u \in [0, T]$

$$- (T - u + \gamma_\alpha)^\alpha E \|X_{T-u} - Y\|^2 \leq C_{3,\alpha}^c \int_{T-u}^T \alpha (t + \gamma_\alpha)^{-1} dt + C_0 \gamma_\alpha \int_{T-u}^T (t + \gamma_\alpha)^{-\alpha} dt$$

$$+ 2\alpha C_{1,\alpha}^c \log(T + \gamma_\alpha) \left\{ \int_{T-u}^T (t + \gamma_\alpha)^{-\alpha} dt + (T + \gamma_\alpha)^{1-2\alpha} \int_{T-u}^T (t + \gamma_\alpha)^{-\alpha-1} dt \right\}$$

$$- 2 \int_{T-u}^T E [f(X_t) - f(Y)] dt$$

$$\leq C_{3,\alpha}^c ((T + \gamma_\alpha)^\alpha - (T - u + \gamma_\alpha)^\alpha) - 2 \int_{T-u}^T E [f(X_t) - f(Y)] dt$$

$$+ (\gamma_\alpha C_0 + 2\alpha C_{1,\alpha}^c (1 - \alpha)^{-1} ((T + \gamma_\alpha)^{1-\alpha} - (T - u + \gamma_\alpha)^{1-\alpha}) \log(T + \gamma_\alpha)$$

$$+ 2C_{1,\alpha}^c \log(T + \gamma_\alpha) \{ (T + \gamma_\alpha)^\alpha - (T - u + \gamma_\alpha)^\alpha \} (T + \gamma_\alpha)^{1-2\alpha}.$$ 

(b) If $\alpha > \alpha^*$, then combining F2-(b), A2-(b), (65) (67) and (68) we obtain for any $u \in [0, T]$

$$- (T - u + \gamma_\alpha)^\alpha E \|X_{T-u} - Y\|^2 \leq C_{3,\alpha}^c \int_{T-u}^T \alpha (t + \gamma_\alpha)^{-1} dt + C_0 \gamma_\alpha \int_{T-u}^T (t + \gamma_\alpha)^{-\alpha} dt$$

$$+ 2\alpha C_{1,\alpha}^c \log(T + \gamma_\alpha) \left\{ \int_{T-u}^T (t + \gamma_\alpha)^{-\alpha} dt + (T + \gamma_\alpha)^{1-2\alpha} \int_{T-u}^T (t + \gamma_\alpha)^{-\alpha-1} dt \right\}$$

$$- 2 \int_{T-u}^T E [f(X_t) - f(Y)] dt$$

$$\leq C_{3,\alpha}^c ((T + \gamma_\alpha)^\alpha - (T - u + \gamma_\alpha)^\alpha) - 2 \int_{T-u}^T E [f(X_t) - f(Y)] dt$$

$$+ (\gamma_\alpha C_0 + 2\alpha C_{1,\alpha}^c (1 - \alpha)^{-1} ((T + \gamma_\alpha)^{1-\alpha} - (T - u + \gamma_\alpha)^{1-\alpha}) \log(T + \gamma_\alpha)$$

$$+ 2C_{1,\alpha}^c \log(T + \gamma_\alpha) \{ (T + \gamma_\alpha)^\alpha - (T - u + \gamma_\alpha)^\alpha \} (T + \gamma_\alpha)^{1-2\alpha}.$$ 

Putting this together we get for any $u \in [0, T]$

$$\int_{T-u}^T E [f(X_t) - f(Y)] dt$$

$$\leq C_1 ((T + \gamma_\alpha)^\alpha - (T - u + \gamma_\alpha)^\alpha)$$

$$+ (1/2)(T - u + \gamma_\alpha)^\alpha E \|X_{T-u} - Y\|^2$$

$$+ C_1 \log(T + \gamma_\alpha) \{ (T + \gamma_\alpha)^\alpha - (T - u + \gamma_\alpha)^\alpha \} (T + \gamma_\alpha)^{1-2\alpha}$$

$$+ C_1 ((T + \gamma_\alpha)^{1-\alpha} - (T - u + \gamma_\alpha)^{1-\alpha}) \log(T + \gamma_\alpha) (1 + \Psi_\alpha(T + \gamma_\alpha)),$$

with $C_1 = \max(C_{3,\alpha}^c/2, (\gamma_\alpha C_0/2 + C_{1,\alpha}^c (1 - \alpha)^{-1})$ and $\Psi_\alpha(t) = 0$ if $\alpha > \alpha^*$ and $t^\beta$ if $\alpha \leq \alpha^*$. ■
We divide the rest of the proof into three parts, to bound the quantities $S(1) - S(T)$, $S(T) - f^*$ and $S(0) - S(1)$.

**Lemma 65** Assume A1, A2-(b), A3, and F2-(b). In addition, assume that (65) holds. Then, for any $\alpha, \gamma \in (0, 1)$ and $T \geq 0$ we have

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

with $S$ given in (66).

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

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

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

$$(T + \gamma_\alpha)^{1-a} - (T + \gamma_\alpha - u)^{1-a}$$

$$= \left( ((T + \gamma_\alpha)^{1-a} - (T + \gamma_\alpha - u)^{1-a})(T + \gamma_\alpha)^a + (T + \gamma_\alpha - u)^a \right)$$

$$\times \left( (T + \gamma_\alpha)^a + (T + \gamma_\alpha - u)^a \right)^{-1}$$

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

$$\times (T + \gamma_\alpha)^{-1-a} \leq 2u/(T + \gamma_\alpha)^{1-a}.$$ (69)

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

$$((T + \gamma_\alpha)^a - (T - u + \gamma_\alpha)^a) \leq ((T + \gamma_\alpha)^a - (T - u + \gamma_\alpha)^a),$$

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

$$(T + \gamma_\alpha)^a - (T + \gamma_\alpha - u)^a$$

$$= \left( ((T + \gamma_\alpha)^a - (T + \gamma_\alpha - u)^a)(T + \gamma_\alpha)^{1-a} + (T + \gamma_\alpha - u)^{1-a} \right)$$

$$\times \left( (T + \gamma_\alpha)^{1-a} + (T + \gamma_\alpha - u)^{1-a} \right)^{-1}$$

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

$$\times (T + \gamma_\alpha)^{-1+a} \leq 2u/(T + \gamma_\alpha)^{1-a}.$$ (69)

Now, plugging $Y = X_{T-u}$ in Lemma 64 we obtain, for all $u \in [0, T]$: 

$$\mathbb{E} \left[ \int_{T-u}^T f(X_t) - f(X_{T-u})dt \right] \leq 2\tilde{C}_1 \log(T + \gamma_\alpha)(T + \gamma_\alpha)^{-\min(\alpha,1-a)}(1 + \Psi_\alpha(T + \gamma_\alpha))u,$$ (70)

with $\tilde{C}_1 = (2C_1 + C_1(1 + \gamma_\alpha^{1-2\alpha})$).

Since $S$ is a differentiable function and using (70), we have for all $u \in (0, T)$,

$$S'(u) = -u^{-2} \int_{T-u}^T \mathbb{E} f(X_t) dt + u^{-1} \mathbb{E} [f(X_{T-u})] = -u^{-1}(S(u) - \mathbb{E} [f(X_{T-u})]).$$

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

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

$\blacksquare$
Lemma 66 Assume A1, A2-(b), A3, and F2-(b). In addition, assume that (65) holds. Then, for any \( \alpha, \gamma \in (0, 1) \) and \( T \geq 0 \) we have

\[
S(T) - f^* \leq 4C_1 T^{-\min(\alpha, 1-\alpha)}(1 + \log(T + \gamma)) \left( 1 + \Psi_\alpha(T + \gamma) \right),
\]

with \( S \) given in (66).

Proof Using Lemma 64 with \( u = T \) and \( Y = x^* \), and \( \|X_0 - x^*\| \leq c_1 \) we obtain

\[
\int_0^T \mathbb{E}[f(X_s)] \, ds - T f^* \leq (C_1/2) \left( (T + \gamma)^{\alpha} - \gamma^\alpha \right) + (1/2)\gamma_0^\alpha \mathbb{E} \left[ \|X_0 - x^*\|^2 \right] \\
+ (C_1/2) \left[ (T + \gamma)\alpha - \gamma_0^\alpha \right] (T + \gamma)^{1-2\alpha} \log(T + \gamma) \\
+ (C_1/2) \log(T + \gamma) \left[ (T + \gamma)^{1-\alpha} - \gamma_0^{1-\alpha} \right] (1 + \Psi_\alpha(T + \gamma)) \\
\leq (C_1/2)(T + \gamma)^{\alpha} + (C_1/2)\gamma_0^\alpha \\
+ (C_1/2)(T + \gamma)^{1-\alpha} \log(T + \gamma) \\
+ (C_1/2) \log(T + \gamma)(T + \gamma)^{1-\alpha} \left( 1 + \Psi_\alpha(T + \gamma) \right).
\]

Using this result we have

\[
S(T) - f^* \leq T^{-1}2C_1(T + \gamma)\max(1-\alpha, \alpha) (1 + \log(T + \gamma)(1 + \Psi_\alpha(T + \gamma)) + 2C_1\gamma_0 T^{-1}/2 \\
\leq 4C_1 T^{-\min(\alpha, 1-\alpha)}(1 + \log(T + \gamma))(1 + \Psi_\alpha(T + \gamma)).
\]

Lemma 67 Assume A1, A2-(b), A3, and F2-(b). In addition, assume that (65) holds. Then, for any \( \alpha, \gamma \in (0, 1) \) and \( T \geq 0 \) we have

\[
S(0) - S(1) \leq c_1 L \log(T + \gamma)(1 + \Psi_\alpha(T + \gamma))(T - 1)^{-2\alpha},
\]

with \( S \) given in (66).

Proof We have

\[
S(0) - S(1) = \mathbb{E}[f(X_T)] - S(1) = \int_{T-1}^T \left( \mathbb{E}[f(X_T)] - \mathbb{E}[f(X_s)] \right) \, ds. \tag{71}
\]

Using Lemma 48 on the stochastic process \( f(X_t)_{t \geq 0}, A1 \) and (65), we have for all \( s \in [T - 1, T] \)

\[
\mathbb{E}[f(X_T)] - \mathbb{E}[f(X_s)] = -\int_s^T (\gamma_0 + t)^{-\alpha} \mathbb{E}[\|\nabla f(X_t)\|^2] \, dt \\
+ (L/2)\gamma_0 \int_s^T (t + \gamma)^{-2\alpha} \mathbb{E}[\text{Tr}(\Sigma(X_t))] \, dt \\
\leq (L/2)\gamma_0 \log(T + \gamma)(1 + \Psi_\alpha(T + \gamma)) \int_s^T (t + \gamma)^{-2\alpha} \, dt \\
\leq c_1 L \log(T + \gamma)(1 + \Psi_\alpha(T + \gamma))(s + \gamma)^{-2\alpha}(T - s).
\]
Plugging this result into (71) yields

\[
S(0) - S(1) \leq c_1 T \log(T + \gamma_\alpha) (1 + \Psi_\alpha(T + \gamma_\alpha)) \int_{T-1}^T (T - s)(s + \gamma_\alpha)^{-2\alpha} ds \\
\leq c_1 T \log(T + \gamma_\alpha) (1 + \Psi_\alpha(T + \gamma_\alpha))(T - 1 + \gamma_\alpha)^{-2\alpha} \\
\leq c_1 T \log(T + \gamma_\alpha) (1 + \Psi_\alpha(T + \gamma_\alpha))(T - 1)^{-2\alpha}.
\]

\[\Box\]

**Theorem 68**  Let \( \alpha, \gamma \in (0, 1) \) and \((X_t)_{t \geq 0}\) be given by (2). Assume \( f \in C^2(\mathbb{R}^d, \mathbb{R})\), \( A1, A2-(b), A3 \) and \( F2-(b) \). Then, there exists \( C \geq 0 \) (explicit and given in the proof) such that for any \( T \geq 1 \)

\[
\mathbb{E} [f(X_T)] - \min_{\mathbb{R}^d} f \leq C(1 + \log(T))^2 / T^{\alpha \wedge (1-\alpha)}.
\]

**Proof**  We begin by proving by induction over \( m \in \mathbb{N}^* \) that the following assertion \( H3(m) \) is true.

\( H3(m) \)  For any \( \alpha > 1/(m+1) \), there exists \( c_\alpha^+ > 0 \) such that for all \( t \geq 0 \),

\[ \mathbb{E} [\text{Tr}(\Sigma(X_t))] \leq c_\alpha^+ \]

In addition, for any \( \alpha \leq 1/(m+1) \), there exists \( c_\alpha^- > 0 \) such that for all \( t \geq 0 \),

\[ \mathbb{E} [\text{Tr}(\Sigma(X_t))] \leq c_\alpha^- (t + \gamma_\alpha)^{1-(m+1)\alpha} \log(t + \gamma_\alpha)^2. \]

For \( m = 1 \), \( H3(1) \) is an immediate consequence of \( A1 \) and Lemma 62. Now, let \( m \in \mathbb{N}^* \) and suppose that \( H3(m) \) holds. Let \( \alpha \in (0, 1) \). Setting \( \alpha^* = 1/(m+1) \) we see that (65) is verified with \( \beta = 1 - (m+1)\alpha \). Consequently, using \( A1, F2-(b), A2-(b) \) we can apply Proposition 63 which shows that, for \( \alpha \leq 1/(m+1) \), there exists \( \tilde{c}^{(c)} > 0 \) such that for all \( T \geq 1 \),

\[
\mathbb{E} [f(X_T)] - f^* \leq \tilde{c}^{(c)} \left\{ \log(T + \gamma_\alpha)^2 / (T + \gamma_\alpha)^{\min(\alpha,1-\alpha)} \Psi_\alpha(T + \gamma_\alpha) \right\} \\
\leq \tilde{c}^{(c)} \left\{ \log(T + \gamma_\alpha)^2 (T + \gamma_\alpha)^{-\alpha} (T + \gamma_\alpha)^{1-(m+1)\alpha} \right\} \\
\leq \tilde{c}^{(c)} \left\{ \log(T + \gamma_\alpha)^2 (T + \gamma_\alpha)^{1-(m+2)\alpha} \right\}.
\]

In particular, if \( \alpha > 1/(m+2) \) we have the existence of \( \tilde{c}_\alpha > 0 \) such that for all \( n \in \{0, \ldots, N\} \),

\[ \mathbb{E} [f(X_n)] - f^* \leq \tilde{c}_\alpha. \]

And using \( A1 \) and Lemma 42 we get that, for all \( t \in [0, T] \),

\[ \mathbb{E} [\text{Tr}(\Sigma(X_t))] \leq L_T (1 + \mathbb{E} [f(X_T) - f^*]) \leq L_T (1 + c_\alpha), \]

Combining this result with (72), we get that \( H3(m+1) \) holds with \( c_\alpha^+ = L_T (1 + \tilde{c}_\alpha) \) and \( c_\alpha^- = \tilde{c}^{(c)} \).

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 \( H3(m) \) shows the existence of \( c_0^+ > 0 \) such that \( \mathbb{E} [\text{Tr}(\Sigma(X_t))] \leq c_0 \) for all \( t \in [0, T] \).

Applying Proposition 63 gives the existence of \( c^{(c)} > 0 \) such that for all \( T \geq 1 \)

\[ \mathbb{E} [f(X_T)] - f^* \leq c^{(c)} (1 + \log(T))^2 / T^{\min(\alpha,1-\alpha)}, \]

concluding the proof.  \[\Box\]
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 A1, F2-(b), A2-(b). Then for any $\alpha, \gamma \in (0, 1)$, there exists $C_{1,\alpha}^{(d)} \geq 0$, $C_{2,\alpha}^{(d)} \geq 0$ and a function $\Phi_{\alpha}^{(d)} : \mathbb{R}_+ \to \mathbb{R}_+$ such that, for any $n \geq 0$,

$$
E \left[ \|X_n - x^*\|^2 \right] \leq C_{1,\alpha}^{(d)} \Phi_{\alpha}^{(d)}(n) + C_{2,\alpha}^{(d)}.
$$

And we have

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

The values of the constants are given by

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

$$
C_{2,\alpha}^{(d)} = \begin{cases} 
  2 \max_{k \leq \lfloor n/2 \rfloor} E \left[ \|X_k - x^*\|^2 \right] & \text{if } \alpha < 1/2, \\
  2 \max_{k \leq \lfloor n/2 \rfloor} E \left[ \|X_k - x^*\|^2 \right] + 2\gamma^2 \eta & \text{if } \alpha = 1/2, \\
  2 \max_{k \leq \lfloor n/2 \rfloor} E \left[ \|X_k - x^*\|^2 \right] + \gamma^2 \eta (2\alpha - 1)^{-1} & \text{if } \alpha > 1/2. 
\end{cases}
$$

**Proof** Let $f : \mathbb{R}^d \to \mathbb{R}$ verifying assumptions A1 and F2-(b). We consider $(X_n)_{n \geq 0}$ satisfying (1). Let $x^* \in \mathbb{R}^d$. We have, using (1), Lemma 41 and F2-(b) that for all $n \geq (\gamma L_T)^{1/\alpha}$,

$$
E[\|X_{n+1} - x^*\|^2 | F_n] = E[\|X_n - x^* - \gamma(n+1)^{-\alpha} \nabla f(X_n, Z_{n+1})\|^2 | F_n]
$$

$$
= \|X_n - x^*\|^2 - 2\gamma/(n+1)^{\alpha} \langle X_n - x^*, \nabla f(X_n, Z_{n+1}) \rangle + 2\gamma^2(n+1)^{-2\alpha} \langle \nabla f(X_n, Z_{n+1}), X_n - x^* \rangle + L_T \gamma^2(n+1)^{-2\alpha} \langle \nabla f(X_n, Z_{n+1}), X_n - x^* \rangle + L_T \gamma^2(n+1)^{-2\alpha}
$$

$$
\leq \|X_n - x^*\|^2 + \gamma^2(n+1)^{\alpha} \langle \nabla f(X_n, X_n - x^*) \rangle + \gamma^2 L_T(n+1)^{-2\alpha}
$$

$$
E[\|X_{n+1} - x^*\|^2] \leq E[\|X_n - x^*\|^2] + \gamma^2 L_T(n+1)^{-2\alpha}.
$$

Summing the previous inequality leads to

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

80
As in the previous proof we now distinguish three cases:

(a) If $\alpha < 1/2$, we have
\[
\mathbb{E} \left[ \|X_n - x^*\|^2 \right] \leq \|X_0 - x^*\|^2 + \gamma^2 L_T (1 - 2\alpha)^{-1} (n + 1)^{1 - 2\alpha} \\
\leq \|X_0 - x^*\|^2 + 2\gamma^2 L_T (1 - 2\alpha)^{-1} n^{1 - 2\alpha}.
\]

(b) If $\alpha = 1/2$, we have $\mathbb{E}[\|X_n - x^*\|^2] \leq \|X_0 - x^*\|^2 + \gamma^2 L_T (\log(n) + 2)$.

(c) If $\alpha > 1/2$, we have $\mathbb{E}[\|X_n - x^*\|^2] \leq \|X_0 - x^*\|^2 + \gamma^2 L_T (2\alpha - 1)^{-1}$.

\[\blacksquare\]

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 A1, F2-(b), A2-(b). In addition, assume that there exists $\alpha^* \in [0, 1/2]$, $\beta > 0$ and $C_0 \geq 0$ such that for all $n \in \{0, \ldots, N\}$
\[
\mathbb{E}[\|\nabla f(X_n, Z)\|^2] \leq \begin{cases} C_0 (n + 1)^\beta \log(n + 1) & \text{if } \alpha \leq \alpha^*, \\ C_0 & \text{if } \alpha > \alpha^*. \end{cases} \quad (73)
\]

Then there exists $\bar{C}_\alpha \geq 0$ such that, for all $N \geq 1$,
\[
\mathbb{E}[f(X_N)] - f^* \leq \bar{C}_\alpha \left\{ (1 + \log(N + 1))^2/(N + 1)^{\min(\alpha, 1 - \alpha)} \Psi_\alpha (N + 1) + 1/(N + 1) \right\},
\]
with
\[
\Psi_\alpha(n) = \begin{cases} n^\beta & \text{if } \alpha \leq \alpha^*, \\ 1 & \text{if } \alpha > \alpha^*. \end{cases}
\]

**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 $\bar{C}_\alpha = 2 \max((2\gamma)^{-1} \|X_0 - x^*\|^2, 2C(d))$,

\[\blacksquare\]

Let $(S_k)_{k \in \{0, \ldots, N\}}$ defined for any $k \in \{0, \ldots, N\}$ by
\[
S_k = (k + 1)^{-1} \sum_{t=N-k}^{N} \mathbb{E}[f(X_t)] .
\]

Note that $\mathbb{E}[f(X_N)] - f^* = (S_N - S_0) + (S_0 - f^*)$. We are now going to control each one of the two terms $(S_0 - S_N)$ and $(S_N - f^*)$ 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 $S(T) - f^*$ (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 A1, A2-(b), and F2-(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 $\mathbb{E}[\|Y - x^*\|^2] \leq C_{1,\alpha}^{(d)} (N + 1)^{1-2\alpha} \log(N + 1) + 4C_{2,\alpha}^{(d)}$ with $C_{1,\alpha}^{(d)}$ and $C_{2,\alpha}^{(d)}$ given in Lemma 69, we have

$$
\mathbb{E} \left[ \sum_{k=N-u}^{N} f(X_k) - f(Y) \right] \leq 2C_{(d)}(u+1)/(N+1)^{\min(\alpha,1-\alpha)} (1 + \log(N + 1)) \Psi_\alpha(N+1) + (2\gamma)^{-1} (N-u+1)^{\alpha} \mathbb{E}[\|X_{N-u} - Y\|^2],
$$

with

$$
\Psi_\alpha(n) = \begin{cases} 
\frac{n^d}{\alpha} & \text{if } \alpha \leq \alpha^*, \\
1 & \text{if } \alpha > \alpha^*, 
\end{cases}
$$

and $C_{(d)} = 4((\gamma/2)C_0)(1-\alpha)^{-1} + (2\gamma)^{-1}(C_{2,\alpha}^{(d)} + 4C_{1,\alpha}^{(d)})$.

**Proof** Let $\ell \in \{0, \ldots, N\}$, let $k \geq \ell$, let $Y \in F_{\ell}$. Using F2-(b) we have

$$
\mathbb{E} \left[ \|X_{k+1} - Y\|^2 \mid F_{k} \right] = \mathbb{E} \left[ \|X_k - Y - \gamma(k+1)^{-\alpha} \nabla \tilde{f}(X_k, Z_{k+1})\|^2 \mid F_{k} \right] \\
= \|X_k - Y\|^2 + \gamma^2(k+1)^{-2\alpha} \mathbb{E} \left[ \|\nabla \tilde{f}(X_k, Z_{k+1})\|^2 \mid F_{k} \right] \\
- 2\gamma(k+1)^{-\alpha} \langle X_k - Y, \nabla \tilde{f}(X_k) \rangle \\
\mathbb{E} \left[ f(X_k) - f(Y) \right] \leq (2\gamma)^{-1}(k+1)^{\alpha} \left( \mathbb{E} \left[ \|X_k - Y\|^2 \right] - \mathbb{E} \left[ \|X_{k+1} - Y\|^2 \right] \right) \\
+ (\gamma/2)(k+1)^{-\alpha} \mathbb{E} \left[ \|\nabla \tilde{f}(X_k, Z_{k+1})\|^2 \mid F_{k} \right] \\
\mathbb{E} \left[ f(X_k) - f(Y) \right] \leq (2\gamma)^{-1}(k+1)^{\alpha} \left( \mathbb{E} \left[ \|X_k - Y\|^2 \right] - \mathbb{E} \left[ \|X_{k+1} - Y\|^2 \right] \right) \\
+ (\gamma/2)(k+1)^{-\alpha} \mathbb{E} \left[ \|\nabla \tilde{f}(X_k, Z_{k+1})\|^2 \right]. \tag{75}
$$

Let $u \in \{0, \ldots, N\}$. Summing now (75) between $k = N-u$ and $k = N$ gives

$$
\mathbb{E} \left[ \sum_{k=N-u}^{N} f(X_k) - f(Y) \right] \leq (2\gamma)^{-1} \sum_{k=N-u+1}^{N} \mathbb{E} \left[ \|X_k - Y\|^2 \right] ((k+1)^{\alpha} - k^\alpha) \\
+ (\gamma/2) \sum_{k=N-u}^{N} \mathbb{E} \left[ \|\nabla \tilde{f}(X_k, Z_{k+1})\|^2 \right] (k+1)^{-\alpha} \\
+ (2\gamma)^{-1} (N-u+1)^{\alpha} \mathbb{E} \left[ \|X_{N-u} - Y\|^2 \right]. \tag{76}
$$
In the following we will take for $Y$ either $x^*$ or $X_m$ for $m \in [0, N]$. We now have to run separate analyses depending on the value of $\alpha$.

(a) If $\alpha \leq \alpha^*$, then (73) gives that

$$\mathbb{E} \left[ \left\| \nabla f(X_k, Z_{k+1}) \right\|^2 \right] \leq C_0 (N + 1)^{\beta} \log(N + 1),$$

and Lemma 69 gives that for all $k \in \{0, \ldots, N\}$,

$$\mathbb{E} \left[ \left\| X_k - Y \right\|^2 \right] \leq 2 \mathbb{E} \left[ \left\| X_k - x^* \right\|^2 \right] + 2 \mathbb{E} \left[ \left\| Y - x^* \right\|^2 \right]$$

$$\leq 2C_{1,\alpha}^{(d)} (k + 1)^{1-2\alpha} \log(k + 1) + 2C_{1,\alpha}^{(d)} (N + 1)^{1-2\alpha} \log(N + 1) + 4C_{2,\alpha}^{(d)}$$

$$\leq 4C_{1,\alpha}^{(d)} (N + 1)^{1-2\alpha} \log(N + 1) + 4C_{2,\alpha}^{(d)}.$$

We define $C_{3,\alpha}^{(d)} = 4C_{2,\alpha}^{(d)}$. Using (76) with $c^{(b)} = ((\gamma/2)C_0) (1 - \alpha)^{-1}$, we get

$$\mathbb{E} \left[ \sum_{k=N-u}^{N} f(X_k) - f(Y) \right] \leq (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \left\| X_{N-u} - Y \right\|^2 \right]$$

$$+ (2\gamma)^{-1} \left( C_{3,\alpha}^{(d)} + 4C_{1,\alpha}^{(d)} (N + 1)^{1-2\alpha} \log(N + 1) \right) ((N + 1)^{\alpha} - (N - u + 1)^{\alpha})$$

$$+ (\gamma/2)C_0 (N + 1)^{\beta} \log(N + 1) ((N + 1)^{1-\alpha} - (N - u)^{1-\alpha})$$

$$\leq C^{(b)} (N + 1)^{\beta} \log(N + 1) ((N + 1)^{1-\alpha} - (N - u)^{1-\alpha})$$

$$+ (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \left\| X_{N-u} - Y \right\|^2 \right]$$

$$+ (2\gamma)^{-1} C_{3,\alpha}^{(d)} ((N + 1)^{\alpha} - (N - u)^{\alpha})$$

$$+ (2\gamma)^{-1} 4C_{1,\alpha}^{(d)} ((N + 1)^{1-\alpha} - (N - u)^{1-\alpha}) \log(N + 1)$$

$$\leq C^{(d)} (N + 1)^{\beta} (1 + \log(N + 1)) ((N + 1)^{1-\alpha} - (N - u)^{1-\alpha})$$

$$+ (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \left\| X_{N-u} - Y \right\|^2 \right],$$

where we used Lemma 50. Similarly to (69) we have

$$(N + 1)^{1-\alpha} - (N - u)^{1-\alpha}$$

$$= \left\{ ((N + 1)^{1-\alpha} - (N - u)^{1-\alpha}) ((N + 1)^{\alpha} + (N - u)^{\alpha}) \right\} ((N + 1)^{\alpha} + (N - u)^{\alpha})^{-1}$$

$$\leq 2(u + 1)/(N + 1)^{\alpha}.$$

(b) If $\alpha \in (\alpha^*, 1/2]$, then Lemma 69 gives that for all $k \in \{0, \ldots, N\}$,

$$\mathbb{E} \left[ \left\| X_k - Y \right\|^2 \right] \leq 2 \mathbb{E} \left[ \left\| X_k - x^* \right\|^2 \right] + 2 \mathbb{E} \left[ \left\| Y - x^* \right\|^2 \right]$$

$$\leq 2C_{1,\alpha}^{(d)} (k + 1)^{1-2\alpha} \log(k + 1) + 2C_{1,\alpha}^{(d)} (N + 1)^{1-2\alpha} \log(N + 1) + 4C_{2,\alpha}^{(d)}$$

$$\leq 4C_{1,\alpha}^{(d)} (N + 1)^{1-2\alpha} \log(N + 1) + 4C_{2,\alpha}^{(d)}.$$
Combining (73) and (76) we have

\[\mathbb{E} \left[ \sum_{k=N-u}^{N} f(X_k) - f(Y) \right] \leq (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \|X_{N-u} - Y\|^2 \right] + (2\gamma)^{-1} \left( C_{3,\alpha}^{(d)} + 4C_{1,\alpha}^{(d)} \log(N+1)(N+1)^{1-2\alpha} \right) ((N+1)^{\alpha} - (N - u + 1)^{\alpha}) \]

\[+ (\gamma/2) C_0(1 - \alpha)^{-1} ((N+1)^{1-\alpha} - (N - u)^{1-\alpha}) \]

\[\leq C^{(b)} ((N+1)^{1-\alpha} - (N - u)^{1-\alpha}) + (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \|X_{N-u} - Y\|^2 \right] + (2\gamma)^{-1} \left( C_{3,\alpha}^{(d)} + 4C_{1,\alpha}^{(d)} \right) (1 + \log(N+1)) ((N+1)^{\alpha} - (N - u)^{\alpha}) \]

\[\leq C^{(d)} (1 + \log(N+1)) ((N+1)^{1-\alpha} - (N - u)^{1-\alpha}) + (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \|X_{N-u} - Y\|^2 \right]. \]

(c) If \( \alpha > 1/2 \), then \( \alpha > \alpha^\star \) and Lemma 69 gives

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

Using Lemma 50, (73) and (76) we have

\[\mathbb{E} \left[ \sum_{k=N-u}^{N} f(X_k) - f(Y) \right] \leq (\gamma C_0/2)(1 - \alpha)^{-1} ((N+1)^{1-\alpha} - (N - u)^{1-\alpha}) \]

\[+ (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \|X_{N-u} - Y\|^2 \right] + (2\gamma)^{-1} C_{3,\alpha}^{(d)} ((N+1)^{\alpha} - (N - u + 1)^{\alpha}) \]

\[\leq C^{(b)} ((N+1)^{1-\alpha} - (N - u)^{1-\alpha}) + (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \|X_{N-u} - Y\|^2 \right] + (2\gamma)^{-1} C_{3,\alpha}^{(d)} ((N+1)^{\alpha} - (N - u)^{\alpha}) \]

\[\leq C^{(d)} ((N+1)^{\alpha} - (N - u)^{\alpha}) + (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \|X_{N-u} - Y\|^2 \right]. \]

Similarly to (69) we have

\[(N+1)^{\alpha} - (N - u)^{\alpha} = \{(N+1)^{\alpha} - (N - u)^{\alpha}\} ((N+1)^{1-\alpha} + (N - u)^{1-\alpha}) \]

\[\times ((N+1)^{1-\alpha} + (N - u)^{1-\alpha})^{-1} \]

\[\leq 2(u+1)/(N+1)^{1-\alpha}. \]

Finally, putting the three cases above together we obtain

\[\mathbb{E} \left[ \sum_{k=N-u}^{N} f(X_k) - f(Y) \right] \leq 2C_{2,\alpha}^{(d)} (u+1)/(N+1)^{\min(\alpha,1-\alpha)} (1 + \log(N+1)) \Psi_\alpha(N+1) \]

\[+ (2\gamma)^{-1} (N - u + 1)^{\alpha} \mathbb{E} \left[ \|X_{N-u} - Y\|^2 \right], \]
CONVERGENCE AND APPROXIMATION FOR SGD

\[ \Psi_\alpha(n) = \begin{cases} n^\beta & \text{if } \alpha \leq \alpha^*, \\ 1 & \text{if } \alpha > \alpha^*. \end{cases} \]

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

**Lemma 72**  Assume A1, A2-(b) and F2-(b). In addition, assume that (73) holds. Then, for any \( \alpha, \gamma \in (0, 1) \) and \( N \in \mathbb{N} \) we have

\[ S_0 - S_N \leq 2c^{(d)}(N + 1)^{-\min(\alpha, 1-\alpha)}(1 + \log(N + 1))^2 \Psi_\alpha(N + 1). \]

with \( S \) given in (74).

**Proof** Let \( u \in \{0, \ldots, N\} \). Using Lemma 71 with the choice \( Y = X_{N-u} \) gives

\[ \mathbb{E} \left[ \sum_{k=N-u}^{N} f(X_k) - f(X_{N-u}) \right] \leq 2c^{(d)}(N + 1)^{-\min(\alpha, 1-\alpha)}(1 + \log(N + 1))\Psi_\alpha(N + 1). \]

And then,

\[ S_u = (u + 1)^{-1} \sum_{k=N-u}^{N} \mathbb{E}[f(X_k)] \leq 2c^{(d)}(N + 1)^{-\min(\alpha, 1-\alpha)}(1 + \log(N + 1))\Psi_\alpha(N + 1) + \mathbb{E}[f(X_{N-u})]. \] (77)

We have now, using (77),

\[ uS_{u-1} = (u + 1)S_u - \mathbb{E}[f(X_{N-u})] \]
\[ = uS_u + S_u - \mathbb{E}[f(X_{N-u})] \leq uS_u + 2c^{(d)}(N + 1)^{-\min(\alpha, 1-\alpha)}(1 + \log(N + 1))\Psi_\alpha(N + 1) \]
\[ S_{u-1} - S_u \leq 2c^{(d)}u^{-1}(N + 1)^{-\min(\alpha, 1-\alpha)}\log(N + 1) \]
\[ S_0 - S_N \leq 2c^{(d)}(N + 1)^{-\min(\alpha, 1-\alpha)}(1 + \log(N + 1))\Psi_\alpha(N + 1) \sum_{u=1}^{N}(1/u) \]
\[ S_0 - S_N \leq 2c^{(d)}(N + 1)^{-\min(\alpha, 1-\alpha)}(1 + \log(N + 1))^2\Psi_\alpha(N + 1). \] (78)

**Lemma 73**  Assume A1, A2-(b) and F2-(b). In addition, assume that (73) holds. Then, for any \( \alpha, \gamma \in (0, 1) \) and \( N \in \mathbb{N} \) we have

\[ S_N - f^* \leq 2c^{(d)}(1 + \log(N + 1))^2(N + 1)^{-\min(\alpha, 1-\alpha)}\Psi_\alpha(N + 1) \]
\[ + (2\gamma)^{-1}(N + 1)^{-1}\|X_0 - x^*\|^2. \] (79)

with \( S \) given in (74).
Proof Using Lemma 71 with the choice \( Y = x^\ast \) and \( u = N \) gives
\[
( N + 1 )^{-1} \mathbb{E} \left[ \sum_{k=0}^{N} f(X_k) - f(x^\ast) \right] \leq 2 \mathcal{C}_d^d (1 + \log(N + 1)) ( N + 1 )^{-\min(\alpha, 1 - \alpha)} \Phi_\alpha ( N + 1 )
\]
\[
+ (2\gamma)^{-1} ( N + 1 )^{-1} \| X_0 - x^\ast \|^2
\]
Therefore,
\[
S_N - f^\ast \leq 2 \mathcal{C}_d (1 + \log(N + 1))^2 ( N + 1 )^{-\min(\alpha, 1 - \alpha)} \Phi_\alpha ( N + 1 )
\]
\[
+ (2\gamma)^{-1} ( N + 1 )^{-1} \| X_0 - x^\ast \|^2 .
\]
(80)

Theorem 74 Let \( \gamma, \alpha \in (0, 1) \) and \( (X_n)_{n \geq 0} \) be given by (1). Assume A1, A2-(b) and F2-(b). Then, there exists \( C \geq 0 \) (explicit and given in the proof) such that for any \( N \geq 1 \),
\[
\mathbb{E} [ f(X_N) ] - \min_{x \in \mathcal{X}} f \leq C (1 + \log(N + 1))^2 / (N + 1)^\alpha (1 - \alpha)
\]

Proof We begin by proving by induction over \( m \in \mathbb{N}^* \) that the following assertion H4(\( m \)) is true.

H4 (m) For any \( \alpha > 1 / (m + 1) \), there exists \( \mathcal{C}_\alpha^+ > 0 \) such that for all \( n \in \mathbb{N} \), \( \mathbb{E}[\| \nabla \tilde{f}(X_n, Z) \|^2] \leq \mathcal{C}_\alpha^+ \). In addition, for any \( \alpha \leq 1 / (m + 1) \), there exists \( \mathcal{C}_\alpha^- > 0 \) such that for all \( n \in \mathbb{N} \), \( \mathbb{E}[\| \nabla \tilde{f}(X_n, Z) \|^2] \leq \mathcal{C}_\alpha^- n^{-(m+1)\alpha}(1 + \log(n))^2 \).

For \( m = 1 \), H4(1) is an immediate consequence of A1 and Lemma 69, with \( \mathcal{C}_\alpha^+ = L^2 \mathcal{C}_{2,\alpha}^d \) and \( \mathcal{C}_\alpha^- = L^2 \max(\mathcal{C}_{1,\alpha}^d, \mathcal{C}_{2,\alpha}^d) \). Now, let \( m \in \mathbb{N}^* \) and assume that H4(\( m \)) holds. Let \( \alpha \in (0, 1) \). Setting \( \alpha^\ast = 1 / m + 1 \) we see that (73) is verified with \( \beta = 1 - (m + 1) \alpha \). Consequently, using A1, F2-(b), A2-(b) we can apply Proposition 70 which shows that, for \( \alpha \leq 1 / (m + 1) \)
\[
\mathbb{E} [ f(X_N) ] - f^\ast \leq \tilde{\mathcal{C}}_\alpha \left\{ (1 + \log(N + 1))^2 / (N + 1)^{\min(\alpha, 1 - \alpha)} \mathcal{Y}_\alpha ( N + 1 ) + 1 / (N + 1) \right\}
\]
\[
\leq \tilde{\mathcal{C}}_\alpha \left\{ (1 + \log(N + 1))^2 (N + 1)^{-\alpha} (N + 1)^{1-(m+1)\alpha} + 1 / (N + 1) \right\}
\]
\[
\leq \tilde{\mathcal{C}}_\alpha \left\{ (1 + \log(N + 1))^2 (N + 1)^{1-(m+2)\alpha} + 1 / (N + 1) \right\} .
\]
(81)

In particular, if \( \alpha > 1 / (m + 2) \) we have the existence of \( \tilde{\mathcal{C}}_\alpha > 0 \) such that for all \( n \in \{0, \ldots, N\} \),
\[
\mathbb{E} [ f(X_n) ] - f^\ast \leq \tilde{\mathcal{C}}_\alpha .
\]
And using A1 and Lemma 41 we get that, for all \( n \in \{0, \ldots, N\} \)
\[
\mathbb{E}[\| \nabla \tilde{f}(X_n, Z) \|^2] \leq L_T(1 + \mathbb{E} [ f(X_n) - f^\ast ] ) \leq L_T(1 + \tilde{\mathcal{C}}_\alpha) ,
\]
Combining this result with (81), we get that H4(\( m + 1 \)) holds with \( \mathcal{C}_\alpha^+ = L_T(1 + \tilde{\mathcal{C}}_\alpha) \) and \( \mathcal{C}_\alpha^- = 2\tilde{\mathcal{C}}_\alpha \). Finally this proves that H4(\( m \)) 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 H4(\( m \)) shows the existence of \( \mathcal{C}_0 > 0 \) such that \( \mathbb{E}[\| \nabla \tilde{f}(X_n, Z) \|^2] \leq \mathcal{C}_0 \) for all \( n \in \mathbb{N}^* \). Applying Proposition 70 gives the existence of \( \mathcal{C}^{(d)} > 0 \) such that for all \( N \geq 1 \)
\[
\mathbb{E} [ f(X_N) ] - f^\ast \leq \mathcal{C}^{(d)} (1 + \log(N + 1))^2 / (N + 1)^{\min(\alpha, 1 - \alpha)} ,
\]
with \( \mathcal{C}^{(d)} = 2\tilde{\mathcal{C}}_\alpha \), 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 $E||X_t - x^*||_{2p}^{1/p}$ 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 $f^* = 0$. 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 $\delta_1$ and $\delta_2$ given in Theorem 10. Using Lemma 48, we have for any $t \geq 0$

$$E\left[f(X_t)\alpha_t^\delta\ell_t^{-\varepsilon}\right] - f(x_0)\gamma_{\alpha}^\delta$$

$$= \int_0^t \left\{-\ell_s^{-\varepsilon}\alpha_s^{\delta-\alpha}E[||\nabla f(X_s)||^2] + (\gamma_{\alpha}/2)\ell_s^{-\varepsilon}\alpha_s^{\delta-2\alpha}E[\langle\nabla^2 f(X_s), \Sigma(X_s)\rangle]ight. + \delta\ell_s^{-\varepsilon}\alpha_s^{\delta-1}E[f(X_s)] - \varepsilon\ell_s^{-\varepsilon-1}\alpha_s^\delta E[f(X_s)]\right\} ds .$$

Define for any $t \geq 0$, $E(t) = E[f(X_t)]\alpha_t^\delta\ell_t^{-\varepsilon}$. $(t \mapsto E(t))$ is differentiable and using A1 and A2 we have for any $t > 0$,

$$dE(t)/dt \leq -\ell_t^{-\varepsilon}\alpha_t^{\delta-\alpha}E[||\nabla f(X_t)||^2] + (\gamma_{\alpha}/2)\ell_t^{-\varepsilon}\alpha_t^{\delta-2\alpha}\eta + \delta a_t^{-1}E(t) .$$

Using, F3 and Hölder’s inequality we have for any $t \geq 0$

$$\tau E[f(X_t)] \leq E[||X_t - x^*||^{r_3r_1}]^{-\varepsilon} E[||\nabla f(X_t)||^2]^{-r_1/2} .$$

Noting that $(r_3r_1)^{-1} = r_1^{-1} - 1/2$, we get for any $t \geq 0$

$$E[||\nabla f(X_t)||^2] \geq \tau^{2r_1^{-1}} E[f(X_t)]^{2r_1^{-1}} E[||X_t - x^*||^{r_3r_1}]^{1-2r_1^{-1}}$$

$$\geq \tau^{2r_1^{-1}} C_{\beta,\varepsilon}^{1-2r_1^{-1}} a_t^{\beta(1-2r_1^{-1})} \ell_t^{\alpha(1-2r_1^{-1})} E[f(X_t)]^{2r_1^{-1}}$$

$$\geq \tau^{2r_1^{-1}} C_{\beta,\varepsilon}^{1-2r_1^{-1}} a_t^{\beta(1-2r_1^{-1})-2r_1^{-1} \delta} \ell_t^{\alpha(1-2r_1^{-1})-2r_1^{-1} - \varepsilon} E(t)^{2r_1^{-1}} .$$

Therefore, we have for any $t \geq 0$

$$dE(t)/dt \leq -\tau^{2r_1^{-1}} C_{\beta,\varepsilon}^{1-2r_1^{-1}} a_t^{\beta(1-2r_1^{-1})(\delta+\beta)-\alpha} E(t)^{2r_1^{-1}} + \gamma_{\alpha}\ell_t^{-\varepsilon} a_t^{\delta-2\alpha}\eta + \delta a_t^{-1}E(t) .$$

Let $D_3 = \max(D_1, D_2)$ with

$$D_1 = \left(\alpha \left[C_{\beta,\varepsilon}^{2r_1^{-1}} \gamma_{\alpha}^{-1} \ell_t^{-2r_1^{-1}} \gamma_{\alpha}^{-1}(2r_1^{-1}-1)(\delta+\beta)+\alpha-1)(2r_1^{-1}-1)\right] \right) ,$$

$$D_2 = \left(\alpha \left[C_{\beta,\varepsilon}^{2r_1^{-1}} \gamma_{\alpha}^{-1} \ell_t^{-2r_1^{-1}} \gamma_{\alpha}^{-1}(2r_1^{-1}-1)(\delta+\beta)+\alpha-1)(2r_1^{-1}-1)\right] \right) .$$

If $E(t) \geq D_3$ then $dE(t)/dt \leq 0$. Let $D = \max(D_3, E(0))$, then for any $t \geq 0$, $E(t) \leq D$, which concludes the proof.
G.2. Technical lemmas

**Lemma 75** Assume that $f$ is continuous, that $x^* \in \arg \min_{x \in \mathbb{R}^d} f(x)$ and that there exist $c, R \geq 0$ such that for any $x \in \mathbb{R}^d$ with $\|x - x^*\| \geq R$ we have $f(x) - f(x^*) \geq c\|x - x^*\|$. Let $p \in \mathbb{N}$, $X$ a $d$-dimensional random variable and $D_4 \geq 1$ such that $\mathbb{E}[(f(X) - f(x^*))^{2p}] \leq D_4$. Then there exists $D_5 \geq 0$ such that
\[
\mathbb{E}\left[\|X - x^*\|^{2p}\right] \leq D_5 D_4 .
\]

**Proof** Since $f$ is continuous there exists $a \geq 0$ such that for any $x \in \mathbb{R}^d$, $f(x) - f(x^*) \geq c\|x - x^*\| - a$. Therefore, using Jensen’s inequality and that $D_4 \geq 1$ we have
\[
\mathbb{E}\left[\|X - x^*\|^{2p}\right] \leq c^{-2p} \sum_{k=0}^{2p} \binom{k}{2p} \mathbb{E}\left[(f(X) - f(x^*))^k\right] a^{2p-k}
\]
\[
\leq c^{-2p} \sum_{k=0}^{2p} \binom{k}{2p} \mathbb{E}\left[(f(X) - f(x^*))^{2p}\right]^{k/(2p)} a^{2p-k}
\]
\[
\leq c^{-2p} \sum_{k=0}^{2p} \binom{k}{2p} D_4^{k/(2p)} a^{2p-k} \leq D_5 D_4 ,
\]
with $D_5 = c^{-2p} \sum_{k=0}^{2p} \binom{k}{2p} a^{2p-k}$.

**Lemma 76** Assume $F_3$ 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
\[
\mathbb{E}\left[\|\nabla f(X)\|^2 (f(X) - f(x^*))^{p-1}\right] \geq \mathbb{E}\left[(f(X) - f(x^*))^{p}\right]^{1+1/p} \mathbb{E}\left[\|X - x^*\|^{2p}\right]^{-1/p} ,
\]

**Proof** Let $p \in \mathbb{N}$ with $p \geq 2$ and let $\varpi = 2p/(p + 1)$. Using $F_3$ we have for any $x \in \mathbb{R}^d$
\[
\|x - x^*\|^\varpi \|\nabla f(x)\|^\varpi (f(x) - f(x^*))^{(p-1)/2} \geq (f(x) - f(x^*))^{\varpi(p+1)/2} \geq (f(x) - f(x^*))^{p} .
\]
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
\[
\mathbb{E}\left[\|X - x^*\|^\varpi \|\nabla f(X)\|^\varpi (f(X) - f(x^*))^{(p-1)/2}\right]
\]
\[
\leq \mathbb{E}\left[\|\nabla f(X)\|^2 (f(X) - f(x^*))^{p-1}\right]^{1/\varsigma} \mathbb{E}\left[\|X - x^*\|^{2p}\right]^{1/\varkappa} .
\]
Since, $\varkappa^{-1} = (1 + p)^{-1}$ we have
\[
\mathbb{E}\left[\|\nabla f(X)\|^2 (f(X) - f(x^*))^{p-1}\right] \geq \mathbb{E}\left[(f(X) - f(x^*))^{p}\right]^{1+1/p} \mathbb{E}\left[\|X - x^*\|^{2p}\right]^{-1/p} ,
\]
which concludes the proof.
Lemma 77  Let $\alpha, \gamma \in (0, 1)$. Assume that $F3b$ holds then for any $p \in \mathbb{N}$, there exists $D_{p,4} \geq 0$ such that for any $t \geq 0$

$$E\left[\|X_t - x^*\|^{2p}\right]^{1/p} \leq D_{p,4} \{1 + (\gamma_\alpha + t)^{-2\alpha}\}.$$

Proof  Let $\alpha, \gamma \in (0, 1)$ and $p \in \mathbb{N}$. Let $\mathcal{E}_{t,p} = E\left[\|X_t - x^*\|^{2p}\right]$. Using Lemma 47 and Lemma 48 we have for any $t > 0$

$$d\mathcal{E}_{t,p}/dt = -2p(\gamma_\alpha + t)^{-\alpha}E\left[\langle \nabla f(X_t), X_t - x^* \rangle \|X_t - x^*\|^{2(p-1)}\right]$$

$$+ p\gamma_\alpha(\gamma_\alpha + t)^{-2\alpha}\left\{E\left[\text{Tr}((X_t - x^*)^\top(x_t - x^*))\right] \|X_t - x^*\|^{2(p-1)}\right\}$$

$$+ 2(p-1)E\left[\|(X_t - x^*)^\top(x_t - x^*)\|\|X_t - x^*\|^{2(p-2)}\right] \leq 2p\gamma_\alpha(2p-1)(\gamma_\alpha + t)^{-2\alpha}E\left[\|X_t - x^*\|^{2(p-1)}\right]$$

$$\leq p\gamma_\alpha(2p-1)(\gamma_\alpha + t)^{-2\alpha}\mathcal{E}_{t,(p-1)}.$$  (82)

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 $D_{p,2} \geq 0$ such that for any $t \geq 0$, $E[\|X_t - x^*\|^{2p}]^{1/p} \leq D_{p,1}\{1 + (\gamma_\alpha + t)^{m-\varphi}\}$. Assume $A1$ and $F3b$ and that there exist $R \geq 0$ and $c > 0$ such that for any $x \in \mathbb{R}^d$, with $\|x\| \geq R$, $f(x) - f(x^*) \geq c\|x - x^*\|$. Then, for any $p \in \mathbb{N}$, there exists $D_{p,2} \geq 0$ such that for any $t \geq 0$,

$$E\left[\|X_t - x^*\|^{2p}\right]^{1/p} \leq D_{p,2}\{1 + (\gamma_\alpha + t)^{m-(1+\varphi)\alpha}\}.$$ 

Proof  If $\alpha \geq m/\varphi$ the proof is immediate since $\sup_{t \geq 0}\{E[\|X_t - x^*\|^{2p}]^{1/p}\} < +\infty$. 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 $t \geq 0$, $\mathcal{E}_{t,p} = (f(X_t) - f(x^*))^{2p}(\gamma_\alpha + t)^{\delta_p}$. Using Lemma 48 we have for any $t > 0$

$$d\mathcal{E}_{t,p}/dt = -2p(\gamma_\alpha + t)^{-\alpha+\delta_p}E\left[\|\nabla f(X_t)\|^2(f(X_t) - f(x^*))^{2p-1}\right]$$

$$+ p\gamma_\alpha(\gamma_\alpha + t)^{-2\alpha+\delta_p}\left\{E\left[\langle \nabla^2 f(X_t), \Sigma(X_t)\rangle(f(X_t) - f(x^*))^{2p-1}\right]\right\}$$

$$+ (2p-1)E\left[\langle \nabla f(X_t) \nabla f(X_t)\top, \Sigma(X_t)\rangle(f(X_t) - f(x^*))^{2p-2}\right] + \delta_p(\gamma_\alpha + t)^{-1}\mathcal{E}_{t,p}.$$  (83)
Combining (83), Lemma 47, Lemma 40, Lemma 76 and the fact that for any $t \geq 0$, $\mathbb{E}[\|X_t - x^*\|^4]^{1/(2p)} \leq D_{p,1}(1 + (\gamma_m + t)^{m-\varphi_0})$ we get

$$
dE_{t,p}/dt \leq -2p(\gamma_m + t)^{-\alpha - \delta_p} \mathbb{E} \left[(f(X_t) - f(x^*))^{2p}\right]^{1+1/(2p)} \mathbb{E} \left[\|X_t - x^*\|^4\right]^{-1/(2p)}
+ p\gamma_m(\gamma_m + t)^{-2\alpha + \delta_p} \mathbb{E} \left[(f(X_t) - f(x^*))^{2p - 1}\right] + L(2p - 1)\gamma_p^2 \mathbb{E} \left[(f(X_t) - f(x^*))^{2p - 1}\right] + \delta_p(\gamma_m + t)^{-1} E_{t,p} \\
\leq -2p(\gamma_m + t)^{-\alpha - \delta_p/(2p)} E_{t,p}^{1+1/(2p)} \mathbb{E} \left[\|X_t - x^*\|^2\right]^{-1/(2p)}
+ p\gamma_m(d + 2p - 1)\mathbb{E} \left[(1 + \eta)\gamma_m + t)^{-2\alpha + \delta_p/(2p)} E_{t,p}^{1-1/(2p)} + \delta_p(\gamma_m + t)^{-1} E_{t,p} \\
\leq -2p(\gamma_m + t)^{(\alpha - \delta_p/(2p)) - m} E_{t,p}^{1+1/(2p)} D_{p,1}^{-1}(1 + (\gamma_m + t)^{m - \varphi_0})^{-1}
+ p\gamma_m(d + 2p - 1)\mathbb{E} \left[(1 + \eta)\gamma_m + t)^{-2\alpha + \delta_p/(2p)} E_{t,p}^{1-1/(2p)} + \delta_p(\gamma_m + t)^{-1} E_{t,p} \\
\leq -pD_{p,1}^{-1}(1 + \gamma_m - \varphi_0)^{-1}(\gamma_m + t)^{(\alpha - \delta_p/(2p)) - m} E_{t,p}^{1+1/(2p)}
+ 2p\gamma_m(d + 2p - 1)\mathbb{E} \left[(1 + \eta)\gamma_m + t)^{-2\alpha + \delta_p/(2p)} E_{t,p}^{1-1/(2p)} + \delta_p(\gamma_m + t)^{-1} E_{t,p} .
$$

Since $m \in [0, 2]$, we have that $1 - m + (\alpha - \delta_p/(2p)) - m \leq 2\alpha + \delta_p/(2p)$, hence, $(1 - \varphi)\alpha - \delta_p/(2p) - m \leq 1$.

Therefore, using Lemma 3, there exists $D_p^{(\alpha)} \geq 1$ such that for any $t \geq 0$, $E_{t,p} \leq D_p^{(\alpha)}$. Hence, for any $t \geq 0$,

$$
\mathbb{E} \left[(f(X_t) - f(x^*))^{2p}\right] \leq D_p^{(\alpha)}(1 + (\gamma_m + t)^{pm - p(1+\varphi_0)}).
$$

Using Lemma 75, there exists $D_5 \geq 0$ such that

$$
\mathbb{E} \left[\|X_t - x^*\|^{2p}\right] \leq D_5(1 + (\gamma_m + t)^{pm - p(1+\varphi_0)},
$$

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 F2 and that arg min_{R \hat{R}} f is bounded. Then, for any $p \geq 0$ and $t \geq 0$,

$$
\mathbb{E} \left[\|X_t - x^*\|^p\right] < +\infty .
$$

**Proof** Without loss of generality we assume that $x^* = 0$ and $f(x^*) = 0$. First, since arg min_{R \hat{R}} f is bounded, there exists $R \geq 0$ such that for any $x \in \mathbb{R}^d$ with $\|x\| \geq R$, $f(x) > 0$. Let $S = \{x \in \mathbb{R}^d, \|x\| = 1\}$ and consider $m : S \rightarrow (0, +\infty)$ such that for any $\theta \in S$, $m(\theta) = f(\hat{R}\theta)$. $m$ is continuous since $f$ is convex and therefore it attains its minimum and there exists $m^* > 0$ such that for any $\theta \in S$, $m(\theta) \geq m^*$. Let $x \in \mathbb{R}^d$ with $\|x\| \geq 2\hat{R}$. Since $f_x : [0, +\infty) \rightarrow \mathbb{R}$ such that $f_x(t) = f(tx)$ is convex we have

$$(f(x) - f(\hat{R}x/\|x\|))(\|x\| - \hat{R}^{-1}) \geq (f(\hat{R}x/\|x\|))\hat{R}^{-1} \geq m^* \hat{R}^{-1} .$$
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 \( D_p \geq 0 \) such that

\[
\mathbb{E}\left[\|X_t - x^*\|^2\right]^{1/p} \leq D_p \left\{ 1 + (\gamma_0 + t)^{m - \lceil\alpha^{-1}\rceil\alpha} \right\} \leq D_p \left\{ 1 + (\gamma_0 + t)^{m - \lceil\alpha^{-1}\rceil\alpha} \right\} \leq D_p \left\{ 1 + \gamma_0 m - \lceil\alpha^{-1}\rceil\alpha \right\},
\]

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 \( t \geq 0 \)

\[
\mathbb{E}\left[\|X_t - x^*\|^2\right] = \|X_0 - x^*\|^2 - \int (\gamma_0 + s)^{-\alpha} \langle f(X_s), X_s - x^* \rangle ds + (\gamma_0/2) \int (\gamma_0 + s)^{-2\alpha} \langle \Sigma(X_s), \nabla^2 f(X_s) \rangle ds . \tag{84}
\]

Let \( \mathcal{E}_t = \mathbb{E}[\|X_t - x^*\|^2] \). Using, (84) we have for any \( t \geq 0 \),

\[
\mathcal{E}_t \leq (\gamma_0 + t)^{-\alpha} \mathbb{E}[\langle \nabla f(X_t), X_t - x^* \rangle] + (\gamma_0 \Delta / 2) (\gamma_0 + t)^{-2\alpha} . \tag{85}
\]

We divide the proof into three parts.

(a) First, assume that \( F_{3b} \) holds. Combining this result and (85), we get that for any \( t \geq 0 \),

\[
\mathcal{E}_t \leq \gamma_0 \Delta t^d (\gamma_0 + t)^{-2\alpha} . \]

Therefore, there exist \( \beta, \varepsilon \geq 0 \) and \( C_{\beta, \varepsilon} \geq 0 \) such that \( \mathbb{E}[\|X_t - x^*\|^2] < C_{\beta, \varepsilon} (\gamma_0 + t)^{-\beta} (1 + \log(1 + \gamma_0^{-1} t))^{\varepsilon} \) 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 \( D_p \geq 0 \) such that

\[
\mathbb{E}\left[\|X_t - x^*\|^2\right]^{1/p} \leq D_p \left\{ 1 + (\gamma_0 + t)^{m - \lceil\alpha^{-1}\rceil\alpha} \right\} \leq D_p \left\{ 1 + (\gamma_0 + t)^{m - \lceil\alpha^{-1}\rceil\alpha} \right\} \leq D_p \left\{ 1 + \gamma_0 m - \lceil\alpha^{-1}\rceil\alpha \right\},
\]

which concludes the proof.

(c) Finally, assume that there exists \( R \geq 0 \) such that for any \( x \in \mathbb{R}^d \) with \( \|x\| \geq R, \langle \nabla f(x), x - x^* \rangle \geq m \|x - x^*\|^2 \). Therefore, since \( (x \mapsto \nabla f(x)) \) is continuous, there exists \( a \geq 0 \) such that for any \( x \in \mathbb{R}^d, \langle \nabla f(x), x - x^* \rangle \geq m \|x - x^*\|^2 - a \). Combining this result and (85), we get that for any \( t \geq 0 \),

\[
\mathcal{E}_t \leq -m (\gamma_0 + t)^{-\alpha} \mathcal{E}_t + (\gamma_0 + t)^{-\alpha} a + \gamma_0 \Delta \mathcal{E}_t (\gamma_0 + t)^{-2\alpha} .
\]

Hence, if \( \mathcal{E}_0 \geq \max(a/m, \Delta) \) we have that \( \mathcal{E}_t \leq 0 \) and for any \( t \geq 0, \mathcal{E}_t \leq \max(a/m, \Delta, \mathcal{E}_0) \) and is bounded. Therefore, there exist \( \beta, \varepsilon \geq 0 \) and \( C_{\beta, \varepsilon} \geq 0 \) such that \( \mathbb{E}[\|X_t - x^*\|^2] < C_{\beta, \varepsilon} (\gamma_0 + t)^{-\beta} (1 + \log(1 + \gamma_0^{-1} t))^{\varepsilon} \) 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 A1, A2. Then we have:

(a) if $\mathbf{F3b}$ holds then, there exists $D \geq 0$ such that for any $N \in \mathbb{N}^*$
\[
\mathbb{E}[f(X_N)] - f^* \leq D \left[ N^{(1-3\alpha)/2} + N^{-\alpha/2} + N^{\alpha-1} \right],
\]

(b) if $\mathbf{F3}$ holds and if there exists $R \geq 0$ such that for any $x \in \mathbb{R}^d$ with $\|x\| \geq R$, $\langle \nabla f(x), x-x^* \rangle \geq m \|x-x^*\|^2$, then there exists $D \geq 0$ such that for any $N \in \mathbb{N}^*$
\[
\mathbb{E}[f(X_N)] - f^* \leq D \left[ N^{-\alpha/2} + N^{\alpha-1} \right].
\]

Proof Let $\alpha, \gamma \in (0, 1)$ and $x_0 \in \mathbb{R}^d$. We have for any $n \in \mathbb{N}$,
\[
\mathbb{E}\left[ \|X_{n+1} - x^*\|^2 \right] = \mathbb{E}\left[ \|X_n - x^*\|^2 \right] + 2\mathbb{E}\left[ \langle X_n - x^*, X_{n+1} - X_n \rangle \right] + \mathbb{E}\left[ \|X_{n+1} - X_n\|^2 \right] \\
\leq \mathbb{E}\left[ \|X_n - x^*\|^2 \right] - 2\gamma(n+1)^{-\alpha} \mathbb{E}\left[ \langle X_n - x^*, \nabla f(X_n) \rangle \right] \\
+ 2\gamma^2(n+1)^{-2\alpha} \mathbb{E}\left[ \|\nabla f(X_n)\|^2 \right] + 2\gamma(n+1)^{-2\alpha} \eta. \tag{86}
\]

We now divide the proof into two parts.

(a) Using $\mathbf{F3b}$ and Lemma 40 we have for any $x \in \mathbb{R}^d$,
\[
\langle \nabla f(x), x - x^* \rangle \geq \tau(f(x) - f(x^*)) \geq \tau \|\nabla f(x)\|^2 / (2L). \tag{87}
\]

Using A1, (86) and (87) we have for any $n \geq (4\gamma L/\tau)^{1/\alpha}$
\[
\mathbb{E}\left[ \|X_{n+1} - x^*\|^2 \right] \leq \mathbb{E}\left[ \|X_n - x^*\|^2 \right] + 2\gamma(n+1)^{-\alpha} (-\tau/(2L) + \gamma(n+1)^{-\alpha}) \mathbb{E}\left[ \|\nabla f(X_n)\|^2 \right] \\
+ 2\gamma(n+1)^{-2\alpha} \eta \\
\leq \mathbb{E}\left[ \|X_n - x^*\|^2 \right] + 2\gamma(n+1)^{-2\alpha} \eta.
\]

Therefore, there exist $\beta, \varepsilon \geq 0$ and $C_{\beta, \varepsilon} \geq 0$ such that $\mathbb{E}[\|X_n - x^*\|^2] < C_{\beta, \varepsilon}(n+1)^{-\beta}(1+\log(1+n))^\varepsilon$ 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 $\|x\| \geq R$, $\langle \nabla f(x), x - x^* \rangle \geq m \|x-x^*\|^2$. Therefore, since $(x \mapsto \nabla f(x))$ is continuous, there exists $a \geq 0$ such that for any $x \in \mathbb{R}^d$, $\langle \nabla f(x), x - x^* \rangle \geq m \|x-x^*\|^2 - a$. Combining this result and (86) we get that for any $n \in \mathbb{N}$ such that $n \geq (2/\gamma)^{\alpha-1}$
\[
\mathbb{E}\left[ \|X_{n+1} - x^*\|^2 \right] \leq (1 - \gamma(n+1)^{-\alpha}) \mathbb{E}\left[ \|X_n - x^*\|^2 \right] + 2\gamma(n+1)^{-\alpha} a + 2\gamma^2(n+1)^{-2\alpha} \eta.
\]

Hence, if $n \geq (2/\gamma)^{-\alpha-1}$ and $\mathbb{E}[\|X_n - x^*\|^2] \geq \max(2a, 2\gamma \eta)$ then $\mathbb{E}[\|X_{n+1} - x^*\|^2] \leq \mathbb{E}[\|X_n - x^*\|^2]$. Therefore, we obtain by recursion that for any $n \in \mathbb{N}$, that $(\mathbb{E}[\|X_n - x^*\|^2])_{n \in \mathbb{N}}$ is bounded which concludes the proof by applying Theorem 12. \qed
G.6. Proof of Theorem 12

Without loss of generality, we assume that \( f^* = 0 \). Let \( \alpha, \gamma \in (0, 1) \), \( x_0 \in \mathbb{R}^d \). Let \( \delta = \min(\delta_1, \delta_2) \), with \( \delta_1, \delta_2 \) given in Theorem 12 and let \( (E_k)_{k \in \mathbb{N}} \) such that for any \( k \in \mathbb{N} \), \( E_k = (k + 1)^\delta \mathbb{E}[f(X_k)](1 + \log(k + 1))^{-\varepsilon} \). 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 A2 we have for any \( n \in \mathbb{N} \) such that \( n \geq (2L\gamma)^{1/\alpha} \)

\[
\mathbb{E}[f(X_{n+1})|\mathcal{F}_n] \leq f(X_n) - \gamma(n + 1)^{-\alpha} \mathbb{E}[\langle \nabla f(X_n), H(X_n, Z_{n+1}) \rangle|\mathcal{F}_n]
+ (L/2)\gamma^2(n + 1)^{-2\alpha} \mathbb{E}[\|H(X_n, Z_{n+1})\|^2|\mathcal{F}_n],
\]

\[
\mathbb{E}[f(X_{n+1})] \leq \mathbb{E}[f(X_n)] - (n + 1)^{-\alpha} \mathbb{E}[\|\nabla f(X_n)\|^2]
+ L\gamma^2(n + 1)^{-2\alpha} \mathbb{E}[\|\nabla f(X_n)\|^2] + L\gamma^2(n + 1)^{-2\alpha} \eta
\leq \mathbb{E}[f(X_n)] - (n + 1)^{-\alpha} \left\{ 1 - L\gamma(n + 1)^{-\alpha} \right\} \mathbb{E}[\|\nabla f(X_n)\|^2] + L\gamma^2(n + 1)^{-2\alpha} \eta
\leq \mathbb{E}[f(X_n)] - (n + 1)^{-\alpha} \mathbb{E}[\|\nabla f(X_n)\|^2] / 2 + L\gamma^2(n + 1)^{-2\alpha} \eta. \tag{89}
\]

Combining (88) and (89) we get for any \( n \in \mathbb{N} \) such that \( n \geq (2L\gamma)^{1/2} \)

\[
E_{n+1} - E_n = (n + 2)^\delta \mathbb{E}[f(X_{n+1})](1 + \log(n + 2))^{-\varepsilon} - (n + 1)^\delta \mathbb{E}[f(X_n)](1 + \log(n + 1))^{-\varepsilon}
\leq (1 + \log(n + 1))^{-\varepsilon} \left\{(n + 2)^\delta - (n + 1)^\delta\right\} (\mathbb{E}[f(X_{n+1})])
\leq (1 + \log(n + 1))^{-\varepsilon} \left\{(n + 2)^\delta - (n + 1)^\delta\right\} (\mathbb{E}[f(X_n)] + L\gamma^2(n + 1)^{-2\alpha} \eta)
\leq (1 + \log(n + 1))^{-\varepsilon} \left\{c_\delta(n + 1)^{\delta-1}(\mathbb{E}[f(X_n)] + 2\gamma^2(n + 1)^{-2\alpha} \eta)
\leq c_\delta E_n + 2L\gamma^2(1 + c_\delta)(n + 1)^{\delta-2\alpha}(1 + \log(n + 1))^{-\varepsilon} \eta
\leq c_\delta E_n + 2L\gamma^2 \mathbb{E}[\|\nabla f(X_n)\|^2] / 2. \tag{90}
\]

Using (3) and the fact that for any \( k \in \mathbb{N} \), \( \mathbb{E}[\|X_k - x^*\|^2 r_3^i] \leq C_{\beta, \varepsilon}(k + 1)^{\beta}(1 + \log(1 + k))^\varepsilon \) and Hölder’s inequality and that \( r_1 r_3 = 2(2r_1^{-1} - 1)^{-1} \), we have for any \( k \in \mathbb{N} \)

\[
\mathbb{E}[\|\nabla f(X_k)\|^2] \geq \mathbb{E}[f(X_k)]^{2r_1^{-1}} C_{\beta, \varepsilon}^{-(2r_1^{-1} - 1)^{-1}} \tau^{2r_1^{-1}}(k + 1)^{-\beta(2r_1^{-1} - 1)(1 + \log(k + 1))^{-\varepsilon(2r_1^{-1} - 1)}. \tag{91}
\]
Combining (90) and (91) we get that for any \( n \in \mathbb{N} \) with \( n \geq (4\gamma)^{1/\alpha} \)

\[
E_{n+1} - E_n \leq c_\delta E_n + 2L\gamma^2(1 + c_\delta)(n + 1)^{\delta - 2\alpha}(1 + \log(n + 1))^{-\varepsilon}\eta \\
- \gamma(n + 1)^{\delta - \alpha - \beta}(2\tau_1^{-1} - 1)\mathbb{E}[f(X_n)]^{2\tau_1^{-1}} C_{\beta,\varepsilon}^{-(2\tau_1^{-1} - 1) - 1} \tau^{2\tau_1^{-1}} (1 + \log(n + 1))^{-\varepsilon 2\tau_1^{-1}} / 2 \\
\leq c_\delta E_n + 2L\gamma^2(1 + c_\delta)(n + 1)^{\delta - 2\alpha}(1 + \log(n + 1))^{-\varepsilon}\eta \\
- \gamma(n + 1)^{\alpha - (\delta + \beta)(2\tau_1^{-1} - 1)} E_n^{2\tau_1^{-1}} C_{\beta,\varepsilon}^{-(2\tau_1^{-1} - 1) - 1} \tau^{2\tau_1^{-1}} / 2.
\]

Let \( D_3 = \max(D_1, D_2) \) with

\[
\begin{cases}
D_1 = (2|c_\delta|C_{\beta,\varepsilon}^{2\tau_1^{-1} - 1} \tau^{-2\tau_1^{-1}})^{2\tau_1^{-1} - 1}, \\
D_2 = (4L\gamma^2(1 + c_\delta)C_{\beta,\varepsilon}^{2\tau_1^{-1} - 1} \tau^{-2\tau_1^{-1}})^{1/2}.
\end{cases}
\]

If \( E_n \geq D_3 \) and \( n \geq (4\gamma)^{1/\alpha} \) then \( E_{n+1} \leq E_n \). Therefore, we obtain by recursion that \( E_n \leq D \) with \( D = \max(E_0, \ldots, E_{\left\lceil (2L\gamma)^{1/\alpha} \right\rceil}, D_3) \).