



HAL
open science

Asymptotic error analysis for stochastic gradient optimization schemes with first and second order modified equations

Charles-Édouard Bréhier, Marc Dambrine, Nassim En-Nebbazi

► **To cite this version:**

Charles-Édouard Bréhier, Marc Dambrine, Nassim En-Nebbazi. Asymptotic error analysis for stochastic gradient optimization schemes with first and second order modified equations. 2024. hal-04773810

HAL Id: hal-04773810

<https://hal.science/hal-04773810v1>

Preprint submitted on 8 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ASYMPTOTIC ERROR ANALYSIS FOR STOCHASTIC GRADIENT OPTIMIZATION SCHEMES WITH FIRST AND SECOND ORDER MODIFIED EQUATIONS

CHARLES-EDOUARD BRÉHIER, MARC DAMBRINE, AND NASSIM EN-NEBBAZI

ABSTRACT. We consider a class of stochastic gradient optimization schemes. Assuming that the objective function is strongly convex, we prove weak error estimates which are uniform in time for the error between the solution of the numerical scheme, and the solutions of continuous-time modified (or high-resolution) differential equations at first and second orders, with respect to the time-step size. At first order, the modified equation is deterministic, whereas at second order the modified equation is stochastic and depends on a modified objective function. We go beyond existing results where the error estimates have been considered only on finite time intervals and were not uniform in time. This allows us to then provide a rigorous complexity analysis of the method in the large time and small time step size regimes.

1. INTRODUCTION

The analysis of numerical optimization methods, whether deterministic or stochastic, has been the subject of intense research activity. Those methods are widely applied in many situations, in particular recently in machine learning. In order to approximate minimizers of an objective function $F : \mathbb{R}^d \rightarrow \mathbb{R}$, the simplest method is the gradient descent algorithm

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

with a constant time-step size h . To study the asymptotic behavior of x_N when $N \rightarrow \infty$, it is fruitful to make a connection with the continuous-time gradient dynamics

$$x'(t) = -\nabla F(x(t)).$$

In fact, the gradient descent algorithm can be interpreted as the standard explicit Euler scheme applied to the ordinary differential equation above with time-step size h .

Connections between (discrete-time) optimization schemes and continuous-time dynamics have been studied in the literature, in deterministic and stochastic settings, not only for gradient algorithms but also for other methods such as the Polyak's heavy-ball method [9] and Nesterov's acceleration method [8]. The latter methods intend to accelerate the convergence to minimizers by introducing inertia in the scheme, and as a result the related continuous-time dynamics are second-order differential equations. In the last year, the celebrated work [13] on Nesterov's acceleration method has driven the attention of many researchers on that connection. Instead of considering leading order approximations, better understanding on the discrete-time dynamics can be obtained by the identification of more accurate continuous-time dynamics, which depend on the time-step size. In the numerical analysis community for ordinary and stochastic differential equations, such more accurate approximations are referred to as modified equations, see for instance [2, 4, 6, 14]. In the optimization community, they are referred to as high-resolution equations, see for instance [1, 11, 12]. In this article, we use the terminology modified equations.

2020 *Mathematics Subject Classification.* 60H35;65C30;65K05;65K10.

Key words and phrases. stochastic optimization schemes; modified equations; weak and strong error analysis.

The connection between the discrete and continuous time dynamics motivates the splitting of the analysis of the asymptotic behavior of optimization schemes into two parts: the analysis of the asymptotic behavior of the continuous-time system, and the analysis of the approximation error between the discrete and continuous time systems. On the one hand, differential calculus techniques can be applied to show convergence of the continuous-time dynamics when time T goes to infinity, by the identification of so-called Lyapunov functionals. For instance, if the objective function F is strongly convex, the square of the norm is a Lyapunov function, and applying Grönwall's inequality it is straightforward to show that the solution of the continuous-time gradient dynamics converges exponentially fast to the unique minimizer of F . More complex Lyapunov functions dedicated to objective functions satisfying weaker conditions may be constructed. On the other hand, when standard numerical analysis techniques are applied to study the approximation error between the discrete and continuous time systems, usually the error bounds with respect to the time-step size h depend on the final time $T = Nh$. The dependence obtained when discrete Grönwall inequalities are employed may be exponential. Instead, it is desirable to obtain uniform in time error estimates, i.e. with upper bounds independent of the final time T . Combining the two types of error bounds then enables to analyze the complexity of the algorithm, i.e. to determine how to choose the parameters of the algorithm h and N to ensure that the error is below a given threshold.

In this article, we analyze the behavior of the following stochastic gradient optimization scheme

$$\begin{cases} X_{n+1} = X_n - h\nabla F(X_n) + h\sigma(t_n, X_n)\gamma_{n+1}, & \forall n \in \mathbb{N}, \\ X_0 = x_0, \end{cases}$$

where $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is the objective function. Compared with the deterministic version above, the additional term $h\sigma(t_n, X_n)\gamma_{n+1}$ describes random perturbations, where $(\gamma_n)_{n \geq 1}$ are independent standard \mathbb{R}^d -valued Gaussian random variables. We assume that F is strongly convex, thus F admits a unique minimizer denoted by x^* . Our objective is to analyze the convergence of X_N to x^* when $N \rightarrow \infty$ and $h \rightarrow 0$. The stochastic gradient optimization scheme above has been considered previously in the literature, see for instance [1] for the proof of strong error estimates and in [6] for the proof of weak error estimates. Let us recall that when dealing with approximation of stochastic systems and associated error bounds, one may consider either strong approximation and error bounds, concerned with $\mathbb{E}[\|X_N - x^*\|]$ or with $(\mathbb{E}[\|X_N - x^*\|^2])^{1/2}$, or weak approximation and error bounds, concerned with $|\mathbb{E}[\varphi(X_N)] - \varphi(x^*)|$. We refer for instance the reference books [5, 7] on numerical methods for stochastic (partial) differential equations.

The objective of this article is to fill an important gap in the analysis of the stochastic gradient optimization scheme above: we establish uniform in time strong and weak error estimates between the solution X_N of the stochastic optimization scheme and the solution at time $T = Nh$ of associated modified equations. Indeed, the error bounds obtained in previous works such as [1, 6] are not uniform with respect to time T . As explained above, combined with relevant results on the large time of the modified equations, we are able to establish complexity results for the stochastic gradient optimization scheme. We consider two modified equations, which provide first and second order approximation with respect to h .

For the first order approximation, we exploit the connection between the stochastic gradient optimization scheme and the deterministic gradient system, i.e. the ordinary differential equation

$$\frac{d\mathcal{X}^0(t)}{dt} = -\nabla F(\mathcal{X}^0(t)), \quad \forall t \geq 0.$$

As already recalled above, assuming that F is strongly convex, $\mathcal{X}^0(T)$ converges to x^* exponentially fast when $T \rightarrow \infty$. Our first main result is the proof of weak error estimates

$$|\mathbb{E}[\varphi(X_N)] - \varphi(x^*)| = \mathcal{O}(e^{-\mu Nh} + h),$$

see Theorem 1 for the precise statement. This follows from uniform in time weak error estimates

$$\sup_{N \in \mathbb{N}} |\mathbb{E}[\varphi(X_N)] - \varphi(\mathcal{X}^0(Nh))| = \mathcal{O}(h).$$

Moreover, applying the weak error estimates with $\varphi = F$ and noting that the residual $F(X_N) - F(x^*)$ dominates $\|X_N - x^*\|^2$, one obtains directly strong error estimates

$$\mathbb{E}[\|X_N - x^*\|^2] = \mathcal{O}(e^{-\mu Nh} + h),$$

see Corollary 1. This is not a standard result in stochastic numerics, since usually weak error bounds may be deduced from strong error bounds. To prove weak error estimates, which are uniform in time, we revisit the arguments usually employed to study approximations of invariant distributions for stochastic differential equations, in particular we need to study the regularity properties of an auxiliary mapping u^0 (defined by Equation (48)), which is the solution of the (Kolmogorov) partial differential equation (52) associated with the deterministic gradient system. Precisely we prove the exponential decay with respect to t for the first and second order spatial partial derivatives $\nabla_x u^0(t, x)$ and $\nabla_x^2 u^0(t, x)$, see Lemma 5.1.

The considered stochastic gradient optimization scheme can be interpreted as the Euler–Maruyama scheme applied to the stochastic differential equation

$$d\mathcal{X}^h(t) = -\nabla F(\mathcal{X}^h(t))dt + \sqrt{h}\sigma(t, \mathcal{X}^h(t))dB(t), \quad \forall t \geq 0,$$

which can indeed be seen as a modified equation since it depends on the time-step size h . However that simple connection is not sufficient to obtain higher order approximation. Instead, one needs to consider the following modified stochastic differential equation

$$d\mathcal{Y}^h(t) = -\nabla F^h(\mathcal{Y}^h(t))dt + \sqrt{h}\sigma(t, \mathcal{Y}^h(t))dB(t), \quad \forall t \geq 0,$$

which has the same diffusion coefficient $\sqrt{h}\sigma$ but depends on the modified objective function

$$F^h = F + \frac{h}{4}\|\nabla F\|^2.$$

The modified stochastic differential equation and objective function above have also appeared in the recent articles [1, 6], see also [2, 3, 4, 10] for other examples of applications of modified/high-resolution equations in the context of deterministic and stochastic optimization. In this article, we prove uniform in time weak error estimates

$$\sup_{N \in \mathbb{N}} |\mathbb{E}[\varphi(X_N)] - \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))]| = \mathcal{O}(h^2),$$

under stronger conditions on the modified objective function F : we assume that the modified objective function F^h is strongly convex uniformly with respect to h (in an interval $(0, h_0)$). We refer to Assumption 2.3 for a rigorous statement. To the best of our knowledge, this condition has not been exhibited before in the literature. That assumption plays a crucial role in the analysis. First, it allows us to prove the exponential decay with respect to t of spatial derivatives of the solutions v_T^h (defined by Equation (59)) of associated (backward) Kolmogorov equations (60), uniformly with respect to h . In addition, it allows us to prove error bounds for $\mathbb{E}[\|\mathcal{Y}^h(T) - x^*\|^2]$ in the large time regime.

Our second main result is thus the proof of weak error estimates

$$|\mathbb{E}[\varphi(X_N)] - \varphi(x^*)| = \mathcal{O}\left(e^{-\mu Nh} (1 + h\rho(Nh))^{\frac{1}{2}} + h^2\right),$$

where the mapping ρ is defined by (30) and depends on the diffusion coefficient σ . We refer to Theorem 2 for a precise statement. Like for the first order approximation discussed above, one obtains strong error estimates

$$\mathbb{E}[\|X_N - x^*\|^2] = \mathcal{O}\left(e^{-\mu Nh} (1 + h\rho(Nh))^{\frac{1}{2}} + h^2\right),$$

see Corollary 2.

The comparison between the first and second order approximation results in the large time regime is not as straightforward as it may seem. Indeed, to obtain the first order approximation result it is sufficient to assume that some upper bounds on $\sigma(t, \cdot)$ which are uniform with respect to t . It turns out that the second order approximation result reduces the complexity only if $\sigma(t, \cdot)$ converges to 0 when t goes to infinity fast enough. We refer to Section 3 for a precise discussion on the computational complexity of the stochastic gradient optimization scheme.

Let us mention several directions for future works. It would be interesting to study the behavior of the stochastic gradient optimization scheme under weaker conditions on the objective function F . One may also investigate the behavior of other stochastic optimization schemes, such as stochastic versions of the celebrated heavy ball and Nesterov algorithms. The analysis may also be extended in an infinite dimensional setting.

The article is organized as follows. Section 2 is devoted to the description of the setting and the statement of the fundamental assumptions on the objective function F and the diffusion coefficient σ . The main results of this article, Theorem 1 and Corollary 1 for the first order approximation, and Theorem 2 and Corollary 2 for the second order approximation, are stated in Section 3, which also contains a discussion on the computational complexity. Section 4 presents some results on the stochastic gradient optimization scheme, including uniform moment bounds given by Lemma 4.1. The proof of the first order approximation result, Theorem 1, is provided in Section 5.2, and the proof of the second order approximation result, Theorem 2, is provided in Section 6.4. Appendix 6.5 contains the proof of technical but fundamental results, on the regularity properties of the solutions of the Kolmogorov equations associated with the first and second order modified equations.

2. SETTING

2.1. Notation. We denote by $\mathbb{N}^* = \{1, \dots\}$ and $\mathbb{N} = \{0, 1, \dots\}$ the sets of positive and nonnegative integers respectively. Given $d \in \mathbb{N}^*$ be an arbitrary positive integer, the Euclidean space \mathbb{R}^d is equipped with the standard inner product and norm denoted by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ respectively. Let $\mathcal{M}_d(\mathbb{R})$ denote the space of real-valued square matrices with d rows and columns. The identity matrix is denoted by I_d . For any matrix $M \in \mathcal{M}_d(\mathbb{R})$, M^* is the transpose of M and $\text{Tr}(M)$ is the trace of M . The space $\mathcal{M}_d(\mathbb{R})$ is equipped with the norm $\| \cdot \|$ defined by

$$\|M\| = \left(\sum_{1 \leq i, j \leq d} |M_{i,j}|^2 \right)^{\frac{1}{2}}, \quad \forall M \in \mathcal{M}_d(\mathbb{R}).$$

If $M, N \in \mathcal{M}_d(\mathbb{R})$ are two matrices, let

$$M : N = \sum_{1 \leq i, j \leq d} M_{i,j} N_{i,j} = \text{Tr}(MN^*).$$

If $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a real-valued function of class \mathcal{C}^1 , then $\nabla\phi(x) = (\partial_{x_i}\phi(x))_{1 \leq i \leq d} \in \mathbb{R}^d$ denotes the gradient of ϕ at $x \in \mathbb{R}^d$. Similarly, if $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a function of class \mathcal{C}^2 , then $\nabla^2\phi(x) = (\partial_{x_i}\partial_{x_j}\phi(x))_{1 \leq i, j \leq d} \in \mathcal{M}_d(\mathbb{R})$ denotes the Hessian matrix of ϕ at $x \in \mathbb{R}^d$.

Given $d_1, d_2, m \in \mathbb{N}^*$, if $\phi : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$ is a mapping of class \mathcal{C}^m , then $D^m\phi(x).(k_1, \dots, k_m) \in \mathbb{R}^{d_2}$ denotes the differential of order m of ϕ at $x \in \mathbb{R}^{d_1}$ in the directions $k_1, \dots, k_m \in \mathbb{R}^{d_1}$. In the case $d_1 = d$ and $d_2 = 1$, if $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ is of class \mathcal{C}^2 , with the notation introduced above one has

$$D\varphi(x).(k) = \langle \nabla\varphi(x), k \rangle, \quad D^2\varphi(x).(k_1, k_2) = \langle \nabla^2\varphi(x)k_1, k_2 \rangle.$$

For any $m \in \mathbb{N}^*$, if $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ is of class \mathcal{C}^m , then set

$$\|\varphi\|_m = \sup_{x \in \mathbb{R}^d} \sup_{k_1, \dots, k_m \in \mathbb{R}^d \setminus \{0\}} \frac{|D^m\varphi(x).(k_1, \dots, k_m)|}{\|k_1\| \dots \|k_m\|}.$$

In the sequel we consider functions φ of class \mathcal{C}^2 with bounded second order derivative. Note that its first order derivative satisfies the following upper bound: for all $x \in \mathbb{R}^d$ and $k \in \mathbb{R}^d$ one has

$$(1) \quad |D\varphi(x).k| \leq \left(\|\nabla\varphi(x^*)\| + \|\varphi\|_2 \|x - x^*\| \right) \|k\|.$$

Finally, let $(B(t))_{t \geq 0}$ be a standard \mathbb{R}^d -valued Wiener process defined on a probability space denoted by $(\Omega, \mathcal{F}, \mathbb{P})$, and which is adapted to a filtration $(\mathcal{F}_t)_{t \geq 0}$ which satisfies the usual conditions. If X is an integrable real-valued random variable, its expectation is denoted by $\mathbb{E}[X]$. In addition, given a σ -field $\mathcal{G} \subset \mathcal{F}$, $\mathbb{E}[X|\mathcal{G}]$ denotes the conditional expectation of X given \mathcal{G} .

In the analysis, the value of (deterministic) positive real numbers $C \in (0, \infty)$ may vary from line to line. Unless specified, the value of C is uniform with respect to variables such as time $t \in \mathbb{R}^+ = [0, \infty)$, the state $x \in \mathbb{R}^d$ or the time-step size h .

2.2. The considered stochastic gradient optimization scheme. Let $x_0 \in \mathbb{R}^d$ be an arbitrary initial value assumed to be deterministic. Considering \mathcal{F}_0 -measurable random initial values satisfying appropriate integrability properties could be considered with straightforward minor modifications of the statements.

Let $h > 0$ denote the time-step size of the considered stochastic gradient optimization scheme. Without loss of generality, it is assumed that $h \in (0, h_{\max})$ for some positive real number h_{\max} . For any nonnegative integer $n \in \mathbb{N}$, set $t_n = nh$. Let $(\gamma_n)_{n \in \mathbb{N}^*}$ be a sequence of independent standard centered \mathbb{R}^d -valued Gaussian random variables.

The objective of this article is to study the behavior of the stochastic gradient optimization scheme

$$(2) \quad \begin{cases} X_{n+1} = X_n - h\nabla F(X_n) + h\sigma(t_n, X_n)\gamma_{n+1}, & \forall n \in \mathbb{N}, \\ X_0 = x_0, \end{cases}$$

where the mapping $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is the objective function and the mapping $\sigma : [0, \infty) \times \mathbb{R}^d \rightarrow \mathcal{M}_d(\mathbb{R})$ is the diffusion coefficient. We refer to Section 2.3 for precise assumptions.

Let $\Delta B_{n+1} = B(t_{n+1}) - B(t_n)$ denote the increments of the Wiener process on the interval $[t_n, t_{n+1}]$. Recall that $(h^{-\frac{1}{2}}\Delta B_{n+1})_{n \geq 0}$ are independent standard centered Gaussian \mathbb{R}^d -valued random variables. Thus the sequences of random variables $(\sqrt{h}\gamma_n)_{n \in \mathbb{N}^*}$ and $(\Delta B_n)_{n \in \mathbb{N}^*}$ are equal in distribution, and as a result, the scheme (2) can be interpreted using the alternative formulation

$$(3) \quad \begin{cases} X_{n+1} = X_n - h\nabla F(X_n) + \sqrt{h}\sigma(t_n, X_n)\Delta B_{n+1}, & \forall n \in \mathbb{N}, \\ X_0 = x_0. \end{cases}$$

Let us stress that the schemes (2) and (3) can be considered as being equivalent since in this work we only study weak error estimates. The formulation (3) suggests to interpret the scheme as a numerical method applied to a stochastic differential equation, as will be explained below.

2.3. Assumptions. This section provides basic assumptions on the objective function F and on the diffusion coefficient σ , which are necessary to justify rigorously the main properties of the stochastic gradient optimization scheme in the large time and small time-step regime. More precisely, Assumptions 2.1 and 2.2 are required to prove the first main result of this paper, stated in Theorem 1. Stronger additional conditions are imposed in Assumptions 2.3 and 2.4 below in order to prove the second main result of this paper, stated in Theorem 2, we refer to Section 2.5 for their statements.

Let us first state the basic assumptions on the objective function F .

Assumption 2.1 (Basic assumptions on the objective function F). *The mapping $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is of class \mathcal{C}^3 , and it satisfies the following properties:*

(i) $\nabla F : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is globally Lipschitz continuous, i.e. there exists a positive real number $L_F \in (0, \infty)$ such that one has

$$(4) \quad \|\nabla F(x_2) - \nabla F(x_1)\| \leq L_F \|x_2 - x_1\|, \quad \forall x_1, x_2 \in \mathbb{R}^d.$$

(ii) F is a μ -convex function, i.e. there exists a positive real number $\mu \in (0, \infty)$ such that one has

$$(5) \quad \langle \nabla F(x_2) - \nabla F(x_1), x_2 - x_1 \rangle \geq \mu \|x_2 - x_1\|^2, \quad \forall x_1, x_2 \in \mathbb{R}^d.$$

(iii) The third order derivative of F is bounded.

Example 2.1. Assumption 2.1 is satisfied if F is defined as

$$F(x) = \|x\|^2 + \epsilon G(x), \quad \forall x \in \mathbb{R}^d,$$

where the mapping $G : \mathbb{R}^d \rightarrow \mathbb{R}$ is of class \mathcal{C}^3 with bounded first, second and third order derivatives, and for sufficiently small ϵ .

Let us state several elementary consequences of Assumption 2.1.

When the objective function F satisfies Assumption 2.1, it is straightforward to check that $\min(F)$ exists and is attained for a unique minimizer denoted by x^* . Moreover the minimizer x^* is characterized by the first-order optimality condition $\nabla F(x^*) = 0$.

Note that the second order derivative $\nabla^2 F$ satisfies the property

$$(6) \quad \mu \|k\|^2 \leq \langle \nabla^2 F(x)k, k \rangle \leq L_F \|k\|^2, \quad \forall k \in \mathbb{R}^d,$$

where the upper bound follows from the Lipschitz continuity (4) of ∇F and the lower bound is a consequence of the μ -convexity property (5). In addition, the global Lipschitz continuity condition (4) implies that ∇F has at most polynomial growth: for all $x \in \mathbb{R}^d$,

$$(7) \quad \|\nabla F(x)\| \leq L_F \|x - x^*\|.$$

Let us now describe the basic assumptions on the diffusion coefficient σ .

Assumption 2.2 (Assumptions on the diffusion coefficient σ). *The mapping $\sigma : [0, \infty) \times \mathbb{R}^d \rightarrow \mathcal{M}_d(\mathbb{R})$ satisfies the following properties:*

(i) *The mapping $\sigma : [0, \infty) \times \mathbb{R}^d \rightarrow \mathcal{M}_d(\mathbb{R})$ is continuous.*

(ii) *For all $t \in [0, \infty)$, the mapping $\sigma(t, \cdot) : \mathbb{R}^d \rightarrow \mathcal{M}_d(\mathbb{R})$ is globally Lipschitz continuous and grows at most linearly : for all $t \geq 0$, there exists $\varsigma(t) \in [0, \infty)$ such that one has*

$$(8) \quad \|\sigma(t, x_2) - \sigma(t, x_1)\| \leq \varsigma(t) \|x_2 - x_1\|, \quad \forall x_1, x_2 \in \mathbb{R}^d,$$

$$(9) \quad \sup_{x \in \mathbb{R}^d} \frac{\|\sigma(t, x)\|}{1 + \|x - x^*\|} \leq \varsigma(t).$$

(iii) *The mapping $\varsigma : [0, \infty) \rightarrow [0, \infty)$ is assumed to be bounded. Let $\|\varsigma\|_\infty = \sup_{t \geq 0} |\varsigma(t)|$.*

Finally, let us introduce the auxiliary mapping $a : [0, \infty) \times \mathbb{R}^d \rightarrow \mathcal{M}_d(\mathbb{R})$ defined by

$$(10) \quad a(t, x) = \sigma(t, x)\sigma(t, x)^*, \quad \forall t \geq 0, x \in \mathbb{R}^d.$$

2.4. First order approximate ordinary and stochastic differential equations. In this subsection, we consider that the basic assumptions on F and σ , i.e. Assumptions 2.1 and 2.2 respectively, are satisfied.

Let us first study the deterministic situation: when $\sigma = 0$, the scheme (3) is the explicit Euler scheme applied to the deterministic gradient system

$$(11) \quad \begin{cases} \frac{d\mathcal{X}^0(t)}{dt} = -\nabla F(\mathcal{X}^0(t)), & \forall t \geq 0, \\ \mathcal{X}^0(0) = x_0. \end{cases}$$

The analysis of the gradient system (11) under Assumption 2.1 is straightforward and well-known. Owing to the μ -convexity condition (5) from Assumption 2.1 on the objective function F , and due to the first-order optimality condition $\nabla F(x^*) = 0$, one obtains for all $t \geq 0$

$$\begin{aligned} \frac{1}{2} \frac{d\|\mathcal{X}^0(t) - x^*\|^2}{dt} &= \left\langle \frac{d(\mathcal{X}^0(t) - x^*)}{dt}, \mathcal{X}^0(t) - x^* \right\rangle = -\langle \nabla F(\mathcal{X}^0(t)) - \nabla F(x^*), \mathcal{X}^0(t) - x^* \rangle \\ &\leq -\mu \|\mathcal{X}^0(t) - x^*\|^2. \end{aligned}$$

Therefore, applying the Grönwall lemma, for all $t \geq 0$ one obtains

$$(12) \quad \|\mathcal{X}^0(t) - x^*\| \leq e^{-\mu t} \|x_0 - x^*\|.$$

This means that, for any initial value x_0 , the solution $\mathcal{X}^0(t)$ at time t of the deterministic gradient system (11) converges exponentially fast to the unique minimizer x^* of the objective function F in the large time regime $t \rightarrow \infty$, and that the rate of convergence is given by the parameter μ .

Let us now take into account the stochastic perturbation. The stochastic gradient optimization scheme (2), using the equivalent formulation (3), can be interpreted as the Euler–Maruyama scheme with time-step size h applied to the stochastic differential equation

$$(13) \quad \begin{cases} d\mathcal{X}^h(t) = -\nabla F(\mathcal{X}^h(t))dt + \sqrt{h}\sigma(t, \mathcal{X}^h(t))dB(t), & \forall t \geq 0, \\ \mathcal{X}^h(0) = x_0. \end{cases}$$

The noise in the stochastic differential equation (13) is interpreted in the Itô sense. Observe that the (13) depends on the time-step size h , the diffusion coefficient is $\sqrt{h}\sigma$. Let us mention that the stochastic differential equation (13) is globally well-posed, since the mappings ∇F and $\sigma(t, \cdot)$ are globally Lipschitz continuous, see Assumptions 2.1 and 2.2: for any initial value $x_0 \in \mathbb{R}^d$, there exists a unique global solution to the stochastic differential equation (13).

The deterministic system (11) and the stochastic system (13) are first-order approximations, in the sense of weak convergence, of the stochastic gradient optimization scheme (2) in the small time-step size limit $h \rightarrow 0$, for arbitrary fixed value of $T \in (0, \infty)$. More precisely, for any time $T \in (0, \infty)$ and any sufficiently smooth function $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$, there exists $C(T, x_0, \varphi) \in (0, \infty)$ such that for all $h \in (0, h_{\max})$ one has

$$\sup_{nh \leq T} |\mathbb{E}[\varphi(X_n)] - \mathbb{E}[\varphi(\mathcal{X}^0(nh))]| + \sup_{nh \leq T} |\mathbb{E}[\varphi(X_n)] - \mathbb{E}[\varphi(\mathcal{X}^h(nh))]| \leq C(T, x_0, \varphi)h.$$

One of the objectives of this work is to show that under appropriate conditions the weak error estimates are uniform with respect to $T \in (0, \infty)$, which is of tremendous importance in the context of optimization where one needs to consider the large time regime $T \rightarrow \infty$. In other words, it will be proved that the weak error estimates above hold, with $\sup_{T \in (0, \infty)} C(T, x_0, \varphi) < \infty$, we refer to

Theorem 1 for a more precise statement.

Note that in order to understand the behavior of the numerical scheme (2) considering the stochastic differential equation (13) instead of the deterministic gradient system (11) is not relevant

since the weak order of convergence is equal to 1 even when the noise is introduced. The objective of the next section is to introduce a stochastic differential equation depending on a modified objective function F^h , depending on the time-step size h , in order to obtain a second order weak approximation. As already mentioned, stronger conditions will be imposed on F and σ to prove this result.

2.5. Second order approximate stochastic differential equation. Following the works [6, 1], instead of considering the stochastic differential equation (13), we introduce the family of stochastic differential equations depending on the time-step size h defined by

$$(14) \quad \begin{cases} d\mathcal{Y}^h(t) = -\nabla(F + \frac{h}{4}\|\nabla F\|^2)(\mathcal{Y}^h(t))dt + \sqrt{h}\sigma(t, \mathcal{Y}^h(t))dB(t), & \forall t \geq 0, \\ \mathcal{Y}^h(0) = x_0. \end{cases}$$

This family provides second order weak approximation for the stochastic gradient optimization scheme (2): for any time $T \in (0, \infty)$ and any sufficiently smooth function $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$, there exists $C(T, x_0, \varphi) \in (0, \infty)$ such that one has

$$\sup_{nh \leq T} |\mathbb{E}[\varphi(X_n)] - \mathbb{E}[\varphi(\mathcal{Y}^h(nh))]| \leq C(T, x_0, \varphi)h^2.$$

Like for the first order approximations described in Section 2.4, the main objective of this work is to show that under appropriate conditions the weak error estimates are uniform with respect to $T \in (0, \infty)$. We refer to Theorem 2 for a precise statement.

Note that the stochastic differential equations (13) and (14) have the same diffusion coefficient $\sqrt{h}\sigma$, but they do not have the same drift: the drift in (13) is $-\nabla F$ where F is the objective function, whereas the drift in (14) is given by $-\nabla F^h$ where $F^h : \mathbb{R}^d \rightarrow \mathbb{R}$ is the so-called modified objective function defined by

$$(15) \quad F^h(x) = F(x) + \frac{h}{4}\|\nabla F(x)\|^2, \quad \forall x \in \mathbb{R}^d.$$

It is worth mentioning that x^* is the unique minimizer of F^h for any time-step size h . Indeed, one has $F^h(x^*) = F(x^*)$ and $F^h(x) > F(x)$ for all $x \in \mathbb{R}^d \setminus \{x^*\}$.

Following from straightforward computations, note that $\nabla\|\nabla F\|^2(x^*) = 0$, and that there exists a positive real number $C \in (0, \infty)$ such that for all $x \in \mathbb{R}^d$ one has

$$(16) \quad \|\nabla\|\nabla F\|^2(x)\| \leq C\|x - x^*\|, \quad \|\nabla^2\|\nabla F\|^2(x)\| \leq C(1 + \|x - x^*\|).$$

As a consequence, there exists a positive real number $L_{F,2} \in (0, \infty)$ such that, for all $x_1, x_2 \in \mathbb{R}^d$, one has

$$(17) \quad \|\nabla\|\nabla F\|^2(x_2) - \nabla\|\nabla F\|^2(x_1)\| \leq L_{F,2}(1 + \|x_1 - x^*\| + \|x_2 - x^*\|)\|x_2 - x_1\|.$$

The conditions imposed on F in Assumption 2.1 are not sufficient for the analysis, additional conditions are given in Assumption 2.3 below. Note that as usual higher order regularity conditions on F and σ are needed to justify the second order weak approximation result stated above, compared with the first order weak approximation result. A more restrictive condition needs to be imposed: the mapping F^h is assumed to be μ -convex for any (sufficiently small) time-step size h , and the positive real number μ is independent of h . This assumption is non trivial and is discussed below. To the best of our knowledge, this type of condition has not been presented so far in the literature.

Assumption 2.3 (Strengthened assumptions on the objective function F). *Assume that the objective function F satisfies the basic Assumption 2.1.*

Furthermore, assume that the mapping $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is of class \mathcal{C}^5 , and that it satisfies the following properties:

(i) There exist positive real numbers $h_0 \in (0, h_{\max})$ and $\mu \in (0, \infty)$, such that for all $h \in (0, h_0)$ the mapping F^h is μ -convex:

$$(18) \quad \langle \nabla F^h(x_2) - \nabla F^h(x_1), x_2 - x_1 \rangle \geq \mu \|x_2 - x_1\|^2, \quad \forall x_1, x_2 \in \mathbb{R}^d, \quad \forall h \in (0, h_0).$$

(ii) The fourth and fifth order derivatives of F are bounded.

Let us give an example of a function such that Assumption 2.3 is satisfied.

Example 2.2. Let F be given by $F(x) = \|x\|^2 + \epsilon G(x)$ for all $x \in \mathbb{R}^d$, where $G : \mathbb{R}^d \rightarrow \mathbb{R}$ is a mapping of class C^5 with bounded derivatives of any order and ϵ is chosen to ensure that the μ -convexity condition from Assumption 2.1 is satisfied. Then Assumption 2.3 is satisfied if and only if there exists $c_F \in (0, \infty)$ such that

$$\inf_{x \in \mathbb{R}^d} D^3 G(x) \cdot (x, k, k) \geq -c_F \|k\|^2, \quad \forall k \in \mathbb{R}^d,$$

is satisfied.

Remark 2.1. The mapping $-\nabla F^h : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is only locally Lipschitz continuous, owing to the inequality (17), following from the basic Assumption 2.1: there exists $C \in (0, \infty)$ such that for all $x_1, x_2 \in \mathbb{R}^d$ one has

$$(19) \quad \|\nabla F^h(x_2) - \nabla F^h(x_1)\| \leq (L_F + hL_{F,2})(1 + \|x_1 - x^*\| + \|x_2 - x^*\|)\|x_2 - x_1\|.$$

This local Lipschitz continuity property ensures local well-posedness of the stochastic differential equation (14). Owing to the strengthened Assumption 2.3, F^h satisfies the μ -convexity condition (18), which implies that $-\nabla F^h$ satisfies a one-sided Lipschitz continuity property. This provides the global well-posedness of (14) by standard arguments. Moreover, the μ -convexity condition (18) is employed to prove moment bounds on the solution of the stochastic differential equation (14), see Lemma 6.1 in Section 6, which are uniform with respect to $t \geq 0$.

Finally, let us describe additional regularity conditions imposed on the diffusion coefficient σ . These assumptions are less restrictive than those imposed on F in Assumption 2.3 above.

Assumption 2.4 (Strengthened assumptions on the diffusion coefficient σ). Assume that the diffusion coefficient σ satisfies the basic Assumption 2.2.

Furthermore, assume that for all $t \in [0, \infty)$ the mapping $\sigma(t, \cdot) : \mathbb{R}^d \rightarrow \mathcal{M}_d(\mathbb{R})$ is of class C^3 . Moreover, the second and the third derivatives of $\sigma(t, \cdot)$ are assumed to be bounded, uniformly with respect to $t \in [0, \infty)$: one has

$$\sup_{t \in [0, \infty)} \sup_{x \in \mathbb{R}^d} \sup_{k_1, k_2 \in \mathbb{R}^d \setminus \{0\}} \frac{\|D^2 \sigma(t, x) \cdot (k_1, k_2)\|}{\|k_1\| \|k_2\|} + \sup_{t \in [0, \infty)} \sup_{x \in \mathbb{R}^d} \sup_{k_1, k_2, k_3 \in \mathbb{R}^d \setminus \{0\}} \frac{\|D^3 \sigma(t, x) \cdot (k_1, k_2, k_3)\|}{\|k_1\| \|k_2\| \|k_3\|} < \infty.$$

Finally, assume that for all $x \in \mathbb{R}^d$, the mapping $\sigma(\cdot, x) : [0, \infty) \rightarrow \mathcal{M}_d(\mathbb{R})$ is globally Lipschitz continuous, uniformly with respect to x : there exists a positive real number $L_\sigma \in (0, \infty)$ such that for all $x \in \mathbb{R}^d$ one has

$$(20) \quad \|\sigma(t_2, x) - \sigma(t_1, x)\| \leq L_\sigma |t_2 - t_1|, \quad \forall t_1, t_2 \in [0, \infty).$$

We conclude this section by remarks on the modified objective function F^h .

Remark 2.2. The gradient ∇F^h of the modified objective function can be written as

$$\nabla F^h(x) = (I + \frac{h}{2} \nabla^2 F(x)) \nabla F(x), \quad \forall x \in \mathbb{R}^d.$$

This observation suggests to propose alternative stochastic differential equations which provide a second order weak approximation instead of (14). For instance, one can consider the stochastic differential equations

$$d\bar{Y}^h(t) = \bar{b}^h(\bar{Y}^h(t))dt + \sqrt{h}\sigma(t, \bar{Y}^h(t))dB(t)$$

with $\bar{b}^h(x) = -(I - \frac{h}{2}\nabla^2 F(x))^{-1}\nabla F(x)$ (under the condition $h < 2L_F$ to ensure that the matrix is invertible) or with $\bar{b}^h(x) = -\exp(\frac{h}{2}\nabla^2 F(x))\nabla F(x)$. Note that in general it does not seem possible to write $\bar{b}^h(x)$ given in the two examples above as a gradient. It can be proved that second order weak error estimates are still valid with $\mathcal{Y}^h(nh)$ replaced by $\bar{\mathcal{Y}}^h(nh)$, for $nh \leq T$.

It would be interesting to exhibit a choice of alternative drift coefficient \bar{b}^h such that one could obtain uniform in time weak error estimates at second order in h under weaker conditions on the objective function F . For the two examples mentioned above this does not seem to be the case. We leave investigations in this direction for future works.

3. MAIN RESULTS

This section is devoted to the full and rigorous statements of the main results of this article. We also discuss the relevance of the results. The proofs of Theorems 1 and 2 are technical and require some auxiliary results, they are postponed to Sections 5 and 6 respectively.

Let us first state Theorem 1, which provides weak error estimates, i.e. upper bounds for $\mathbb{E}[\varphi(X_N)] - \varphi(x^*)$ for sufficiently smooth functions φ . The error is the sum of two contributions: one term which converges exponentially fast to 0 when $Nh \rightarrow \infty$, and one term which converges to 0 with order 1 when $h \rightarrow 0$.

Theorem 1. *Assume that the objective function F and that the diffusion coefficient σ satisfy the basic Assumption 2.1 and the basic Assumption 2.2 respectively.*

There exist positive real numbers $H \in (0, h_{\max})$ and $C \in (0, \infty)$ such that for any mapping $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ of class \mathcal{C}^2 with bounded second order derivative, for any initial value $x_0 \in \mathbb{R}^d$, for all $N \in \mathbb{N}^$ and $h \in (0, H)$, one has*

$$(21) \quad |\mathbb{E}[\varphi(X_N)] - \varphi(x^*)| \leq C_1(\varphi, x_0)(e^{-\mu Nh} + h),$$

where the positive real number $C_1(\varphi, x_0)$ is given by

$$(22) \quad C_1(\varphi, x_0) = C(\|\nabla\varphi(x^*)\| + \|\varphi\|_2)(1 + \|x_0 - x^*\|^3).$$

Applying the result of Theorem 1 with the function $\varphi = F$, and applying the inequality

$$(23) \quad F(x) - F(x^*) \geq \frac{\mu}{2}\|x - x^*\|^2, \quad \forall x \in \mathbb{R}^d,$$

which follows from the μ -convexity condition (5) of F and the optimality condition $\nabla F(x^*)$, one obtains the following error estimates for the residual $F(X_N) - F(x^*)$ and for the error $\|X_N - x^*\|$.

Corollary 1. *Let the setting of Theorem 1 be satisfied. For any initial value $x_0 \in \mathbb{R}^d$, for all $N \in \mathbb{N}^*$ and $h \in (0, H)$, for the residual one has the upper bound*

$$(24) \quad \mathbb{E}[F(X_N) - F(x^*)] \leq C(1 + \|x_0 - x^*\|^3)(e^{-\mu Nh} + h),$$

and one has strong error estimates

$$(25) \quad \mathbb{E}[\|X_N - x^*\|] \leq (\mathbb{E}[\|X_N - x^*\|^2])^{\frac{1}{2}} \leq C(1 + \|x_0 - x^*\|^{\frac{3}{2}})(e^{-\frac{\mu Nh}{2}} + h^{\frac{1}{2}}).$$

Observe that the upper bound (25) is a strong error estimate for the error $\|X_N - x^*\|$, contrary to (21) and (24) which are weak error estimates. It is not common in stochastic numerics to deduce strong error estimates from weak error estimates. However it is common (but it is not always the case) to observe that the weak order of convergence is twice the strong order of convergence with respect to a discretization parameter h . From the strong error estimates (25), it is straightforward to deduce error bounds associated with convergence in probability.

Let us perform some complexity analysis as a consequence of the weak and strong error estimates stated in Theorem 1. Let the initial value x_0 be fixed, and assume also that the function φ is fixed

when dealing with errors in the weak sense. Given an arbitrarily small positive real number $\varepsilon \in (0, 1)$, the objective is to identify small enough parameters $h \in (0, H)$ and large enough parameters $N \in \mathbb{N}^*$ such that

$$(26) \quad |\mathbb{E}[\varphi(X_N)] - \varphi(x^*)| \leq C\varepsilon,$$

when one is concerned with weak error estimates, or such that

$$(27) \quad (\mathbb{E}[\|X_N - x^*\|^2])^{\frac{1}{2}} \leq C\varepsilon,$$

when one is concerned with strong error estimates. In the above $C \in (0, \infty)$ is a positive real number which is independent of N , h and ε .

If one is interested in weak errors, owing to (21), N and h should be chosen such that $h \approx \varepsilon$ and such that $Nh \approx \log(\varepsilon^{-1})$ to ensure that (26) holds. The complexity is given by the computational cost, which is proportional to

$$(28) \quad N \approx \varepsilon^{-1} \log(\varepsilon^{-1}) =: N_{1,w}.$$

If one is interested in strong error estimates, owing to (25), N and h should be chosen such that $\sqrt{h} \approx \varepsilon$ and such that $Nh \approx |\log(\varepsilon)|$ to ensure that (27) holds. The complexity is given by

$$(29) \quad N \approx \varepsilon^{-2} \log(\varepsilon^{-1}) =: N_{1,s}.$$

Let us now state Theorem 2. The main achievement is to obtain a second order convergence with respect to h in the weak error estimates for $\mathbb{E}[\varphi(X_N)] - \varphi(x^*)$, instead of the first order convergence obtained in Theorem 1. In order to obtain this result, it is necessary to impose the strengthened Assumptions 2.3 and 2.4 on the objective function F and on the diffusion coefficient σ . Let us emphasize that the strengthened Assumption 2.4 contains only a few additional regularity conditions on σ , whereas the strengthened Assumption 2.3 is more restrictive and more fundamental, since it is needed to ensure that the modified objective function F^h defined by (15) satisfies a μ -convexity property, uniformly with respect to $h \in (0, h_0)$.

The statement of Theorem 2 requires to introduce

$$(30) \quad \rho(T) = \int_0^T e^{2\mu s} \zeta(s)^2 ds,$$

for all $T \in (0, \infty)$.

Theorem 2. *Assume that the objective function F and that the diffusion coefficient σ satisfy the strengthened Assumption 2.3 and the strengthened Assumption 2.4 respectively.*

There exist positive real numbers $H \in (0, h_{\max})$ and $C \in (0, \infty)$ such that for any mapping $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ of class \mathcal{C}^3 with bounded second and third order derivatives, for any initial value $x_0 \in \mathbb{R}^d$, for all $N \in \mathbb{N}^$ and $h \in (0, H)$, one has*

$$(31) \quad |\mathbb{E}[\varphi(X_N)] - \varphi(x^*)| \leq C_2(\varphi, x_0) \left(e^{-\mu Nh} (1 + h\rho(Nh))^{\frac{1}{2}} + h^2 \right),$$

where the positive real number $C_2(\varphi, x_0)$ is given by

$$(32) \quad C_2(\varphi, x_0) = C(\|\nabla\varphi(x^*)\| + \|\varphi\|_2 + \|\varphi\|_3)(1 + \|x_0 - x^*\|^5).$$

Similarly as Corollary 1 which is deduced from Theorem 1, we obtain Corollary 2 as a consequence of Theorem 2.

Corollary 2. *Let the setting of Theorem 2 be satisfied. For any initial value $x_0 \in \mathbb{R}^d$, for all $N \in \mathbb{N}^*$ and $h \in (0, H)$, for the residual one has the upper bound*

$$(33) \quad \mathbb{E}[F(X_N) - F(x^*)] \leq C(1 + \|x_0 - x^*\|^5) \left(e^{-\mu Nh} (1 + h\rho(Nh))^{\frac{1}{2}} + h^2 \right),$$

and one has strong error estimates

$$(34) \quad \mathbb{E}[\|X_N - x^*\|] \leq (\mathbb{E}[\|X_N - x^*\|^2])^{\frac{1}{2}} \leq C(1 + \|x_0 - x^*\|^{\frac{5}{2}}) \left(e^{-\frac{\mu Nh}{2}} (1 + h\rho(Nh))^{\frac{1}{4}} + h \right).$$

To perform a complexity analysis and a comparison with the results obtained in Theorem 1 and 1, one needs to study the behavior of the mapping ρ defined by (30) above. Recall that the mapping ς is assumed to be bounded, see the basic Assumption 2.2 on the diffusion coefficient σ . Therefore one obtains the upper bound

$$\rho(T) \leq \frac{\|\varsigma\|_{\infty}^2}{2\mu} e^{2\mu T}, \quad \forall T \in (0, \infty).$$

Plugging that inequality in the error bounds stated in Theorem 2 and Corollary 2 is not satisfactory, since the error bounds in Theorem 1 and Corollary 1 would be more precise. Improved error bounds are obtained if one assumes that

$$\varsigma(s) \xrightarrow{s \rightarrow \infty} 0.$$

Then the upper bounds on $\rho(T)$ depend on the speed of convergence. Let us consider the cases of exponential and polynomial convergence.

On the one hand, assume that there exist positive real numbers $c \in (0, \infty)$ and $\nu \in (0, \infty)$ such that for all $s \geq 0$ one has $\varsigma(s) \leq ce^{-\nu s}$. It is straightforward to check that for all $T \geq 0$ one obtains

$$\rho(T) \leq \begin{cases} C, & \text{if } \nu > \mu, \\ CT, & \text{if } \nu = \mu, \\ Ce^{2(\mu-\nu)T}, & \text{if } \nu < \mu. \end{cases}$$

Therefore, when the convergence of ς to 0 is exponential one has

$$e^{-2\mu T} \rho(T) \leq Ce^{-2\mu' T}, \quad \forall T \geq 0,$$

for some positive real numbers $C \in (0, \infty)$ and $\mu' \in (0, \mu]$. As a result, the analysis of the computational cost can be performed as above, but one can take advantage of the second order convergence with respect to h to reduce the computational cost. Owing to (31), to ensure that (26) holds, when one is concerned with weak error estimates, N and h should be chosen such that $h \approx \sqrt{\varepsilon}$ and such that $Nh \approx \log(\varepsilon^{-1})$. The complexity is then given by

$$(35) \quad N \approx \varepsilon^{-\frac{1}{2}} \log(\varepsilon^{-1}) =: N_{2,w,e}.$$

If one is interested in strong error estimates, owing to (34), to ensure that (27) holds, N and h should be chosen such that $h \approx \varepsilon$ and such that $Nh \approx \log(\varepsilon^{-1})$. The complexity is then given by

$$(36) \quad N \approx \varepsilon^{-1} \log(\varepsilon^{-1}) =: N_{2,s,e}.$$

On the other hand, assume that there exist positive real numbers $c \in (0, \infty)$ and $\alpha \in (0, \infty)$ such that for all $s \geq 0$ one has $\varsigma(s) \leq c(1 + s^\alpha)^{-1}$. Then one checks that for all $T \geq 0$ one obtains

$$\rho(T) \leq \frac{Ce^{2\mu T}}{1 + T^{2\alpha}}.$$

As a result, when the convergence of ς to 0 is polynomial one has

$$e^{-2\mu T} \rho(T) \leq \frac{C}{1 + T^{2\alpha}}, \quad \forall T \geq 0,$$

for some positive real number $C \in (0, \infty)$. As a result, the analysis of the computational cost needs more care. Owing to (31), to ensure that (26) holds, when one is concerned with weak error

estimates, N and h should be chosen such that $h \approx \sqrt{\varepsilon}$ and such that $\frac{h}{1+(Nh)^{2\alpha}} \approx \varepsilon^2$. The complexity is then given by

$$(37) \quad N \approx \varepsilon^{-\frac{1}{2}-\frac{3}{4\alpha}} =: N_{2,w,p}.$$

If one is interested in strong error estimates, owing to (34), to ensure that (27) holds, N and h should be chosen such that $h \approx \varepsilon$ and such that $\frac{h}{1+(Nh)^{2\alpha}} \approx \varepsilon^4$. The complexity is then given by

$$(38) \quad N \approx \varepsilon^{-1-\frac{3}{2\alpha}} =: N_{2,s,p}.$$

It remains to compare the complexity results obtained above. Note first that the computational cost related to the strong error estimates is in all cases the square of the computational cost related to the weak error estimates, namely one has $N_{1,s} = N_{1,w}^2$, $N_{2,s,e} = N_{2,w,e}^2$ and $N_{2,s,p} = N_{2,w,p}^2$. This is consistent with the fact (which is common in stochastic numerics and is observed in the results stated above) that the weak rate of convergence is twice the strong rate of convergence. To compare the results, it is thus sufficient in the sequel to deal with the weak error estimates. It is necessary to treat separately the cases where ς converges either exponentially or polynomially fast to 0 at infinity. When the convergence is exponentially fast, one obtains $N_{2,w,e} = \sqrt{N_{1,w}}$, meaning that Theorem 2 implies a lower computational cost – but recall that this requires to impose the strengthened Assumption 2.3 on the objective function F . When the convergence is only polynomially fast, it is not always true that $N_{2,w,p}$ is smaller than $N_{1,w}$, i.e. that Theorem 2 provides a reduction of the computational cost compared with Theorem 1. This reduction only occurs when α is sufficiently large, namely that $\alpha > 3/2$. When $\alpha \leq 3/2$, considering Theorem 1 seems to provide better results than using Theorem 2.

Remark 3.1. *The reason why $\rho(T)$ with $T = Nh$ appears in Theorem 2 and Corollary 2, whereas it does not appear in Theorem 1 and 1 is due to the use of different decompositions of the weak error*

$$\begin{aligned} \mathbb{E}[\varphi(X_N)] - \varphi(x^*) &= \varphi(\mathcal{X}^0(Nh)) - \varphi(x^*) + \mathbb{E}[\varphi(X_N)] - \varphi(\mathcal{X}^0(Nh)) \\ &= \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))] - \varphi(x^*) + \mathbb{E}[\varphi(X_N)] - \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))]. \end{aligned}$$

On the one hand, the first decomposition (47) (see Section 5) is the primary step in the proof of Theorem 1, and exploits the large time behavior of the solution $\mathcal{X}^0(T)$ of the deterministic gradient system (11). The diffusion coefficient σ does not appear in this description. On the other hand, the second decomposition (57) (see Section 6) is the primary step in the proof of Theorem 2. To deal with the first term of the decomposition, it is necessary to study the large time behavior of the solution $\mathcal{Y}^h(T)$ of the stochastic differential equation (14), see Proposition 6.1.

4. AUXILIARY RESULTS ON THE STOCHASTIC GRADIENT OPTIMIZATION SCHEME

This section is devoted to studying some properties of the discrete-time stochastic process $(X_n)_{n \in \mathbb{N}}$ defined by the stochastic gradient optimization scheme (2). First, Lemma 4.1 states moment bounds, which are uniform with respect to the time-step size h , and with respect to time $n \in \mathbb{N}$. Second, for any time-step size h an auxiliary continuous time process $(\tilde{X}^h(t))_{t \geq 0}$ is introduced and some properties are given in Lemma 4.2.

4.1. Moment bounds on the scheme.

Lemma 4.1. *Assume that the objective F satisfies the basic Assumption 2.1 and that the diffusion coefficient σ satisfies the basic Assumption 2.2. There exist a non-increasing sequence $(H_p)_{p \in \mathbb{N}^*}$, with $H_1 \in (0, h_{\max})$, and a non-decreasing sequence $(C_p)_{p \in \mathbb{N}^*}$, such that for any initial value $x_0 \in \mathbb{R}^d$ and for any $p \in \mathbb{N}^*$ one has*

$$(39) \quad \sup_{h \in (0, H_p)} \sup_{n \in \mathbb{N}} \mathbb{E}[\|X_n - x^*\|^{2p}] \leq C_p(1 + \|x_0 - x^*\|^{2p}).$$

In the proof of Lemma 4.1 given below, it will be obtained that H_p tends to 0 when p tends to infinity. In the error analysis below, it will be sufficient to apply the result of Lemma 4.1 with $p = 3$, therefore the condition $h \in (0, H_3)$ on the time-step size will be sufficient.

Remark 4.1. *Lemma 4.1 is valid under the basic Assumptions 2.1 and 2.2 on the objective function F and the diffusion coefficient σ , in other words it is not necessary to impose the strengthened Assumptions 2.3 and 2.4.*

Proof. Let us first introduce some auxiliary notation.

For any positive integer $n \in \mathbb{N}^*$, define the σ -field $\mathcal{F}_n = \sigma(\gamma_1, \dots, \gamma_n)$. Set also $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Recall the notation $\mathbb{E}[\cdot | \mathcal{F}_n]$ for the conditional expectation.

For all $n \in \mathbb{N}$, set

$$\Delta X_{n+1} = X_{n+1} - X_n = -h\nabla F(X_n) + h\sigma(t_n, X_n)\gamma_{n+1},$$

Observe that the random variable X_n is \mathcal{F}_n -measurable for all $n \in \mathbb{N}$. In addition, one has

$$(40) \quad X_{n+1} - x^* = X_n - x^* + \Delta X_{n+1}, \quad \forall n \in \mathbb{N}.$$

The proof then proceeds in two steps, in order to treat the cases $p = 1$ and $p \geq 2$ separately.

Step 1: the case $p = 1$. Using the identity (40) above, one obtains for all $n \in \mathbb{N}$

$$\|X_{n+1} - x^*\|^2 = \|X_n - x^*\|^2 + 2\langle X_n - x^*, \Delta X_{n+1} \rangle + \|\Delta X_{n+1}\|^2.$$

Let $n \in \mathbb{N}$. Since the random variable X_n is \mathcal{F}_n -measurable, one has

$$\mathbb{E}[\|X_n - x^*\|^2 | \mathcal{F}_n] = \|X_n - x^*\|^2.$$

In addition, the random variable γ_{n+1} is independent of \mathcal{F}_n , thus one has

$$\mathbb{E}[\gamma_{n+1} | \mathcal{F}_n] = \mathbb{E}[\gamma_{n+1}] = 0.$$

As a result, using the definition of ΔX_{n+1} and the properties of conditional expectation, one has

$$\begin{aligned} \mathbb{E}[\langle X_n - x^*, \Delta X_{n+1} \rangle | \mathcal{F}_n] &= -h\langle X_n - x^*, \nabla F(X_n) \rangle + h\langle X_n - x^*, \sigma(t_n, X_n)\mathbb{E}[\gamma_{n+1} | \mathcal{F}_n] \rangle \\ &= -h\langle X_n - x^*, \nabla F(X_n) \rangle. \end{aligned}$$

Using the μ -convexity of F (5) in Assumption 2.1, one obtains

$$\langle X_n - x^*, \nabla F(X_n) \rangle = \langle X_n - x^*, \nabla F(X_n) - \nabla F(x^*) \rangle \geq \mu\|X_n - x^*\|^2,$$

since $\nabla F(x^*) = 0$ by the the first-order optimality condition. One obtains the the upper bound

$$\mathbb{E}[\langle X_n - x^*, \Delta X_{n+1} \rangle | \mathcal{F}_n] = -h\langle X_n - x^*, \nabla F(X_n) \rangle \leq -\mu h\|X_n - x^*\|^2.$$

Finally, using the definition of ΔX_{n+1} and similar arguments one has

$$\begin{aligned} \mathbb{E}[\|\Delta X_{n+1}\|^2 | \mathcal{F}_n] &= h^2\mathbb{E}[\|\nabla F(X_n)\|^2 | \mathcal{F}_n] + h^2\mathbb{E}[\|\sigma(t_n, X_n)\gamma_{n+1}\|^2 | \mathcal{F}_n] \\ &\quad - 2h^2\mathbb{E}[\langle \nabla F(X_n), \sigma(t_n, X_n)\gamma_{n+1} \rangle | \mathcal{F}_n] \\ &= h^2\|\nabla F(X_n)\|^2 + h^2\|\sigma(t_n, X_n)\|^2. \end{aligned}$$

Recall that the mappings ∇F and σ have linear growth (see (7) and (9)), and that the mapping ς is bounded (by Assumption 2.2). Then one obtains for all $n \in \mathbb{N}$ and all $h \in (0, h_{\max})$

$$\mathbb{E}[\|\Delta X_{n+1}\|^2 | \mathcal{F}_n] \leq L_F^2 h^2\|X_n - x^*\|^2 + h^2\varsigma(t_n)^2(1 + \|X_n - x^*\|^2)$$

and gathering the upper bounds for $\mathbb{E}[\langle X_n - x^*, \Delta X_{n+1} \rangle | \mathcal{F}_n]$ and $\mathbb{E}[\|\Delta X_{n+1}\|^2 | \mathcal{F}_n]$ one has

$$\mathbb{E}[\|X_{n+1} - x^*\|^2 | \mathcal{F}_n] \leq \left(1 - 2\mu h + (L_F^2 + 2\|\varsigma\|_\infty^2)h^2\right)\|X_n - x^*\|^2 + 2h^2\|\varsigma\|_\infty^2.$$

For all $h \in (0, h_{\max})$, set

$$\rho_1(h) = 1 - 2\mu h + (L_F^2 + 2\|\varsigma\|_\infty^2)h^2.$$

The tower property of conditional expectation yields the identity

$$\mathbb{E}[\mathbb{E}[\|X_{n+1} - x^\star\|^2 | \mathcal{F}_n]] = \mathbb{E}[\|X_{n+1} - x^\star\|^2],$$

thus taking expectation, for all $n \in \mathbb{N}$ one has

$$\mathbb{E}[\|X_{n+1} - x^\star\|^2] \leq \rho_1(h) \mathbb{E}[\|X_n - x^\star\|^2] + 2h^2 \|\varsigma\|_\infty^2.$$

Set

$$H_1 = \frac{1}{2} \min\left(\frac{1}{2\mu}, \frac{2\mu}{L_F^2 + 2\|\varsigma\|_\infty^2}, h_{\max}\right)$$

and observe that one has $\rho_1(h) \in (0, 1)$ for all $h \in (0, H_1)$. Applying the discrete Grönwall inequality, one obtains for all $n \geq 0$ and for all $h \in (0, H_1)$

$$\begin{aligned} \mathbb{E}[\|X_n - x^\star\|^2] &\leq \rho_1(h)^n \|x_0 - x^\star\|^2 + \frac{2h^2 \|\varsigma\|_\infty^2}{1 - \rho_1(h)} \\ &\leq \|x_0 - x^\star\|^2 + \frac{2h \|\varsigma\|_\infty^2}{2\mu - (L_F^2 + 2\|\varsigma\|_\infty^2)h} \\ &\leq \|x_0 - x^\star\|^2 + \frac{2H_1 \|\varsigma\|_\infty^2}{\mu}. \end{aligned}$$

This concludes the proof of the inequality (39) in the case $p = 1$.

Step 2: the case $p \geq 2$.

We start by a useful elementary remark we shall use also in other proofs.

Fact 1. Define the auxiliary mapping ψ_p by $\psi_p(x) = \|x - x^\star\|^{2p}$. For all $x \in \mathbb{R}^d$ one has

$$(41) \quad D\psi_p(x) \cdot k = 2p \|x - x^\star\|^{2(p-1)} \langle x - x^\star, k \rangle, \quad \forall k \in \mathbb{R}^d,$$

$$(42) \quad |D^2\psi_p(x) \cdot (k_1, k_2)| \leq 2p(2p-1) \|x - x^\star\|^{2(p-1)} \|k_1\| \|k_2\|, \quad \forall k_1, k_2 \in \mathbb{R}^d.$$

Applying the Taylor formula, one obtains

$$\begin{aligned} \|X_{n+1} - x^\star\|^{2p} &= \psi_p(X_{n+1}) = \psi_p(X_n + \Delta X_{n+1}) \\ &= \psi_p(X_n) + D\psi_p(X_n) \cdot \Delta X_{n+1} \\ &\quad + \int_0^1 (1-\theta) D^2\psi_p(X_n + \theta \Delta X_{n+1}) \cdot (\Delta X_{n+1}, \Delta X_{n+1}) d\theta \\ &\leq \|X_n - x^\star\|^{2p} + 2p \|X_n - x^\star\|^{2(p-1)} \langle X_n - x^\star, \Delta X_{n+1} \rangle \\ &\quad + 2p^2 (\|X_n - x^\star\| + \|\Delta X_{n+1}\|)^{2(p-1)} \|\Delta X_{n+1}\|^2. \end{aligned}$$

We first consider the term with the inner product in the first line. Like in the treatment of the case $p = 1$, considering the conditional expectation with respect to \mathcal{F}_n one has

$$\begin{aligned} \mathbb{E}[2p \|X_n - x^\star\|^{2(p-1)} \langle X_n - x^\star, \Delta X_{n+1} \rangle | \mathcal{F}_n] &= 2p \|X_n - x^\star\|^{2(p-1)} \mathbb{E}[\langle X_n - x^\star, \Delta X_{n+1} \rangle | \mathcal{F}_n] \\ &\leq -2p\mu \|X_n - x^\star\|^{2p}. \end{aligned}$$

For the remainder term on the second line, the convexity inequality $(a+b)^q \leq 2^{q-1}(a^q + b^q)$ infers the upper bound

$$(\|X_n - x^\star\| + \|\Delta X_{n+1}\|)^{2(p-1)} \|\Delta X_{n+1}\|^2 \leq 2^{2p-3} (\|X_n - x^\star\|^{2(p-1)} \|\Delta X_{n+1}\|^2 + \|\Delta X_{n+1}\|^{2p}).$$

Employing the linear growth conditions (7) and (9) on ∇F and on σ respectively, then yields the upper bound

$$\|\Delta X_{n+1}\| \leq Ch(1 + \|X_n - x^\star\|)(1 + \|\gamma_{n+1}\|).$$

As a result, one obtains the following upper bounds on conditional expectations: there exists a positive real number $C_p \in (0, \infty)$ such that

$$\begin{aligned}\mathbb{E}[\|X_n - x^*\|^{2(p-1)} \|\Delta X_{n+1}\|^2 | \mathcal{F}_n] &= \|X_n - x^*\|^{2(p-1)} \mathbb{E}[\|\Delta X_{n+1}\|^2 | \mathcal{F}_n] \\ &\leq C_p h^2 (1 + \|X_n - x^*\|^{2p}) \\ \mathbb{E}[\|\Delta X_{n+1}\|^{2p} | \mathcal{F}_n] &\leq C_p h^2 (1 + \|X_n - x^*\|^{2p}).\end{aligned}$$

Combining the results above, for all $n \in \mathbb{N}$ one obtains

$$\mathbb{E}[\|X_{n+1} - x^*\|^{2p} | \mathcal{F}_n] \leq (1 - 2p\mu h + C_p h^2) \|X_n - x^*\|^{2p} + C_p h^2,$$

For all $h \in (0, h_{\max})$ set

$$\rho_p(h) = 1 - 2p\mu h + C_p h^2.$$

The tower property of conditional expectation yields the identity

$$\mathbb{E}[\mathbb{E}[\|X_{n+1} - x^*\|^{2p} | \mathcal{F}_n]] = \mathbb{E}[\|X_{n+1} - x^*\|^{2p}],$$

thus taking expectation, for all $n \in \mathbb{N}$ one has

$$\mathbb{E}[\|X_{n+1} - x^*\|^{2p}] \leq \rho_p(h) \mathbb{E}[\|X_n - x^*\|^{2p}] + C_p h^2,$$

Set

$$H_p = \frac{1}{2} \min\left(\frac{1}{2p\mu}, \frac{2p\mu}{C_p}, h_{\max}\right),$$

then for all $h \in (0, H_p)$ one has $\rho_p(h) \in (0, 1)$. Applying the discrete Grönwall inequality, one obtains for all $n \geq 0$ and for all $h \in (0, H_p)$

$$\begin{aligned}\mathbb{E}[\|X_n - x^*\|^{2p}] &\leq \rho_p(h)^n \|x_0 - x^*\|^{2p} + \frac{C_p h^2}{1 - \rho_p(h)} \\ &\leq \|x_0 - x^*\|^{2p} + \frac{C_p h}{2p\mu - C_p h} \\ &\leq \|x_0 - x^*\|^{2p} + \frac{C_p H_p}{p\mu}.\end{aligned}$$

This concludes the proof of the inequality (39) in the case $p \geq 2$. The proof of Lemma 4.1 is thus completed. \square

4.2. Auxiliary process. Given any time-step size $h \in (0, h_{\max})$, let us introduce an auxiliary stochastic process $(\tilde{X}^h(t))_{t \geq 0}$ defined as follows: for any integer $n \geq 0$, and for all $t \in [t_n, t_{n+1}]$ one has

$$(43) \quad \tilde{X}^h(t) = X_n - \nabla F(X_n)(t - t_n) + \sqrt{h} \sigma(t_n, X_n)(B(t) - B(t_n)).$$

Note that $\tilde{X}^h(t_n) = X_n$ for all $n \in \mathbb{N}$, i.e. the auxiliary process interpolates the stochastic gradient optimization scheme (2) at the grid times $t_n = nh$. In addition, the stochastic process $(\tilde{X}^h(t))_{t \geq 0}$ has continuous trajectories. Finally, for any integer $n \in \mathbb{N}$, on the interval $[t_n, t_{n+1}]$ the process $t \mapsto \tilde{X}^h(t)$ is the solution of the stochastic differential equation

$$(44) \quad \begin{cases} d\tilde{X}^h(t) = -\nabla F(X_n)dt + \sqrt{h} \sigma(t_n, X_n)dB(t), & t \in [t_n, t_{n+1}], \\ \tilde{X}^h(t_n) = X_n. \end{cases}$$

The auxiliary process $(\tilde{X}^h(t))_{t \geq 0}$ plays a technical role in the proofs of the weak error estimates. The following properties, which follow from Lemma 4.1, are employed in the error analysis.

Lemma 4.2. *Assume that the objective F satisfies the basic properties of Assumptions 2.1 and that the diffusion coefficient σ satisfies Assumptions 2.2. For all $p \in \mathbb{N}^*$, there exists $H_p \in (0, h_{\max})$ and $C_p \in (0, \infty)$ such that one has the moment bounds*

$$(45) \quad \sup_{h \in (0, H_p)} \sup_{t \geq 0} \mathbb{E}[\|\tilde{X}^h(t) - x^*\|^{2p}] \leq C_p(1 + \|x_0 - x^*\|^{2p})$$

and such that one has the increment bounds: for all $h \in (0, H_p)$,

$$(46) \quad \sup_{n \in \mathbb{N}} \sup_{t \in [t_n, t_{n+1}]} \mathbb{E}[\|\tilde{X}^h(t) - X_n\|^{2p}] \leq C_p(1 + \|x_0 - x^*\|^{2p})h^{2p}.$$

Proof of Lemma 4.2. Let $p \in \mathbb{N}^*$ and let H_p be given by Lemma 4.1.

Proof of the moments estimates (45). For all $h \in (0, H_p)$, for any integer $n \geq 0$, owing to the definition (43) of the interpolating process $\tilde{X}^h(t)$, one has for all $t \in [t_n, t_{n+1}]$

$$\begin{aligned} (\mathbb{E}[\|\tilde{X}^h(t) - x^*\|^{2p}])^{\frac{1}{2p}} &\leq (\mathbb{E}[\|X_n - x^*\|^{2p}])^{\frac{1}{2p}} + (t - t_n)(\mathbb{E}[\|\nabla F(X_n)\|^{2p}])^{\frac{1}{2p}} \\ &\quad + \sqrt{h}(\mathbb{E}[\|\sigma(t_n, X_n)(B(t) - B(t_n))\|^{2p}])^{\frac{1}{2p}}. \end{aligned}$$

Then by the linear growth conditions (7) and (9),

$$(\mathbb{E}[\|\tilde{X}^h(t) - x^*\|^{2p}])^{\frac{1}{2p}} \leq (1 + L_F)h(\mathbb{E}[\|X_n - x^*\|^{2p}])^{\frac{1}{2p}} + C_p h \|\varsigma\|_\infty \left(1 + (\mathbb{E}[\|X_n - x^*\|^{2p}])^{\frac{1}{2p}}\right),$$

using the property

$$(\mathbb{E}[\|B(t) - B(t_n)\|^{2p}])^{\frac{1}{2p}} = C_p(t - t_n)^{\frac{1}{2}} \leq C_p h^{\frac{1}{2}}.$$

The moment bounds (39) from Lemma 4.1 then imply that for all $h \in (0, H_p)$ one has

$$\begin{aligned} \sup_{t \geq 0} (\mathbb{E}[\|\tilde{X}^h(t) - x^*\|^{2p}])^{\frac{1}{2p}} &= \sup_{n \geq 0} \sup_{t \in [t_n, t_{n+1}]} (\mathbb{E}[\|\tilde{X}^h(t) - x^*\|^{2p}])^{\frac{1}{2p}} \\ &\leq C_p \left(1 + \sup_{n \geq 0} (\mathbb{E}[\|X_n - x^*\|^{2p}])^{\frac{1}{2p}}\right) \\ &\leq C_p (1 + \|x_0 - x^*\|). \end{aligned}$$

This concludes the proof of the moment bounds (45).

Proof of the increments bounds (46). One proceeds by using similar arguments. Let $h \in (0, H_p)$ and $n \geq 0$. Then for all $t \in [t_n, t_{n+1}]$ owing to (43) one has

$$\begin{aligned} (\mathbb{E}[\|\tilde{X}^h(t) - X_n\|^{2p}])^{\frac{1}{2p}} &\leq (t - t_n)(\mathbb{E}[\|\nabla F(X_n)\|^{2p}])^{\frac{1}{2p}} + \sqrt{h}(\mathbb{E}[\|\sigma(t_n, X_n)(B(t) - B(t_n))\|^{2p}])^{\frac{1}{2p}} \\ &\leq L_F h (\mathbb{E}[\|X_n - x^*\|^{2p}])^{\frac{1}{2p}} + C_p h \|\varsigma\|_\infty \left(1 + (\mathbb{E}[\|X_n - x^*\|^{2p}])^{\frac{1}{2p}}\right) \\ &\leq C_p h \left(1 + \sup_{n \geq 0} (\mathbb{E}[\|X_n - x^*\|^{2p}])^{\frac{1}{2p}}\right) \\ &\leq C_p h (1 + \|x_0 - x^*\|). \end{aligned}$$

This concludes the proof of the increment bounds (46). The proof of Lemma 4.2 is thus completed. \square

5. PROOF OF THEOREM 1

This section is devoted to the detailed proof of Theorem 1. In the sequel, it is assumed that the objective function F satisfies the basic Assumption 2.1, and that the diffusion coefficient σ satisfies the basic Assumption 2.2.

Let us explain the strategy of the proof. Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a mapping of class \mathcal{C}^2 . The error is decomposed as

$$(47) \quad \mathbb{E}[\varphi(X_N)] - \varphi(x^*) = \varphi(\mathcal{X}^0(Nh)) - \varphi(x^*) + \mathbb{E}[\varphi(X_N)] - \varphi(\mathcal{X}^0(Nh)),$$

where $(\mathcal{X}^0(t))_{t \geq 0}$ is the solution of the deterministic gradient system (11). On the one hand, the treatment of the first error term $\varphi(\mathcal{X}^0(Nh)) - \varphi(x^*)$ follows from the exponential convergence property (12) of $\mathcal{X}^0(T) - x^*$ when $T \rightarrow \infty$, with $T = t_N = Nh$. On the other hand, the treatment of the second error term $\mathbb{E}[\varphi(X_N)] - \varphi(\mathcal{X}^0(Nh))$ requires more effort. Introduce the auxiliary function $u^0 : [0, \infty) \times \mathbb{R}^d \rightarrow \mathbb{R}$ given by

$$(48) \quad u^0(t, x) = \varphi(\mathcal{X}_x^0(t)),$$

where $(\mathcal{X}_x^0(t))_{t \geq 0}$ denotes the solution of (11) with arbitrary initial value $\mathcal{X}_x^0(0) = x \in \mathbb{R}^d$. As a result, the second error term is written as

$$(49) \quad \mathbb{E}[\varphi(X_N)] - \varphi(\mathcal{X}^0(Nh)) = \mathbb{E}[u^0(0, X_N)] - \mathbb{E}[u^0(t_N, X_0)].$$

Then a standard telescoping sum argument provides the decomposition of the error as

$$\mathbb{E}[u^0(0, X_N)] - \mathbb{E}[u^0(t_N, X_0)] = \sum_{n=0}^{N-1} (\mathbb{E}[u^0(t_N - t_{n+1}, X_{n+1})] - \mathbb{E}[u^0(t_N - t_n, X_n)]).$$

Recalling that $X_n = \tilde{X}^h(t_n)$ for all $n \in \{0, \dots, N\}$, where the auxiliary continuous time process $(\tilde{X}^h(t))_{t \geq 0}$ is defined by (43), one obtains

$$(50) \quad \mathbb{E}[u^0(0, X_N)] - \mathbb{E}[u^0(t_N, X_0)] = \sum_{n=0}^{N-1} \epsilon_n,$$

where for all $n \in \{0, \dots, N-1\}$ the error term ϵ_n is defined by

$$(51) \quad \epsilon_n = \mathbb{E}[u^0(t_N - t_{n+1}, \tilde{X}^h(t_{n+1}))] - \mathbb{E}[u^0(t_N - t_n, \tilde{X}^h(t_n))].$$

Then one proceeds by applying the Itô formula, using (44). Error bounds for ϵ_n are obtained in the proof of Proposition 5.1. To analyze the error terms, it is fundamental to observe that the function u^0 is solution of the linear transport partial differential equation

$$(52) \quad \begin{cases} \partial_t u^0(t, x) = -\langle \nabla u^0(t, x), \nabla F(x) \rangle, & \forall t \geq 0, x \in \mathbb{R}^d, \\ u^0(0, x) = \varphi(x), & \forall x \in \mathbb{R}^d. \end{cases}$$

The remainder of this section is organized as follows. Regularity properties of the function u^0 are stated in Section 5.1, with proofs postponed to Appendix A. The application of those results to obtain upper bounds on the second error term appearing on the right-hand side of (47), following the strategy outlined above, is then presented in Section 5.2. Finally, Section 5.3 combines the upper bounds on the error terms and provides the proof of Theorem 1.

5.1. Regularity properties of the auxiliary function u^0 .

Lemma 5.1. *Assume that the objective function F satisfies the basic properties from Assumptions 2.1. Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ be of class \mathcal{C}^2 with bounded second order derivative.*

For all $t \geq 0$ the mapping $u^0(t, \cdot)$ defined by (48) is of class \mathcal{C}^2 . Moreover, there exists a positive real number $C^0 \in (0, \infty)$ such that the first order derivative satisfies for all $t \geq 0$

$$(53) \quad \sup_{x \in \mathbb{R}^d} \sup_{k \in \mathbb{R}^d \setminus \{0\}} \frac{|Du^0(t, x) \cdot k|}{(1 + \|x - x^*\|)\|k\|} \leq C^0 (\|\nabla \varphi(x^*)\| + \|\varphi\|_2) e^{-\mu t},$$

and such that the second order derivative satisfies for all $t \geq 0$

$$(54) \quad \sup_{x \in \mathbb{R}^d} \sup_{k_1, k_2 \in \mathbb{R}^d \setminus \{0\}} \frac{|D^2 u^0(t, x) \cdot (k_1, k_2)|}{(1 + \|x - x^*\|) \|k_1\| \|k_2\|} \leq C^0 (\|\nabla \varphi(x^*)\| + \|\varphi\|_2) e^{-\mu t}.$$

The proof of Lemma 5.1 is given in Appendix A.

5.2. Proof of weak error estimates. We now focus on the main task of the proof of Theorem 1, which consists in proving error estimates for $\mathbb{E}[\varphi(X_N)] - \varphi(\mathcal{X}^0(Nh))$, uniformly with respect to N .

Proposition 5.1. *Assume that the objective function F satisfies the basic Assumption 2.1, and that the diffusion coefficient σ satisfies the basic Assumption 2.2.*

There exist positive real numbers $H \in (0, h_{\max})$ and $C \in (0, \infty)$ such that for any mapping $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ of class C^2 with bounded second order derivative, for any initial value $x_0 \in \mathbb{R}^d$, and for any $h \in (0, H)$, one has the weak error estimate

$$(55) \quad \sup_{N \in \mathbb{N}^*} |\mathbb{E}[\varphi(X_N)] - \varphi(\mathcal{X}^0(Nh))| \leq C_1(\varphi, x_0)h,$$

where the positive real number $C_1(\varphi, x_0)$ is given by (22).

In order to prove Proposition 5.1, we follow a standard approach employed to prove weak error estimates for numerical methods applied to stochastic differential equations, sketched at the beginning of Section 5 above. Having exponential decrease with respect to t in the bounds (53) and (54) of the first and second derivatives with respect to x of the solution of the partial differential equation (48) is fundamental in order to obtain a right-hand side which is independent of N . The value of H is imposed to apply some moment bounds on the solution of the stochastic optimization gradient scheme $(X_n)_{n \geq 0}$ given by Lemma 4.1 and on the solution $(\tilde{X}^h(t))_{t \geq 0}$ given by Lemma 4.3.

Proof of Proposition 5.1. Let the function $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ and the initial value $x_0 \in \mathbb{R}^d$ be given. Combining the expressions (49) and (50) for the weak error $\mathbb{E}[\varphi(X_N)] - \varphi(\mathcal{X}^0(Nh))$, it is sufficient to prove the following claim for the error terms ϵ_n defined by (51): for all $N \in \mathbb{N}^*$ and $h \in (0, H)$, for all $n \in \{0, \dots, N-1\}$, one has

$$(56) \quad |\epsilon_n| \leq C_1(\varphi, x_0)h \int_{t_n}^{t_{n+1}} e^{-\mu(t_N-t)} dt,$$

with $C_1(\varphi, x_0)$ defined by (22) (up to a modification of the multiplicative constant $C \in (0, \infty)$).

Let us prove the claim (56) above. For all $N \in \mathbb{N}^*$ and $n \in \{0, \dots, N-1\}$, recall that $t \in [t_n, t_{n+1}] \mapsto \tilde{X}^h(t)$ is solution of the stochastic differential equation (44). Therefore, applying Itô's formula on the interval $[t_n, t_{n+1}]$ yields the following decomposition of the error term ϵ_n : one has

$$\begin{aligned} \epsilon_n &= \int_{t_n}^{t_{n+1}} \mathbb{E}[-\partial_t u^0(t_N - t, \tilde{X}^h(t))] dt - \int_{t_n}^{t_{n+1}} \mathbb{E}[\langle \nabla u^0(t_N - t, \tilde{X}^h(t)), \nabla F(X_n) \rangle] dt \\ &\quad + \frac{h}{2} \int_{t_n}^{t_{n+1}} \mathbb{E}[\nabla^2 u^0(t_N - t, \tilde{X}^h(t)) : a(t_n, \tilde{X}^h(t))] dt. \end{aligned}$$

Recall that the function u^0 is the solution of the partial differential equation (52). This allows us to decompose ϵ_n as

$$\epsilon_n = \epsilon_{n,1} + \epsilon_{n,2},$$

where for all $n \in \{0, \dots, N-1\}$ one defines

$$\epsilon_{n,1} = \int_{t_n}^{t_{n+1}} \mathbb{E}[\langle \nabla u^0(t_N - t, \tilde{X}^h(t)), \nabla F(\tilde{X}^h(t)) - \nabla F(X_n) \rangle] dt$$

$$\epsilon_{n,2} = \frac{h}{2} \int_{t_n}^{t_{n+1}} \mathbb{E}[\nabla^2 u^0(t_N - t, \tilde{X}^h(t)) : a(t_n, \tilde{X}^h(t_n))] dt.$$

The claim (56) follows from the bounds for $\epsilon_{n,1}$ and $\epsilon_{n,2}$ which are obtained below.

Upper bounds for $\epsilon_{n,1}$. Recall that $\nabla F : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a Lipschitz continuous mapping (see Equation (4) from Assumption 2.1). Therefore, using the inequality (53) from Lemma 5.1 on the first order derivative of $u(t_N - t, \cdot)$, one has

$$|\epsilon_{n,1}| \leq C(\|\nabla\varphi(x^*)\| + \|\varphi\|_2) \int_{t_n}^{t_{n+1}} e^{-\mu(t_N-t)} \mathbb{E}[(1 + \|\tilde{X}^h(t) - x^*\|) \|\tilde{X}^h(t) - X_n\|] dt.$$

Assume that $H \leq H_1$ where $H_1 \in (0, h_{\max})$ is given by Lemma 4.2. Applying the Cauchy–Schwarz inequality and using the moment bounds (45) and the increment bounds (46) from Lemma 4.2 for the auxiliary process $(\tilde{X}^h(t))_{t \geq 0}$, one obtains the upper bounds

$$\begin{aligned} |\epsilon_{n,1}| &\leq C(\|\nabla\varphi(x^*)\| + \|\varphi\|_2) h (1 + \|x_0 - x^*\|^2) \int_{t_n}^{t_{n+1}} e^{-\mu(t_N-t)} dt \\ &\leq C_1(\varphi, x_0) h \int_{t_n}^{t_{n+1}} e^{-\mu(t_N-t)} dt, \end{aligned}$$

for all $N \in \mathbb{N}^*$, all $n \in \{0, \dots, N-1\}$ and all $h \in (0, H)$.

Upper bounds for $\epsilon_{n,2}$.

Recall that the mapping $(t, x) \mapsto \sigma(t, x)$ has at most linear growth with respect to $x \in \mathbb{R}^d$, uniformly with respect to $t \geq 0$ (see Equation (9) from Assumption 2.2). Recall also that a is defined by (10). Therefore, using the inequality (54) from Lemma 5.1 on the second order derivative of $u(t_N - t, \cdot)$, one has

$$|\epsilon_{n,2}| \leq C(\|\nabla\varphi(x^*)\| + \|\varphi\|_2) h \varsigma(t_n)^2 \int_{t_n}^{t_{n+1}} e^{-\mu(t_N-t)} \mathbb{E}[(1 + \|\tilde{X}^h(t) - x^*\|)(1 + \|\tilde{X}^h(t_n) - x^*\|)^2] dt.$$

Recall that $\varsigma(t_n) \leq \|\varsigma\|_\infty$ for all $n \in \mathbb{N}$. Assume that $H \leq H_2$ where $H_2 \in (0, h_{\max})$ is given by Lemma 4.2. Applying the moment bounds (45) from Lemma 4.2, one obtains the upper bounds

$$\begin{aligned} |\epsilon_{n,2}| &\leq C(\|\nabla\varphi(x^*)\| + \|\varphi\|_2) h (1 + \|x_0 - x^*\|^3) \int_{t_n}^{t_{n+1}} e^{-\mu(t_N-t)} dt \\ &\leq C_1(\varphi, x_0) h \int_{t_n}^{t_{n+1}} e^{-\mu(t_N-t)} dt, \end{aligned}$$

for all $N \in \mathbb{N}^*$, all $n \in \{0, \dots, N-1\}$ and all $h \in (0, H)$.

Conclusion. Gathering the upper bounds on $|\epsilon_{n,1}|$ and $|\epsilon_{n,2}|$ obtained above then yields the claim (56), under the condition $H \leq H_2$.

The proof of the weak error estimate (55) is then straightforward: for all $N \in \mathbb{N}^*$ one has

$$\begin{aligned} |\mathbb{E}[\varphi(X_N)] - \varphi(\mathcal{X}^0(Nh))| &\leq \sum_{n=0}^{N-1} |\epsilon_n| \\ &\leq C_1(\varphi, x_0) h \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} e^{-\mu(t_N-t)} dt \\ &\leq C_1(\varphi, x_0) h \int_0^{t_N} e^{-\mu(t_N-t)} dt \leq \frac{C_1(\varphi, x_0) h}{\mu}. \end{aligned}$$

This concludes the proof of Proposition 5.1. □

5.3. Proof of Theorem 1. We are now in a position to provide the proof of Theorem 1, by a straightforward combination of the inequality (12) and of the weak error estimates (55) from Proposition 5.1.

Proof of Theorem 1. Recall that owing to (47) the error term appearing in the left-hand side of (21) can be decomposed as

$$\mathbb{E}[\varphi(X_N)] - \varphi(x^*) = \varphi(\mathcal{X}^0(Nh)) - \varphi(x^*) + \mathbb{E}[\varphi(X_N)] - \varphi(\mathcal{X}^0(Nh)).$$

On the one hand, using the inequality (1) and owing to (12) one has

$$\begin{aligned} |\varphi(\mathcal{X}^0(Nh)) - \varphi(x^*)| &\leq C(\|\nabla\varphi(x^*)\| + \|\varphi\|_2\|\mathcal{X}^0(Nh) - x^*\|)\|\mathcal{X}^0(Nh) - x^*\| \\ &\leq C_1(\varphi, x_0)e^{-\mu Nh}. \end{aligned}$$

On the other hand, applying the result (55) from Proposition 5.1 one has for all $N \in \mathbb{N}$ and all $h \in (0, H)$

$$|\mathbb{E}[\varphi(X_N)] - \varphi(\mathcal{X}^0(Nh))| \leq C_1(\varphi, x_0)h.$$

Gathering the two upper bounds yields the error bound (21) and concludes the proof of Theorem 1. \square

6. PROOF OF THEOREM 2

This section is devoted to the detailed proof of Theorem 2. In the sequel, it is assumed that the objective function F satisfies the strengthened Assumption 2.3, and that the diffusion coefficient σ satisfies the strengthened Assumption 2.4. Recall that this implies that the basic Assumptions 2.1 and 2.2 are also satisfied.

Let us explain the strategy of the proof, and compare it with the proof of Theorem 1 given in Section 5. Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a mapping of class \mathcal{C}^3 . The error is decomposed as

$$(57) \quad \mathbb{E}[\varphi(X_N)] - \varphi(x^*) = \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))] - \varphi(x^*) + \mathbb{E}[\varphi(X_N)] - \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))],$$

where $(\mathcal{Y}^h(t))_{t \geq 0}$ is the solution of the modified stochastic differential equation (14). On the one hand, the treatment of the first error term $\mathbb{E}[\varphi(\mathcal{Y}^h(Nh))] - \varphi(x^*)$ follows from an upper bound on $\mathbb{E}[\|\mathcal{Y}^h(t) - x^*\|^2]$ which is stated in Proposition 6.1 below. On the other hand, the approach used to treat the second error term $\mathbb{E}[\varphi(X_N)] - \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))]$ follows arguments similar to those used in Section 5, where an auxiliary function u^0 has been introduced. The situation is more complex here: since the diffusion coefficient σ depends on time in general, considering the stochastic differential equation (14) on the time interval $[0, \infty)$, with initial value imposed at time $t = 0$, is not sufficient. Instead one needs to consider the stochastic differential equation

$$(58) \quad \begin{cases} d\mathcal{Y}_y^h(s|t) = -\nabla F^h(\mathcal{Y}_y^h(s|t))ds + \sqrt{h}\sigma(s, \mathcal{Y}_y^h(s|t))dB(s), & \forall s \geq t, \\ \mathcal{Y}_y^h(t|t) = y, \end{cases}$$

on time intervals $[t, \infty)$, with initial value imposed at time $s = t$, for arbitrary $t \in [0, \infty)$. The time variable in (58) and below is denoted by s . The stochastic differential equation (14) is retrieved by choosing $t = 0$. Recall that the modified objective function F^h is assumed to be μ -convex for all $h \in (0, h_0)$ (see Assumption 2.3), and that the diffusion coefficient $\sigma(s, \cdot)$ is globally Lipschitz continuous, uniformly with respect to $s \in [0, \infty)$ (see Assumption 2.2). As a result, the stochastic differential equation (58) is globally well-posed: for any time $t \in [0, \infty)$ and any initial condition $y \in \mathbb{R}^d$, there exists a unique solution of (58) denoted by $(\mathcal{Y}_y^h(s|t))_{s \geq t}$. Note that choosing $t = 0$ and $y = x_0$ in (58), one obtains $\mathcal{Y}^h(T) = \mathcal{Y}_{x_0}^h(T|0)$ for all $T \in [0, \infty)$. Moment bounds for the solutions $(\mathcal{Y}_y^h(s|t))_{s \geq t}$ of (58) are given in Lemma 6.1 below.

Introducing the auxiliary stochastic differential equations (58) indexed by the initial time $t \in [0, \infty)$ allows us to define, for any time $T \in [0, \infty)$, an auxiliary mapping $v_T^h : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$, as follows: for all $t \in [0, T]$ and all $y \in \mathbb{R}^d$, set

$$(59) \quad v_T^h(t, y) = \mathbb{E}[\varphi(\mathcal{Y}_y^h(T|t))].$$

The mapping v_T^h is the solution of the backward Kolmogorov equation

$$(60) \quad \begin{cases} \partial_t v_T^h(t, y) - \langle \nabla v_T^h(t, y), \nabla F^h(y) \rangle + \frac{h}{2} \nabla^2 v_T^h(t, y) : a(t, y) = 0, & \forall t \in [0, T], y \in \mathbb{R}^d, \\ v_T^h(T, y) = \varphi(y), & \forall y \in \mathbb{R}^d. \end{cases}$$

The following remark justifies why it is necessary to introduce (58) instead of considering only (14).

Remark 6.1. *Let us emphasize that the mapping $u^h : [0, \infty) \times \mathbb{R}^d \rightarrow \mathbb{R}$ defined by*

$$u^h(t, y) = \mathbb{E}[\varphi(\mathcal{Y}_y^h(t))]$$

for all $t \in [0, \infty)$ and $y \in \mathbb{R}^d$, where $(\mathcal{Y}_y^h(t))_{t \geq 0}$ denotes the solution of (14) with arbitrary initial value $\mathcal{Y}_y^h(0) = y \in \mathbb{R}^d$, in general does not solve the partial differential equation

$$\begin{cases} \partial_t u^h(t, y) = -\langle \nabla u^h(t, y), \nabla F^h(y) \rangle + \frac{h}{2} \nabla^2 u^h(t, y) : a(t, y), & \forall t \geq 0, y \in \mathbb{R}^d, \\ u^h(0, y) = \varphi(y), & \forall y \in \mathbb{R}^d. \end{cases}$$

This would be the case only if the diffusion coefficient σ and the auxiliary mapping a would be independent of time t .

The auxiliary functions v_T^h play a role in the analysis of the second error term $\mathbb{E}[\varphi(X_N)] - \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))]$ which is similar to the role played by the auxiliary mapping u^0 in Section 5. Assume that $T = Nh$, then owing to the definition (59) of v^h the second error term is written as

$$(61) \quad \mathbb{E}[\varphi(X_N)] - \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))] = \mathbb{E}[v_T^h(T, X_N)] - \mathbb{E}[v_T^h(0, X_0)].$$

Then a standard telescoping sum argument provides the decomposition of the error as

$$\mathbb{E}[v_T^h(T, X_N)] - \mathbb{E}[v_T^h(0, X_0)] = \sum_{n=0}^{N-1} (\mathbb{E}[v_T^h(t_{n+1}, X_{n+1})] - \mathbb{E}[v_T^h(t_n, X_n)]).$$

As in Section 5, recalling that $X_n = \tilde{X}^h(t_n)$ for all $n \in \{0, \dots, N\}$, where the auxiliary continuous time process $(\tilde{X}^h(t))_{t \geq 0}$ is defined by (43), one obtains

$$(62) \quad \mathbb{E}[v_T^h(T, X_N)] - \mathbb{E}[v_T^h(0, X_0)] = \sum_{n=0}^{N-1} \delta_n$$

where for all $n \in \{0, \dots, N-1\}$ the error term δ_n is defined by

$$(63) \quad \delta_n = \mathbb{E}[v_T^h(t_{n+1}, \tilde{X}^h(t_{n+1}))] - \mathbb{E}[v_T^h(t_n, \tilde{X}^h(t_n))].$$

Then one proceeds by applying the Itô formula, using (44). Compared with Section 5, one needs to bound the error terms δ_n as $O(h^3)$ instead of $O(h^2)$ in order to obtain second order convergence with respect to h . This requires different arguments, however similar techniques are used to obtain error bounds which are uniform with respect to time.

The remainder of this section is organized as follows. Moment bounds for the solutions $(\mathcal{Y}_y^h(s|t))_{s \geq t}$ are first studied in Section 6.1. The large time behavior of the solution $\mathcal{Y}^h(T) = \mathcal{Y}_{x_0}^h(T|0)$ when $T \rightarrow \infty$ is then studied in Section 6.2, this is the key ingredient to treat the first error term appearing in the right-hand side of (57). Regularity properties of the functions v_T^h are stated in Section 6.3,

with proofs postponed to Appendix B. The application of those results to obtain upper bounds on the second error term appearing on the right-hand side of (57), following the strategy outlined above, is then presented in Section 6.4. Finally, Section 6.5 combines the upper bounds on the error terms and provides the proof of Theorem 2.

6.1. Moment bounds for the solutions of (58). Let us first provide moment bounds for the solutions $(\mathcal{Y}_y^h(s|t))_{s \geq t}$ of the stochastic differential equations (58) indexed by time $t \in [0, \infty)$ and the time-step size $h \in (0, h_0)$. It is worth mentioning that the moment bounds are uniform with respect to $s \geq t \geq 0$ and with respect to $h \in (0, h_0)$.

Lemma 6.1. *Assume that the objective F satisfies the strengthened Assumption 2.3 and that the diffusion coefficient σ satisfies the basic Assumption 2.2. There exists a non-increasing sequence $(h_p^{(0)})_{p \in \mathbb{N}^*}$, with $h_1^{(0)} \in (0, h_0)$, such that for any $p \in \mathbb{N}^*$, one has*

$$(64) \quad \sup_{h \in (0, h_p^{(0)})} \sup_{s \geq t} \sup_{y \in \mathbb{R}^d} \frac{\mathbb{E}[\|\mathcal{Y}_y^h(s|t) - x^*\|^{2p}]}{1 + \|y - x^*\|^{2p}} < \infty.$$

Proof. Assume that $h \in (0, h_0)$ and let $p \in \mathbb{N}^*$. Let the initial time $t \in [0, \infty)$ and the initial value $y \in \mathbb{R}^d$ be given. Recall that the diffusion coefficient $\sigma(s, \cdot)$ has at most linear growth, uniformly with respect to $s \in [0, \infty)$, see the condition (9) from Assumption 2.2. Using the auxiliary results (41) and (42) (stated in the proof of Lemma 4.1) and applying Itô's formula, there exists a positive real number $C_p^{(0)} \in (0, \infty)$ such that for all $s \geq t$ one has

$$\begin{aligned} \frac{1}{2p} \frac{d\mathbb{E}[\|\mathcal{Y}_y^h(s|t) - x^*\|^{2p}]}{ds} &\leq -\mathbb{E}[\langle \nabla F^h(\mathcal{Y}_y^h(s|t)), \mathcal{Y}_y^h(s|t) - x^* \rangle \|\mathcal{Y}_y^h(s|t) - x^*\|^{2(p-1)}] \\ &\quad + C_p^{(0)} h \mathbb{E}[(1 + \|\mathcal{Y}_y^h(s|t) - x^*\|^2) \|\mathcal{Y}_y^h(s|t) - x^*\|^{2(p-1)}]. \end{aligned}$$

The modified objective function F^h is μ -convex for all $h \in (0, h_0)$, see the condition (18) from Assumption 2.3. Moreover, one has $\nabla F^h(x^*) = 0$ for all $h \in (0, h_0)$. As a consequence, one obtains, for all $s \geq t$,

$$\langle \nabla F^h(\mathcal{Y}_y^h(s|t)), \mathcal{Y}_y^h(s|t) - x^* \rangle \leq -\mu \|\mathcal{Y}_y^h(s|t) - x^*\|^2.$$

Applying Young's inequality, there exists a positive real number $C_p \in (0, \infty)$ such that one has, for all $s \geq t$,

$$\frac{1}{2p} \frac{d\mathbb{E}[\|\mathcal{Y}_y^h(s|t) - x^*\|^{2p}]}{ds} \leq -(\mu - C_p^{(0)} h) \mathbb{E}[\|\mathcal{Y}_y^h(s|t) - x^*\|^{2p}] + C_p h.$$

Set $h_p^{(0)} = \frac{1}{2} \min\left(\frac{\mu}{C_p^{(0)}}, h_0\right)$. For all $h \in (0, h_p^{(0)})$, one has $\mu - C_p^{(0)} h \geq \mu - C_p^{(0)} h_p^{(0)} > 0$. Applying Grönwall's inequality and recalling that $\mathcal{Y}_y^h(t|t) = y$, one obtains the following upper bound: for all $s \geq t$ and all $h \in (0, h_p^{(0)})$ one has

$$\mathbb{E}[\|\mathcal{Y}_y^h(s|t) - x^*\|^{2p}] \leq e^{-(\mu - C_p^{(0)} h_p^{(0)})(s-t)} \|y - x^*\|^{2p} + \frac{C_p h}{\mu - C_p^{(0)} h_p^{(0)}}.$$

Note that $h \in (0, h_p^{(0)})$ and that $e^{-(\mu - C_p^{(0)} h_p^{(0)})(s-t)} \leq 1$. Then one obtains the moment bounds (64), and this concludes the proof of Lemma 6.1. \square

6.2. Large time behavior of the solution of (14).

Proposition 6.1. *Assume that the objective function F satisfies the strengthened Assumption 2.3 and that the diffusion coefficient σ satisfies the basic Assumption 2.2.*

There exists a positive real number $C \in (0, \infty)$ such that for any initial value $x_0 \in \mathbb{R}^d$, any $h \in (0, h_1^{(0)})$ and any $T \in (0, \infty)$, one has

$$(65) \quad \mathbb{E}[\|\mathcal{Y}^h(T) - x^*\|^2] \leq C e^{-2\mu T} (\|x_0 - x^*\|^2 + h\rho(T)(1 + \|x_0 - x^*\|^2)),$$

where we recall that $(\mathcal{Y}^h(t))_{t \geq 0}$ denotes the solution of the stochastic differential equation (14), with initial value $\mathcal{Y}^h(0) = x_0$, and that $\rho(T)$ is defined by (30).

Proof. Applying Itô's formula to the stochastic differential equation (14), one obtains for all $t \geq 0$

$$\begin{aligned} \frac{1}{2} \frac{d\mathbb{E}[\|\mathcal{Y}^h(t) - x^*\|^2]}{dt} &= -\mathbb{E}[\langle \nabla F^h(\mathcal{Y}^h(t)), \mathcal{Y}^h(t) - x^* \rangle] + \frac{h}{2} \mathbb{E}[\|\sigma(t, \mathcal{Y}^h(t))\|^2] \\ &\leq -\mu \mathbb{E}[\|\mathcal{Y}^h(t) - x^*\|^2] + \frac{h}{2} \zeta(t)^2 (1 + \mathbb{E}[\|\mathcal{Y}^h(t) - x^*\|^2]), \end{aligned}$$

using the μ -convexity property (18) of the modified objective function F^h , and the at most linear growth property (9) of the diffusion coefficient $\sigma(t, \cdot)$, for all $t \geq 0$.

Applying the moment bounds stated in Lemma 6.1, there exists a positive real number $C \in (0, \infty)$ such that, for all $h \in (0, h_1^{(0)})$ and all $t \geq 0$ one obtains the inequality

$$\frac{1}{2} \frac{d\mathbb{E}[\|\mathcal{Y}^h(t) - x^*\|^2]}{dt} \leq -\mu \mathbb{E}[\|\mathcal{Y}^h(t) - x^*\|^2] + Ch\zeta(t)^2 (1 + \|x_0 - x^*\|^2).$$

Applying Grönwall's inequality, as a result one has for all $h \in (0, h_1^{(0)})$ and all $t \geq 0$

$$\mathbb{E}[\|\mathcal{Y}^h(t) - x^*\|^2] \leq e^{-2\mu t} \|x_0 - x^*\|^2 + Ch \int_0^t e^{-2\mu(t-r)} \zeta(r)^2 dr (1 + \|x_0 - x^*\|^2).$$

This yields the inequality (65) and concludes the proof of Proposition (6.1). \square

6.3. Regularity properties of the auxiliary functions v_T^h .

Lemma 6.2. *Assume that the objective function F satisfies the strengthened Assumption 2.3, and that the diffusion coefficient σ satisfies the strengthened Assumption 2.4.*

Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a mapping of class \mathcal{C}^3 , with bounded second and third order derivatives. Given any $T \in (0, \infty)$, let the mapping $v_T^h : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ be defined by (59).

For all $t \in [0, T]$ the mapping $v_T^h(t, \cdot)$ is of class \mathcal{C}^3 . Moreover, the first, second and third order derivatives of $v_T^h(t, \cdot)$ satisfy the following upper bounds.

- *There exist positive real numbers $h_1 \in (0, h_0)$ and $C_1 \in (0, \infty)$, satisfying the condition $C_1 h_1 < \mu$, such that the first order derivative satisfies: for all $h \in (0, h_1)$ and all $T \geq t \geq 0$, one has*

$$(66) \quad \sup_{y \in \mathbb{R}^d} \sup_{k \in \mathbb{R}^d \setminus \{0\}} \frac{|Dv_T^h(t, y) \cdot k|}{(1 + \|y - x^*\|)\|k\|} \leq C_1 (\|\nabla \varphi(x^*)\| + \|\varphi\|_2) e^{-(\mu - C_1 h)(T-t)}.$$

- *For any $\lambda \in (0, \mu)$, there exist positive real numbers $h_2 \in (0, h_1)$ and $C_{2,\lambda} \in (C_1, \infty)$, satisfying the condition $C_{2,\lambda} h_2 < \lambda$, such that the second order derivative satisfies: for all $h \in (0, h_2)$ and all $T \geq t \geq 0$, one has*

$$(67) \quad \sup_{y \in \mathbb{R}^d} \sup_{k_1, k_2 \in \mathbb{R}^d \setminus \{0\}} \frac{|D^2 v_T^h(t, y) \cdot (k_1, k_2)|}{(1 + \|y - x^*\|^2)\|k_1\|\|k_2\|} \leq C_{2,\lambda} (\|\nabla \varphi(x^*)\| + \|\varphi\|_2) e^{-(\lambda - C_{2,\lambda} h)(T-t)}.$$

- For any $\lambda \in (0, \mu)$, there exist positive real numbers $h_3 \in (0, h_2)$ and $C_{3,\lambda} \in (C_{2,\lambda}, \infty)$, satisfying the condition $C_{3,\lambda}h_3 < \lambda$, such that the second order derivative satisfies: for all $h \in (0, h_3)$ and all $T \geq t \geq 0$, one has

$$(68) \quad \sup_{y \in \mathbb{R}^d} \sup_{k_1, k_2, k_3 \in \mathbb{R}^d \setminus \{0\}} \frac{|D^3 v_T^h(t, y) \cdot (k_1, k_2, k_3)|}{(1 + \|y - x^*\|^2) \|k_1\| \|k_2\| \|k_3\|} \leq C_{3,\lambda} (\|\nabla \varphi(x^*)\| + \|\varphi\|_2 + \|\varphi\|_3) e^{-(\lambda - C_{3,\lambda}h)(T-t)}.$$

The proof of Lemma 6.2 is postponed to Appendix B.

A straightforward consequence of Lemma 6.2 is Lemma 6.3 stated below, giving bounds on the first order derivative of $\partial_t v_T^h(t, \cdot)$. The proof follows from using the backward Kolmogorov equation (60) and the upper bounds from Lemma 6.2.

Lemma 6.3. *Assume that the objective function F satisfies the strengthened Assumption 2.3, and that the diffusion coefficient σ satisfies the strengthened Assumption 2.4.*

Let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a mapping of class \mathcal{C}^3 , with bounded second and third order derivatives. Given any $T \in (0, \infty)$, let the mapping $v_T^h : [0, T] \times \mathbb{R}^d \rightarrow \mathbb{R}$ be defined by (59).

For all $t \in [0, T]$, the mapping $\partial_t v_T^h(t, \cdot)$ is of class \mathcal{C}^1 . Moreover, for any $\lambda \in (0, \mu)$, there exists positive real numbers $h_4 \in (0, h_0)$ and $C_{4,\lambda} \in (0, \infty)$, satisfying the condition $C_{4,\lambda}h_4 < \lambda$, such that for all $h \in (0, h_4)$ and $T \geq t \geq 0$ one has

$$(69) \quad \sup_{y \in \mathbb{R}^d} \sup_{k \in \mathbb{R}^d \setminus \{0\}} \frac{|D(\partial_t v_T^h)(t, y) \cdot k|}{(1 + \|y - x^*\|^4) \|k\|} \leq C_{4,\lambda} (\|\nabla \varphi(x^*)\| + \|\varphi\|_2 + \|\varphi\|_3) e^{-(\lambda - C_{4,\lambda}h)(T-t)}.$$

The proof of Lemma 6.3 is postponed to Appendix B.

Observe that the rate $\mu - C_1 h$ appearing in the exponential factors in the right-hand sides of (66) can be bounded from below uniformly with respect to $h \in (0, h_1)$: indeed one has $\mu - C_1 h \geq \mu - C_1 h_1 > 0$. Similarly, the rates $\lambda - C_{j,\lambda} h$ for $j \in \{2, 3, 4\}$, (67), (68) and (69) can be bounded from below uniformly with respect to $h \in (0, h_j)$: indeed one has $\lambda - C_{j,\lambda} h \geq \lambda - C_{j,\lambda} h_j > 0$. This means that the upper bounds in Lemma 6.2 and Lemma 6.3 can be interpreted as being uniform with respect to h .

6.4. Proof of weak error estimates. We now focus on the second main task of the proof of Theorem 2, which consists in proving error estimates for $\mathbb{E}[\varphi(X_N)] - \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))]$, uniformly with respect to N .

Proposition 6.2. *Assume that the objective function F satisfies the strengthened Assumption 2.3, and that the diffusion coefficient σ satisfies the strengthened Assumption 2.4.*

There exist positive real numbers $H \in (0, h_0)$ and $C \in (0, \infty)$ such that for any mapping $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ of class \mathcal{C}^3 with bounded second and third order derivatives, for any initial value $x_0 \in \mathbb{R}^d$, and for any $h \in (0, H)$, one has the weak error estimate

$$(70) \quad \sup_{N \in \mathbb{N}^*} |\mathbb{E}[\varphi(X_N)] - \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))]| \leq C_2(\varphi, x_0) h^2,$$

where the positive real number $C_2(\varphi, x_0)$ is given by (32).

In order to prove Proposition 6.2, we follow the same approach as in the proof of Proposition 5.1 in Section 5.2. This approach is sketched at the beginning of Section 6 above, see also Section 5.2 for some comments. Note that in (70) the order of convergence with respect to h is equal to 2, instead of 1 in (55). Obtaining an higher order of convergence requires additional arguments in the analysis. In particular, in the proof one considers the solution v_T^h of the backward Kolmogorov equation (60), associated with the modified stochastic differential equation (14), instead of dealing with the solution u^0 of the partial differential equation (52). It is worth mentioning that the proof

below exploits bounds on derivatives of v_T^h which are uniform with respect to h , given by Lemma 6.2 and Lemma 6.3.

Proof of Proposition 6.2. Let the function $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ and the initial value x_0 be given. Combining the expressions (61) and (62) for the weak error $\mathbb{E}[\varphi(X_N)] - \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))]$, it is sufficient to prove the following claim for the error terms δ_n defined by (63): for all $N \in \mathbb{N}^*$ and $h \in (0, H)$, for all $n \in \{0, \dots, N-1\}$, one has

$$(71) \quad |\delta_n| \leq C_2(\varphi, x_0) h^2 \int_{t_n}^{t_{n+1}} e^{-\frac{\mu(t_N-t)}{4}} dt,$$

with $C_2(\varphi, x_0)$ defined by (32) (up to a modification of the multiplicative constant $C \in (0, \infty)$).

Let us prove the claim (71) above. For all $N \in \mathbb{N}^*$ and $n \in \{0, \dots, N-1\}$, recall that $t \in [t_n, t_{n+1}] \mapsto \tilde{X}^h(t)$ is solution of the stochastic differential equation (44). Therefore, applying Itô's formula on the interval $[t_n, t_{n+1}]$ yields the following decomposition of the error term δ_n : one has

$$\begin{aligned} \delta_n &= \int_{t_n}^{t_{n+1}} \mathbb{E}[\partial_t v_T^h(t, \tilde{X}^h(t))] dt \\ &\quad - \int_{t_n}^{t_{n+1}} \mathbb{E}[\langle \nabla v_T^h(t, \tilde{X}^h(t)), \nabla F(\tilde{X}^h(t_n)) \rangle] dt \\ &\quad + \frac{h}{2} \int_{t_n}^{t_{n+1}} \mathbb{E}[\nabla^2 v_T^h(t, \tilde{X}^h(t)) : a(t_n, \tilde{X}^h(t_n))] dt. \end{aligned}$$

Recall that the function v_T^h is the solution of the backward Kolmogorov equation (59). This allows us to decomposed δ_n as

$$(72) \quad \delta_n = \delta_{n,1} + \delta_{n,2},$$

where for all $n \in \{0, \dots, N-1\}$ one defines

$$\begin{aligned} \delta_{n,1} &= \int_{t_n}^{t_{n+1}} \mathbb{E}[\langle \nabla v_T^h(t, \tilde{X}^h(t)), \nabla F^h(\tilde{X}^h(t)) - \nabla F(\tilde{X}^h(t_n)) \rangle] dt, \\ \delta_{n,2} &= \frac{h}{2} \int_{t_n}^{t_{n+1}} \mathbb{E}[\nabla^2 v_T^h(t, \tilde{X}^h(t)) : (a(t_n, \tilde{X}^h(t_n)) - a(t, \tilde{X}^h(t)))] dt. \end{aligned}$$

The analysis of the error term $\delta_{n,1}$ is more subtle than the analysis of the error term $\epsilon_{n,1}$ in the proof of Proposition 5.1, since for $\delta_{n,1}$ one needs to obtain second order convergence with respect to h . Let us decompose the error term $\delta_{n,1}$ as

$$(73) \quad \delta_{n,1} = \delta_{n,1,1} + \delta_{n,1,2},$$

where the auxiliary error terms $\delta_{n,1,1}$ and $\delta_{n,1,2}$ are defined for all $n \in \{0, \dots, N-1\}$ as

$$\begin{aligned} \delta_{n,1,1} &= \int_{t_n}^{t_{n+1}} \mathbb{E}[\langle \nabla v_T^h(t, \tilde{X}^h(t)) - \nabla v_T^h(t_n, \tilde{X}^h(t_n)), \nabla F^h(\tilde{X}^h(t)) - \nabla F(\tilde{X}^h(t_n)) \rangle] dt, \\ \delta_{n,1,2} &= \int_{t_n}^{t_{n+1}} \mathbb{E}[\langle \nabla v_T^h(t_n, \tilde{X}^h(t_n)), \nabla F^h(\tilde{X}^h(t)) - \nabla F(\tilde{X}^h(t_n)) \rangle] dt. \end{aligned}$$

The claim (71) follows from the bounds for $\delta_{n,1,1}$, $\delta_{n,1,2}$ and $\delta_{n,2}$ which are obtained below.

Upper bounds for $\delta_{n,1,1}$.

On the one hand, one has

$$\begin{aligned} \nabla v_T^h(t, \tilde{X}^h(t)) - \nabla v_T^h(t_n, \tilde{X}^h(t_n)) &= \nabla v_T^h(t, \tilde{X}^h(t)) - \nabla v_T^h(t_n, \tilde{X}^h(t)) \\ &\quad + \nabla v_T^h(t_n, \tilde{X}^h(t)) - \nabla v_T^h(t_n, \tilde{X}^h(t_n)). \end{aligned}$$

On the other hand, by the definition (15) of the modified objective function F^h , one has

$$\nabla F^h(\tilde{X}^h(t)) - \nabla F(\tilde{X}^h(t_n)) = \nabla F(\tilde{X}^h(t)) - \nabla F(\tilde{X}^h(t_n)) + \frac{h}{4} \nabla(\|\nabla F\|^2)(\tilde{X}^h(t)),$$

and recall that ∇F is globally Lipschitz continuous (see the condition (4) from Assumption 2.1) and that $\nabla(\|\nabla F\|^2)$ satisfies the inequality (16).

Using the inequality (69) from Lemma 6.3 on the first order derivative of $\partial_t v_T^h(t, \cdot)$ and the inequality (67) from Lemma 6.2 on the second order derivative of $v_T^h(t, \cdot)$ with $\lambda = \frac{\mu}{2}$, for all $h \in (0, h_4)$ one has

$$\begin{aligned} |\delta_{n,1,1}| &\leq C(\varphi) \int_{t_n}^{t_{n+1}} e^{-(\frac{\mu}{2}-Ch)(t_N-t)}(t-t_n) \mathbb{E}[(1 + \|\tilde{X}^h(t) - x^*\|^4) \|\tilde{X}^h(t) - \tilde{X}^h(t_n)\|] dt \\ &+ C(\varphi)h \int_{t_n}^{t_{n+1}} e^{-(\frac{\mu}{2}-Ch)(t_N-t)}(t-t_n) \mathbb{E}[(1 + \|\tilde{X}^h(t) - x^*\|^5)] dt \\ &+ C(\varphi) \int_{t_n}^{t_{n+1}} e^{-(\frac{\mu}{2}-Ch)(t_N-t)} \mathbb{E}[(1 + \|\tilde{X}^h(t_n) - x^*\| + \|\tilde{X}^h(t) - x^*\|)^2 \|\tilde{X}^h(t) - \tilde{X}^h(t_n)\|^2] dt \\ &+ C(\varphi)h \int_{t_n}^{t_{n+1}} e^{-(\frac{\mu}{2}-Ch)(t_N-t)} \mathbb{E}[(1 + \|\tilde{X}^h(t_n) - x^*\| + \|\tilde{X}^h(t) - x^*\|)^3 \|\tilde{X}^h(t) - \tilde{X}^h(t_n)\|] dt, \end{aligned}$$

where $C(\varphi) = C(\|\nabla\varphi(x^*)\| + \|\varphi\|_2 + \|\varphi\|_3)$ for some positive real number $C \in (0, \infty)$. Moreover, choose $H \in (0, h_4)$ such that $\frac{\mu}{2} - Ch \geq \frac{\mu}{4}$ for all $h \in (0, H)$.

Applying the Cauchy-Schwarz inequality, the moment bounds (45) and the increment bounds (46) from Lemma 4.2, if H is chosen such that $H \leq H_4$ where H_4 is given by Lemma 4.2, one obtains the upper bound

$$(74) \quad |\delta_{n,1,1}| \leq C_2(\varphi, x_0)h^2 \int_{t_n}^{t_{n+1}} e^{-\frac{\mu(t_N-t)}{4}} dt,$$

for all $N \in \mathbb{N}^*$, all $n \in \{0, \dots, N-1\}$ and all $h \in (0, H)$.

Upper bounds for $\delta_{n,1,2}$.

Given the filtration $(\mathcal{F}_t)_{t \geq 0}$ introduced in Section 2.1, for any $n \in \{0, \dots, N-1\}$, the random variable $\tilde{X}^h(t_n) = X_n$ is \mathcal{F}_{t_n} measurable. Considering the conditional expectation $\mathbb{E}[\cdot | \mathcal{F}_{t_n}]$ then yields the identity

$$\delta_{n,1,2} = \int_{t_n}^{t_{n+1}} \mathbb{E}[\langle \nabla v_T^h(t_n, \tilde{X}^h(t_n)), \mathbb{E}[\nabla F^h(\tilde{X}^h(t)) | \mathcal{F}_{t_n}] - \nabla F(\tilde{X}^h(t_n)) \rangle] dt.$$

One thus needs to study the conditional expectation $\mathbb{E}[\nabla F^h(\tilde{X}^h(t)) - \nabla F(\tilde{X}^h(t_n)) | \mathcal{F}_{t_n}]$ for all $t \in [t_n, t_{n+1}]$, which owing to the definition (15) of the modified objective function F^h can be decomposed as

$$\begin{aligned} \mathbb{E}[\nabla F^h(\tilde{X}^h(t)) | \mathcal{F}_{t_n}] - \nabla F(\tilde{X}^h(t_n)) &= \mathbb{E}[\nabla F(\tilde{X}^h(t)) | \mathcal{F}_{t_n}] - \nabla F(\tilde{X}^h(t_n)) \\ &+ \frac{h}{4} \mathbb{E}[\nabla(\|\nabla F\|^2)(\tilde{X}^h(t)) | \mathcal{F}_{t_n}] \\ &= \mathbb{E}[\nabla F(\tilde{X}^h(t)) | \mathcal{F}_{t_n}] - \nabla F(\tilde{X}^h(t_n)) \\ &+ \frac{h}{4} \nabla(\|\nabla F\|^2)(\tilde{X}^h(t_n)) \\ &+ \frac{h}{4} \mathbb{E}[\nabla(\|\nabla F\|^2)(\tilde{X}^h(t)) | \mathcal{F}_{t_n}] - \frac{h}{4} \nabla(\|\nabla F\|^2)(\tilde{X}^h(t_n)). \end{aligned}$$

Let us consider the first term of the right-hand side above. Recall that $t \in [t_n, t_{n+1}] \mapsto \tilde{X}^h(t)$ is solution of the stochastic differential equation (44). Therefore, for all $t \in [t_n, t_{n+1}]$ applying Itô's formula on the interval $[t_n, t]$ gives

$$\begin{aligned} \mathbb{E}[\nabla F(\tilde{X}^h(t)) | \mathcal{F}_{t_n}] - \nabla F(\tilde{X}^h(t_n)) &= -\mathbb{E}\left[\int_{t_n}^t \nabla^2 F(\tilde{X}^h(s)) \nabla F(X_n) ds | \mathcal{F}_{t_n}\right] \\ &\quad + \frac{h}{2} \sum_{k, \ell=1}^d \mathbb{E}\left[\int_{t_n}^t D^3 F(\tilde{X}^h(s)) \cdot (e_\ell, \sigma(t_n, X_n) e_k, \sigma(t_n, X_n) e_k) e_\ell ds | \mathcal{F}_{t_n}\right]. \end{aligned}$$

Moreover, the first term may be written as

$$\begin{aligned} -\mathbb{E}\left[\int_{t_n}^t (\nabla^2 F(\tilde{X}^h(s)) \nabla F(X_n) ds | \mathcal{F}_{t_n})\right] &= -\mathbb{E}\left[\int_{t_n}^t (\nabla^2 F(\tilde{X}^h(s)) - \nabla^2 F(\tilde{X}^h(t_n))) \nabla F(X_n) ds | \mathcal{F}_{t_n}\right] \\ &\quad - (t - t_n) \nabla^2 F(\tilde{X}^h(t_n)) \nabla F(X_n). \end{aligned}$$

Observe that one has the identity

$$\nabla^2 F(\tilde{X}^h(t_n)) \nabla F(X_n) = \nabla^2 F(\tilde{X}^h(t_n)) \nabla F(\tilde{X}^h(t_n)) = \frac{1}{2} \nabla(\|\nabla F\|^2)(\tilde{X}^h(t_n)).$$

As a result, using the tower property of conditional expectation, the error term $\delta_{n,1,2}$ is decomposed as

$$(75) \quad \delta_{n,1,2} = \delta_{n,1,2,1} + \delta_{n,1,2,2} + \delta_{n,1,2,3} + \delta_{n,1,2,4},$$

with the auxiliary error terms defined by

$$\begin{aligned} \delta_{n,1,2,1} &= \int_{t_n}^{t_{n+1}} \left(\frac{h}{4} - \frac{(t-t_n)}{2}\right) dt \mathbb{E}[\langle \nabla v_T^h(t_n, \tilde{X}^h(t_n)), \nabla(\|\nabla F\|^2)(\tilde{X}^h(t_n)) \rangle], \\ \delta_{n,1,2,2} &= - \int_{t_n}^{t_{n+1}} \mathbb{E}[\langle \nabla v_T^h(t_n, \tilde{X}^h(t_n)), \int_{t_n}^t (\nabla^2 F(\tilde{X}^h(s)) - \nabla^2 F(\tilde{X}^h(t_n))) \nabla F(X_n) ds \rangle] dt, \\ \delta_{n,1,2,3} &= \frac{h}{4} \int_{t_n}^{t_{n+1}} \mathbb{E}[\langle \nabla v_T^h(t_n, \tilde{X}^h(t_n)), \nabla(\|\nabla F\|^2)(\tilde{X}^h(t)) - \nabla(\|\nabla F\|^2)(\tilde{X}^h(t_n)) \rangle] dt, \\ \delta_{n,1,2,4} &= \frac{h}{2} \sum_{k, \ell=1}^d \int_{t_n}^{t_{n+1}} \int_{t_n}^t \mathbb{E}[D^3 F(\tilde{X}^h(s)) \cdot (\nabla v_T^h(t_n, \tilde{X}^h(t_n)), \sigma(t_n, X_n) e_k, \sigma(t_n, X_n) e_k)] ds dt. \end{aligned}$$

The key observation is that the first auxiliary error term $\delta_{n,1,2,1}$ vanishes, indeed it is straightforward to check that

$$\int_{t_n}^{t_{n+1}} \frac{(t-t_n)}{2} dt = \frac{(t_{n+1}-t_n)^2}{4} = \frac{h^2}{4} = \int_{t_n}^{t_{n+1}} \frac{h}{4} dt.$$

The other auxiliary error terms appearing in the right-hand side of the decomposition (75) of $\delta_{n,1,2}$ can then be treated as follows.

First, recall that the third order derivative of the objective function F is bounded, owing to the basic Assumption 2.1, and ∇F has at most linear growth, see (7). As a result, using the inequality (66) from Lemma 6.2 on the first order derivative of $v_T^h(t, \cdot)$, for all $h \in (0, h_1)$ one has

$$|\delta_{n,1,2,2}| \leq C(\varphi) \int_{t_n}^{t_{n+1}} \int_{t_n}^{t_n} e^{-(\mu-Ch)(t_N-t)} \mathbb{E}[(1 + \|\tilde{X}^h(t_n) - x^*\|) \|\tilde{X}^h(s) - \tilde{X}^h(t_n)\| \|X_n - x^*\|] ds dt,$$

where $C(\varphi) = C(\|\nabla \varphi(x^*)\| + \|\varphi\|_2)$ for some positive real number $C \in (0, \infty)$. Moreover, choose $H \in (0, h_1)$ such that $\mu - Ch \geq \frac{h}{4}$ for all $h \in (0, H)$.

Applying the Cauchy–Schwarz inequality, the moment bounds (45) and the increment bounds (46) from Lemma 4.2, if H is chosen such that $H \leq H_2$ where H_2 is given by Lemma 4.2, one obtains the upper bound

$$(76) \quad |\delta_{n,1,2,2}| \leq C_2(\varphi, x_0) h^2 \int_{t_n}^{t_{n+1}} e^{-\frac{\mu(t_N-t)}{4}} dt,$$

for all $N \in \mathbb{N}^*$, all $n \in \{0, \dots, N-1\}$ and all $h \in (0, H)$.

Second, recall that the mapping $\nabla(\|\nabla F\|^2)$ satisfies the local Lipschitz continuity property (17). As a result, using the inequality (66) from Lemma 6.2 on the first order derivative of $v_T^h(t, \cdot)$, for all $h \in (0, h_1)$ one has

$$|\delta_{n,1,2,3}| \leq C(\varphi) h \int_{t_n}^{t_{n+1}} e^{-(\mu-Ch)(t_N-t_n)} \mathbb{E}[(1 + \|\tilde{X}^h(t_n) - x^*\| + \|\tilde{X}^h(t) - x^*\|)^2 \|\tilde{X}^h(s) - \tilde{X}^h(t_n)\|] dt,$$

where $C(\varphi)$ is defined as above. Moreover, choose $H \in (0, h_1)$ such that $\mu - Ch \geq \frac{\mu}{4}$ for all $h \in (0, H)$.

Applying the Cauchy–Schwarz inequality, the moment bounds (45) and the increment bounds (46) from Lemma 4.2, if H is chosen such that $H \leq H_2$ where H_2 is given by Lemma 4.2, one obtains the upper bound

$$(77) \quad |\delta_{n,1,2,3}| \leq C_2(\varphi, x_0) h^2 \int_{t_n}^{t_{n+1}} e^{-\frac{\mu(t_N-t)}{4}} dt,$$

for all $N \in \mathbb{N}^*$, all $n \in \{0, \dots, N-1\}$ and all $h \in (0, H)$.

Finally, recall that the third order derivative of the objective function F is bounded, owing to the basic Assumption 2.1, and the diffusion coefficient σ has at most linear growth, see (9) from the basic Assumption 2.2. As a result, using the inequality (66) from Lemma 6.2 on the first order derivative of $v_T^h(t, \cdot)$, for all $h \in (0, h_1)$ one has

$$|\delta_{n,1,2,4}| \leq C(\varphi) h^2 \int_{t_n}^{t_{n+1}} e^{-(\mu-Ch)(t_N-t_n)} \mathbb{E}[(1 + \|X_n - x^*\|)^3] dt,$$

where $C(\varphi)$ is defined as above. Moreover, choose $H \in (0, h_1)$ such that $\mu - Ch \geq \frac{\mu}{4}$ for all $h \in (0, H)$.

Applying the moment bounds (39) from Lemma 4.1, if H is chosen such that $H \leq H_2$ where H_2 is given by Lemma 4.2, one obtains the upper bound

$$(78) \quad |\delta_{n,1,2,4}| \leq C_2(\varphi, x_0) h^2 \int_{t_n}^{t_{n+1}} e^{-\frac{\mu(t_N-t)}{4}} dt,$$

for all $N \in \mathbb{N}^*$, all $n \in \{0, \dots, N-1\}$ and all $h \in (0, H)$.

Combining the upper bounds (76), (77) and (78) on the error terms $\delta_{n,1,2,2}$, $\delta_{n,1,2,3}$ and $\delta_{n,1,2,4}$, and recalling that the error term $\delta_{n,1,2,1} = 0$ vanishes, owing to the decomposition (75) one obtains the following upper bound for the error term $\delta_{n,1,2}$: one has

$$(79) \quad |\delta_{n,1,2}| \leq C_2(\varphi, x_0) h^2 \int_{t_n}^{t_{n+1}} e^{-\frac{\mu(t_N-t)}{4}} dt,$$

for all $N \in \mathbb{N}^*$, all $n \in \{0, \dots, N-1\}$ and all $h \in (0, H)$.

Upper bounds for $\delta_{n,1}$.

Combining the upper bounds (74) and (79) on the error terms $\delta_{n,1,1}$ and $\delta_{n,1,2}$, owing to the decomposition (73) of the error term $\delta_{n,1}$ one obtains the following upper bound for the error term

$\delta_{n,1}$: one has

$$(80) \quad |\delta_{n,1}| \leq C_2(\varphi, x_0) h^2 \int_{t_n}^{t_{n+1}} e^{-\frac{\mu(t_N-t)}{4}} dt,$$

for all $N \in \mathbb{N}^*$, all $n \in \{0, \dots, N-1\}$ and all $h \in (0, H)$.

Upper bounds for $\delta_{n,2}$.

Let e_1, \dots, e_d denote an arbitrary orthonormal system of \mathbb{R}^d . The error term $\delta_{n,2}$ is written as

$$\begin{aligned} \delta_{n,2} &= \frac{h}{2} \int_{t_n}^{t_{n+1}} \mathbb{E} \left[\sum_{j=1}^d D^2 v_T^h(t, \tilde{X}^h(t)) \cdot (\sigma(t_n, \tilde{X}^h(t_n)) e_j, \sigma(t_n, \tilde{X}^h(t_n)) e_j) \right] dt \\ &\quad - \frac{h}{2} \int_{t_n}^{t_{n+1}} \mathbb{E} \left[\sum_{j=1}^d D^2 v_T^h(t, \tilde{X}^h(t)) \cdot (\sigma(t, \tilde{X}^h(t)) e_j, \sigma(t, \tilde{X}^h(t)) e_j) \right] dt \\ &= \frac{h}{2} \int_{t_n}^{t_{n+1}} \mathbb{E} \left[\sum_{j=1}^d D^2 v_T^h(t, \tilde{X}^h(t)) \cdot ((\sigma(t_n, \tilde{X}^h(t_n)) - \sigma(t, \tilde{X}^h(t))) e_j, (\sigma(t_n, \tilde{X}^h(t_n)) + \sigma(t, \tilde{X}^h(t))) e_j) \right] dt. \end{aligned}$$

Observe that one has

$$\sigma(t_n, \tilde{X}^h(t_n)) - \sigma(t, \tilde{X}^h(t)) = \sigma(t_n, \tilde{X}^h(t_n)) - \sigma(t, \tilde{X}^h(t_n)) + \sigma(t, \tilde{X}^h(t_n)) - \sigma(t, \tilde{X}^h(t)).$$

Recall, from Assumption 2.2 on the diffusion coefficient, that for all $x \in \mathbb{R}^d$ the mapping $\sigma(\cdot, x)$ is globally Lipschitz continuous, and that for all $t \in [0, \infty)$ the mapping $\sigma(t, \cdot)$ is globally Lipschitz continuous. Moreover, recall the linear growth property (9) for σ .

Using the inequality (67) from Lemma 6.2 on the second order derivative of $v_T^h(t, \cdot)$ with $\lambda = \frac{\mu}{2}$, for all $h \in (0, h_2)$ one has

$$\begin{aligned} |\delta_{n,2}| &\leq C(\varphi) h \int_{t_n}^{t_{n+1}} e^{-(\frac{\mu}{2}-Ch)(t_N-t)} (t-t_n) \mathbb{E}[G_1(\tilde{X}^h(t_n), \tilde{X}^h(t))] dt \\ &\quad + C(\varphi) h \int_{t_n}^{t_{n+1}} e^{-(\frac{\mu}{2}-Ch)(t_N-t)} \mathbb{E}[G_2(\tilde{X}^h(t_n), \tilde{X}^h(t))] dt, \end{aligned}$$

with auxiliary functions G_1 and G_2 defined by

$$\begin{aligned} G_1(x_1, x_2) &= (1 + \|x_2 - x^*\|^2)(1 + \|x_1 - x^*\| + \|x_2 - x^*\|), \\ G_2(x_1, x_2) &= (1 + \|x_2 - x^*\|^2) \|x_2 - x_1\| (1 + \|x_1 - x^*\| + \|x_2 - x^*\|), \end{aligned}$$

where $C(\varphi) = C(\|\nabla \varphi(x^*)\| + \|\varphi\|_2)$ for some positive real number $C \in (0, \infty)$. Moreover, choose $H \in (0, h_2)$ such that $\frac{\mu}{2} - Ch \geq \frac{\mu}{4}$ for all $h \in (0, H)$.

Applying the Cauchy–Schwarz inequality, the moment bounds (45) and the increment bounds (46) from Lemma 4.2, if H is chosen such that $H \leq H_2$ where H_2 is given by Lemma 4.2, one obtains the following upper bound for the error term $\delta_{n,2}$: one has

$$(81) \quad |\delta_{n,2}| \leq C_2(\varphi, x_0) h^2 \int_{t_n}^{t_{n+1}} e^{-\frac{\mu(t_N-t)}{4}} dt,$$

for all $N \in \mathbb{N}^*$, all $n \in \{0, \dots, N-1\}$ and all $h \in (0, H)$.

Conclusion. Gathering the upper bounds (80) and (81) on the error terms $|\delta_{n,1}|$ and $|\delta_{n,2}|$ and recalling the decomposition (72) of δ_n then yields the claim (71).

The proof of the weak error estimate (70) is then straightforward: for all $N \in \mathbb{N}^*$ one has

$$|\mathbb{E}[\varphi(X_N)] - \varphi(\mathcal{Y}^h(Nh))| \leq \sum_{n=0}^{N-1} |\delta_n|$$

$$\begin{aligned}
&\leq C_2(\varphi, x_0)h^2 \sum_{n=0}^{N-1} \int_{t_n}^{t_{n+1}} e^{-\frac{\mu(t_N-t)}{4}} dt \\
&\leq C_2(\varphi, x_0)h^2 \int_0^{t_N} e^{-\frac{\mu(t_N-t)}{4}} dt \leq \frac{4C_2(\varphi, x_0)h^2}{\mu}.
\end{aligned}$$

This concludes the proof of Proposition 6.2. \square

6.5. Proof of Theorem 2. We are now in a position to provide the proof of Theorem 2. It is a straightforward combination of the inequality (65) from Proposition 6.1 and of the weak error estimates (70) from Proposition 6.2.

Proof of Theorem 2. Recall that owing to (57) the error term appearing in the left-hand side of (31) can be decomposed as

$$\mathbb{E}[\varphi(X_N)] - \varphi(x^*) = \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))] - \varphi(x^*) + \mathbb{E}[\varphi(X_N)] - \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))].$$

On the one hand, using the inequality (1) and owing to the inequality (65) from Proposition 6.1 one has

$$\begin{aligned}
|\mathbb{E}[\varphi(\mathcal{Y}^h(Nh))] - \varphi(x^*)| &\leq C\mathbb{E}[(\|\nabla\varphi(x^*)\| + \|\varphi\|_2\|\mathcal{Y}^h(Nh) - x^*\|)\|\mathcal{Y}^h(Nh) - x^*\|] \\
&\leq e^{-\mu Nh} (1 + h\rho(Nh))^{\frac{1}{2}} (1 + \|x_0 - x^*\|).
\end{aligned}$$

On the other hand, applying the result (70) from Proposition 6.2 one has for all $N \in \mathbb{N}^*$ and all $h \in (0, H)$

$$|\mathbb{E}[\varphi(X_N)] - \mathbb{E}[\varphi(\mathcal{Y}^h(Nh))]| \leq C_2(\varphi, x_0)h^2.$$

Gathering the two upper bounds yields the error bound (31) and concludes the proof of Theorem 2. \square

APPENDIX A. PROOF OF LEMMA 5.1

The objective of this section is to prove the estimates (53) and (54) on the first and second order derivatives of the mapping $u^0(t, \cdot)$ defined by (48). Let us start with some preliminary results.

Repeating the proof of the inequality (12) with an arbitrary initial value $x \in \mathbb{R}^d$, it is straightforward to obtain the following bound on the solution $(\mathcal{X}_x^0(t))_{t \geq 0}$ of the ordinary differential equation (11) with arbitrary initial value $\mathcal{X}_x^0(0) = x \in \mathbb{R}^d$: one has

$$(82) \quad \sup_{t \geq 0} \sup_{x \in \mathbb{R}^d} \frac{\|\mathcal{X}_x^0(t)\|}{1 + \|x - x^*\|} < \infty.$$

In this section, we let $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ be a mapping of class \mathcal{C}^2 with a bounded second order derivative. Recall that its first order derivative is not bounded but it satisfies the inequality (1).

To simplify the notation in the proof below, let $b(x) = -\nabla F(x)$ for all $x \in \mathbb{R}^d$. As a result one has $Db(x).k = -\nabla^2 F(x)k$ for all $x \in \mathbb{R}^d$ and $k \in \mathbb{R}^d$. Applying the equivalent formulation (6) of the μ -convexity condition (5) from Assumption 2.1, then one has the inequality

$$(83) \quad \langle Db(x).k, k \rangle \leq -\mu\|k\|^2, \quad \forall x \in \mathbb{R}^d, k \in \mathbb{R}^d.$$

The proof of Lemma 5.1 proceeds by proving the inequalities (53) and (54).

Proof of the inequality (53). Owing to the definition (48) of the mapping u^0 , for all $t \geq 0$, $x \in \mathbb{R}^d$ and $k \in \mathbb{R}^d$, one has

$$Du^0(t, x).k = D\varphi(\mathcal{X}_x^0(t)).\eta_k^0(t),$$

where $t \geq 0 \mapsto \eta_k^0(t) \in \mathbb{R}^d$ is the solution of the linear differential equation

$$\begin{cases} \frac{d\eta_k^0(t)}{dt} = Db(\mathcal{X}_x^0(t)) \cdot \eta_k^0(t), & \forall t \geq 0, \\ \eta_k^0(0) = k. \end{cases}$$

Therefore owing to the inequality (1), one obtains for all $t \geq 0$

$$(84) \quad |Du^0(t, x) \cdot k| \leq \left(\|\nabla\varphi(x^*)\| + \|\varphi\|_2 \|\mathcal{X}_x^0(t) - x^*\| \right) \|\eta_k^0(t)\|,$$

and one thus needs to prove an upper bound for $\|\eta_k^0(t)\|$. Employing the inequality (83) yields, for all $t \geq 0$,

$$\frac{1}{2} \frac{d\|\eta_k^0(t)\|^2}{dt} = \left\langle \frac{d\eta_k^0(t)}{dt}, \eta_k^0(t) \right\rangle = \langle Db(\mathcal{X}_x^0(t)) \cdot \eta_k^0(t), \eta_k^0(t) \rangle \leq -\mu \|\eta_k^0(t)\|^2,$$

and applying the Grönwall lemma, one then obtains the upper bound

$$(85) \quad \|\eta_k^0(t)\| \leq e^{-\mu t} \|k\|, \quad \forall t \geq 0.$$

Combining the inequalities (82) and (85) with the upper bound (84) for $|Du^0(t, x) \cdot k|$ above, there exists a positive real number $C \in (0, \infty)$ such that for all $x, k \in \mathbb{R}^d$ and all $t \geq 0$ one has

$$|Du^0(t, x) \cdot k| \leq C \left(\|\nabla\varphi(x^*)\| + \|\varphi\|_2 \right) (1 + \|x - x^*\|) e^{-\mu t} \|k\|.$$

The proof of the inequality (53) is thus completed. \square

Proof of the inequality (54). Owing to the definition (48) of the mapping u^0 , for all $t \geq 0$, $x \in \mathbb{R}^d$ and $k_1, k_2 \in \mathbb{R}^d$, one has

$$D^2u^0(t, x) \cdot (k_1, k_2) = D\varphi(\mathcal{X}_x^0(t)) \cdot \zeta_{k_1, k_2}^0(t) + D^2\varphi(\mathcal{X}_x^0(t)) \cdot (\eta_{k_1}^0(t), \eta_{k_2}^0(t)),$$

where $t \geq 0 \mapsto \zeta_{k_1, k_2}^0(t) \in \mathbb{R}^d$ is the solution of the linear differential equation

$$\begin{cases} \frac{d\zeta_{k_1, k_2}^0(t)}{dt} = Db(\mathcal{X}_x^0(t)) \cdot \zeta_{k_1, k_2}^0(t) + D^2b(\mathcal{X}_x^0(t)) \cdot (\eta_{k_1}^0(t), \eta_{k_2}^0(t)), & \forall t \geq 0, \\ \zeta_{k_1, k_2}^0(0) = 0. \end{cases}$$

Since φ is of class \mathcal{C}^2 with bounded second order derivative, using the inequality (1), one obtains

$$(86) \quad |D^2u^0(t, x) \cdot (k_1, k_2)| \leq \left(\|\nabla\varphi(x^*)\| + \|\varphi\|_2 \|\mathcal{X}_x^0(t) - x^*\| \right) \|\zeta_{k_1, k_2}^0(t)\| + \|\varphi\|_2 \|\eta_{k_1}^0(t)\| \|\eta_{k_2}^0(t)\|.$$

Upper bounds for $\|\mathcal{X}_x^0(t) - x^*\|$ and $\|\eta_{k_j}^0(t)\|$ ($j = 1, 2$) are given by (82) and (85) respectively. It thus remains to prove an upper bound for $\|\zeta_{k_1, k_2}^0(t)\|$.

For all $t \geq 0$, one has

$$\begin{aligned} \frac{1}{2} \frac{d\|\zeta_{k_1, k_2}^0(t)\|^2}{dt} &= \left\langle \frac{d\zeta_{k_1, k_2}^0(t)}{dt}, \zeta_{k_1, k_2}^0(t) \right\rangle \\ &= \langle Db(\mathcal{X}_x^0(t)) \cdot \zeta_{k_1, k_2}^0(t), \zeta_{k_1, k_2}^0(t) \rangle + \langle D^2b(\mathcal{X}_x^0(t)) \cdot (\eta_{k_1}^0(t), \eta_{k_2}^0(t)), \zeta_{k_1, k_2}^0(t) \rangle. \end{aligned}$$

Using the inequality (83), the boundedness of the second order derivative of b (the third order derivative of F is bounded, owing to Assumption 2.1), and the Cauchy–Schwarz and Young inequalities, there exists a positive real number $C \in (0, \infty)$ such that for all $t \geq 0$ one has

$$\begin{aligned} \frac{1}{2} \frac{d\|\zeta_{k_1, k_2}^0(t)\|^2}{dt} &\leq -\mu \|\zeta_{k_1, k_2}^0(t)\|^2 + C \|\eta_{k_1}^0(t)\| \|\eta_{k_2}^0(t)\| \|\zeta_{k_1, k_2}^0(t)\| \\ &\leq -\frac{\mu}{2} \|\zeta_{k_1, k_2}^0(t)\|^2 + C \|\eta_{k_1}^0(t)\|^2 \|\eta_{k_2}^0(t)\|^2. \end{aligned}$$

As a result, applying the Grönwall inequality and using the inequality (85), for all $t \geq 0$ one obtains

$$\|\zeta_{k_1, k_2}^0(t)\|^2 \leq C \int_0^t e^{-\mu(t-s)} e^{-4\mu s} ds \|k_1\|^2 \|k_2\|^2,$$

which gives the following inequality: for all $t \geq 0$

$$\|\zeta_{k_1, k_2}^0(t)\| \leq C e^{-\frac{\mu}{2}t} \|k_1\| \|k_2\|.$$

Using this inequality and the inequality (85), one then obtains another upper bound: for all $t \geq 0$ one has

$$\begin{aligned} \frac{1}{2} \frac{d\|\zeta_{k_1, k_2}^0(t)\|^2}{dt} &\leq -\mu \|\zeta_{k_1, k_2}^0(t)\|^2 + C \|\eta_{k_1}^0(t)\| \|\eta_{k_2}^0(t)\| \|\zeta_{k_1, k_2}^0(t)\| \\ &\leq -\mu \|\zeta_{k_1, k_2}^0(t)\|^2 + C e^{-(2\mu + \frac{\mu}{2})t} \|k_1\|^2 \|k_2\|^2. \end{aligned}$$

Applying again the Grönwall inequality, for all $t \geq 0$ one obtains

$$\|\zeta_{k_1, k_2}^0(t)\|^2 \leq C \int_0^t e^{-2\mu(t-s)} e^{-2(\mu + \frac{\mu}{2})s} ds \|k_1\|^2 \|k_2\|^2 = C e^{-2\mu t} \int_0^t e^{-\mu s} ds \|k_1\|^2 \|k_2\|^2.$$

As a result one obtains the following inequality: for all $t \geq 0$

$$(87) \quad \|\zeta_{k_1, k_2}^0(t)\| \leq C e^{-\mu t} \|k_1\| \|k_2\|.$$

Combining the inequalities (82), (85) and (87) with the upper bound (86) for $|D^2 u^0(t, x).(k_1, k_2)|$ above, as a result there exists a positive real number $C \in (0, \infty)$ such that for all $x, k_1, k_2 \in \mathbb{R}^d$ one has

$$|D^2 u^0(t, x).(k_1, k_2)| \leq C (\|\nabla \varphi(x^*)\| + \|\varphi\|_2) (1 + \|x - x^*\|) e^{-\mu t} \|k_1\| \|k_2\|.$$

The proof of the inequality (54) is thus completed. \square

APPENDIX B. PROOF OF LEMMAS 6.2 AND 6.3

B.1. Proof of Lemma 6.2. Like in Section A, let us start with some preliminary results.

In this proof, let $b^h = -\nabla F^h$. As a result one has $Db^h(y).k = -\nabla^2 F^h(y)k$ for all $y \in \mathbb{R}^d$ and $k \in \mathbb{R}^d$. Since the modified objective function F^h is assumed to be μ -convex for all $h \in (0, h_0)$ owing to Assumption 2.3, one has the inequality

$$(88) \quad \langle Db^h(y).k, k \rangle \leq -\mu \|k\|^2, \quad \forall y \in \mathbb{R}^d, k \in \mathbb{R}^d, h \in (0, h_0).$$

Moreover, the second and third order derivatives of b^h have at most polynomial growth, uniformly with respect to $h \in (0, h_0)$: there exists a positive real number $C \in (0, \infty)$ such that for all $h \in (0, h_0)$ and $y \in \mathbb{R}^d$ one has

$$(89) \quad \|D^2 b^h(y).(k_1, k_2)\| \leq C(1 + \|y - x^*\|) \|k_1\| \|k_2\|, \quad \forall k_1, k_2 \in \mathbb{R}^d,$$

$$(90) \quad \|D^3 b^h(y).(k_1, k_2, k_3)\| \leq C(1 + \|y - x^*\|) \|k_1\| \|k_2\| \|k_3\|, \quad \forall k_1, k_2, k_3 \in \mathbb{R}^d.$$

To simplify the notation below, given $k_1, k_2, k_3 \in \mathbb{R}^d$, define the set

$$(91) \quad \mathcal{K}(k_1, k_2, k_3) = \{(k_1, k_2, k_3), (k_2, k_3, k_1), (k_3, k_1, k_2)\}.$$

The proof below proceeds as the proof of Lemma 5.1 in Section A, with two main differences: one needs to deal with stochastic processes, thus Itô's formula is employed, and one needs to prove bounds up to the third order derivative of $v^h(t, \cdot)$. In addition, the main effort is devoted to prove bounds which are independent of the parameter h in a range $(0, h_1)$.

Proof of the inequality (66). Owing to the definition (59) of the mapping v_T^h , for all $T \geq t \geq 0$, $y \in \mathbb{R}^d$ and $k \in \mathbb{R}^d$, one has

$$Dv_T^h(t, y).k = \mathbb{E}[D\varphi(\mathcal{Y}_y^h(T|t)).\eta_k^h(T|t)],$$

where the process $s \geq t \mapsto \eta_k^h(s|t)$ is the solution of the linear stochastic differential equation

$$(92) \quad \begin{cases} d\eta_k^h(s|t) = Db^h(\mathcal{Y}_y^h(s|t))\eta_k^h(s|t) ds + \sqrt{h}D\sigma(s, \mathcal{Y}_y^h(s|t))\eta_k^h(s|t) dB(s), \\ \eta_k^h(t|t) = k. \end{cases}$$

Using the inequality (1) and the Cauchy–Schwarz inequality, one obtains for all $T \geq t \geq 0$

$$(93) \quad |Dv_T^h(t, y).k| \leq \left(\|\nabla\varphi(x^*)\| + \|\varphi\|_2 (\mathbb{E}[\|\mathcal{Y}_y^h(T|t) - x^*\|^2])^{\frac{1}{2}} \right) (\mathbb{E}[\|\eta_k^h(T|t)\|^2])^{\frac{1}{2}}.$$

Moment bounds for $\mathcal{Y}_y^h(T|t) - x^*$ are given in Lemma 6.1, it thus remains to prove moment bounds for $\eta_k^h(T|t)$. Even if it would be sufficient to deal with second order moments to deal with (93) above, moment bounds at arbitrary order are proved below since they are instrumental in the analysis of second and third order derivatives of v_T^h .

Recall that from Assumption 2.2 the mapping $\sigma(s, \cdot)$ is globally Lipschitz continuous on \mathbb{R}^d , uniformly with respect to $s \in \mathbb{R}^+$. For all $p \in \mathbb{N}^*$, applying the Itô formula to the process $s \geq t \mapsto \|\eta_k^h(s|t)\|^{2p}$, there exists a positive real number $C_p^{(1)} \in (0, \infty)$ such that for all $s \geq t$ one has

$$\frac{1}{2p} \frac{d\mathbb{E}[\|\eta_k^h(s|t)\|^{2p}]}{ds} \leq \mathbb{E}[\langle Db^h(\mathcal{Y}_y^h(s|t)).\eta_k^h(s|t), \eta_k^h(s|t) \rangle \|\eta_k^h(s|t)\|^{2(p-1)}] + C_p^{(1)} h \mathbb{E}[\|\eta_k^h(s|t)\|^{2p}].$$

Applying the inequality (88) which follows from the μ -convexity of F^h , one obtains the following upper bound: for all $s \geq t$ one has

$$\frac{1}{2p} \frac{d\mathbb{E}[\|\eta_k^h(s|t)\|^{2p}]}{ds} \leq -(\mu - C_p^{(1)}h) \mathbb{E}[\|\eta_k^h(s|t)\|^{2p}].$$

Let $h_p^{(1)} \in (0, \frac{\mu}{C_p^{(1)}})$, then for all $h \in (0, h_p^{(1)}]$ one has $\mu - C_p^{(1)}h \geq \mu - C_p^{(1)}h_p^{(1)} > 0$. Applying the Gronwall inequality and recalling that $\eta_k^h(t|t) = k$ one obtains the following inequality: for all $s \geq t$ and all $h \in (0, h_p^{(1)}]$ one has

$$(94) \quad \mathbb{E}[\|\eta_k^h(s|t)\|^{2p}] \leq e^{-2p(\mu - C_p^{(1)}h)(s-t)} \|k\|^{2p} \leq e^{-2p(\mu - C_p^{(1)}h_p^{(1)})(s-t)} \|k\|^{2p}.$$

It thus suffices to combine the moment bounds (64) and (94) with $p = 1$, with the inequality (93), to obtain the inequality (66): there exist positive real number $h_1 \in (0, h_0)$ and $C_1 \in (0, \infty)$ such that for all $y, k \in \mathbb{R}^d$, all $T \geq t \geq 0$, and all $h \in (0, h_1)$ one has

$$|Dv_T^h(t, y).k| \leq C_1 (\|\nabla\varphi(x^*)\| + \|\varphi\|_2) (1 + \|y - x^*\|) e^{-(\mu - C_1 h_1)(T-t)} \|k\|.$$

The proof of the inequality (66) is thus completed. \square

Proof of the inequality (67). Owing to the definition (59) of the mapping v_T^h , for all $T \geq t$, $y \in \mathbb{R}^d$, and $k_1, k_2 \in \mathbb{R}^d$, one has

$$D^2v_T^h(t, x).(k_1, k_2) = \mathbb{E}[D\varphi(\mathcal{Y}_y^h(T|t)).\zeta_{k_1, k_2}^h(T|t)] + \mathbb{E}[D^2\varphi(\mathcal{Y}_y^h(T|t)).(\eta_{k_1}^h(T|t), \eta_{k_2}^h(T|t))],$$

where the processes $s \geq t \mapsto \eta_{k_1}^h(T|t)$ and $s \geq t \mapsto \eta_{k_2}^h(T|t)$ are given by (92) with $k = k_1$ and $k = k_2$ respectively, and where the process $s \geq t \mapsto \zeta_{k_1, k_2}^h(s|t)$ is the solution of the stochastic differential

equation

$$(95) \quad \begin{cases} d\zeta_{k_1, k_2}^h(s|t) = Db^h(\mathcal{Y}_y^h(s|t)) \cdot \zeta_{k_1, k_2}^h(s) ds + \sqrt{h} D\sigma(s, \mathcal{Y}_y^h(s|t)) \cdot \zeta_{k_1, k_2}^h(s|t) dB(s) \\ + D^2b^h(\mathcal{Y}_y^h(s|t)) \cdot (\eta_{k_1}^h(s|t), \eta_{k_2}^h(s|t)) ds + \sqrt{h} D^2\sigma(s, \mathcal{Y}_y^h(s|t)) \cdot (\eta_{k_1}^h(s|t), \eta_{k_2}^h(s|t)) dB(s), \\ \zeta_{k_1, k_2}^h(t|t) = 0. \end{cases}$$

Using the inequality (1) and the Cauchy–Schwarz inequality, one obtains for all $T \geq t \geq 0$

$$(96) \quad |D^2v_T^h(t, y) \cdot (k_1, k_2)| \leq \left(\|\nabla\varphi(x^*)\| + \|\varphi\|_2 (\mathbb{E}[\|\mathcal{Y}_y^h(T|t) - x^*\|^2])^{\frac{1}{2}} \right) (\mathbb{E}[\|\zeta_{k_1, k_2}^h(T|t)\|^2])^{\frac{1}{2}} \\ + \|\varphi\|_2 (\mathbb{E}[\|\eta_{k_1}^h(T|t)\|^2])^{\frac{1}{2}} (\mathbb{E}[\|\eta_{k_2}^h(T|t)\|^2])^{\frac{1}{2}}.$$

Moment bounds for $\mathcal{Y}_y^h(T|t) - x^*$ are given by Lemma 6.1, whereas moment bounds for $\eta_{k_1}^h(T|t)$ and $\eta_{k_2}^h(T|t)$ are given by (94) above. It thus remains to prove moment bounds for $\zeta_{k_1, k_2}^h(T|t)$.

Recall that the first and second order derivatives of the mapping $\sigma(s, \cdot)$ are bounded owing to Assumption 2.4, uniformly with respect to $s \in [0, \infty)$. Moreover, the second order derivative of the auxiliary mapping b^h has at most linear growth, uniformly with respect to $h \in (0, h_0)$, see (89). For all $p \in \mathbb{N}^*$, applying the Itô formula to the process $s \geq t \mapsto \|\zeta_{k_1, k_2}^h(s|t)\|^{2p}$, there exists a positive real number $C_p \in (0, \infty)$ such that for all $s \geq t$ one has

$$\begin{aligned} \frac{1}{2p} \frac{d}{ds} \mathbb{E}[\|\zeta_{k_1, k_2}^h(s|t)\|^{2p}] &\leq \mathbb{E}[\langle Db^h(\mathcal{Y}_y^h(s|t)) \cdot \zeta_{k_1, k_2}^h(s|t), \zeta_{k_1, k_2}^h(s|t) \rangle \|\zeta_{k_1, k_2}^h(s|t)\|^{2(p-1)}] \\ &+ C_p h \mathbb{E}[\|\zeta_{k_1, k_2}^h(s|t)\|^{2p}] \\ &+ C_p \mathbb{E}[(1 + \|\mathcal{Y}_y^h(s|t) - x^*\|) \|\eta_{k_1}^h(s|t)\| \|\eta_{k_2}^h(s|t)\| \|\zeta_{k_1, k_2}^h(s|t)\|^{2p-1}] \\ &+ C_p h \mathbb{E}[\|\eta_{k_1}^h(s|t)\|^2 \|\eta_{k_2}^h(s|t)\|^2 \|\zeta_{k_1, k_2}^h(s|t)\|^{2(p-1)}]. \end{aligned}$$

Assume that $\lambda \in (0, \mu)$. Using the μ -convexity property (88) of the modified objective function F^h , one has

$$\mathbb{E}[\langle Db^h(\mathcal{Y}_y^h(s|t)) \cdot \zeta_{k_1, k_2}^h(s|t), \zeta_{k_1, k_2}^h(s|t) \rangle \|\zeta_{k_1, k_2}^h(s|t)\|^{2(p-1)}] \leq -\mu \mathbb{E}[\|\zeta_{k_1, k_2}^h(s|t)\|^{2p}].$$

In addition, using the Hölder and Young inequalities, one obtains the upper bounds

$$\begin{aligned} &\mathbb{E}[(1 + \|\mathcal{Y}_y^h(s|t) - x^*\|) \|\eta_{k_1}^h(s|t)\| \|\eta_{k_2}^h(s|t)\| \|\zeta_{k_1, k_2}^h(s|t)\|^{2p-1}] \\ &\leq (\mathbb{E}[(1 + \|\mathcal{Y}_y^h(s|t) - x^*\|)^{2p} \|\eta_{k_1}^h(s|t)\|^{2p} \|\eta_{k_2}^h(s|t)\|^{2p}])^{\frac{1}{2p}} (\mathbb{E}[\|\zeta_{k_1, k_2}^h(s|t)\|^{2p}])^{\frac{2p-1}{2p}} \\ &\leq C_p \mathbb{E}[(1 + \|\mathcal{Y}_y^h(s|t) - x^*\|)^{2p} \|\eta_{k_1}^h(s|t)\|^{2p} \|\eta_{k_2}^h(s|t)\|^{2p}] + (\mu - \lambda) \mathbb{E}[\|\zeta_{k_1, k_2}^h(s|t)\|^{2p}], \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}[\|\eta_{k_1}^h(s|t)\|^2 \|\eta_{k_2}^h(s|t)\|^2 \|\zeta_{k_1, k_2}^h(s|t)\|^{2(p-1)}] \\ &\leq (\mathbb{E}[\|\eta_{k_1}^h(s|t)\|^{2p} \|\eta_{k_2}^h(s|t)\|^{2p}])^{\frac{1}{p}} (\mathbb{E}[\|\zeta_{k_1, k_2}^h(s|t)\|^{2p}])^{\frac{p-1}{p}} \\ &\leq \frac{1}{p} \mathbb{E}[\|\eta_{k_1}^h(s|t)\|^{2p} \|\eta_{k_2}^h(s|t)\|^{2p}] + \frac{p-1}{p} \mathbb{E}[\|\zeta_{k_1, k_2}^h(s|t)\|^{2p}]. \end{aligned}$$

Combining the upper bounds above, there exists a positive real number $C_p^{(2)} \in (0, \infty)$ such that for all $s \geq t$ one has

$$\begin{aligned} \frac{1}{2p} \frac{d}{ds} \mathbb{E}[\|\zeta_{k_1, k_2}^h(s|t)\|^{2p}] &\leq -(\lambda - C_p^{(2)} h) \mathbb{E}[\|\zeta_{k_1, k_2}^h(s|t)\|^{2p}] \\ &+ C_p^{(2)} \mathbb{E}[(1 + \|\mathcal{Y}_y^h(s|t) - x^*\|)^{2p} \|\eta_{k_1}^h(s|t)\|^{2p} \|\eta_{k_2}^h(s|t)\|^{2p}]. \end{aligned}$$

In addition, applying the Hölder inequality and using the moment bounds (64) and (94), for all $s \geq t$ one has

$$\mathbb{E}[(1 + \|\mathcal{Y}_y^h(s|t) - x^*\|)^{2p} \|\eta_{k_1}^h(s|t)\|^{2p} \|\eta_{k_2}^h(s|t)\|^{2p}] \leq C_p^{(2)} (1 + \|y - x^*\|)^{2p} e^{-4p(\mu - C_{2p}^{(1)}h)(s-t)} \|k_1\|^{2p} \|k_2\|^{2p}.$$

Without loss of generality, it may be assumed that $C_p^{(2)} \geq C_{2p}^{(1)}$. Let $h_p^{(2)} \in (0, \frac{\lambda}{C_p^{(2)}})$, then for all $h \in (0, h_p^{(2)})$ one has $\lambda - C_p^{(2)}h \geq \lambda - C_p^{(2)}h_p^{(2)} > 0$. Applying the Grönwall inequality and recalling that $\zeta_{k_1, k_2}^h(t|t) = 0$, one obtains for all $s \geq t$ and all $h \in (0, h_p^{(2)})$

$$(97) \quad \mathbb{E}[\|\zeta_{k_1, k_2}^h(s|t)\|^{2p}] \leq C_p^{(2)} e^{-2p(\lambda - C_p^{(2)}h)(s-t)} (1 + \|y - x^*\|)^{2p} \|k_1\|^{2p} \|k_2\|^{2p}.$$

It thus suffices to combine the moment bounds (64), (94) and (97) with $p = 1$, with the inequality (96), to obtain the inequality (67): for all $\lambda \in (0, \mu)$, there exist positive real numbers $h_2 \in (0, h_1)$ and $C_2 \in (0, \infty)$ such that for all $y, k_1, k_2 \in \mathbb{R}^d$, all $T \geq t \geq 0$, and all $h \in (0, h_2)$ one has

$$|D^2 v_T^h(t, y) \cdot (k_1, k_2)| \leq C_2 (\|\nabla \varphi(x^*)\| + \|\varphi\|_2) (1 + \|y - x^*\|^2) e^{-(\mu - C_2 h_2)(T-t)} \|k_1\| \|k_2\|.$$

The proof of the inequality (67) is thus completed. \square

Proof of the inequality (68). Owing to the definition (59) of the mapping v_T^h , for all $T \geq t \geq 0$, $y \in \mathbb{R}^d$, and $k_1, k_2, k_3 \in \mathbb{R}^d$, one has

$$\begin{aligned} D^3 v_T^h(t, x) \cdot (k_1, k_2, k_3) &= \mathbb{E}[D\varphi(\mathcal{Y}_y^h(T|t)) \xi_{k_1, k_2, k_3}^h(T|t)] \\ &\quad + \sum_{(\ell_1, \ell_2, \ell_3) \in \mathcal{K}} \mathbb{E}[D^2 \varphi(\mathcal{Y}_y^h(T|t)) (\eta_{\ell_1}^h(T|t), \zeta_{\ell_2, \ell_3}^h(T|t))] \\ &\quad + \mathbb{E}[D^3 \varphi(\mathcal{Y}_y^h(T|t)) (\eta_{k_1}^h(T|t), \eta_{k_2}^h(T|t), \eta_{k_3}^h(T|t))], \end{aligned}$$

where the set $\mathcal{K} = \mathcal{K}(k_1, k_2, k_3)$ is defined by (91), where the processes $s \geq t \mapsto \eta_\ell^h(T|t)$ with $\ell \in \{k_1, k_2, k_3\}$ are given by (92), where the processes $s \geq t \mapsto \zeta_{\ell_1, \ell_2}^h(T|t)$ with indices $(\ell_1, \ell_2) \in \{(k_2, k_3), (k_3, k_1), (k_1, k_2)\}$ are given by (95), and where the process $s \geq t \mapsto \xi_{k_1, k_2, k_3}^h(s|t)$ is the solution of the stochastic differential equation

$$(98) \quad \left\{ \begin{aligned} d\xi_{k_1, k_2, k_3}^h(s|t) &= Db^h(\mathcal{Y}_y^h(s|t)) \cdot \xi_{k_1, k_2, k_3}^h(s|t) ds + \sqrt{h} D\sigma(s, \mathcal{Y}_y^h(s|t)) \cdot \xi_{k_1, k_2, k_3}^h(s|t) dB(s) \\ &\quad + \sum_{(\ell_1, \ell_2, \ell_3) \in \mathcal{K}} D^2 b^h(\mathcal{Y}_y^h(s|t)) \cdot (\eta_{\ell_1}^h(s|t), \zeta_{\ell_2, \ell_3}^h(s|t)) ds \\ &\quad + \sum_{(\ell_1, \ell_2, \ell_3) \in \mathcal{K}} \sqrt{h} D^2 \sigma(s, \mathcal{Y}_y^h(s|t)) \cdot (\eta_{\ell_1}^h(s|t), \zeta_{\ell_2, \ell_3}^h(s|t)) dB(s) \\ &\quad + D^3 b^h(\mathcal{Y}_y^h(s|t)) \cdot (\eta_{k_1}^h(s|t), \eta_{k_2}^h(s|t), \eta_{k_3}^h(s|t)) ds \\ &\quad + \sqrt{h} D^3 \sigma(s, \mathcal{Y}_y^h(s|t)) \cdot (\eta_{k_1}^h(s|t), \eta_{k_2}^h(s|t), \eta_{k_3}^h(s|t)) dB(s), \\ \xi_{k_1, k_2, k_3}^h(t|t) &= 0. \end{aligned} \right.$$

Using the inequality (1) and the Cauchy–Schwarz and Hölder inequality, one obtains for all $s \geq t$

$$(99) \quad \begin{aligned} |D^3 v_T^h(t, y) \cdot (k_1, k_2, k_3)| &\leq \left(\|\nabla \varphi(x^*)\| + \|\varphi\|_2 (\mathbb{E}[\|\mathcal{Y}_y^h(T|t) - x^*\|^2])^{\frac{1}{2}} \right) (\mathbb{E}[\|\xi_{k_1, k_2, k_3}^h(T|t)\|^2])^{\frac{1}{2}} \\ &\quad + \|\varphi\|_2 \sum_{(\ell_1, \ell_2, \ell_3) \in \mathcal{K}} (\mathbb{E}[\|\eta_{\ell_1}^h(T|t)\|^2])^{\frac{1}{2}} (\mathbb{E}[\|\zeta_{\ell_2, \ell_3}^h(T|t)\|^2])^{\frac{1}{2}} \\ &\quad + \|\varphi\|_3 (\mathbb{E}[\|\eta_{k_1}^h(T|t)\|^3])^{\frac{1}{3}} (\mathbb{E}[\|\eta_{k_2}^h(T|t)\|^3])^{\frac{1}{3}} (\mathbb{E}[\|\eta_{k_3}^h(T|t)\|^3])^{\frac{1}{3}}. \end{aligned}$$

Moment bounds for $\mathcal{Y}_y^h(T|t) - x^\star$ are given by Lemma 6.1, moment bounds for $\eta_{k_1}^h(T|t)$, $\eta_{k_2}^h(T|t)$ and $\eta_{k_3}^h(T|t)$ are given by (94) above, and moment bounds for $\zeta_{k_1, k_2}^h(T|t)$, $\zeta_{k_2, k_3}^h(T|t)$ and $\zeta_{k_3, k_1}^h(T|t)$ are given by (97) above. It thus remains to prove moment bounds for $\xi_{k_1, k_2, k_3}^h(T|t)$. It suffices to consider second order moments, i.e. $p = 1$.

Recall that the first, second and third order derivatives of the mapping $\sigma(s, \cdot)$ are bounded owing to Assumption 2.4, uniformly with respect to $s \in (0, \infty)$. Moreover, the second and third order derivatives of the auxiliary mapping b^h have at most linear growth, see (89) and (90), uniformly with respect to $h \in (0, h_0)$. Applying the Itô formula to the process $s \geq t \mapsto \|\xi_{k_1, k_2, k_3}^h(s|t)\|^2$, there exists a positive real number $C \in (0, \infty)$ such that for all $s \geq t$ one has

$$\begin{aligned} \frac{1}{2} \frac{d \mathbb{E}[\|\xi_{k_1, k_2, k_3}^h(s|t)\|^2]}{ds} &\leq \mathbb{E}[\langle Db^h(\mathcal{Y}_y^h(s|t)) \cdot \xi_{k_1, k_2, k_3}^h(s|t), \xi_{k_1, k_2, k_3}^h(s|t) \rangle] \\ &\quad + Ch \mathbb{E}[\|\xi_{k_1, k_2, k_3}^h(s|t)\|^2] \\ &\quad + C \sum_{(\ell_1, \ell_2, \ell_3) \in \mathcal{K}} \mathbb{E}[(1 + \|\mathcal{Y}_y^h(s|t) - x^\star\|) \|\eta_{\ell_1}^h(s|t)\| \|\zeta_{\ell_2, \ell_3}^h(s|t)\| \|\xi_{k_1, k_2, k_3}^h(s|t)\|] \\ &\quad + Ch \sum_{(\ell_1, \ell_2, \ell_3) \in \mathcal{K}} \mathbb{E}[\|\eta_{\ell_1}^h(s|t)\|^2 \|\zeta_{\ell_2, \ell_3}^h(s|t)\|^2] \\ &\quad + C \mathbb{E}[(1 + \|\mathcal{Y}_y^h(s|t) - x^\star\|) \|\eta_{k_1}^h(s|t)\| \|\eta_{k_2}^h(s|t)\| \|\eta_{k_3}^h(s|t)\| \|\xi_{k_1, k_2, k_3}^h(s|t)\|] \\ &\quad + Ch \mathbb{E}[\|\eta_{k_1}^h(s|t)\|^2 \|\eta_{k_2}^h(s|t)\|^2 \|\eta_{k_3}^h(s|t)\|^2]. \end{aligned}$$

Assume that $\lambda \in (0, \mu)$. Using the μ -convexity property (88) of the modified objective function F^h , one has

$$\mathbb{E}[\langle Db^h(\mathcal{Y}_y^h(s|t)) \cdot \xi_{k_1, k_2, k_3}^h(s|t), \xi_{k_1, k_2, k_3}^h(s|t) \rangle] \leq -\mu \mathbb{E}[\|\xi_{k_1, k_2, k_3}^h(s|t)\|^2].$$

Using the Hölder and Young inequalities, there exists a positive real number $C_1^{(3)} \in (0, \infty)$ such that for all $s \geq t$ one has

$$\begin{aligned} \frac{1}{2} \frac{d \mathbb{E}[\|\xi_{k_1, k_2, k_3}^h(s|t)\|^2]}{ds} &\leq -(\lambda - C_1^{(3)} h) \mathbb{E}[\|\xi_{k_1, k_2, k_3}^h(s|t)\|^2] \\ &\quad + C_1^{(3)} \sum_{(\ell_1, \ell_2, \ell_3) \in \mathcal{K}} \mathbb{E}[(1 + \|\mathcal{Y}_y^h(s|t) - x^\star\|)^2 \|\eta_{\ell_1}^h(s|t)\|^2 \|\zeta_{\ell_2, \ell_3}^h(s|t)\|^2] \\ &\quad + C_1^{(3)} \mathbb{E}[(1 + \|\mathcal{Y}_y^h(s|t) - x^\star\|)^2 \|\eta_{k_1}^h(s|t)\|^2 \|\eta_{k_2}^h(s|t)\|^2 \|\eta_{k_3}^h(s|t)\|^2] \end{aligned}$$

Applying the Hölder inequality and using the moment bounds (64), (94) and (97), one thus obtains the upper bound

$$\begin{aligned} \frac{1}{2} \frac{d \mathbb{E}[\|\xi_{k_1, k_2, k_3}^h(s|t)\|^2]}{ds} &\leq -(\lambda - C_1^{(3)} h) \mathbb{E}[\|\xi_{k_1, k_2, k_3}^h(s|t)\|^2] \\ &\quad + C_1^{(3)} e^{-2(\mu + \lambda - (C_3^{(1)} + C_3^{(2)})h)(s-t)} (1 + \|y - x^\star\|)^2 \|k_1\|^2 \|k_2\|^2 \|k_3\|^2 \\ &\quad + C_1^{(3)} e^{-3(\mu - C_4^{(1)}h)(s-t)} (1 + \|y - x^\star\|)^2 \|k_1\|^2 \|k_2\|^2 \|k_3\|^2. \end{aligned}$$

Without loss of generality it may be assumed that $C_1^{(3)} \geq \max(C_3^{(1)}, C_3^{(2)}, C_4^{(1)})$. Let $h_1^{(3)} \in (0, \frac{\lambda}{C_1^{(3)}})$, then for all $h \in (0, h_1^{(3)})$ one has $\lambda - C_1^{(3)} h \geq \lambda - C_1^{(3)} h_1^{(3)} > 0$. Applying the Grönwall inequality and recalling that $\xi_{k_1, k_2, k_3}^h(t|t) = 0$, one obtains for all $h \in (0, h_1^{(3)})$

$$(100) \quad \mathbb{E}[\|\xi_{k_1, k_2, k_3}^h(s|t)\|^2] \leq C_1^{(3)} e^{-2(\lambda - C_1^{(3)} h)(s-t)} (1 + \|y - x^\star\|)^2 \|k_1\|^2 \|k_2\|^2 \|k_3\|^2.$$

It suffices to combine the moment bounds (64), (94), (97) and (100) with the inequality (99) to obtain the inequality (68): for all $\lambda \in (0, \mu)$, there exist positive real numbers $h_3 \in (0, h_2)$ and $C_3 \in (0, \infty)$ such that for all $y, k_1, k_2 \in \mathbb{R}^d$, all $T \geq t \geq 0$, and all $h \in (0, h_3)$ one has

$$|D^3 v_T^h(t, y) \cdot (k_1, k_2, k_3)| \leq C_2 (\|\nabla \varphi(x^*)\| + \|\varphi\|_2 + \|\varphi\|_3) (1 + \|y - x^*\|)^2 e^{-(\mu - C_3 h_3)(T-t)} \|k_1\| \|k_2\| \|k_3\|.$$

The proof of the inequality (68) is thus completed. \square

B.2. Proof of Lemma 6.3. Lemma 6.3 provides upper bounds for the first order derivative $D(\partial_t v_T^h)(t, \cdot)$. To prove them, it suffices to express the temporal derivative $\partial_t v_T^h(t, y)$ using the backward Kolmogorov equation (60), to differentiate with respect to y , and to apply the upper bounds on the derivatives $D^j v_T^h(t, \cdot)$ of order $j \in \{1, 2, 3\}$ stated in Lemma 6.3.

Proof of Lemma 6.3. Owing to the backward Kolmogorov equation (60), for all $t \in [0, T]$, $y \in \mathbb{R}^d$, one has

$$\partial_t v_T^h(t, y) = Dv_T^h(t, y) \cdot \nabla F^h(y) - \frac{h}{2} \sum_{j=1}^d D^2 v_T^h(t, y) \cdot (\sigma(t, y) e_j, \sigma(t, y) e_j),$$

where e_1, \dots, e_n denotes an arbitrary orthonormal system of \mathbb{R}^d .

Differentiating with respect to $y \in \mathbb{R}^d$, for all $k \in \mathbb{R}^d$ one obtains

$$\begin{aligned} D(\partial_t v_T^h)(t, y) \cdot k &= D^2 F^h(y) \cdot (\nabla v_T^h(t, y), k) + D^2 v_T^h(t, y) \cdot (\nabla F^h(y), k) \\ &\quad - h \sum_{j=1}^d D^2 v_T^h(t, y) \cdot ((D\sigma(t, y) \cdot k) e_j, \sigma(t, y) e_j) - \frac{h}{2} \sum_{j=1}^d D^3 v_T^h(t, y) \cdot (k, \sigma(t, y) e_j, \sigma(t, y) e_j). \end{aligned}$$

It remains to prove upper bounds for the four terms appearing in the right-hand side above. Since the results of Lemma 6.2 are employed, it is assumed that $h \in (0, h_3)$ below.

For the first term, using the local Lipschitz continuity property (19) of ∇F^h and to the inequality (66) on the first order derivative of $v_T^h(t, \cdot)$, one obtains the upper bounds

$$\begin{aligned} |D^2 F^h(y) \cdot (\nabla v_T^h(t, y), k)| &\leq C(1 + \|y - x^*\|) \|\nabla v_T^h(t, y)\| \|k\| \\ &\leq C(\|\nabla \varphi(x^*)\| + \|\varphi\|_2) (1 + \|y - x^*\|^2) e^{-(\mu - Ch)(T-t)} \|k\|. \end{aligned}$$

For the second term, observe that combining (7) and the first inequality from (16), one has $\|\nabla F^h(y)\| \leq C\|y - x^*\|$. As a result, using the inequality (67) on the second order derivative of $v_T^h(t, \cdot)$, one obtains the upper bound

$$|D^2 v_T^h(t, y) \cdot (\nabla F^h(y), k)| \leq C(\|\nabla \varphi(x^*)\| + \|\varphi\|_2) e^{-(\lambda - Ch)(T-t)} (1 + \|y - x^*\|^3) \|k\|.$$

For the third term, using the linear growth and Lipschitz continuity properties of the diffusion coefficient $\sigma(t, \cdot)$, see (9) and (8), and the inequality (67) on the second order derivative of $v_T^h(t, \cdot)$, one obtains the upper bounds

$$\begin{aligned} & \left| \sum_{j=1}^d D^2 v_T^h(t, y) \cdot ((D\sigma(t, y) \cdot k) e_j, \sigma(t, y) e_j) \right| \\ & \leq C(\|\nabla \varphi(x^*)\| + \|\varphi\|_2) e^{-(\lambda - Ch)(T-t)} (1 + \|y - x^*\|^2) \|D\sigma(t, y) \cdot k\| \|\sigma(t, y)\| \\ & \leq C\|\varsigma\|_\infty^2 (\|\nabla \varphi(x^*)\| + \|\varphi\|_2) e^{-(\lambda - Ch)(T-t)} (1 + \|y - x^*\|^3). \end{aligned}$$

For the fourth term, using the linear growth property (9) of the diffusion coefficient $\sigma(t, \cdot)$, and the inequality (68) on the third order derivative of $v_T^h(t, \cdot)$, one obtains the upper bounds

$$\left| \sum_{j=1}^d D^3 v_T^h(t, y) \cdot (k, \sigma(t, y) e_j, \sigma(t, y) e_j) \right|$$

$$\begin{aligned} &\leq C(\|\nabla\varphi(x^*)\| + \|\varphi\|_2 + \|\varphi\|_3)e^{-(\lambda-Ch)(T-t)}(1 + \|y - x^*\|^2)\|\sigma(t, y)\|^2 \\ &\leq C(\|\nabla\varphi(x^*)\| + \|\varphi\|_2 + \|\varphi\|_3)e^{-(\lambda-Ch)(T-t)}(1 + \|y - x^*\|^4). \end{aligned}$$

Combining the four upper bounds then yields the inequality (69), for a sufficiently large positive real number $C_4 \in (0, \infty)$ and a sufficiently small positive real number $(0, h_0)$. The proof of Lemma 6.3 is thus completed. \square

REFERENCES

- [1] M. Dambrine, Ch. Dossal, B. Puig, and A. Rondepierre. Stochastic Differential Equations for Modeling First Order Optimization Methods. *SIAM J. Optim.*, 34(2):1402–1426, 2024.
- [2] S. Di Giovacchino, D.J. Higham, and K. Zygalkis. Backward error analysis and the qualitative behaviour of stochastic optimization algorithms: Application to stochastic coordinate descent. *Arxiv preprint*, 2023.
- [3] Y. Feng, T. Gao, Lei Li, J.-G. Liu, and Y. Lu. Uniform-in-time weak error analysis for stochastic gradient descent algorithms via diffusion approximation. *Commun. Math. Sci.*, 18(1):163–188, 2020.
- [4] B. Gess, R. S. Gvalani, and V. Konarovskyi. Stochastic modified flows, mean-field limits and dynamics of stochastic gradient descent. *Arxiv preprint*, 2023.
- [5] P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*, volume 23 of *Stochastic Modelling and Applied Probability*. Springer, Berlin, Heidelberg, 1992.
- [6] Q. Li, C. Tai, and W. E. Stochastic modified equations and dynamics of stochastic gradient algorithms I: mathematical foundations. *J. Mach. Learn. Res.*, 20:Paper No. 40, 47, 2019.
- [7] G. J. Lord, C. E. Powell, and T. Shardlow. *An Introduction to Computational Stochastic PDEs*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2014.
- [8] Y. Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27:372–376, 1983.
- [9] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.
- [10] J. M. Sanz Serna and K. C. Zygalkis. The connections between Lyapunov functions for some optimization algorithms and differential equations. *SIAM J. Numer. Anal.*, 59(3):1542–1565, 2021.
- [11] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Math. Program.*, 195(1–2):79–148, September 2022.
- [12] B. Shi, S. S. Du, W. J. Su, and M. I. Jordan. Acceleration via symplectic discretization of high-resolution differential equations. volume 32, 2019.
- [13] W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling Nesterov’s accelerated gradient method: theory and insights. *J. Mach. Learn. Res.*, 17:Paper No. 153, 43, 2016.
- [14] K. C. Zygalkis. On the existence and the applications of modified equations for stochastic differential equations. *SIAM Journal on Scientific Computing*, 33(1):102–130, 2011.

UNIVERSITE DE PAU ET DES PAYS DE L’ADOUR, E2S UPPA, CNRS, LMAP, PAU, FRANCE
Email address: charles-edouard.brehier@univ-pau.fr

UNIVERSITE DE PAU ET DES PAYS DE L’ADOUR, E2S UPPA, CNRS, LMAP, PAU, FRANCE
Email address: marc.dambrine@univ-pau.fr

UNIVERSITE DE PAU ET DES PAYS DE L’ADOUR, E2S UPPA, CNRS, LMAP, PAU, FRANCE
Email address: nenebbazi@univ-pau.fr