ORIGINAL PAPER

Multi-index antithetic stochastic gradient algorithm

Mateusz B. Majka[1] · **Marc Sabate-Vidales[3]** · **Łukasz Szpruch[2,3]**

Abstract

Stochastic Gradient Algorithms (SGAs) are ubiquitous in computational statistics, machine learning and optimisation. Recent years have brought an influx of interest in SGAs, and the non-asymptotic analysis of their bias is by now well-developed. However, relatively little is known about the optimal choice of the random approximation (e.g mini-batching) of the gradient in SGAs as this relies on the analysis of the variance and is problem specific. While there have been numerous attempts to reduce the variance of SGAs, these typically exploit a particular structure of the sampled distribution by requiring a priori knowledge of its density's mode. In this paper, we construct a Multi-index Antithetic Stochastic Gradient Algorithm (MASGA) whose implementation is independent of the structure of the target measure. Our rigorous theoretical analysis demonstrates that for log-concave targets, MASGA achieves performance on par with Monte Carlo estimators that have access to unbiased samples from the distribution of interest. In other words, MASGA is an optimal estimator from the mean square error-computational cost perspective within the class of Monte Carlo estimators. To illustrate the robustness of our approach, we implement MASGA also in some simple non-log-concave numerical examples, however, without providing theoretical guarantees on our algorithm's performance in such settings.

1 Introduction

Variations of Stochastic Gradient Algorithms (SGAs) are central in many modern machine learning applications such as large scale Bayesian inference (Welling and Teh 2011), variational inference (Hoffman et al. 2013), generative adversarial networks (Goodfellow et al. 2014), variational autoencoders (Kingma and Welling 2013) and deep reinforcement learning (Mnih et al. 2015). Statistical sampling perspective provides a unified framework to study non-asymptotic behaviour of these algorithms, which is the main topic of this work. More precisely, consider a data set $D = (\xi_i)_{i=1}^m \subset \mathbb{R}^n$,

✉ Mateusz B. Majka
  m.majka@hw.ac.uk

  Marc Sabate-Vidales
  m.sabate-vidales@sms.ed.ac.uk

  Łukasz Szpruch
  l.szpruch@ed.ac.uk

[1] School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK

[2] The Alan Turing Institute, London, UK

[3] School of Mathematics, University of Edinburgh, Edinburgh, UK

with $m \in \mathbb{N} \cup \{\infty\}$ and the corresponding empirical measure $\nu^m := \frac{1}{m} \sum_{i=1}^m \delta_{\xi_i}$, where δ is a Dirac measure. Denote by $\mathcal{P}(\mathbb{R}^n)$ the space of all probability measures on $\mathbb{R}^n$ and consider a potential $V : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$. We are then interested in the problem of sampling from the (data-dependent) probability distribution π on $\mathbb{R}^d$, given by

$$\pi(x) \propto \exp\left(-\frac{2}{\beta^2} V(x, \nu^m)\right) dx \qquad (1.1)$$

for some fixed parameter $\beta > 0$. Under some mild assumptions on V, the measure π is a stationary distribution of the (overdamped) Langevin stochastic differential equation (SDE) and classical Langevin Monte Carlo (Dalalyan 2017) algorithms utilise discrete-time counterparts of such SDEs to provide tools for approximate sampling from π, which, however, require access to exact evaluations of $\nabla V(\cdot, \nu^m)$. On the other hand, SGAs take as input a noisy evaluation $\nabla V(\cdot, \nu^s)$ for some $s \in \{1, \dots m\}$. The simplest example of $\nabla V(\cdot, \nu^s)$ utilizes the subsampling with replacement method. Namely, consider a sequence of i.i.d. uniformly distributed random variables $\tau_i^k \sim \text{Unif}(\{1, \cdots, m\})$ for $k \geq 0$ and $1 \leq i \leq s$ and define a sequence of random data batches $D_s^k := (\xi_{\tau_i^k})_{i=1}^s$ and corresponding random measures $\nu_s^k := \frac{1}{s} \sum_{i=1}^s \delta_{\xi_{\tau_i^k}}$ for

$k \geq 0$. Fix a learning rate (time-step) $h > 0$. The corresponding algorithm to sample from (1.1) is given by

$$X_{k+1} = X_k - h\nabla_x V(X_k, \nu_s^k) + \beta\sqrt{h}Z_{k+1}, \qquad (1.2)$$

where $(Z_i)_{i=1}^{\infty}$ are i.i.d. random variables with the standard normal distribution. This method in its simplest form, without mini-batching (i.e., when we use the exact evaluation $\nabla V(\cdot, \nu^m)$) is known in computational statistics as the Unadjusted Langevin Algorithm (ULA) (Durmus and Moulines 2017), but it has numerous more sophisticated variants and alternatives (Brosse et al. 2018; Chatterji et al. 2018; Cornish et al. 2019; Ma et al. 2019; Majka et al. 2020).

Numerical methods based on Euler schemes with inaccurate (randomised) drifts such as (1.2) have recently become an object of considerable interest in both the computational statistics and the machine learning communities (Aicher et al. 2019; Chau and Rasonyi 2019; Deniz Akyildiz and Sabanis 2020; Gao et al. 2018; Ma et al. 2015; Nemeth and Fearnhead 2019; Raginsky et al. 2017; Zou et al. 2019). In particular, the Stochastic Gradient Langevin Dynamics (SGLD) method for approximate sampling from invariant measures of Langevin SDEs has been studied e.g. in Welling and Teh (2011), Teh et al. (2016), Vollmer et al. (2016), Dubey et al. (2016), Brosse et al. (2018), Barkhagen et al. (2020), Chau et al. (2019), Zhang et al. (2019). Furthermore, recall that under some mild assumptions on V, when $\beta \to 0$, the measure π concentrates on the set of minima of V, i.e., on $\{x \in \mathbb{R}^d : x = \arg\inf V(\cdot, \nu^m)\}$, cf. (Hwang 1980). Remarkably, no convexity of V is required for this to be true, which makes SGAs good candidates for a tool for non-convex optimisation. We would like to stress that throughout the paper, in our analysis of algorithm (1.2), we allow for $\beta = 0$ and hence we cover also the Stochastic Gradient Descent (SGD) (Moulines and Bach 2011; Dalalyan and Karagulyan 2019).

Despite the great success of algorithm (1.2) and its various extensions, relatively little is known about how to optimally choose s and whether sub-sampling (or mini-batching) is a good idea at all (Nagapetyan et al. 2017). One of the reasons is that performance analysis appears to be problem-specific as we shall demonstrate on a simple example below. It is clear that subsampling increases the variance of the estimator and induces an additional non-asymptotic bias, see e.g. (Brosse et al. 2018; Nagapetyan et al. 2017). Therefore it is not clear that the reduced computational cost of running the algorithm compensates for these adverse effects. On the other hand, in a big data regime it may be computationally infeasible to use all data points at every step of the gradient algorithm and hence subsampling becomes a necessity.

In the present paper we propose a solution to these challenges by constructing a novel Multi-index Antithetic Stochastic Gradient Algorithm (MASGA). In settings where the measure π in (1.1) is log-concave, we will rigorously

demonstrate that MASGA performs on par with Monte Carlo estimators having access to unbiased samples from the target measure, even though it consists of biased samples. Remarkably, our numerical results in Sect. 3 demonstrate that a good performance can be achieved even in some simple non-log-concave settings. To our knowledge, all current state-of-the-art SGAs (Baker et al. 2019; Cornish et al. 2019) require the user to a priori determine the mode of the target distribution and hence it is not clear how to implement them in non-log-concave settings. This problem is absent with MASGA, whose implementation is independent of the structure of the target measure. Moreover, the analysis in Cornish et al. (2019) is based on the Bernstein-von Mises phenomenon, which describes the asymptotic behaviour of the target measure as $m \to \infty$, and hence their algorithm is aimed explicitly at the big data regime, see Sect. 3 therein. Meanwhile, as we will discuss below, MASGA works well irrespectively of the size of the dataset.

Mean-square error analysis. In the present paper we are studying the problem of computing

$$(f, \pi) := \int_{\mathbb{R}^d} f(x)\pi(dx),$$

for some $f \in L^2(\pi)$. This framework covers the tasks of approximating minima of possibly non-convex functions or the computation of normalising constants in statistical sampling. To this end, the Markov chain specified by (1.2) is used to approximate (f, π) with $\mathbb{E}[f(X_k)]$ for large $k > 0$. More precisely, one simulates $N > 1$ independent copies $(X_k^i)_{k=0}^{\infty}$, for $i \in \{1, \ldots, N\}$, of (1.2), to compute the empirical measure $\mu_{N,k} := \frac{1}{N}\sum_{i=1}^{N} \delta_{X_k^i}$. The usual metric for measuring the performance of such algorithms is the (root) mean square error (MSE). Namely, for any $f \in L^2(\pi)$, $k \geq 1$ and $N \geq 1$, we define

$$\text{MSE}(\mathcal{A}^{f,k,N}) := \left(\mathbb{E}\left|(f, \pi) - (f, \mu_{N,k})\right|^2\right)^{1/2},$$

where $\mathcal{A}^{f,k,N}$ is the algorithm specified by the estimator $(f, \mu_{N,k})$. Then, for a given $\varepsilon > 0$, we look for the optimal number k of steps and the optimal number N of simulated paths, such that for any fixed integrable function f we have $\text{MSE}(\mathcal{A}^{f,k,N}) < \varepsilon$. Note that

$$\text{MSE}(\mathcal{A}^{f,k,N}) \leq |(f, \pi) - (f, \text{Law}(X_k))| + \left(N^{-1}\mathbb{V}[f(X_k)]\right)^{1/2}. \qquad (1.3)$$

If f is Lipschitz with a Lipschitz constant L, then the Kantorovich duality representation of the L^1-Wasserstein distance W_1, see (Villani 2009, Remark 6.5), allows us to upper bound the first term of the right hand side by $LW_1(\text{Law}(X_k), \pi)$. Hence it is possible to control the bias by using the vast literature on such Wasserstein bounds (see

e.g. Brosse et al. 2018; Cheng et al. 2018; Durmus and Eberle 2021; Majka et al. 2020 and the references therein). Controlling the variance, however, is a more challenging task. Before we proceed, let us consider an example.

Motivational example In the context of Bayesian inference, one is interested in sampling from the posterior distribution π on $\mathbb{R}^d$ given by

$$\pi(dx) \propto \pi_0(dx) \prod_{i=1}^{m} \pi(\xi_i|x),$$

where the measure $\pi_0(dx) = \pi_0(x)dx$ is called the prior and $(\xi_i)_{i=1}^m$ are i.i.d. data points with densities $\pi(\xi_i|x)$ for $x \in \mathbb{R}^d$. Note that in this example, for convenience, we assume that the data are conditionally independent given the parameters. See Sect. 4 for more general settings. Taking $\beta = \sqrt{2}$, the potential V in (1.1) becomes

$$V(x, \nu^m) := -\log \pi_0(x) - m \int \log \pi(y|x) \nu^m(dy)$$

$$= -\log \pi_0(x) - \sum_{i=1}^{m} \log \pi(\xi_i|x). \quad (1.4)$$

In stochastic gradient algorithms, one replaces the exact gradient $\nabla V(\cdot, \nu^m)$ in (1.2) with an approximate gradient, constructed by sampling only $s \ll m$ terms from the sum in (1.4). This amounts to considering $V(x, \nu^s) = -\log \pi_0(x) - m \int \log \pi(y|x) \nu^s(dy) = -\log \pi_0(x) - \frac{m}{s} \sum_{i=1}^{s} \log \pi(\xi_{\tau_i^k}|x)$, where $\tau_i^k \sim \text{Unif}(\{1, \ldots, m\})$ for $i = 1, \ldots, s$ are i.i.d. random variables uniformly distributed on $\{1, \ldots, m\}$. Note that for any $x \in \mathbb{R}^d$ the noisy gradient $\nabla_x V(x, \nu^s)$ is an unbiased estimator of $\nabla_x V(x, \nu^m)$. If we choose $\beta = \sqrt{2}$ and the time-step $h/(2m)$ in (1.2), we arrive at the Markov chain

$$X_{k+1} = X_k + \frac{1}{2}h\left(\frac{1}{m}\nabla \log \pi_0(X_k)\right.$$

$$\left. + \frac{1}{s}\sum_{i=1}^{s} \nabla \log \pi(\xi_{\tau_i^k}|X_k)\right) + \frac{1}{\sqrt{m}}\sqrt{h}Z_{k+1}. \quad (1.5)$$

Note that the term in brackets in (1.5) is an unbiased estimator of $-\frac{1}{m}\nabla V$.

The main difficulty in quantifying the cost of Monte Carlo algorithms based on (1.5) stems from the fact that the variance is problem-specific and depends substantially on the interplay between the parameters m, s, h and N. Hence, one may obtain different costs (and thus different answers to the question of profitability of using mini-batching) for different models and different data regimes (Nagapetyan et al. 2017).

In Figs. 1 and 2 we present a numerical experiment for a simple example of a Monte Carlo estimator $(f, \mu_{N,k})$ based on the chain (1.5) with the densities $\pi(\xi_i|x)$ specified by a Bayesian logistic regression model, cf. Sect. 3 for details.

We take $m = 512$, $h = 10^{-2}$ and we simulate up to $t = 2$, hence we have $k = 200$ iterations. On the left-hand side of both figures we can see the estimated MSE for different numbers of paths N and different numbers of samples $s \leq m$, for two different functions f. On the right hand side we can see how the variance changes with s. Evidently, subsampling/mini-batching works better for $f(x) = \log|x|$ than for $f(x) = \exp(x)$, since in the former case it allows us to obtain a small MSE by using just two samples, while in the latter the minimal reasonable number of samples seems to be 16. However, even in this simple example we see that the optimal choice of s and N is far from obvious and very much problem-specific. For an additional discussion on this subject, see Appendix 2.

It has been observed in Nagapetyan et al. (2017), that in some specific regimes there is no benefit of employing the subsampling scheme. The authors of Nagapetyan et al. (2017) also observed that a subsampling scheme utilizing control variates can exhibit improved performance, but again only in some specific regimes (see also Baker et al. 2019; Brosse et al. 2018 for related ideas). Moreover, in order to implement their scheme one has to know the mode of the sampled distribution in advance and hence it is not clear how to adapt it to non-convex settings.

The fact that the analysis of the variance of SGLD (and hence of the computational cost of the algorithm $\mathcal{A}^{f,k,N}$) becomes cumbersome even in the simplest possible examples, clearly demonstrates the need for developing a different algorithm, for which the benefits of subsampling could be proven in a rigorous way for a reasonably large class of models. To this end, we turn towards the Multi-level Monte Carlo (MLMC) technique.

2 Main result

In order to approximate the measure π, we consider a family of (possibly random) measures $(\pi^{\boldsymbol{\ell}})_{\boldsymbol{\ell}\in\mathbb{N}^r}$, where $\mathbb{N}^r$ for $r \geq 1$ is the set of multi-indices $\boldsymbol{\ell} = (\ell_1, \cdots, \ell_r)$, where each $\ell_i \geq 0$ corresponds to a different type of approximation. In this work we focus on the case $r = 2$, with ℓ_1 dictating the number of subsamples at each step of the algorithm and ℓ_2 the time discretisation error. Note that for any $i \in \mathbb{N}$, each possible value of ℓ_i corresponds to a different level of approximation within the given type of approximation, i.e., in our main example $\ell_1 = k$ will correspond to the k-th level of approximation with respect to the number of subsamples, where we use, say, s_k subsamples out of the total number of m samples. Note that for some types of approximation, ℓ_i can have the highest possible value (corresponding to the best possible approximation), such as $\ell_1 = m$ for the number of subsamples, making the total possible number of approximation levels finite. However, for other types of approximation
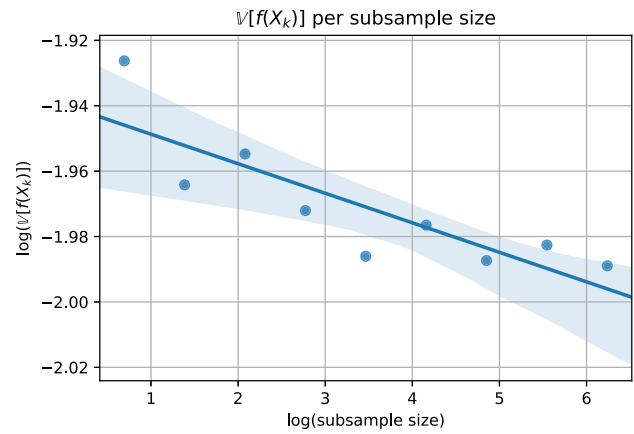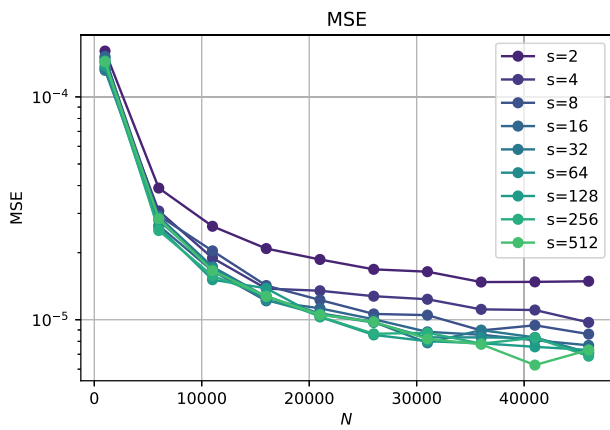
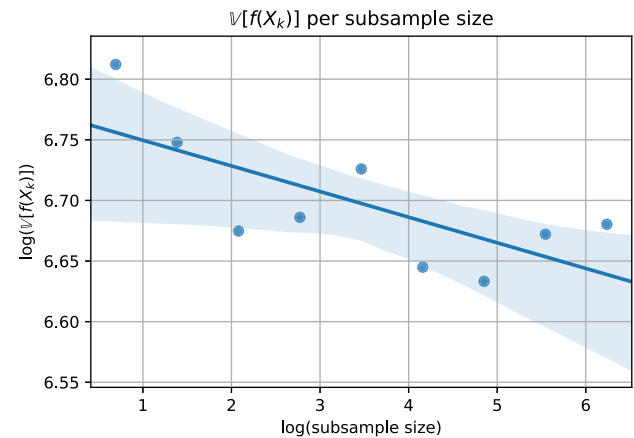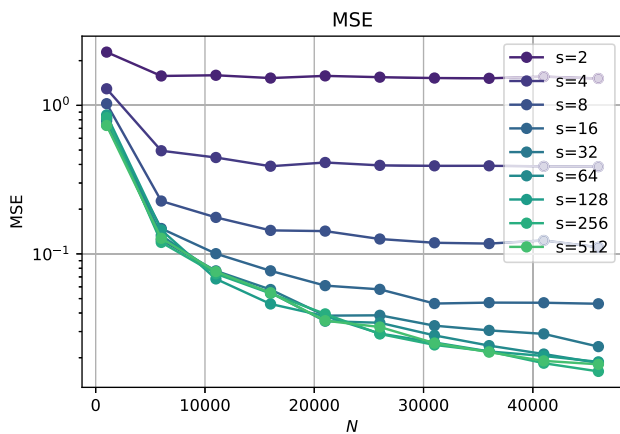**Fig. 1** MSE and Variance $\mathbb{V}[f(X_k)]$ with $f(x) = \log|x|$

Fig. 2 MSE and Variance $\mathbb{V}[f(X_k)]$ with $f(x) = \exp(x)$

(such as with respect to the time discretisation error), we can keep refining the approximation parameter indefinitely and hence have an infinite number of approximation levels. While for a fixed ℓ the samples are biased (if at least one type of approximation has an infinite number of possible approximation levels), we work with methods that are asymptotically unbiased in the sense that

$$\lim_{\ell \to \infty} \mathbb{E}(f, \pi^\ell) = (f, \pi), \tag{2.1}$$

where $\ell \to \infty$ means that $\min_{i \in 1,\dots,r} \ell_i \to \infty$. Note that we need to take the expectation in (2.1) since the measures π^ℓ can be random. Here we use the convention that if for a given type of approximation $\ell_i = (\ell_i^j)_{j=1}^\infty$, there is only a finite number, say K, of possible approximation levels, then for any $j \geq K$ we have $\ell_i^j = \infty$, since then the bias with respect to that particular type of approximation is eliminated. We also remark that as the coordinates of ℓ increase and the bias decreases, the corresponding computational cost of MLMC increases. One therefore faces an optimisation problem and tries to obtain the minimal computational cost for

a prescribed accuracy (or, equivalently, to minimise the bias for a fixed computational budget).

It turns out, perhaps surprisingly, that a Multi-level Monte Carlo estimator that consists of a hierarchy of biased approximations can achieve computational efficiency of vanilla Monte Carlo built from directly accessible unbiased samples (Giles 2008, 2015). In order to explain this approach, let us define backward difference operators

$$\Delta_s \pi^\ell := \pi^\ell - \pi^{\ell - e_s}, \quad \boldsymbol{\Delta} \pi^\ell := \left(\prod_{s=1}^r \Delta_s \right) \pi^\ell,$$

where e_s is the unit vector in the direction $s \in \{1 \dots, r\}$ and $\prod_{s=1}^r \Delta_s$ denotes the concatenation of operators. The core idea of MLMC is to observe that thanks to (2.1),

$$(f, \pi) = \mathbb{E} \sum_{\ell \in \mathbb{N}^r} (f, \boldsymbol{\Delta} \pi^\ell), \tag{2.2}$$

where we set $\pi^{\mathbf{0} - e_s} := 0$ for all unit vectors e_s, with $\mathbf{0} = (0, \dots, 0)$. The original MLMC (Giles 2015) has been developed for $r = 1$, and the extension to an arbitrary r (named Multi-index Monte Carlo, or MIMC) has been developed in Haji-Ali et al. (2016). In MIMC we approximate

each term π^ℓ on the right-hand side of (2.2) using mutually independent, unbiased Monte Carlo estimators π^{ℓ,N_ℓ} with $N_\ell \geq 1$ samples each, and we choose a finite set of multi-indices $\mathcal{L} \subset \mathbb{N}^r$ to define

$$\mathcal{A}^{\mathcal{L}}(f) := \sum_{\ell \in \mathcal{L}} (f, \boldsymbol{\Delta}\pi^{\ell,N_\ell}). \qquad (2.3)$$

Clearly, $\mathbb{E}(f, \boldsymbol{\Delta}\pi^{\ell,N_\ell}) = \mathbb{E}(f, \boldsymbol{\Delta}\pi^\ell)$ and hence $\mathcal{A}^{\mathcal{L}}(f)$ is an asymptotically unbiased estimator of (f, π) when $\mathcal{L}$ increases to $\mathbb{N}^r$. Moreover, we note that due to the independence of π^{ℓ,N_ℓ} across levels, the variance of the MIMC estimator satisfies $\mathbb{V}[\mathcal{A}^{\mathcal{L}}(f)] = \sum_{\ell \in \mathcal{L}} \mathbb{V}[(f, \boldsymbol{\Delta}\pi^{\ell,N_\ell})]$.

In this work we develop an antithetic extension of the MIMC algorithm. To this end we define pairs $(\pi^{+,\ell}, \pi^{-,\ell})$ of copies of π^ℓ in the sense that $\mathbb{E}[(f, \pi^\ell)] = \mathbb{E}[(f, \pi^{+,\ell})] = \mathbb{E}[(f, \pi^{-,\ell})]$ for all Lipschitz functions f, and

$$\Delta_s^A \pi^\ell := \pi^\ell - \tfrac{1}{2}(\pi^{+,\ell-e_s} + \pi^{-,\ell-e_s}), \quad \boldsymbol{\Delta}^A \pi^\ell := \left(\prod_{s=1}^r \Delta_s^A\right) \pi^\ell. \qquad (2.4)$$

The corresponding Antithetic MIMC estimator is given by

$$\mathcal{A}^{A,\mathcal{L}}(f) := \sum_{\ell \in \mathcal{L}} (f, \boldsymbol{\Delta}^A \pi^{\ell,N_\ell}). \qquad (2.5)$$

As we will see in the sequel, the reduction of the variance of $\mathcal{A}^{A,\mathcal{L}}(f)$ in comparison to $\mathcal{A}^{\mathcal{L}}(f)$ can be achieved as a consequence of an appropriately chosen coupling between π^ℓ, $\pi^{+,\ell-e_s}$ and $\pi^{-,\ell-e_s}$ for each $\ell \in \mathcal{L}$ and $s \in \{1, \dots, r\}$. Using the notation introduced in (1.2), we will apply (2.5) to the specific case of $\ell = (\ell_1, \ell_2)$ and π^ℓ given as the law of $X_k^{\ell_1,\ell_2}$ for some fixed $k \geq 1$, where

$$X_{k+1}^{\ell_1,\ell_2} = X_k^{\ell_1,\ell_2} - h_{\ell_2} \nabla_x V(X_k^{\ell_1,\ell_2}, \nu^{s_{\ell_1}}) + \beta\sqrt{h_{\ell_2}} Z_{k+1}. \qquad (2.6)$$

In this setting, we call the algorithm $\mathcal{A}^{A,\mathcal{L}}(f)$ specified by (2.5) the Multi-index Antithetic Stochastic Gradient Algorithm (MASGA). Note that for a fixed $t > 0$ such that $t = k h_{\ell_2}$ for some $k \geq 1$, the chain (2.6) can be interpreted as a discrete-time approximation, both in time with parameter h but also in data with parameter s, of the SDE

$$dY_t = -\nabla_x V(Y_t, \nu^m) dt + \beta dW_t, \qquad (2.7)$$

where $(W_t)_{t \geq 0}$ is the standard Brownian motion in $\mathbb{R}^d$. Then, MASGA with π^ℓ given as the law of $X_k^{\ell_1,\ell_2}$, for any finite subset $\mathcal{L} \subset \mathbb{N}^2$ provides a biased estimator of (f, π_t) (where $\pi_t := \mathrm{Law}(Y_t)$ with Y_t given by (2.7)), where the bias stems from the use of a finite number of levels. However, since π given by (1.1) is the limiting stationary distribution of (2.7), $\mathcal{A}^{A,\mathcal{L}}(f)$ can be also interpreted as a biased estimator of (f, π), with an additional bias coming from the difference

between (f, π) and (f, π_t) due to the simulation up to a finite time.

Note that the construction of MASGA does not require any knowledge of the structure of the target measure, such as the location of its modes (Baker et al. 2019; Brosse et al. 2018; Nagapetyan et al. 2017), or any properties of the potential V. However, in order to formulate the main result of this paper, we will use the following set of assumptions.

Assumptions 2.1 Let the potential $V : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ be of the form $V(x, \nu^m) := v_0(x) + \int_{\mathbb{R}^k} v(x, y)\nu^m(dy) = v_0(x) + \frac{1}{m}\sum_{i=1}^m v(x, \xi_i)$, where $(\xi_i)_{i=1}^m \subset \mathbb{R}^k$ is the data and the functions $v_0 : \mathbb{R}^d \to \mathbb{R}$ and $v : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ are such that

i) For all $\xi \in \mathbb{R}^k$ we have $\nabla v(\cdot, \xi), \nabla v_0(\cdot) \in C_b^2(\mathbb{R}^d; \mathbb{R}^d)$, i.e., the gradients of v and v_0 are twice continuously differentiable with all partial derivatives of the first and second order bounded (but the gradients themselves are not necessarily bounded).

ii) There exists a constant $K > 0$ such that for all $\xi \in \mathbb{R}^k$ and for all $x, y \in \mathbb{R}^d$ we have

$$\langle x - y, \nabla_x v(x, \xi) - \nabla_y v(y, \xi)\rangle \geq K|x - y|^2 \quad \text{and}$$
$$\langle x - y, \nabla v_0(x) - \nabla v_0(y)\rangle \geq K|x - y|^2.$$

Note that the first condition above in particular implies that the gradient $\nabla V(\cdot, \nu^m)$ of the potential is globally Lipschitz. Furthermore, as a consequence of Assumption 2.1(i), for all $\xi \in \mathbb{R}^k$ the gradients $\nabla v(\cdot, \xi)$, as well as the gradient $\nabla v_0(\cdot)$ satisfy a standard linear growth condition $|\nabla_x v(x, \xi)| \leq c(1 + |x|)$ and $|\nabla v_0(x)| \leq c(1 + |x|)$ with a constant $c > 0$. Using a standard inequality $(a + b)^4 \leq 8a^4 + 8b^4$, we can easily conclude that there exists a constant $C > 0$ such that for all $\xi \in \mathbb{R}^k$ and for all $x \in \mathbb{R}^d$ we have

$$|\nabla_x v(x, \xi)|^4 \leq C(1 + |x|^4) \quad \text{and} \quad |\nabla v_0(x)|^4 \leq C(1 + |x|^4).$$

This growth condition will be important in our proofs in Sect. 4. We further remark that Assumption 2.1(ii) implies that π given via (1.1) is log-concave. Note also that the Bayesian inference example given in (1.5) satisfies Assumptions 2.1 if the functions $x \mapsto -\nabla \log \pi(\xi|x)$ satisfy all the respective regularity conditions. We remark that we formulate our main result in this section only for the specific form of V given above just for convenience. Our result holds also for a much more general class of potentials, but the assumptions for the general case are more cumbersome to formulate and hence we postpone their presentation to Sect. 4. Moreover, we stress that assuming convexity of V is not necessary for the construction of our algorithm and it is a choice we made solely to simplify the proofs. By combining our approach with the coupling techniques from (Majka et al. 2020), it

should be possible to extend our results to the non-convex case. This, however, falls beyond the scope of the present paper and is left for future work. We have the following result.

Theorem 2.2 *Let Assumptions 2.1 hold. Let $\mathcal{A}^{A,\mathcal{L}}$ be the MASGA estimator defined in (2.5) with $\pi^{\boldsymbol{\ell}}$ given as the law of $X_k^{\ell_1,\ell_2}$, defined via (2.6) for some fixed $k \geq 1$. Then there exists a set of indices $\mathcal{L}$ and a sequence $(N_{\boldsymbol{\ell}})_{\boldsymbol{\ell} \in \mathcal{L}}$ such that for any $\varepsilon > 0$ and for any Lipschitz function f, the estimator $\mathcal{A}^{A,\mathcal{L}}(f)$ requires the computational cost of order ε^{-2} to achieve mean square error ε for approximating (f, π_t) with $t = kh_{\ell_2}$, where $\pi_t := \mathrm{Law}(Y_t)$ with Y_t given by (2.7).*

The proof of Theorem 2.2 and an explicit description of the sequence $(N_{\boldsymbol{\ell}})_{\boldsymbol{\ell} \in \mathcal{L}}$, as well as all the details of the relaxed assumptions that can be imposed on V, will be given in Sect. 4. Note that requiring computational cost of order ε^{-2} to achieve MSE $< \varepsilon$ is the best performance that one can expect from Monte Carlo methods, cf. (Giles 2015).

The two quantities that feature in the optimization problem formulated in Theorem 2.2 are the MSE and the computational cost (i.e., we want to optimize the cost given the constraint MSE $< \varepsilon$ for a fixed $\varepsilon > 0$). Note that the cost is explicitly defined to be

$$\mathrm{cost}(\mathcal{A}^{A,\mathcal{L}}(f)) := \sum_{\boldsymbol{\ell} \in \mathcal{L}} N_{\boldsymbol{\ell}} C_{\boldsymbol{\ell}}, \quad C_{\boldsymbol{\ell}} := t h_{\ell_2}^{-1} s_{\ell_1}, \qquad (2.8)$$

where $C_{\boldsymbol{\ell}}$ is the cost of each path at the level $\boldsymbol{\ell}$, i.e., the product of $k = t h_{\ell_2}^{-1}$ steps and s_{ℓ_1} subsamples for a given level $\boldsymbol{\ell} = (\ell_1, \ell_2)$. On the other hand, from our discussion on MSE (1.3) it is evident that in order to control $\mathrm{MSE}(\mathcal{A}^{A,\mathcal{L}}(f))$, it is crucial to find upper bounds on the variance $\mathbb{V}[(f, \boldsymbol{\Delta}^A \pi^{\boldsymbol{\ell}})]$ and the bias of the estimator at each level $\boldsymbol{\ell}$. In our proof we will in fact rely on the complexity analysis from (Haji-Ali et al. 2016) (see also Giles 2015 and Theorem 4.12 below for more details), which is concerned exactly with such optimization problems.

For convenience, from now on we will assume that in our MASGA estimator (2.5), both the number of subsamples and the discretisation parameter are rescaled by two when moving between levels. More precisely, for a fixed $s_0 \in \mathbb{N}_+$ and $h_0 > 0$, we assume that $s_{\ell_1} = 2^{\ell_1} s_0$ and $h_{\ell_2} = 2^{-\ell_2} h_0$ for all $\boldsymbol{\ell} = (\ell_1, \ell_2) \in \mathbb{N}^2$. In this setting, the complexity analysis from (Haji-Ali et al. 2016) tells us, roughly speaking (with more details in Sect. 4 below) that in order to obtain $\mathrm{MSE}(\mathcal{A}^{A,\mathcal{L}}(f)) < \varepsilon$ with $\mathrm{cost}(\mathcal{A}^{A,\mathcal{L}}(f)) < \varepsilon^{-2}$, we need to have for each $\boldsymbol{\ell} \in \mathbb{N}^2$,

$$\mathbb{E}[(f, \boldsymbol{\Delta}^A \pi^{\boldsymbol{\ell}})^2] \lesssim 2^{-\langle \boldsymbol{\beta}, \boldsymbol{\ell} \rangle} \qquad \text{and} \qquad C_{\boldsymbol{\ell}} \lesssim 2^{\langle \boldsymbol{\gamma}, \boldsymbol{\ell} \rangle}, \qquad (2.9)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2)$ and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2) \in \mathbb{R}^2$ are such that $\gamma_1 < \beta_1$ and $\gamma_2 < \beta_2$. Note that the bound on $\mathbb{E}[(f, \boldsymbol{\Delta}^A \pi^{\boldsymbol{\ell}})^2]$ trivially implies a bound on $\mathbb{V}[(f, \boldsymbol{\Delta}^A \pi^{\boldsymbol{\ell}})]$ of the same order (i.e., with the same $\boldsymbol{\beta}$), as well as the bound $\mathbb{E}[(f, \boldsymbol{\Delta}^A \pi^{\boldsymbol{\ell}})] \lesssim$

$2^{-\langle \boldsymbol{\alpha}, \boldsymbol{\ell} \rangle}$ with $\boldsymbol{\alpha} = \boldsymbol{\beta}/2$, which turns out to be crucial in the complexity analysis from (Haji-Ali et al. 2016) and is the reason why it suffices to verify (2.9), cf. also Theorem 2 in Giles (2015) and Theorem 4.12 below. Since, straight from the definition of $C_{\boldsymbol{\ell}}$, it is clear that in our setting we have $\boldsymbol{\gamma} = (1, 1)$ (recall that $C_{\boldsymbol{\ell}} \lesssim s_{\ell_1} h_{\ell_2}^{-1} \lesssim 2^{\ell_1 + \ell_2}$), we can infer that in order to prove Theorem 2.2 all we have to do is to find an upper bound on $\mathbb{E}[(f, \boldsymbol{\Delta}^A \pi^{\boldsymbol{\ell}})^2]$ proportional to $2^{-\langle \boldsymbol{\beta}, \boldsymbol{\ell} \rangle}$ with $\boldsymbol{\beta} = (\beta_1, \beta_2)$ such that $\beta_1 > 1$ and $\beta_2 > 1$. In the proof of Theorem 2.2 we will in fact obtain $\boldsymbol{\beta} = (2, 2)$. However, the crucial difficulty in our argument will be to ensure that our upper bound is indeed of the product form, i.e., that we obtain $\mathbb{E}[(f, \boldsymbol{\Delta}^A \pi^{\boldsymbol{\ell}})^2] \lesssim s_{\ell_1}^{-2} h_{\ell_2}^2 \lesssim 2^{-2\ell_1 - 2\ell_2}$.

Further extensions

The assertion of Theorem 2.2 states that $\mathcal{A}^{A,\mathcal{L}}(f)$, interpreted as an estimator of (f, π_t), requires computational cost of order ε^{-2} to achieve mean square error ε. Since the difference between (f, π_t) and (f, π) is of order $\mathcal{O}(e^{-\lambda t})$ for some $\lambda > 0$ (cf. the discussion in Appendix 2), this means that $\mathcal{A}^{A,\mathcal{L}}(f)$ interpreted as an estimator of (f, π), requires computational cost of order $\varepsilon^{-2} \log(\varepsilon^{-1})$ to achieve mean square error ε. However, $\mathcal{A}^{A,\mathcal{L}}(f)$ could be further modified in order to remove the log term, by employing the MLMC in terminal time technique introduced in Giles et al. (2020), cf. Sect. 2.3 and Remark 3.8 therein. This would involve taking $r = 3$ in (2.5) and modifying the definition as

$$\mathcal{A}^{A,\mathcal{L}}(f) := \sum_{\boldsymbol{\ell} \in \mathcal{L}} (f, \Delta_1^A \Delta_2^A \Delta_3 \pi^{\boldsymbol{\ell}, N_{\boldsymbol{\ell}}}), \qquad (2.10)$$

i.e., we would take $\pi^{\boldsymbol{\ell}}$ with $\boldsymbol{\ell} = (\ell_1, \ell_2, \ell_3)$ to be the law of $X_{k_{\ell_3}}^{\ell_1, \ell_2}$ given by (2.6), hence we would introduce a sequence of terminal times $t := k_{\ell_3} h_{\ell_2}$ for the chain (2.6), changing at each level. However, in the definition (2.10) of $\mathcal{A}^{A,\mathcal{L}}(f)$ we would use the antithetic difference operators Δ^A only with respect to the subsampling level parameter ℓ_1 and the discretisation level parameter ℓ_2, while applying the plain difference operator Δ to the terminal time level parameter ℓ_3. The details of how to construct the sequence of terminal times $t := k_{\ell_3} h_{\ell_2}$ can be found in Giles et al. (2020). We skip this modification in the present paper in the attempt to try to keep the notation as simple as possible.

In the setting where the computational complexity is ε^{-2}, (as it is for MASGA), it is possible to easily modify the biased estimator $\mathcal{A}^{A,\mathcal{L}}(f)$ to obtain its unbiased counterpart. Indeed, let $\boldsymbol{M} = (M_1, \ldots, M_r)$ be a random variable on $\mathbb{N}^r$, independent of $(\boldsymbol{\Delta}^A \pi^{\boldsymbol{\ell}, N_{\boldsymbol{\ell}}})_{\boldsymbol{\ell} \in \mathcal{L}}$. Define $\mathcal{L}^{\boldsymbol{M}} := \{\boldsymbol{\ell} \in \mathbb{N}^r : \ell_1 \leq M_1, \cdots, \ell_r \leq M_r\}$ and

$$\mathcal{A}^{UA, \mathcal{L}^{\boldsymbol{M}}}(f) := \sum_{\boldsymbol{\ell} \in \mathcal{L}^{\boldsymbol{M}}} \frac{(f, \boldsymbol{\Delta}^A \pi^{\boldsymbol{\ell}, N_{\boldsymbol{\ell}}})}{\mathbb{P}(\boldsymbol{M} \geq \boldsymbol{\ell})} = \sum_{\boldsymbol{\ell} \in \mathbb{N}^r} 1_{\{\boldsymbol{M} \geq \boldsymbol{\ell}\}} \frac{(f, \boldsymbol{\Delta}^A \pi^{\boldsymbol{\ell}, N_{\boldsymbol{\ell}}})}{\mathbb{P}(\boldsymbol{M} \geq \boldsymbol{\ell})},$$

where $\boldsymbol{M} \geq \boldsymbol{\ell}$ is understood component-wise. One can see that $\mathcal{A}^{UA, \mathcal{L}^{\boldsymbol{M}}}(f)$ is then an unbiased estimator of (f, π).

Indeed,

$$\mathbb{E}\big[\mathcal{A}^{UA,\mathcal{L}^M}(f)\big] = \mathbb{E}\big[\textstyle\sum_{\boldsymbol{\ell}\in\mathbb{N}^r}(f, \boldsymbol{\Delta}^A\pi^{\boldsymbol{\ell},N_{\boldsymbol{\ell}}})\big] = (f, \pi),$$

due to (2.2). It turns out that in order for the variance of $\mathcal{A}^{UA,\mathcal{L}^M}(f)$ to be finite, we need to be in the regime where the computational complexity of the original estimator is ε^{-2}. We refer the reader to (Crisan et al. 2018; Rhee and Glynn 2015) for more details and recipes for constructing $\boldsymbol{M}$. The methods from (Rhee and Glynn 2015) have recently been extended to more general classes of MCMC algorithms in Jacob et al. (2020), which contains further discussion of the benefits and costs of debiasing.

Literature review The idea of employing MLMC apparatus to improve efficiency of stochastic gradient algorithms has been studied before. In Giles et al. (2020) we introduced a Multi-level Monte Carlo (MLMC) method for SGLD in the global convexity setting, based on a number of decreasing discretisation levels. We proved that under certain assumptions on the variance of the estimator of the drift, this technique can indeed improve the overall performance of the Monte Carlo method with stochastic gradients. In Majka et al. (2020) we extended our approach to cover the non-convex setting, allowing for sampling from probability measures that satisfy a weak log-concavity at infinity condition. However, the computational complexity of such algorithms is sub-optimal. As we will observe in Sect. 3 with numerical experiments, the crucial insight of the present paper (and a novelty compared to Giles et al. 2020; Majka et al. 2020) is the application of MLMC with respect to the subsampling parameter. Note that, in a different context, the idea to apply MLMC to stochastic approximation algorithms has been studied in Frikha (2016), see also (Dereich 2021; Dereich and Müller-Gronbach 2019).

At the core of our analysis of Multi-level Monte Carlo estimators lies the problem of constructing the right couplings between Euler schemes (2.6) on different discretisation levels. For Euler schemes with standard (accurate) drifts this is done via a one-step analysis by coupling the driving noise in a suitable way, cf. Sects. 2.2 and 2.5 in Majka et al. (2020) and Sect. 2.4 in Eberle and Majka (2019). However, in the case of SGLD, one is faced with an additional problem of coupling the drift estimators used on different discretisation levels. In both (Giles et al. 2020) and (Majka et al. 2020) we addressed this issue in the simplest possible way, by coupling the drift estimators independently. In the present paper we show that by employing a non-trivial coupling we can substantially reduce the variance of MLMC and thus obtain the required bound $\mathbb{V}[(f, \boldsymbol{\Delta}^A\pi^{\boldsymbol{\ell}})] \lesssim 2^{-2\ell_1-2\ell_2}$ as explained above. We achieve this by utilising the antithetic approach to MLMC as defined in (2.4). Related ideas for antithetic multi-level estimators have been used e.g. in Giles and Szpruch (2013), Giles and Szpruch (2014), Szpruch and

Tse (2019). However, in the present paper we apply this concept for the first time to Euler schemes with inaccurate drifts. We also remark that, due to our bounds on second moments, we can easily derive confidence intervals for MASGA using Chebyshev's inequality. However, it would also be possible to derive a Central Limit Theorem and corresponding concentration inequalities, in the spirit of Ben Alaya and Kebaier (2015), Ben Alaya et al. (2020), Jourdain and Kebaier (2019), Kebaier (2005).

The remaining part of this paper is organised as follows. In Sect. 3 we present numerical experiments confirming our theoretical findings. In Sect. 4 we present a more general framework for the MASGA estimator, we explain the intuition behind the antithetic approach to MLMC in more detail (see Example 4.2) and we formulate the crucial Lemma 4.13. We also explain how to prove Theorem 2.2 based on Lemma 4.13. In Sect. 5 we prove Lemma 4.13 in a few steps: we first discuss the antithetic estimator with respect to the discretisation parameter, which corresponds to taking $r = 1$ and $\boldsymbol{\ell} = \ell_2$ in (2.5), then we consider the antithetic estimator with respect to the subsampling parameter, which corresponds to taking $r = 1$ and $\boldsymbol{\ell} = \ell_1$ in (2.5) and, finally, we explain how these approaches can be combined in a multi-index estimator with $r = 2$ and $\boldsymbol{\ell} = (\ell_1, \ell_2)$ to prove Lemma 4.13. Several technical proofs are postponed to Appendices.

3 Numerical experiments

We showcase the performance of the MASGA estimator in Bayesian inference problems, combining different Bayesian models and priors. The code for all the numerical experiments can be found at https://github.com/msabvid/MLMC-MIMC-SGD.

In Sect. 3.1 we compare the MASGA estimator introduced in Sect. 2 with an Antithetic Multi-level Monte Carlo (AMLMC) estimator with respect to the subsampling parameter, corresponding to taking $r = 1$ and $\boldsymbol{\ell} = \ell_1$ in (2.5). We demonstrate that MASGA indeed achieves the optimal computational complexity. Both these estimators are also compared to a standard Monte Carlo estimator for reference. As we shall see, while the performance of MASGA in our experiments is always better than that of AMLMC, the difference is not substantial. This suggests that, from the practical standpoint, the crucial insight of this paper is the application of the antithetic MLMC approach with respect to the subsampling parameter. Hence in our subsequent experiments in Sects. 3.2 and 3.3, we will focus on the AMLMC estimator, which is easier to implement than MASGA.

Note that this phenomenon is partially explained by our computations, which show that the variance of AMLMC with respect to subsampling is already of the optimal order in both

parameters, i.e., $\mathcal{O}(h^2 s^{-2})$, see Sect. 5 for details. However, based on the theory of MIMC, in order to prove the optimal computational complexity, we need to use AMLMC with respect to subsampling in conjunction with AMLMC with respect to discretisation as a multi-index estimator.

In Sect. 3.2 we will check the performance of AMLMC with respect to subsampling in both convex and non-convex settings, whereas in Sect. 3.3 we will compare it to the Stochastic Gradient Langevin Dynamics with Control Variates (SGLD-CV) method introduced in Baker et al. (2019). More precisely, in Sect. 3.3 we present an example of a convex setting in which AMLMC outperforms SGLD-CV. It is worth pointing out, that the latter method can be applied only to convex settings (similarly as the method from (Cornish et al. 2019)), whereas AMLMC is free of such limitations.

We would like to remark that our main point in implementing AMLMC in non-convex settings is to demonstrate the robustness of our approach and show that AMLMC can perform well in some simple examples where other methods (Baker et al. 2019; Cornish et al. 2019) are not even implementable. However, we do not guarantee (and do not expect) the optimal performance of AMLMC in such settings. We believe that in order to ensure the optimal performance of our algorithms in non-convex settings, we would have to re-design MASGA by employing a different coupling. More precisely, our current algorithm is based on coupling multiple chains in a synchronous way, which is well-suited to the convex setting, but not necessarily to non-convex settings. Adapting our method to non-convex settings would require a re-design of the algorithm by utilising a combination of the maximal and the reflection couplings, as described for ULA and SGLD in Eberle and Majka (2019), Majka et al. (2020). This is, however, highly non-trivial (since MASGA relies on coupling of multiple chains and not just two), and will require a lot of additional work, hence falls beyond the scope of the present paper.

3.1 MASGA and AMLMC with respect to subsampling

Let us begin by introducing the Bayesian logistic regression setting that we will use for our simulations. The data is modelled by

$$p(y_i | \iota_i, x) = g(y_i x^T \iota_i) \tag{3.1}$$

with $g(z) = 1/(1 + e^{-z})$, $z \in \mathbb{R}$, where $x \in \mathbb{R}^d$ are the parameters of the model that need to be sampled from their posterior distribution, ι_i denotes an observation of the predictive variable in the data, and y_i the binary target variable. Given a dataset of size m, by Bayes' rule, the posterior density of x satisfies

$$\pi(x) \propto \pi_0(x) \prod_{i=1}^m g(y_i x^T \iota_i). \tag{3.2}$$

In our experiments, we will consider two different priors π_0, namely

i) a Gaussian prior $\pi_0 \sim \mathcal{N}(0, I)$.
ii) a mixture of two Gaussians $\pi_0 \sim \frac{1}{2}\mathcal{N}(0, I) + \frac{1}{2}\mathcal{N}(1, I)$.

We use Algorithm 1 to empirically calculate the cost of approximating $\mathbb{E}(f(X))$, for a function $f : \mathbb{R}^d \to \mathbb{R}$, where the law of X is the posterior π, by the MASGA estimator $\mathcal{A}_{MASGA}^{A,\mathcal{L}}(f) = \mathcal{A}^{A,\mathcal{L}}(f)$ defined in (2.5), such that its MSE is under some threshold ε. Recall that in our notation $(f, \boldsymbol{\Delta}^A \pi^{\boldsymbol{\ell}})$ denotes the integral of f with respect to the antithetic measure $\Delta^A \pi^{\boldsymbol{\ell}}$ given by (2.4), where $\pi^{\boldsymbol{\ell}}$ is specified by the law of the Markov chain (2.6) with the potential V determined from (3.2) in an analogous way as in the Bayesian inference example (1.4). More explicitly, we have

$$V(x, \nu^m) = -\log \pi_0(x) - \sum_{i=1}^m \log g(y_i x^T \iota_i)$$

and

$$
\begin{aligned}
X_{k+1}^{\ell_1, \ell_2} = {} & X_k^{\ell_1, \ell_2} + \tfrac{1}{2} h_{\ell_2} \left(\tfrac{1}{m} \nabla \log \pi_0(X_k^{\ell_1, \ell_2}) \right. \\
& \left. + \frac{1}{s_{\ell_1}} \sum_{i=1}^{s_{\ell_1}} \nabla \log g \left(y_{\tau_i^k} (X_k^{\ell_1, \ell_2})^T \iota_{\tau_i^k} \right) \right) + \frac{1}{\sqrt{m}} \sqrt{h_{\ell_2}} Z_{k+1},
\end{aligned}
\tag{3.3}
$$

where τ_i^k for $k \geq 1$ and $i \in \{1, \ldots, s_{\ell_1}\}$ can correspond to subsampling either with or without replacement (in our simulations we choose the latter).

Below we present the results of our experiments for $f(x) = |x|^2$ (we would like to remark that we obtained similar conclusions for $f(x) = |x|$ and hence we skip the latter example to save space).

Furthermore, for the Bayesian Logistic regressions we use the covertype dataset (Blackard and Dean 1999) which has 581 012 observations, and 54 columns[1]. We create a training set containing 20% of the original observations.

On the other hand, for AMLMC with respect to subsampling (denoted below by $\mathcal{A}_{MLMC}^{A,\mathcal{L}}(f)$) we take $r = 1$ and $\ell = \ell_1$ in (2.5). This corresponds to using a fixed discretisation parameter h and applying the antithetic MLMC estimator only with respect to the subsampling parameter.

Note that in our experiments we apply the estimators $\mathcal{A}_{MASGA}^{A,\mathcal{L}}(f)$ and $\mathcal{A}_{MLMC}^{A,\mathcal{L}}(f)$ to approximate (f, π_t^h),

---

[1] In order to perform a Bayesian logistic regression, the categorical variable specifying the forest type designation is aggregated into a binary variable and is used as the target variable y_i in the model.

Algorithm 1 MASGA

Initialisation: n, L, ε

Calculate n samples $(f, \mathbf{\Delta}^A \pi^{\boldsymbol{\ell}})^{(j)}$ of $(f, \mathbf{\Delta}^A \pi^{\boldsymbol{\ell}})$ independently, $j = 1, \ldots, n$ for each multi-index level in $\mathcal{L} := \{\boldsymbol{\ell} = (\ell_1, \ell_2) : \ell_1, \ell_2 = 0, \ldots, L\}$.

Calculate the MASGA estimator $\mathcal{A}^{A,\mathcal{L}}(f)$ as in (2.5) with $N_{\boldsymbol{\ell}} = n$, i.e., $\mathcal{A}^{A,\mathcal{L}}(f) = \sum_{\boldsymbol{\ell} \in \mathcal{L}} (f, \mathbf{\Delta}^A \pi^{\boldsymbol{\ell},n})$, where for each $\boldsymbol{\ell} \in \mathcal{L}$ we take $(f, \mathbf{\Delta}^A \pi^{\boldsymbol{\ell},n}) := n^{-1} \sum_{j=1}^{n} (f, \mathbf{\Delta}^A \pi^{\boldsymbol{\ell}})^{(j)}$.

while MSE $> \varepsilon$ **do**

    Calculate the number of paths in each level ℓ given by

$$N_{\ell_1,\ell_2} = \varepsilon^{-2} \left(\sqrt{\frac{\mathbb{V}[(f, \mathbf{\Delta}^A \pi^{(\ell_1,\ell_2)})]}{C_{(\ell_1,\ell_2)}}} \sum_{\boldsymbol{\ell} \in \mathcal{L}} \sqrt{\mathbb{V}[(f, \mathbf{\Delta}^A \pi^{\boldsymbol{\ell}})] C_{\boldsymbol{\ell}}} \right)$$

    (see Theorem 4.11 for details) and calculate extra $N_{\ell_1,\ell_2} - n$ samples in each multi-index level.

    Update $\mathcal{A}^{A,\mathcal{L}}(f)$ and estimate its bias (see Giles 2015; Haji-Ali et al. 2016).

    if bias estimate is less than $\varepsilon/2$ **then**

        Set convergence=True, and stop.

    else

        Set $L := L + 1$, and calculate n samples of $(f, \mathbf{\Delta}^A \pi^{\boldsymbol{\ell}})$ for $\boldsymbol{\ell} = (\ell_1, L), \ell_1 \leq L$ and for $\boldsymbol{\ell} = (L, \ell_2), \ell_2 \leq L$.

    end if

end while

return $\mathcal{A}^{A,\mathcal{L}}(f)$ and the number of samples N_{ℓ_1,ℓ_2} in each level.

---

where π_t^h is the law given by the chain X_k with $t = kh$, defined by

$$X_{k+1} = X_k + \frac{1}{2} h \left(\frac{1}{m} \nabla \log \pi_0(X_k) + \frac{1}{m} \sum_{i=1}^{m} \nabla \log g(y_i X_k^T \iota_i) \right)$$
$$+ \frac{1}{\sqrt{m}} \sqrt{h} Z_{k+1}, \tag{3.4}$$

with a fixed discretisation parameter h (i.e., we do not take into account the error between (f, π_t^h) and (f, π_t) when calculating the MSE, where π_t is the law of the SDE (2.7)). The value of this h is determined by the final level used in MASGA, i.e., $h = h_L$.

The experiment is organised as follows:

i) let $L \geq 1$, and $(\ell_s, \ell_d) = (L, L)$ be the highest multi-indices used in the calculation of $\mathcal{A}_{MASGA}^{A,\mathcal{L}}(f)$.

ii) We measure the bias of $\mathcal{A}_{MASGA}^{A,\mathcal{L}}(f)$ and set $\varepsilon := \sqrt{2} \left(\mathbb{E} \left[|(f, \pi_t^{h_L}) - \mathcal{A}_{MASGA}^{A,\mathcal{L}}(f)|^2 \right] \right)^{1/2}$.

iii) We then compare the cost of $\mathcal{A}_{MASGA}^{A,\mathcal{L}}(f)$ against the cost of $\mathcal{A}_{MLMC}^{A,\mathcal{L}}(f)$ with fixed discretisation parameter $h = h_{\ell_d} = h_L$ satisfying $\mathbb{E}[((f, \pi_t^{h_L}) - \mathcal{A}_{MLMC}^{A,\mathcal{L}}(f))^2] < \varepsilon^2$.

We repeat the above three steps for $L = 1, \ldots, 7$, in order to measure the cost for different values of ε.

We perform this comparison on two data regimes: first on the covertype dataset with 100K observations and 54 covariates, and second on a smaller synthetic dataset with 1K observations and 5 covariates. Results are shown in

Fig. 3, where each ε corresponds to different values of L (see Table 1).

As expected, the higher the accuracy (the lower the ε) the better the cost of $\mathcal{A}_{MASGA}^{A,\mathcal{L}}(f)$ compared to the cost of $\mathcal{A}_{MLMC}^{A,\mathcal{L}}(f)$. Depending on the dataset size, it is necessary to reach different levels of accuracy of the MSE to notice an improvement on the cost. This comes from the amount of variance added by the noise added in the chain $X_k^{\ell_1,\ell_2}$ (3.3) that will decrease as the dataset size m increases.

3.2 AMLMC with respect to subsampling in convex and non-convex settings

In the subsequent experiments, we study the AMLMC estimator with respect to subsampling, i.e., with a fixed discretisation parameter h. We simulate 10×2^4 steps of the chain (3.3). We take X_0 to be an approximation of the mode of the posterior that we pre-compute using Stochastic Gradient Descent to replace the burn-in phase of the Markov chain, cf. (Baker et al. 2019). The number of steps and the step size are chosen so as to be consistent with the finest discretisation level of the MASGA experiment provided in the previous section.

A summary of the AMLMC setting is provided in Table 2.

Plots in Fig. 4 correspond to the results where π_t^h is the approximation of the posterior of a Bayesian Logistic Regression with Gaussian prior. Plots in Fig. 5 use a mixture of Gaussians for the prior. The left plot shows the variance of $(f, \mathbf{\Delta}^A \pi^{\boldsymbol{\ell}})$ and $(f, \pi^{\boldsymbol{\ell}})$ per subsampling level. The right plot displays the computational cost multiplied by ε^2.

These figures indicate that the total cost of approximating (f, π_t^h) by $\mathcal{A}^{A,\mathcal{L}}(f)$ as described above, is $\mathcal{O}(\varepsilon^{-2})$, even when the prior is not log-concave as is the case of a Mixture of two Gaussians (Fig. 5).

3.2.1 Bayesian mixture of Gaussians

For our next experiment, we use the setting from Example 5.1 in Welling and Teh (2011) to consider a Bayesian mixture of two Gaussians on a 2-dimensional dataset, in order to make the posterior multi-modal. Given a dataset of size m, by Bayes' rule

$$\pi(x) \propto \pi_0(x) \prod_{i=1}^{m} g(\iota_1, \iota_2 | x), \tag{3.5}$$

where $x = (x_1, x_2)$, g is the joint density of (ι_1, ι_2) where each $\iota_i \sim \frac{1}{2} \mathcal{N}(x_1, 5) + \frac{1}{2} \mathcal{N}(x_1 + x_2, 5)$. For the experiment, we consider a Gaussian prior $\mathcal{N}(0, I)$ for $\pi_0(x)$, and we create a synthetic dataset with 200 observations, by sampling from $\iota_i \sim \frac{1}{2} \mathcal{N}(0, 5) + \frac{1}{2} \mathcal{N}(1, 5)$.

In this experiment we again take $r = 1$ and $\boldsymbol{\ell} = \ell_1$ in (2.5), which corresponds to using a fixed discretisation parameter h and applying the antithetic MLMC estimator only
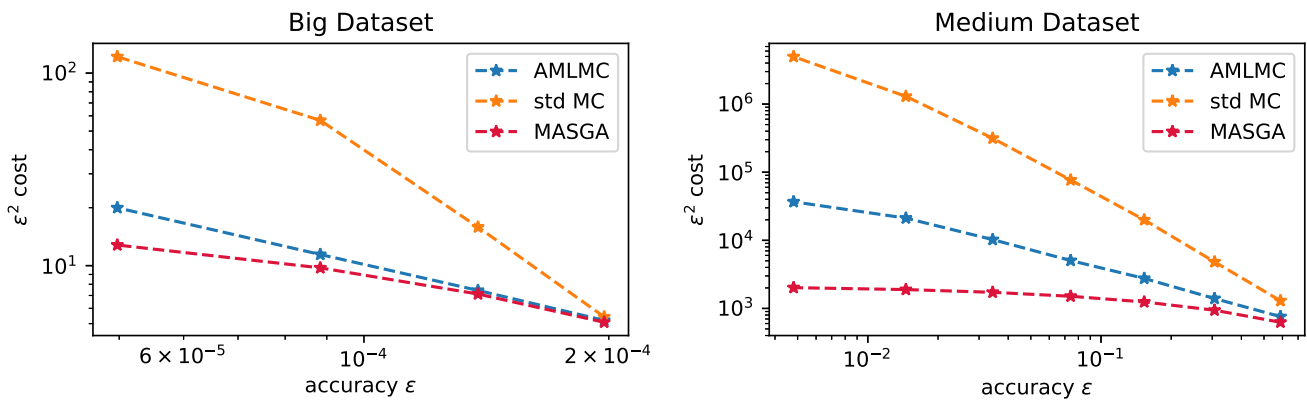
Fig. 3 Comparison of MASGA cost against AMLMC cost for the standard Gaussian prior

Table 1 MASGA setting for Bayesian Logistic Regression

AMLMC parameter	Value
h_0 (initial discretisation step size)	0.005
Number of steps in initial discretisation level	100
(dataset size, dataset dim)	(116 202, 54), (1 000, 5)
s_0 (initial subsample size)	4
X_0	Approximation of the mode of the posterior

Table 2 AMLMC setting for Bayesian Logistic Regression

AMLMC parameter	Value
h (fixed discretisation step size)	0.005×2^{-4}
Number of steps	100×2^4
(dataset size, dataset dim)	(116 202, 54)
s_0 (initial subsample size)	4
X_0	Approximation of the mode of the posterior

with respect to the subsampling parameter. We then use the same setting as before: we apply our estimator $\mathcal{A}^{A,\mathcal{L}}_{MLMC}(f)$ to approximate (f, π_t^h), where π_t^h is the law given by the chain X_k defined in (3.4), with a fixed discretisation parameter $h = 1$ (i.e., we do not take into account the error between (f, π_t^h) and (f, π_t) when calculating the MSE). We simulate

2×10^5 steps of the chain (2.6), starting from $X_0 = 0$ (see Table 3).

In this example there is the additional difficulty that the posterior has two modes (Fig. 6). It is therefore necessary to ensure that the number of steps is high enough so that the chain has explored all the space (Fig. 7).
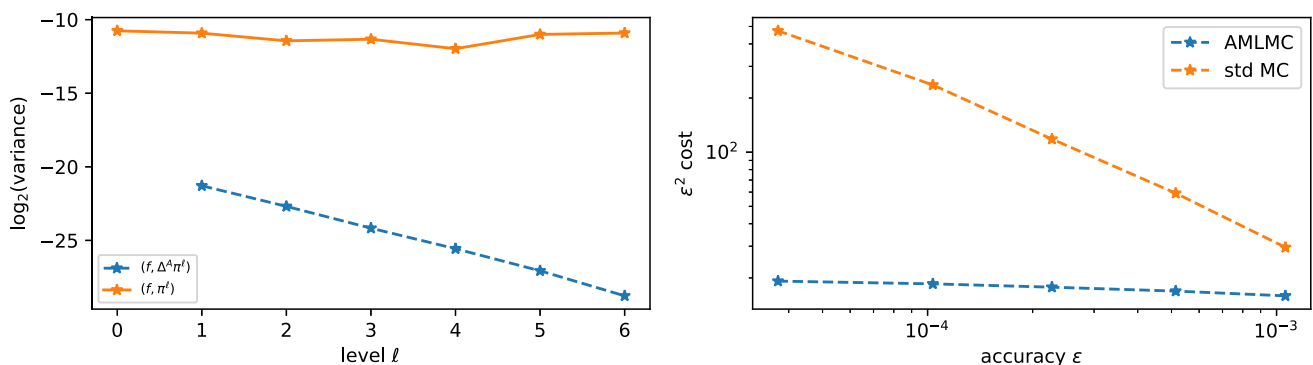

Fig. 4 AMLMC estimator $\mathcal{A}^{A,\mathcal{L}}(f)$ with respect to subsampling for $f(x) = |x|^2$, where the prior is the standard Gaussian
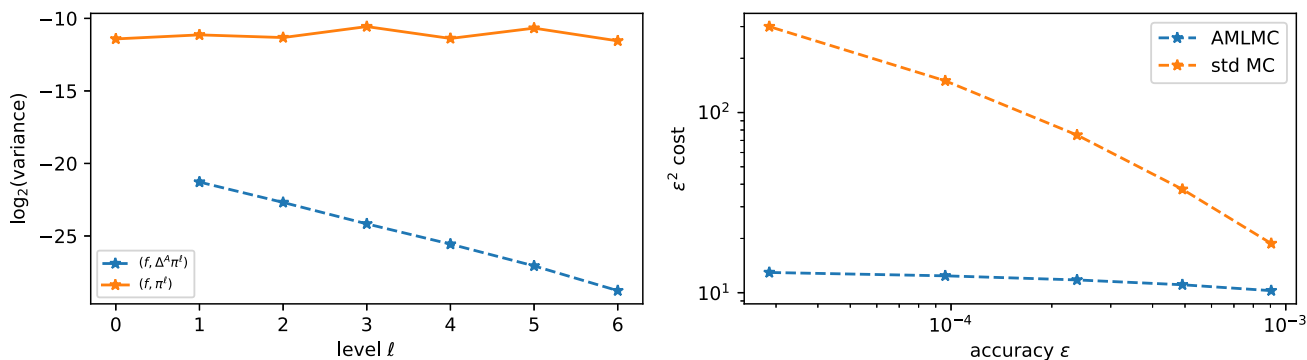
Fig. 5 AMLMC estimator $\mathcal{A}^{A,\mathcal{L}}(f)$ with respect to subsampling for $f(x) = |x|^2$, where the prior is a mixture of two Gaussians

Table 3 AMLMC setting for Bayesian Mixture of Gaussians

AMLMC parameter	Value
h (fixed discretisation step size)	1
Number of steps	200 000
(dataset size, dataset dim)	(200, 2)
s_0 (initial subsample size)	2
X_0	(0, 0)

Results for this experiment are shown in Fig. 6, indicating that the total cost of approximating (f, π_t) by the MASGA estimator $\mathcal{A}^{A,\mathcal{L}}(f)$ is $\mathcal{O}(\varepsilon^{-2})$. We obtain the following rates of decay of the variance and the absolute mean of $\mathbf{\Delta}^A \pi^\ell$:

$i)$ $\mathbb{E}[(f, \mathbf{\Delta}^A \pi^\ell)]| \lesssim 2^{-1.01\ell}$, $ii)$ $\mathbb{V}[(f, \mathbf{\Delta}^A \pi^\ell)] \lesssim 2^{-1.82\ell}$.

3.3 AMLMC and SGLD with control variates

In this subsection we compare the AMLMC estimator with respect to subsampling against the Stochastic Gradient Langevin Dynamics method with control variates (SGLD-CV) from Baker et al. (2019), Nagapetyan et al. (2017). We work with the standard Gaussian prior π_0. For SGLD-CV,

for a fixed time step size h, and for a fixed subsample size s_{ℓ_1}, instead of the process (2.6) we use

$$X_{k+1}^{\ell_1} = X_k^{\ell_1} - h \left(\nabla_x V(\hat{x}, \nu^m) + \left(\nabla_x V(X_k^{\ell_1}, \nu^{s_{\ell_1}}) - \nabla_x V(\hat{x}, \nu^{s_{\ell_1}}) \right) \right)$$
$$+ \beta \sqrt{h_{\ell_2}} Z_{k+1}, \tag{3.6}$$

where $\beta = 1/\sqrt{m}$ and $\hat{x}$ is a fixed value denoting an estimate of the mode of the posterior $\pi(x)$. We undertake the following steps:

1. We estimate the mode of the posterior $\hat{x}$ by using stochastic gradient descent.
2. For each considered accuracy ε, we run AMLMC with respect to subsampling (as described in Sect. 3.2) to get the maximum subsample size s_{ℓ_1} necessary to achieve an estimator $\mathcal{A}_{MLMC}^{A,\mathcal{L}}(f)$ such that $\mathrm{MSE}(\mathcal{A}_{MLMC}^{A,\mathcal{L}}(f)) \lesssim \varepsilon$.
3. We use each pair $(\varepsilon, s_{\ell_1})$ from the previous step to calculate the cost of SGLD-CV.

The AMLMC setting values are listed in Table 4 and results are shown in Fig. 8.

The SGLD-CV method has been shown in Baker et al. (2019), Nagapetyan et al. (2017) to reduce the variance (and
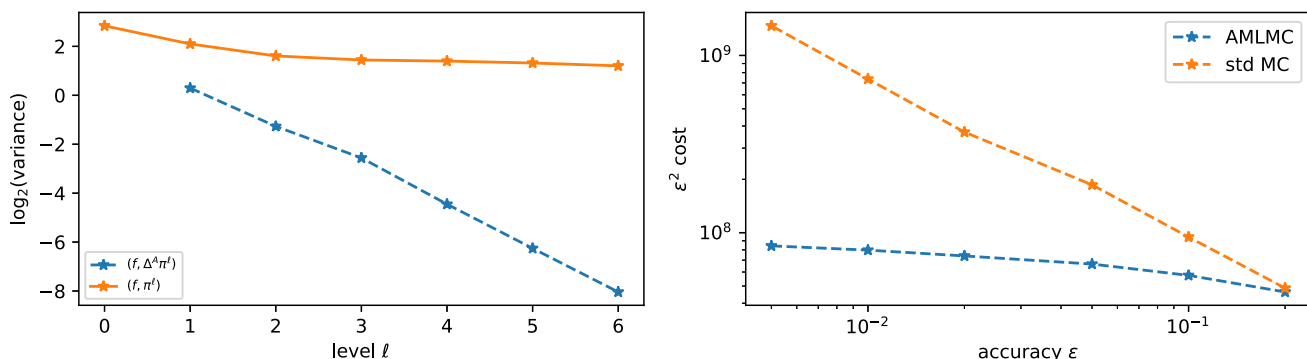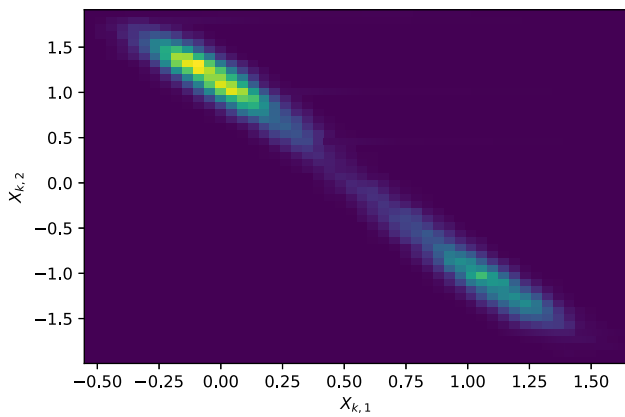

Fig. 6 AMLMC estimator $\mathcal{A}^{A,\mathcal{L}}(f)$ with respect to subsampling for $f(x) = |x|^2$, for a multi-modal posterior, where the prior π_0 is the standard Gaussian

Fig. 7 Histogram of $X_k = (X_{k,1}, X_{k,2})$ sampled from (2.6)

Fig. 8 Comparison of AMLMC with respect to subsampling for $f(x) = |x|^2$ vs SGLD with control variate

hence improve the performance) compared to the standard SGLD and the standard Monte Carlo methods. However, in some numerical examples, as demonstrated in this subsection, this gain can be relatively small compared to the gain from using AMLMC.

4 General setting for MASGA

We will work in a more general setting than the one presented in Sect. 2. Namely, we consider an SDE

$$dX_t = a(X_t)dt + \beta dW_t, \qquad (4.1)$$

where $a : \mathbb{R}^d \to \mathbb{R}^d$, $\beta \in \mathbb{R}_+$ and $(W_t)_{t \geq 0}$ is a d-dimensional Brownian motion. Furthermore, let $(Z_k)_{k=1}^{\infty}$ be i.i.d. random variables in $\mathbb{R}^d$ with $Z_k \sim \mathcal{N}(0, I)$ for $k \geq 1$. For a fixed discretisation parameter $h > 0$ we consider a discretisation of (4.1) given by $X_{k+1} = X_k + ha(X_k) + \beta\sqrt{h}Z_{k+1}$, as well as its inaccurate drift counterpart
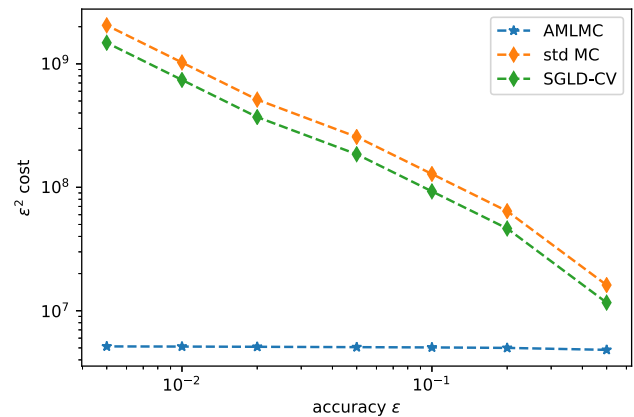
$$X_{k+1} = X_k + hb(X_k, U_k) + \beta\sqrt{h}Z_{k+1}. \qquad (4.2)$$

Here $b : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}^d$ is an unbiased estimator of a in the sense that $(U_k)_{k=0}^{\infty}$ are mutually independent $\mathbb{R}^n$-valued random variables independent of $(Z_k)_{k=1}^{\infty}$ such that for any $k \geq 0$ we have

$$\mathbb{E}[b(x, U_k)] = a(x) \text{ for all } x \in \mathbb{R}^d. \qquad (4.3)$$

Note that for each k, the random variable X_k is independent of U_k and that $\mathbb{E}[b(X_k, U_k)|X_k] = a(X_k)$. Moreover, note that the framework where the drift estimator $b(x, U)$ depends on a random variable U is obviously a generalisation of (1.2), since as a special case of (4.2) we can take $b(x, U) = -\nabla_x V(x, \text{Law}(U))$, where $\text{Law}(U)$ denotes the law of U. We use the name Stochastic Gradient Langevin Dynamics (SGLD) (Dalalyan and Karagulyan 2019; Majka et al. 2020; Nagapetyan et al. 2017) to describe (4.2) even in the general abstract setting where b and a are not necessarily of gradient form.

This setting, besides having the obvious advantage of being more general than the one presented in Assumptions 2.1, allows us also to reduce the notational complexity by replacing sums of gradients with general abstract functions a and b. As a motivation for considering such a general framework, let us discuss an example related to generative models.

Example 4.1 Let ν denote an unknown data measure, supported on $\mathbb{R}^D$, and let ν^m be its empirical approximation. While D is typically very large, in many applications ν can be well approximated by a probability distribution supported on a lower dimensional space, say $\mathbb{R}^d$, with $d \ll D$. The aim of generative models (Goodfellow et al. 2014) is to map samples from some basic distribution μ supported on $\mathbb{R}^d$, into samples from ν. More precisely, one considers a parametrised map $G : \mathbb{R}^d \times \Theta \to \mathbb{R}^D$, with a parameter space $\Theta \subseteq \mathbb{R}^p$,

Table 4 AMLMC setting for Bayesian Logistic Regression

AMLMC parameter	Value
h (fixed discretisation step size)	0.5
Number of steps	100
(dataset size, dataset dim)	(116 202, 54)
s_0 (initial subsample size)	4
X_0	Approximation of the mode of the posterior

that transports μ into $G(\theta)_{\#}\mu := \mu(G(\theta)^{-1}(B))$, $B \in \mathbb{R}^D$. One then seeks θ such that $G(\theta)_{\#}\mu$ is a good approximation of ν with respect to a user-specified metric. In this example we consider $f : \mathbb{R}^d \to \mathbb{R}$ and $\Phi : \mathbb{R} \to \mathbb{R}$ and define

$$\text{dist}(G(\theta)_{\#}\mu, \nu^m) := \Phi\left(\int f(x)(G(\theta)_{\#}\mu)(dx) - \int f(x)\nu^m(dx)\right).$$

A popular choice for the generator G is a neural network (Goodfellow et al. 2014). In the case of a one-hidden-layer network with $\theta = (\alpha, \beta) \in \mathbb{R}^p \times \mathbb{R}^p$ with the activation function $\psi : \mathbb{R}^d \to \mathbb{R}^d$, one takes

$$G(x, \theta) := \frac{1}{p}\sum_{i=1}^{p} \beta_i \psi(\alpha_i x).$$

With this choice of G, the authors of Hu et al. (2019) derived a gradient flow equation that minimises suitably regularised $\text{dist}(G(\theta)_{\#}\mu, \nu^m)$. The gradient flow identified in Hu et al. (2019), when discretised, is given by

$$\theta_{k+1} = \theta_k - h\left(b(\theta_k, \nu^m) - \frac{\sigma^2}{2}\nabla U(\theta_k)\right) + \sigma\sqrt{h}Z_{k+1}, \tag{4.4}$$

with $U : \mathbb{R}^d \to \mathbb{R}$ being a regulariser, $\sigma > 0$ a regularisation parameter, $(Z_k)_{k=1}^{\infty}$ a sequence of i.i.d. random variables with the standard normal distribution, and

$$b(\theta, \nu^m) = \nabla_x \Phi\left(\int (f \circ G)(x, \theta)\mu(dx) - \int f(x)\nu^m(dx)\right)$$
$$\int \nabla_\theta (f \circ G)(x, \theta)\mu(dx).$$

We refer the reader to Hu et al. (2019) for more details and to Jabir et al. (2019) for an extension to deep neural networks. One can see that b may depend on the data in a non-linear way and hence the general setting of (4.2) becomes necessary for the analysis of the stochastic gradient counterpart of (4.4). An application of MASGA to the study of generative models will be further developed in a future work.

In order to analyse the MASGA estimator (2.5), we need to interpret the Markov chain

$$X_{k+1}^{s,h} = X_k^{s,h} + hb(X_k^{s,h}, U_k^s) + \beta\sqrt{h}Z_{k+1} \tag{4.5}$$

as characterized by two parameters: the discretisation parameter $h > 0$ and the drift estimation parameter $s \in \mathbb{N}$ that corresponds to the quality of approximation of $a(x)$ by $b(x, U_k^s)$, for some mutually independent random variables $(U_k^s)_{k=0}^{\infty}$. We will now carefully explain how to implement the antithetic MLMC framework from Sect. 2 in this setting.

To this end, suppose that we have a decreasing sequence $(h_\ell)_{\ell=0}^{L} \subset \mathbb{R}_+$ of discretisation parameters and an increasing sequence $(s_\ell)_{\ell=0}^{L} \subset \mathbb{N}_+$ of drift estimation parameters for some $L \geq 1$. For any $\ell_1, \ell_2 \in \{1, \dots, L\}$ and any function $f : \mathbb{R}^d \to \mathbb{R}$, we define

$$\Delta\Phi_{f,k}^{s_{\ell_1}, h_{\ell_2}} := \left(f(X_k^{s_{\ell_1}, h_{\ell_2}}) - f(X_k^{s_{\ell_1}, h_{\ell_2-1}})\right)$$
$$- \left(f(X_k^{s_{\ell_1-1}, h_{\ell_2}}) - f(X_k^{s_{\ell_1-1}, h_{\ell_2-1}})\right), \tag{4.6}$$

and we also put $\Delta\Phi_{f,k}^{s_0, h_0} := f(X_k^{s_0, h_0})$, $\Delta\Phi_{f,k}^{s_0, h_1} := f(X_k^{s_0, h_1}) - f(X_k^{s_0, h_0})$ and $\Delta\Phi_{f,k}^{s_1, h_0} := f(X_k^{s_1, h_0}) - f(X_k^{s_0, h_0})$. Then we can define a Multi-index Monte Carlo estimator

$$\mathcal{A} := \sum_{\ell_1=0}^{L}\sum_{\ell_2=0}^{L} \frac{1}{N_{\ell_1, \ell_2}} \sum_{j=1}^{N_{\ell_1, \ell_2}} \Delta\Phi_{f,k}^{s_{\ell_1}, h_{\ell_2}, (j)}, \tag{4.7}$$

where $\Delta\Phi_{f,k}^{s_{\ell_1}, h_{\ell_2}, (j)}$ for $j = 1, \dots, N_{\ell_1, \ell_2}$ are independent copies of $\Delta\Phi_{f,k}^{s_{\ell_1}, h_{\ell_2}}$. Here N_{ℓ_1, ℓ_2} is the number of samples at the (doubly-indexed) level $\ell = (\ell_1, \ell_2)$. Note that (4.7) corresponds to the regular (non-antithetic) MLMC estimator defined in (2.3) with $r = 2$ and with L levels for both parameters.

We will now explain how to obtain the MASGA estimator (2.5) by modifying (4.7) by replacing the difference operator $\Delta\Phi_{f,k}^{s_{\ell_1}, h_{\ell_2}}$ with its antithetic counterpart. To this end, we will need to take a closer look at the relation between the chains (4.5) on different levels. From now on, we focus on sequences of parameters $h_\ell := 2^{-\ell}h_0$ and $s_\ell := 2^\ell s_0$ for $\ell \in \{1, \dots, L\}$ and some fixed $h_0 > 0$ and $s_0 \in \mathbb{N}$. Then, we observe that for a fixed s_{ℓ_1}, the chain $X^{s_{\ell_1}, h_{\ell_2}}$ has twice as many steps as the chain $X^{s_{\ell_1}, h_{\ell_2-1}}$, i.e., for any $k \geq 0$ we have

$$X_{k+2}^{s_{\ell_1}, h_{\ell_2}} = X_{k+1}^{s_{\ell_1}, h_{\ell_2}} + h_{\ell_2}b(X_{k+1}^{s_{\ell_1}, h_{\ell_2}}, U_{k+1}^{s_{\ell_1}, h_{\ell_2}}) + \beta\sqrt{h_{\ell_2}}Z_{k+2}$$
$$X_{k+1}^{s_{\ell_1}, h_{\ell_2}} = X_k^{s_{\ell_1}, h_{\ell_2}} + h_{\ell_2}b(X_k^{s_{\ell_1}, h_{\ell_2}}, U_k^{s_{\ell_1}, h_{\ell_2}}) + \beta\sqrt{h_{\ell_2}}Z_{k+1}$$
$$X_{k+2}^{s_{\ell_1}, h_{\ell_2-1}} = X_k^{s_{\ell_1}, h_{\ell_2-1}} + 2h_{\ell_2}b(X_k^{s_{\ell_1}, h_{\ell_2-1}}, U_k^{s_{\ell_1}, h_{\ell_2-1}})$$
$$+ \beta\sqrt{2h_{\ell_2}}(Z_{k+2} + Z_{k+1}). \tag{4.8}$$

Throughout the paper, we will refer to $(X_k^{s_{\ell_1}, h_{\ell_2}})_{k \in \mathbb{N}}$ as the fine chain, and to $(X_k^{s_{\ell_1}, h_{\ell_2-1}})_{k \in 2\mathbb{N}}$ as the coarse chain. Note that for the chain $(X_k^{s_{\ell_1}, h_{\ell_2-1}})_{k \in 2\mathbb{N}}$ we could in principle use a sequence of standard Gaussian random variables $(\hat{Z}_k)_{k \in 2\mathbb{N}}$ completely unrelated to the one that we use for $(X_k^{s_{\ell_1}, h_{\ell_2}})_{k \in \mathbb{N}}$ (which is $(Z_k)_{k \in \mathbb{N}}$). However, we choose $\hat{Z}_{k+2} := (Z_{k+1} + Z_{k+2})/\sqrt{2}$ for all $k \geq 0$, which corresponds to using the synchronous coupling between levels (which turns out to be

a good choice in the global convexity setting as in Assumptions 2.1, cf. (Giles et al. 2020)). Moreover, note that since the chain $X^{s_{\ell_1}, h_{\ell_2}}$ moves twice as frequently as $X^{s_{\ell_1}, h_{\ell_2}-1}$, it needs twice as many random variables $U^{s_{\ell_1}, h_{\ell_2}}$ as $X^{s_{\ell_1}, h_{\ell_2}-1}$ needs $U^{s_{\ell_1}, h_{\ell_2}-1}$. This can be interpreted as having to choose how to estimate the drift a twice as frequently (at each step of the chain).

The idea of the antithetic estimator (with respect to the discretisation parameter) involves replacing $f(X_k^{s_{\ell_1}, h_{\ell_2}-1})$ in (4.6) with a mean of its two independent copies, i.e., with the quantity given by $\frac{1}{2}\left(f(X^{s_{\ell_1}, h_{\ell_2}-1-}) + f(X^{s_{\ell_1}, h_{\ell_2}-1+})\right)$, where the first copy $X^{s_{\ell_1}, h_{\ell_2}-1-}$ uses $U_k^{s_{\ell_1}, h_{\ell_2}}$ and the other copy $X^{s_{\ell_1}, h_{\ell_2}-1+}$ uses $U_{k+1}^{s_{\ell_1}, h_{\ell_2}}$ to estimate the drift, instead of drawing their own independent copies of $U_k^{s_{\ell_1}, h_{\ell_2}-1}$, i.e.,

$$
\begin{aligned}
X_{k+2}^{s_{\ell_1}, h_{\ell_2}-1-} &= X_k^{s_{\ell_1}, h_{\ell_2}-1-} + 2h_{\ell_2} b(X_k^{s_{\ell_1}, h_{\ell_2}-1-}, U_k^{s_{\ell_1}, h_{\ell_2}}) \\
&\quad + \beta\sqrt{2h_{\ell_2}}(Z_{k+2} + Z_{k+1}) \\
X_{k+2}^{s_{\ell_1}, h_{\ell_2}-1+} &= X_k^{s_{\ell_1}, h_{\ell_2}-1+} + 2h_{\ell_2} b(X_k^{s_{\ell_1}, h_{\ell_2}-1+}, U_{k+1}^{s_{\ell_1}, h_{\ell_2}}) \\
&\quad + \beta\sqrt{2h_{\ell_2}}(Z_{k+2} + Z_{k+1}).
\end{aligned}
$$

Hence the term $\left(f(X_k^{s_{\ell_1}, h_{\ell_2}}) - f(X_k^{s_{\ell_1}, h_{\ell_2}-1})\right)$ appearing in (4.6) would be replaced with the antithetic term $\left(f(X_k^{s_{\ell_1}, h_{\ell_2}}) - \frac{1}{2}\left(f(X^{s_{\ell_1}, h_{\ell_2}-1-}) + f(X^{s_{\ell_1}, h_{\ell_2}-1+})\right)\right)$ and the same can be done for any fixed s_{ℓ_1}. Let us explain the intuition behind this approach on a simple example with $f(x) = x$ and a state-independent drift a.

Example 4.2 We fix s_{ℓ_1} and suppress the dependence on ℓ_1 in the notation, in order to focus only on MLMC via discretisation parameter. Let $\xi = (\xi_1, \ldots, \xi_m)$ be a collection of m data points and let us consider a state-independent drift $a = \frac{1}{m}\sum_{i=1}^m \xi_i$ and its unbiased estimator $b(U^s) := \frac{1}{s}\sum_{i=1}^s \xi_{U_i^s}$ for $s \leq m$, where $U^s = (U_1^s, \ldots U_s^s)$ is such that $U_j^s \sim \text{Unif}\{1, \ldots, m\}$ for all $j \in \{1, \ldots, s\}$ (i.e., we sample with replacement s data points from the data set ξ of size m). Consider the standard MLMC estimator with the fine X^f and the coarse X^c schemes defined as

$$
\begin{aligned}
X_{k+1}^f &= X_k^f + hb(U_k^{s,f}) + \beta\sqrt{h}Z_{k+1}, \\
X_{k+2}^f &= X_{k+1}^f + hb(U_{k+1}^{s,f}) + \beta\sqrt{h}Z_{k+2}, \\
X_{k+2}^c &= X_k^c + 2hb(U_k^{s,c}) + \beta\sqrt{2h}\hat{Z}_{k+2},
\end{aligned}
$$

where $\hat{Z}_{k+2} := (Z_{k+1} + Z_{k+2})/\sqrt{2}$, $\beta = 1/\sqrt{m}$, with $U_k^{s,f}$, $U_{k+1}^{s,f}$ and $U_k^{s,c}$ being independent copies of U^s for any $k \geq 0$. Note that $(X_k^f)_{k\in\mathbb{N}}$ corresponds to $(X_k^{s_{\ell_1}, h_{\ell_2}})_{k\in\mathbb{N}}$ in (4.8) whereas $(X_k^c)_{k\in2\mathbb{N}}$ corresponds to $(X_k^{s_{\ell_1}, h_{\ell_2}-1})_{k\in2\mathbb{N}}$ for some fixed ℓ_1, ℓ_2. Recall from the discussion in Sect. 2

that our goal is to find a sharp upper bound on the variance (or the second moment) of $X_k^f - X_k^c$ for any $k \geq 1$ (which corresponds to bounding the variance of the standard, non-antithetic MLMC estimator (4.7) for a Lipschitz function f, cf. the difference (4.6) taken only with respect to the time-discretisation parameter h, with fixed s). We have

$$
\begin{aligned}
\mathbb{E}\left|X_{k+2}^f - X_{k+2}^c\right|^2 &= \mathbb{E}\left|X_k^f - X_k^c\right|^2 + \mathbb{E}\left\langle X_k^f - X_k^c, hb(U_k^{s,f}) \right. \\
&\quad \left. + hb(U_{k+1}^{s,f}) - 2hb(U_k^{s,c}) \right\rangle \\
&\quad + \mathbb{E}\left|hb(U_k^{s,f}) + hb(U_{k+1}^{s,f}) - 2hb(U_k^{s,c})\right|^2 = \mathbb{E}\left|X_k^f - X_k^c\right|^2 \\
&\quad + h^2\mathbb{E}\left|b(U_k^{s,f}) + b(U_{k+1}^{s,f}) - 2b(U_k^{s,c})\right|^2,
\end{aligned}
$$

where in the second step we used conditioning and the fact that b is an unbiased estimator of a. Hence we can show that, if we choose $X_0^f = X_0^c$, then for all $k \geq 1$ we have $\mathbb{E}\left|X_k^f - X_k^c\right|^2 \leq Ch$ for some $C > 0$ and we get a variance contribution of order h. On the other hand, if we want to apply the antithetic approach as in (2.5), we can define

$$
\begin{aligned}
X_{k+2}^{c-} &= X_k^{c-} + 2hb(U_k^{s,f}) + \beta\sqrt{2h}\hat{Z}_{k+2}, \\
X_{k+2}^{c+} &= X_k^{c+} + 2hb(U_{k+1}^{s,f}) + \beta\sqrt{2h}\hat{Z}_{k+2}
\end{aligned}
$$

with $\beta = 1/\sqrt{m}$, and, putting $\bar{X}_k^c := \frac{1}{2}\left(X_k^{c-} + X_k^{c+}\right)$, we obtain

$$
\bar{X}_{k+2}^c = \frac{1}{2}\left(X_k^{c-} + X_k^{c+}\right) + h\left(b(U_k^{s,f}) + b(U_{k+1}^{s,f})\right) + \beta\sqrt{2h}\hat{Z}_{k+2}.
$$

Then we have $\mathbb{E}\left|X_{k+2}^f - \bar{X}_{k+2}^c\right|^2 = \mathbb{E}\left|X_k^f - \bar{X}_k^c\right|^2$ and, choosing $X_0^f = X_0^{c-} = X_0^{c+}$, the variance contribution vanishes altogether.

In the general case, $b(x, U^s)$ is a nonlinear function of the data and also depends on the state x. Therefore one should not expect that the variance of the drift estimator can be completely mitigated. Nonetheless, careful analysis will allow us to conclude that the application of the antithetic difference operators on all levels in our MASGA estimator allows us to obtain a desired upper bound on the variance as described in Sect. 2.

Having explained the motivation behind the antithetic approach to MLMC, let us now focus on the antithetic estimator with respect to the drift estimation parameter s. To this end, let us now fix h_{ℓ_2} and observe that for the chain $X^{s_{\ell_1}, h_{\ell_2}}$, the value of the drift estimation parameter $s_{\ell_1} = 2s_{\ell_1-1}$ is twice the value for the chain $X^{s_{\ell_1-1}, h_{\ell_2}}$. In the context of the subsampling drift as in Assumptions 2.1, this corresponds to the drift estimator in $X^{s_{\ell_1}, h_{\ell_2}}$ using twice as many samples as the drift estimator in $X^{s_{\ell_1-1}, h_{\ell_2}}$. Hence, instead of

using independent samples for $X^{s_{\ell_1}-1,h_{\ell_2}}$, we can consider two independent copies of $X^{s_{\ell_1}-1,h_{\ell_2}}$, the first of which uses the first half $U_k^{s_{\ell_1},h_{\ell_2},1}$ of samples of $X^{s_{\ell_1},h_{\ell_2}}$ and the other uses the second half $U_k^{s_{\ell_1},h_{\ell_2},2}$, namely,

$$
\begin{aligned}
X_{k+1}^{s_{\ell_1}-1-,h_{\ell_2}} &:= X_k^{s_{\ell_1}-1-,h_{\ell_2}} + h_{\ell_2} b(X_k^{s_{\ell_1}-1-,h_{\ell_2}}, U_k^{s_{\ell_1},h_{\ell_2},1}) \\
&\quad + \beta\sqrt{h_{\ell_2}} Z_{k+1} \\
X_{k+1}^{s_{\ell_1}-1+,h_{\ell_2}} &:= X_k^{s_{\ell_1}-1+,h_{\ell_2}} + h_{\ell_2} b(X_k^{s_{\ell_1}-1+,h_{\ell_2}}, U_k^{s_{\ell_1},h_{\ell_2},2}) \\
&\quad + \beta\sqrt{h_{\ell_2}} Z_{k+1}.
\end{aligned}
\tag{4.9}
$$

Hence, using the antithetic approach, we could replace $\left(f(X_k^{s_{\ell_1},h_{\ell_2}}) - f(X_k^{s_{\ell_1}-1,h_{\ell_2}}) \right)$ in (4.6) with the difference $\left(f(X_k^{s_{\ell_1},h_{\ell_2}}) - \frac{1}{2}\left(f(X_k^{s_{\ell_1}-1-,h_{\ell_2}}) + f(X_k^{s_{\ell_1}-1+,h_{\ell_2}}) \right) \right)$.

Combining the ideas of antithetic estimators both with respect to the parameter s and h, we arrive at a nested antithetic difference $\Delta\,\mathrm{Ant}\,\Phi_{f,k}^{s_{\ell_1},h_{\ell_2}}$, defined for any $\ell_1, \ell_2 \in \{1, \dots L\}$ and any $k \geq 1$ as

$$
\begin{aligned}
\Delta\,\mathrm{Ant}\,\Phi_{f,k}^{s_{\ell_1},h_{\ell_2}} &:= \left[f(X_k^{s_{\ell_1},h_{\ell_2}}) - \frac{1}{2}\left(f(X_k^{s_{\ell_1},h_{\ell_2}-1-}) \right.\right.\\
&\qquad\left.\left. + f(X_k^{s_{\ell_1},h_{\ell_2}-1+}) \right) \right] \\
&\quad - \frac{1}{2}\left[\left(f(X_k^{s_{\ell_1}-1-,h_{\ell_2}}) - \frac{1}{2}\left(f(X_k^{s_{\ell_1}-1-,h_{\ell_2}-1-}) \right.\right.\right.\\
&\qquad\left.\left. + f(X_k^{s_{\ell_1}-1-,h_{\ell_2}-1+}) \right) \right) \\
&\quad + \left(f(X_k^{s_{\ell_1}-1+,h_{\ell_2}}) - \frac{1}{2}\left(f(X_k^{s_{\ell_1}-1+,h_{\ell_2}-1-}) \right.\right.\\
&\qquad\left.\left.\left. + f(X_k^{s_{\ell_1}-1+,h_{\ell_2}-1+}) \right) \right) \right],
\end{aligned}
\tag{4.10}
$$

with the same convention as in (4.6) for the case of $\ell_1 = 0$ or $\ell_2 = 0$. We can now plug this difference into the definition of a Multi-index Monte Carlo estimator (4.7) to obtain

$$
\mathrm{Ant}\,\mathcal{A} := \sum_{\ell_1=0}^{L} \sum_{\ell_2=0}^{L} \frac{1}{N_{\ell_1,\ell_2}} \sum_{j=1}^{N_{\ell_1,\ell_2}} \Delta\,\mathrm{Ant}\,\Phi_{f,k}^{s_{\ell_1},h_{\ell_2},(j)}. \tag{4.11}
$$

Note that this corresponds to the Antithetic MIMC estimator introduced in (2.5), based on the chain (2.6), with ℓ_1 corresponding to the number of samples s and ℓ_2 to the discretisation parameter h, but with a more general drift a and its estimator b.

In order to formulate our result for the general setting presented in this section, we need to specify the following assumptions.

Assumptions 4.3 (Lipschitz condition and global contractivity of the drift) The drift function $a : \mathbb{R}^d \to \mathbb{R}^d$ satisfies the following conditions:

i) Lipschitz condition: there is a constant $L > 0$ such that

$$
|a(x) - a(y)| \leq L|x - y| \qquad \text{for all } x, y \in \mathbb{R}^d. \tag{4.12}
$$

ii) Global contractivity condition: there exists a constant $K > 0$ such that

$$
\langle x - y, a(x) - a(y) \rangle \leq -K|x - y|^2 \qquad \text{for all } x, y \in \mathbb{R}^d. \tag{4.13}
$$

iii) Smoothness: $a \in C_b^2(\mathbb{R}^d; \mathbb{R}^d)$ (where $C_b^2(\mathbb{R}^d; \mathbb{R}^d)$ is defined as in Assumptions 2.1). In particular, there exist constants $C_{a^{(1)}}, C_{a^{(2)}} > 0$ such that

$$
|D^\alpha a(x)| \leq C_{a^{(|\alpha|)}} \tag{4.14}
$$

for all $x \in \mathbb{R}^d$ and for all multiindices α with $|\alpha| = 1, 2$.

We remark that condition (4.14) could be easily removed by approximating a non-smooth drift a with suitably chosen smooth functions. This, however, would create additional technicalities in the proof and hence we decided to work with (4.14).

We now impose the following assumptions on the estimator b of the drift a.

Assumption 4.4 (Lipschitz condition of the estimator) There is a constant $\bar{L} > 0$ such that for all $x, y \in \mathbb{R}^d$ and all random variables U^s such that $\mathbb{E}[b(x, U^s)] = a(x)$ for all $x \in \mathbb{R}^d$, we have

$$
\mathbb{E}|b(x, U^s) - b(y, U^s)| \leq \bar{L}|x - y|. \tag{4.15}
$$

Assumption 4.5 (Variance of the estimator) There exists a constant $\sigma > 0$ of order $\mathcal{O}(s^{-1})$ such that for any $x \in \mathbb{R}^d$ and any random variable U^s such that $\mathbb{E}[b(x, U^s)] = a(x)$ for all $x \in \mathbb{R}^d$, we have

$$
\mathbb{E}\left| b(x, U^s) - a(x) \right|^2 \leq \sigma^2(1 + |x|^2). \tag{4.16}
$$

Assumption 4.6 (Fourth centered moment of the estimator) There exists a constant $\sigma^{(4)} \geq 0$ of order $\mathcal{O}(s^{-2})$ such that for any $x \in \mathbb{R}^d$ and for any random variable U^s such that $\mathbb{E}[b(x, U^s)] = a(x)$ for all $x \in \mathbb{R}^d$, we have

$$
\mathbb{E}\left| b(x, U^s) - a(x) \right|^4 \leq \sigma^{(4)}(1 + |x|^4). \tag{4.17}
$$

Note that obviously Assumption 4.6 implies Assumption 4.5. However, we formulate these conditions separately

in order to keep track of the constants in the proofs. Moreover, with the same constant $\sigma^{(4)}$ as in (4.17), we impose

Assumption 4.7 The estimator $b(x, U)$ as a function of x is twice continuously differentiable for any U and, for any $x \in \mathbb{R}^d$, we have $\mathbb{E} |\nabla b(x, U^s) - \nabla a(x)|^4 \leq \sigma^{(4)}(1 + |x|^4)$.

Note that $\nabla a(x), \nabla b(x, U^s) \in \mathbb{R}^{d \times d}$ and we use the matrix norm $|\nabla a(x)|^2 := \sum_{i,j=1}^d |\partial_i a_j(x)|^2$, where $a(x) = (a_1(x), \ldots, a_d(x))$.

Assumptions 4.8 Partial derivatives of the estimator of the drift are estimators of the corresponding partial derivatives of the drift. More precisely, for any multi-index α with $|\alpha| \leq 2$ and for any random variable U^s such that $\mathbb{E}[b(x, U^s)] = a(x)$ for all $x \in \mathbb{R}^d$, we have $\mathbb{E}[D^\alpha b(x, U^s)] = D^\alpha a(x)$ for any $x \in \mathbb{R}^d$.

Assumption 4.9 (Growth of the drift) There exists a constant $L_0^{(4)} > 0$ such that for all $x \in \mathbb{R}^d$ we have

$$|a(x)|^4 \leq L_0^{(4)} \left(1 + |x|^4\right). \tag{4.18}$$

Finally, we have the following condition that specifies the behaviour of the drift estimator b with respect to the random variables $U_k^{s_{\ell_1}, h_{\ell_2}, 1}$ and $U_k^{s_{\ell_1}, h_{\ell_2}, 2}$ introduced in (4.9).

Assumption 4.10 For any $x \in \mathbb{R}^d$ we have $b(x, U_k^{s_{\ell_1}, h_{\ell_2}}) = \frac{1}{2} b(x, U_k^{s_{\ell_1}, h_{\ell_2}, 1}) + \frac{1}{2} b(x, U_k^{s_{\ell_1}, h_{\ell_2}, 2})$.

Even though this set of conditions is long, the assumptions are in fact rather mild and it is an easy exercise to verify that when

$$a(x) := \frac{1}{m} \sum_{i=1}^m \nabla_x v(x, \xi_i) \quad \text{and} \quad b(x, U^s) := \frac{1}{s} \sum_{i=1}^s \nabla_x v(x, \xi_{U_i^s}),$$

where $U_i^s \sim \text{Unif}(\{1, \ldots, m\})$ for $i \in \{1, \ldots, s\}$ are i.i.d. random variables, uniformly distributed on $\{1, \ldots, m\}$, whereas $v : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ is the function satisfying Assumptions 2.1, then a and b satisfy all Assumptions 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9 and 4.10. The only conditions that actually require some effort to be checked are Assumptions 4.5 and 4.6, however, they can be verified by extending the argument from Example 2.15 in Majka et al. (2020), where a similar setting was considered. As it turns out, these conditions hold also for the case of subsampling without replacement. All the details are provided in Appendix 1.

We have the following result.

Theorem 4.11 *Under Assumptions 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9 and 4.10 on the drift a and its estimator b, the MASGA estimator (4.11) achieves the mean square error smaller than*

$\varepsilon > 0$ *at the computational cost* ε^{-2}. *Here, at each level* $(\ell_1, \ell_2) \in \{0, \ldots, L\}^2$, *the number of paths* N_{ℓ_1, ℓ_2} *is given by*

$$\varepsilon^{-2} \left(\sqrt{\frac{\mathbb{V}[(f, \mathbf{\Delta}^A \pi^{(\ell_1, \ell_2)})]}{C_{(\ell_1, \ell_2)}}} \sum_{\boldsymbol{\ell} \in [L]^2} \sqrt{\mathbb{V}[(f, \mathbf{\Delta}^A \pi^{\boldsymbol{\ell}})] C_{\boldsymbol{\ell}}} \right), where$$

$[L]^2 := \{0, \ldots, L\}^2$, $C_{\boldsymbol{\ell}}$ *is defined via (2.8) and* $(f, \mathbf{\Delta}^A \pi^{(\ell_1, \ell_2)}) := \Delta \, \text{Ant} \, \Phi_{f,k}^{s_{\ell_1}, h_{\ell_2}}$ *given by (4.10)*.

As we explained in Sect. 2, the proof of Theorem 2.2 (and its generalisation Theorem 4.11) relies on the MIMC complexity analysis from (Haji-Ali et al. 2016) (see also Giles 2015).

Theorem 4.12 *Fix $\varepsilon \in (0, e^{-1})$. Let $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \in \mathbb{R}^{3r}$ be a triplet of vectors in $\mathbb{R}^r$ such that for all $k \in \{1, \ldots, r\}$ we have $\alpha_k \geq \frac{1}{2}\beta_k$. Assume that for each $\boldsymbol{\ell} \in \mathbb{N}^r$*

$$i) \ |\mathbb{E}[(f, \mathbf{\Delta}^A \pi^{\boldsymbol{\ell}})]| \lesssim 2^{-\langle \boldsymbol{\alpha}, \boldsymbol{\ell} \rangle}, \quad ii) \ \mathbb{V}[(f, \mathbf{\Delta}^A \pi^{\boldsymbol{\ell}})] \lesssim 2^{-\langle \boldsymbol{\beta}, \boldsymbol{\ell} \rangle}$$
$$iii) \ C_{\boldsymbol{\ell}} \lesssim 2^{\langle \boldsymbol{\gamma}, \boldsymbol{\ell} \rangle}.$$

If $\max_{k \in [1, \ldots, r]} \frac{(\gamma_k - \beta_k)}{\alpha_k} < 0$, then there exists a set $\mathcal{L} \subset \mathbb{N}^r$ and a sequence $(N_{\boldsymbol{\ell}})_{\boldsymbol{\ell} \in \mathcal{L}}$ such that the MLMC estimator $\mathcal{A}^{A, \mathcal{L}}(f)$ defined in (2.5) satisfies

$$\mathbb{E}[((f, \pi) - \mathcal{A}^{A, \mathcal{L}}(f))^2] < \varepsilon^2,$$

with the computational cost ε^{-2}.

The key challenge in constructing and analysing MIMC estimators is to establish conditions i)-iii) in Theorem 4.12 i.e., to show that the leading error bounds for the bias, variance and cost can be expressed in the product form. In fact, there are very few results in the literature that present the analysis giving (i)–(iii), with the exception of (Giles 2015, Sect. 9) and Giles and Haji-Ali (2019). The bulk of the analysis in this paper is devoted to the analysis of ii). We remark that the optimal choice of $\mathcal{L} = [L]^2$ is dictated by the relationship between $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, see (Haji-Ali et al. 2016).

The following lemma will be crucial for the proof of Theorem 4.11.

Lemma 4.13 *Let Assumptions 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9 and 4.10 hold. Then there is a constant $h_0 > 0$ and a constant $C > 0$ (independent of s and h) such that for any Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$, for any $s \geq 1$, $h \in (0, h_0)$ and for any $k \geq 1$,*

$$\mathbb{E}|\Delta \, \text{Ant} \, \Phi_{f,k}^{s,h}|^2 \leq Ch^2/s^2.$$

As we already indicated in Sect. 2, once we have an upper bound on the second moment (and thus on the variance) of $\Delta \, \text{Ant} \, \Phi_{f,k}^{s,h}$ such as in Lemma 4.13, the proof of Theorem 4.11 becomes rather straightforward.

Proof of Theorem 4.11 Note that we have $\mathbb{V}[(f, \boldsymbol{\Delta}^A \pi^\ell)] \leq \mathbb{E}[(f, \boldsymbol{\Delta}^A \pi^\ell)^2] \lesssim h_{\ell_2}^2/s_{\ell_1}^2$ due to Lemma 4.13. Moreover, the number of time-steps and the number of subsamples at each level of MIMC is doubled (i.e., we have $s_{\ell_1} = 2^{\ell_1} s_0$ and $h_{\ell_2} = 2^{-\ell_2} h_0$ for all $\ell_1, \ell_2 \in \{0, \ldots, L\}$) and hence $\boldsymbol{\gamma} = (1, 1)$ and $\boldsymbol{\beta} = (2, 2)$ in the assumption of Theorem 4.12 (recall that $C_\ell \lesssim s_{\ell_1} h_{\ell_2}^{-1}$). Finally, we have

$$|\mathbb{E}(f, \boldsymbol{\Delta}^A \pi^\ell)| \leq \left(\mathbb{E}|(f, \boldsymbol{\Delta}^A \pi^\ell)|^2\right)^{1/2} \lesssim h_{\ell_2}/s_{\ell_1},$$

which implies that $\boldsymbol{\alpha} = (1, 1)$. Hence the assumptions of Theorem 4.12 are satisfied and the overall complexity of MIMC is indeed ε^{-2}. $\qquad \square$

Proof of Theorem 2.2 Under Assumptions 2.1, the function $a(x) := -\nabla v_0(x) - \frac{1}{m} \sum_{i=1}^m \nabla_x v(x, \xi_i)$ and its estimator $b(x, U^s) := -\nabla v_0(x) - \frac{1}{s} \sum_{i=1}^s \nabla_x v(x, \xi_{U_i^s})$, where U_i^s are mutually independent random variables, uniformly distributed on $\{1, \ldots m\}$, satisfy all Assumptions 4.3, 4.4, 4.5, 4.7, 4.8, 4.9, 4.6 and 4.10. Hence we can just apply Theorem 4.11 to conclude. $\qquad \square$

5 Analysis of AMLMC

The estimator we introduced in (2.5), see also (4.11), can be interpreted as built from two building blocks: the antithetic MLMC estimator with respect to the discretisation parameter, which corresponds to taking $r = 1$ and $\ell = \ell_2$ in (2.5), and the antithetic MLMC estimator with respect to subsampling, which corresponds to taking $r = 1$ and $\ell = \ell_1$ in (2.5). Let us begin our analysis by focusing on the former.

5.1 AMLMC via discretisation

We will analyse one step of the MLMC algorithm, for some fixed level ℓ. To this end, let us first introduce the following fine $(X_k^f)_{k \in \mathbb{N}}$ and coarse $(X_k^c)_{k \in 2\mathbb{N}}$ chains

$$
\begin{aligned}
X_{k+1}^f &= X_k^f + h b(X_k^f, U_k^f) + \beta \sqrt{h} Z_{k+1}, \\
X_{k+2}^f &= X_{k+1}^f + h b(X_{k+1}^f, U_{k+1}^f) + \beta \sqrt{h} Z_{k+2} \\
X_{k+2}^c &= X_k^c + 2h b(X_k^c, U_k^c) + \beta \sqrt{2h} \hat{Z}_{k+2},
\end{aligned}
\tag{5.1}
$$

where $h > 0$ is fixed, $(U_k^f)_{k=0}^\infty$ and $(U_k^c)_{k=0}^\infty$ are mutually independent random variables such that for $U \in \{U_k^f, U_k^c\}$ we have $\mathbb{E}[b(x, U)] = a(x)$ for all $x \in \mathbb{R}^d$ and all $k \geq 0$ and $(Z_k)_{k=1}^\infty$ are i.i.d. random variables with $Z_k \sim \mathcal{N}(0, I)$. We also have $\hat{Z}_{k+2} := \frac{1}{\sqrt{2}} (Z_{k+1} + Z_{k+2})$.

In order to analyse the antithetic estimator, we also need to introduce two auxiliary chains

$$
\begin{aligned}
X_{k+2}^{c+} &= X_k^{c+} + 2h b(X_k^{c+}, U_k^f) + \beta \sqrt{2h} \hat{Z}_{k+2} \\
X_{k+2}^{c-} &= X_k^{c-} + 2h b(X_k^{c-}, U_{k+1}^f) + \beta \sqrt{2h} \hat{Z}_{k+2}.
\end{aligned}
\tag{5.2}
$$

Furthermore, we denote $\bar{X}_k^c = \frac{1}{2} \left(X_k^{c+} + X_k^{c-}\right)$.

Before we proceed, let us list a few simple consequences of Assumptions 4.3. We have the following bounds:

$$|a(x)| \leq L_0(1 + |x|) \text{ for all } x \in \mathbb{R}^d, \text{ where } L_0 := \max(L, |a(0)|). \tag{5.3}$$

Let $M_2 := 4L|a(0)|^2/K^2 + 2|a(0)|^2/K$ and $M_1 := K/2$. Then we have

$$\langle x, a(x) \rangle \leq M_2 - M_1 |x|^2 \text{ for all } x \in \mathbb{R}^d. \tag{5.4}$$

Finally, for all random variables U satisfying (4.3), we have

$$\mathbb{E}|b(x, U)|^2 \leq \bar{L}_0(1 + |x|^2) \text{ for all } x \in \mathbb{R}^d, \tag{5.5}$$

where $\bar{L}_0 := \sigma^2 h^\alpha + 2 \max(L^2, |a(0)|^2)$. Note that (5.3) is an immediate consequence of (4.12), (5.4) follows easily from (4.12) and (4.13) (cf. the proof of Lemma 2.11 in Majka et al. (2020)), whereas (5.5) is implied by (4.12) and (4.16), cf. (2.40) in Majka et al. (2020). Throughout our proofs, we will also use uniform bounds on the second and the fourth moments of Euler schemes with time steps h and $2h$, i.e., we have

$$\mathbb{E}|X_k^f|^2 \leq C_{IEul}, \quad \mathbb{E}|X_k^{c-}|^2 \leq C_{IEul}^{(2h)}, \quad \mathbb{E}|X_k^f|^4 \leq C_{IEul}^{(4)} \quad \text{and} \quad \mathbb{E}|X_k^{c-}|^4 \leq C_{IEul}^{(4),(2h)},$$

where the exact formulas for the constants C_{IEul}, $C_{IEul}^{(2h)}$, $C_{IEul}^{(4)}$ and $C_{IEul}^{(4),(2h)}$ can be deduced from Lemma 7.1 in the Appendix (for C_{IEul}, $C_{IEul}^{(2h)}$ see also Lemma 2.17 in Majka et al. (2020)).

We now fix $g \in C_b^2(\mathbb{R}^d; \mathbb{R})$. Denote by $C_{g^{(1)}}, C_{g^{(2)}}$ positive constants such that $|D^\alpha g(x)| \leq C_{g^{(|\alpha|)}}$ for all $x \in \mathbb{R}^d$ and all multiindices α with $|\alpha| = 1, 2$.

We begin by presenting the crucial idea of our proof. We will use the Taylor formula to write

$$
\begin{aligned}
g(\bar{X}_k^c) - g(X_k^{c-}) = -\Big[&\sum_{|\alpha|=1} D^\alpha g(\bar{X}_k^c) \left(\bar{X}_k^c - X_k^{c-}\right)^\alpha \\
&+ \sum_{|\alpha|=2} \int_0^1 (1-t) D^\alpha g\left(\bar{X}_k^c + t\left(\bar{X}_k^c - X_k^{c-}\right)\right) dt \left(\bar{X}_k^c - X_k^{c-}\right)^\alpha \Big].
\end{aligned}
$$

We also express $g(\bar{X}_k^c) - g(X_k^{c+})$ in an analogous way. Note that

$$\bar{X}_k^c - X_k^{c-} = \frac{1}{2}X_k^{c+} - \frac{1}{2}X_k^{c-} = -(\bar{X}_k^c - X_k^{c+}) \qquad (5.6)$$

and hence we have

$$\begin{aligned}
g(X_k^f) &- \frac{1}{2}\left(g(X_k^{c-}) + g(X_k^{c+})\right) = g(X_k^f) - g(\bar{X}_k^c) \\
&+ \frac{1}{2}\left(g(\bar{X}_k^c) - g(X_k^{c-}) + g(\bar{X}_k^c) - g(X_k^{c+})\right) \\
&= g(X_k^f) - g(\bar{X}_k^c) - \frac{1}{2}\sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha \\
&\quad g\left(\bar{X}_k^c + t\left(\bar{X}_k^c - X_k^{c-}\right)\right)dt\left(\bar{X}_k^c - X_k^{c-}\right)^\alpha \\
&\quad - \frac{1}{2}\sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha \\
&\quad g\left(\bar{X}_k^c + t\left(\bar{X}_k^c - X_k^{c+}\right)\right)dt\left(\bar{X}_k^c - X_k^{c+}\right)^\alpha,
\end{aligned}$$

i.e., the first order terms in the Taylor expansions cancel out. Thus, using the inequalities $\sum_{|\alpha|=2} D^\alpha g(x)\left(\bar{X}_k^c - X_k^{c+}\right)^\alpha \leq \|\nabla^2 g\|_{\text{op}}|\bar{X}_k^c - X_k^{c+}|^2$ and $\|\nabla^2 g\|_{\text{op}} \leq C_{g^{(2)}}$ for all $x \in \mathbb{R}^d$, where $\|\nabla^2 g\|_{\text{op}}$ is the operator norm of the Hessian matrix, we see that

$$\begin{aligned}
\mathbb{E}\left|g(X_k^f) - \frac{1}{2}\left(g(X_k^{c-}) + g(X_k^{c+})\right)\right|^2 &\leq 2\mathbb{E}\left|g(X_k^f) - g(\bar{X}_k^c)\right|^2 \\
&+ \frac{1}{2}C_{g^{(2)}}\left(\mathbb{E}\left|\bar{X}_k^c - X_k^{c-}\right|^4 + \mathbb{E}\left|\bar{X}_k^c - X_k^{c+}\right|^4\right) \qquad (5.7) \\
&= 2\mathbb{E}\left|g(X_k^f) - g(\bar{X}_k^c)\right|^2 + \frac{1}{16}C_{g^{(2)}}\mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4.
\end{aligned}$$

Note that we purposefully introduced the term $\mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4$, since it will provide us with an improved rate in h. Indeed, we have the following result.

Lemma 5.1 *Let Assumptions 4.3, 4.5, 4.6 and 4.9 hold. If $X_0^{c+} = X_0^{c-}$, then for all $k \geq 1$ and for all $h \in (0, h_0)$ we have*

$$\mathbb{E}|X_k^{c+} - X_k^{c-}|^4 \leq \frac{\bar{C}_1}{\bar{c}_1}h^2, \qquad (5.8)$$

where $\bar{C}_1 := 72\frac{1}{\varepsilon}h_0^{2\alpha}\left(1 + 2C_{IEul}^{(2h)} + C_{IEul}^{(4),(2h)}\right)\sigma^4 + 32\left(432\sqrt{2} + 648 + 27h_0\right)\sigma^{(4)}(1 + C_{IEul}^{(4),(2h)})$ and $\bar{c}_1, h_0, \varepsilon > 0$ are chosen such that

$$\begin{aligned}
-8K + 72\varepsilon &+ 72h_0L^2 + 432h_0^3L^4 + 32h_0^2\left(\left((1 + 432L^4)216L^4\right)^{1/2}\right. \\
&\left. + \frac{1 + 1296L^4}{4}\right) \leq -\bar{c}_1.
\end{aligned}$$

Note that the constant $\bar{C}_1$ above is of order $\mathcal{O}(s^{-2})$ due to Assumptions 4.5 and 4.6. Hence the bound in (5.8) is in fact of order $\mathcal{O}(h^2 s^{-2})$, which is exactly what is needed in Lemma 4.13.

We remark that, in principle, it would be now possible to bound also the first term on the right hand side of (5.7) and hence to obtain a bound on the variance of the antithetic MLMC estimator with respect to discretisation, corresponding to taking $r = 1$ and $\ell = \ell_2$ in (2.5). However, such an estimator does not perform on par with the MASGA estimator (even though it is better than the standard MLMC) and hence we skip its analysis. In this subsection, we present only the derivation of the inequality (5.7) and we formulate the lemma about the bound on the term $\mathbb{E}|X_k^{c+} - X_k^{c-}|^4$, since they will be needed in our analysis of MASGA.

We remark that the proof of Lemma 5.1 is essentially identical to the proof of Lemma 5.2 below (with different constants) and is therefore skipped. The latter proof can be found in the Appendix.

5.2 AMLMC via subsampling

As the second building block of our MASGA estimator, we discuss a multi-level algorithm for subsampling that involves taking different drift estimators (different numbers of samples) at different levels, but with constant discretisation parameter across the levels.

We fix $s_0 \in \mathbb{N}_+$ and for $\ell \in \{0, \dots, L\}$ we define $s^\ell := 2^\ell s_0$ and we consider chains

$$\begin{aligned}
X_{k+1}^f &= X_k^f + hb(X_k^f, U_k^f) + \beta\sqrt{h}Z_{k+1}, \\
X_{k+1}^c &= X_k^c + hb(X_k^c, U_k^c) + \beta\sqrt{h}Z_{k+1}, \qquad (5.9)
\end{aligned}$$

where U_k^f is from a higher level (in parameter s) than U_k^c, i.e., we have $b(x, U_k^f) = \frac{1}{2}b(x, U_k^{f,1}) + \frac{1}{2}b(x, U_k^{f,2})$ and $U_k^{f,1}, U_k^{f,1}$ are both from the same level (in parameter s) as U_k^c, cf. Assumption 4.10. In the special case of subsampling, we have

$$b(x, U_k^f) := \frac{1}{2s}\sum_{i=1}^{2s}\hat{b}(x, \theta_{(U_k^f)_i}) \qquad \text{and}$$

$$b(x, U_k^c) := \frac{1}{s}\sum_{i=1}^{s}\hat{b}(x, \theta_{(U_k^c)_i}),$$

whereas $b(x, U_k^{f,1}) := \frac{1}{s}\sum_{i=1}^{s}\hat{b}(x, \theta_{(U_k^f)_i})$ and $b(x, U_k^{f,2}) := \frac{1}{s}\sum_{i=s}^{2s}\hat{b}(x, \theta_{(U_k^f)_i})$, for some kernels $\hat{b} : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^d$, where $(U_k^f)_i, (U_k^c)_i \sim \text{Unif}(\{1, \dots, m\})$. In order to introduce the antithetic counterpart of (5.9), we will replace the random variable U_k^c taken on the coarse level with the

two components of U_k^f. Namely, let us denote

$$
\begin{aligned}
X_{k+1}^{c-} &= X_k^{c-} + hb(X_k^{c-}, U_k^{f,1}) + \beta\sqrt{h}Z_{k+1}, \\
X_{k+1}^{c+} &= X_k^{c+} + hb(X_k^{c+}, U_k^{f,2}) + \beta\sqrt{h}Z_{k+1}.
\end{aligned}
\tag{5.10}
$$

We also put $\bar{X}_k^c := \frac{1}{2}\left(X_k^{c-} + X_k^{c+}\right)$.

Following our calculations from the beginning of Sect. 5.1 we see that for any Lipschitz function $g \in C_b^2(\mathbb{R}^d; \mathbb{R})$,

$$
\begin{aligned}
\mathbb{E}&\left|g(X_{k+1}^f) - \frac{1}{2}\left(g(X_{k+1}^{c-}) + g(X_{k+1}^{c+})\right)\right|^2 \\
&\le 2\mathbb{E}\left|g(X_{k+1}^f) - g(\bar{X}_{k+1}^c)\right|^2 + \frac{1}{16}C_{g^{(2)}}^2\mathbb{E}\left|X_{k+1}^{c+} - X_{k+1}^{c-}\right|^4
\end{aligned}
\tag{5.11}
$$

and we need a similar bound as in Lemma 5.1.

Lemma 5.2 *Let Assumptions 4.3, 4.5, 4.6 and 4.9 hold. For X_k^{c+} and X_k^{c-} defined in (5.10), if $X_0^{c+} = X_0^{c-}$, then for any $k \ge 1$ and any $h \in (0, h_0)$ we have*

$$
\mathbb{E}|X_k^{c+} - X_k^{c-}|^4 \le \frac{C_1}{c_1}h^2,
$$

where $C_1 := 18\frac{1}{\varepsilon}\left(1 + 2C_{IEul}^{(2h)} + C_{IEul}^{(4),(2h)}\right)\sigma^4 + 4\left(27\sqrt{2} + 54h_0\right)\sigma^{(4)}(1 + C_{IEul}^{(4),(2h)})$ and c_1, ε and h_0 are chosen so that

$$
\begin{aligned}
&-4h_0 K + 18h_0\varepsilon + 18h_0^2 L^2 + 27h_0^4 L^4 + 4h_0^3\left(\left(\frac{1 + 27L^4}{2}27L^4\right)^{1/2}\right. \\
&\left.+\frac{1 + 81L^4}{4}\right) \le -c_1 h_0.
\end{aligned}
$$

Using the Lipschitz property of g, we see that in order to deal with the first term on the right hand side of (5.11), we need to bound $\mathbb{E}|X_k^f - \bar{X}_k^c|^2$. Indeed, we have the following result.

Lemma 5.3 *Let Assumptions 4.3, 4.5, 4.6, 4.7 and 4.9 hold. For X_k^f and $\bar{X}_k^c$ defined in (5.10), if $X_0^{c+} = X_0^{c-}$, then for any $k \ge 1$ and any $h \in (0, h_0)$ we have*

$$
\mathbb{E}|X_k^f - \bar{X}_k^c|^2 \le \frac{C_2}{c_2}h^2,
$$

where $C_2 := \frac{3}{4}\sigma^{(4)}(1 + C_{IEul}^{(4)}) + \frac{C_1}{c_1}\left(\frac{1}{4}dC_{a^{(2)}}\frac{1}{\varepsilon_1} + \frac{3}{4} + \frac{3}{8}h_0 C_{b^{(2)}}^2\right)$ with C_1 and c_1 given in Lemma 5.2, whereas c_2, ε_1 and h_0 are chosen so that $-h_0 K + \frac{1}{4}dC_{a^{(2)}}\varepsilon_1 h_0 + \frac{3}{2}h_0^2\bar{L}^2 \le -c_2 h_0$.

Note that both C_1 and C_2 in Lemma 5.2 and Lemma 5.3 are of order $\mathcal{O}(s^{-2})$, which follows from the dependence

of both these constants on the parameters $\sigma^{(4)}$ and σ^2 and Assumptions 4.5 and 4.6. The proofs of both these Lemmas can be found in Appendix 4.

5.3 Proof of Lemma 4.13

Similarly as in Sects. 5.1 and 5.2, we will analyse our estimator step-by-step. To this end, we first need to define nine auxiliary Markov chains. In what follows, we will combine the ideas for antithetic estimators with respect to the discretisation parameter and with respect to the subsampling parameter. We will therefore need to consider fine and coarse chains with respect to both parameters. We use the notational convention $X^{\text{subsampling, discretisation}}$, hence e.g. $X^{f,c}$ would be a chain that behaves as a fine chain with respect to the subsampling parameter and as a coarse chain with respect to the discretisation parameter. We define three chains that move as fine chains with respect to the discretisation parameter

$$
\begin{cases}
\Delta X_{k+2}^{f,f} = hb(X_{k+1}^{f,f}, U_{k+1}^f) + \beta\sqrt{h}Z_{k+2} \\
\Delta X_{k+1}^{f,f} = hb(X_k^{f,f}, U_k^f) + \beta\sqrt{h}Z_{k+1}
\end{cases}
$$

$$
\begin{cases}
\Delta X_{k+2}^{c-,f} = hb(X_{k+1}^{c-,f}, U_{k+1}^{f,1}) + \beta\sqrt{h}Z_{k+2} \\
\Delta X_{k+1}^{c-,f} = hb(X_k^{c-,f}, U_k^{f,1}) + \beta\sqrt{h}Z_{k+1}
\end{cases}
$$

$$
\begin{cases}
\Delta X_{k+2}^{c+,f} = hb(X_{k+1}^{c+,f}, U_{k+1}^{f,2}) + \beta\sqrt{h}Z_{k+2} \\
\Delta X_{k+1}^{c+,f} = hb(X_k^{c+,f}, U_k^{f,2}) + \beta\sqrt{h}Z_{k+1}
\end{cases}
$$

and six chains that move as coarse chains

$$
\begin{aligned}
\Delta X_{k+2}^{f,c-} &= 2hb(X_k^{f,c-}, U_k^f) + \beta\sqrt{2h}\hat{Z}_{k+2}, \\
\Delta X_{k+2}^{f,c+} &= 2hb(X_k^{f,c+}, U_{k+1}^f) + \beta\sqrt{2h}\hat{Z}_{k+2} \\
\Delta X_{k+2}^{c-,c-} &= 2hb(X_k^{c-,c-}, U_k^{f,1}) + \beta\sqrt{2h}\hat{Z}_{k+2}, \\
\Delta X_{k+2}^{c-,c+} &= 2hb(X_k^{c-,c+}, U_{k+1}^{f,1}) + \beta\sqrt{2h}\hat{Z}_{k+2} \\
\Delta X_{k+2}^{c+,c-} &= 2hb(X_k^{c+,c-}, U_k^{f,2}) + \beta\sqrt{2h}\hat{Z}_{k+2}, \\
\Delta X_{k+2}^{c+,c+} &= 2hb(X_k^{c+,c+}, U_{k+1}^{f,2}) + \beta\sqrt{2h}\hat{Z}_{k+2}.
\end{aligned}
$$

Here $\Delta X_{k+2}^{f,f} = X_{k+2}^{f,f} - X_{k+1}^{f,f}$, $\Delta X_{k+1}^{f,f} = X_{k+1}^{f,f} - X_k^{f,f}$, $\Delta X_{k+2}^{f,c-} = X_{k+2}^{f,c-} - X_k^{f,c-}$ and likewise for other chains, whereas $\sqrt{2h}\hat{Z}_{k+2} = \sqrt{h}Z_{k+1} + \sqrt{h}Z_{k+2}$. In order to prove Lemma 4.13, we will first show that for any $k \ge 1$ we have $\mathbb{E}|\Delta\,\text{Ant}\,\Phi_{g,k}^{s,h}|^2 \le \widetilde{C}\mathbb{E}|\Delta\,\text{Ant}\,\Phi_k^{s,h}|^2 + \bar{C}h^2$, where $\widetilde{C}, \bar{C}$ are positive constants and $\bar{C}$ is of order $\mathcal{O}(s^{-2})$. Here $\Delta\,\text{Ant}\,\Phi_k^{s,h}$ corresponds to taking $\Delta\,\text{Ant}\,\Phi_{g,k}^{s,h}$ with $g(x) = x$ the identity function.

Then we will show that for all $k \ge 1$ we have $\mathbb{E}|\Delta\,\text{Ant}\,\Phi_{k+2}^{s,h}|^2 \le (1 - ch)\mathbb{E}|\Delta\,\text{Ant}\,\Phi_k^{s,h}|^2 + Ch^3$ for some constants $c, C > 0$, where C is of order $\mathcal{O}(s^{-2})$. Finally, we

will conclude that for all $k \geq 1$ we have $\mathbb{E}|\Delta \text{ Ant } \Phi_k^{s,h}|^2 \leq C_1 h^2/s^2$ for some $C_1 > 0$, which will finish the proof.

In order to simplify the notation, we define here one step of the MASGA estimator as

$$
\Psi_k := \left[X_k^{f,f} - \frac{1}{2} \left(X_k^{f,c-} + X_k^{f,c+} \right) \right]
$$
$$
- \frac{1}{2} \left[\left(X_k^{c-,f} - \frac{1}{2} \left(X_k^{c-,c-} + X_k^{c-,c+} \right) \right) \right.
$$
$$
\left. + \left(X_k^{c+,f} - \frac{1}{2} \left(X_k^{c+,c-} + X_k^{c+,c+} \right) \right) \right].
$$

Let us first focus on the analysis of

$$
\mathbb{E}|\Psi_k^g|^2 := \mathbb{E} \left| \left[g(X_k^{f,f}) - \frac{1}{2} \left(g(X_k^{f,c-}) + g(X_k^{f,c+}) \right) \right] \right.
$$
$$
- \frac{1}{2} \left[\left(g(X_k^{c-,f}) - \frac{1}{2} \left(g(X_k^{c-,c-}) + g(X_k^{c-,c+}) \right) \right) \right.
$$
$$
\left. \left. + \left(g(X_k^{c+,f}) - \frac{1}{2} \left(g(X_k^{c+,c-}) + g(X_k^{c+,c+}) \right) \right) \right] \right|^2
$$

for $g : \mathbb{R}^d \to \mathbb{R}$ Lipschitz. To this end, we will introduce three additional chains

$$
\bar{X}_k^{f,c} = \frac{1}{2} \left(X_k^{f,c-} + X_k^{f,c+} \right), \quad \bar{X}_k^{c-,c} = \frac{1}{2} \left(X_k^{c-,c-} + X_k^{c-,c+} \right) \quad \text{and}
$$
$$
\bar{X}_k^{c+,c} = \frac{1}{2} \left(X_k^{c+,c-} + X_k^{c+,c+} \right).
$$

Observe that in the expression $\mathbb{E}|\Psi_{kh}^g|^2$ above we have three rows of the same structure as the antithetic estimator via discretisation, hence we can proceed exactly as in Sect. 5.1 by adding and subtracting $g\left(\bar{X}_k^{f,c}\right)$, $g\left(\bar{X}_k^{c-,c}\right)$ and $g\left(\bar{X}_k^{c+,c}\right)$, respectively in each row, and then applying Taylor's formula in points $\bar{X}_k^{f,c}$, $\bar{X}_k^{c-,c}$ and $\bar{X}_k^{c+,c}$, respectively in each row (note that the first order terms will cancel out) to get

$$
\mathbb{E}|\Psi_k^g|^2 = \mathbb{E} \left| \left[g(X_k^{f,f}) - g\left(\bar{X}_k^{f,c}\right) + T\left(X_k^{f,c-} - X_k^{f,c+}\right) \right] \right.
$$
$$
- \frac{1}{2} \left[\left(g(X_k^{c-,f}) - g\left(\bar{X}_k^{c-,c}\right) + T\left(X_k^{c-,c-} - X_k^{c-,c+}\right) \right) \right.
$$
$$
\left. \left. + \left(g(X_k^{c+,f}) - g\left(\bar{X}_k^{c+,c}\right) + T\left(X_k^{c+,c-} - X_k^{c+,c+}\right) \right) \right] \right|^2,
$$

where

$$
T(X_k^{f,c-} - X_k^{f,c+}) := -\frac{1}{4} \sum_{|\alpha|=2} \int_0^1 (1-t) D^\alpha g\left(\bar{X}_k^{f,c} + t\left(X_k^{f,c-} - \bar{X}_k^{f,c}\right)\right) dt
$$
$$
\left(X_k^{f,c-} - X_k^{f,c+}\right)^\alpha
$$
$$
-\frac{1}{4} \sum_{|\alpha|=2} \int_0^1 (1-t) D^\alpha g\left(\bar{X}_k^{f,c} + t\left(X_k^{f,c+} - \bar{X}_k^{f,c}\right)\right) dt \left(X_k^{f,c+} - X_k^{f,c-}\right)^\alpha,
$$

where we used $X_k^{f,c+} - \bar{X}_k^{f,c} = \frac{1}{2}\left(X_k^{f,c+} - X_k^{f,c-}\right) = -\left(X_k^{f,c-} - \bar{X}_k^{f,c}\right)$ and T is defined analogously for other terms, and hence $|T(x)|^2 \leq C|x|^4$ for some $C > 0$ and for any $x \in \mathbb{R}^d$. Using $(a+b)^2 \leq 2a^2 + 2b^2$ for $\mathbb{E}|\Psi_{kh}^g|^2$ we can separate the terms involving $T(\cdot)$ and, due to Lemma 5.1, we see that we have the correct order in h and s for all the terms expressed as $T(\cdot)$. Hence the remaining term whose dependence on s and h we still need to check can be expressed as

$$
\mathbb{E} \left| \left[g(X_k^{f,f}) - \frac{1}{2} \left(g(X_k^{c-,f}) + g(X_k^{c+,f}) \right) \right] \right.
$$
$$
\left. - \left[g\left(\bar{X}_k^{f,c}\right) - \frac{1}{2} \left(g\left(\bar{X}_k^{c-,c}\right) + g\left(\bar{X}_k^{c+,c}\right) \right) \right] \right|^2.
$$

We will now need yet another auxiliary chain $\bar{X}_k^{c,f} := \frac{1}{2}\left(X_k^{c-,f} + X_k^{c-,f}\right)$. We repeat our argument from the previous step by adding and subtracting $g\left(\bar{X}_k^{c,f}\right)$ and $g\left(\frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right)$, respectively, to the first and the second term in square brackets above, respectively, and applying Taylor's formula in points $\bar{X}_k^{c,f}$ and $\frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)$, respectively (the first order terms again cancel out), to obtain

$$
\mathbb{E} \left| \left[g(X_k^{f,f}) - g\left(\bar{X}_k^{c,f}\right) + T\left(X_k^{c-,f} - X_k^{c+,f}\right) \right] \right.
$$
$$
\left. - \left[g\left(\bar{X}_k^{f,c}\right) - g\left(\frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right) + T\left(\bar{X}_k^{c-,c} - \bar{X}_k^{c+,c}\right) \right] \right|^2,
$$

where again $|T(x)|^2 \leq C|x|^4$ for some $C > 0$ and for any $x \in \mathbb{R}^d$. Due to Lemma 5.2 we see that $\mathbb{E}\left|X_k^{c-,f} - X_k^{c+,f}\right|^4$ has the correct order in s and h. Moreover, we have $\bar{X}_k^{c-,c} - \bar{X}_k^{c+,c} = \frac{1}{2}\left(X_k^{c-,c-} - X_k^{c+,c-}\right) + \frac{1}{2}\left(X_k^{c-,c+} - X_k^{c+,c+}\right)$ and hence, after using $(a+b)^4 \leq 8a^4 + 8b^4$, Lemma 5.2 applies also to $\mathbb{E}\left|\bar{X}_k^{c-,c} - \bar{X}_k^{c+,c}\right|^4$. Hence we only need to deal with

$$
I := \mathbb{E} \left| g(X_k^{f,f}) - g\left(\bar{X}_k^{c,f}\right) - g\left(\bar{X}_k^{f,c}\right) \right.
$$
$$
\left. + g\left(\frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right) \right|^2.
$$

We can now add and subtract $g\left(\bar{X}_k^{f,c} + \bar{X}_k^{c,f} - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right)$, and, using the Lipschitz property of g (which we assume it satisfies with a Lipschitz constant, say, $L_g > 0$), we see that

$$
I \leq 3 L_g \mathbb{E} \left| X_k^{f,f} - \bar{X}_k^{f,c} - \bar{X}_k^{c,f} + \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right) \right|^2
$$

$$+6L_g \mathbb{E}\left|\bar{X}_k^{f,c} - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right|^2.$$

However, observe that $\mathbb{E}\left|X_k^{f,f} - \bar{X}_k^{f,c} - \bar{X}_k^{c,f} + \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right|^2 = \mathbb{E}|\Psi_k|^2$ and, moreover,

$$\bar{X}_k^{f,c} - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right) = \frac{1}{2}\left(X_k^{f,c-} + X_k^{f,c+}\right)$$
$$- \frac{1}{4}\left(X_k^{c-,c-} + X_k^{c-,c+} + X_k^{c+,c-} + X_k^{c+,c+}\right)$$
$$= \frac{1}{2}\left(X_k^{f,c-} - \frac{1}{2}\left(X_k^{c-,c-} + X_k^{c+,c-}\right)\right)$$
$$+ \frac{1}{2}\left(X_k^{f,c+} - \frac{1}{2}\left(X_k^{c-,c+} + X_k^{c+,c+}\right)\right).$$

Note that both terms on the right hand side above correspond to antithetic estimators via subsampling, hence from Lemma 5.3 we infer that $\mathbb{E}\left|\bar{X}_k^{f,c} - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right|^2$ has the correct order in s and h. We have thus demonstrated that for any $k \geq 1$ we have $\mathbb{E}|\Psi_k^g|^2 \leq C_1\mathbb{E}|\Psi_k|^2 + C_2 h^2/s^2$ for some constants $C_1, C_2 > 0$. Therefore, in order to finish the proof, it remains to be shown that $\mathbb{E}|\Psi_k|^2$ has the correct order in h and s. As explained above, this will be achieved by proving that there exist constants $c, C > 0$ (with C being of order $\mathcal{O}(s^{-2})$) such that for any $k \geq 1$ we have $\mathbb{E}|\Psi_{k+2}|^2 \leq (1-ch)\mathbb{E}|\Psi_k|^2 + Ch^3$. The idea for dealing with $\mathbb{E}|\Psi_{k+2}|^2$ is to group the terms in a specific way, add and subtract certain drifts in order to set up appropriate combinations of drifts for a Taylor's formula application and then to cancel out some first order terms. As we have already seen, the remaining second order terms should then be of the correct order in h and s. To this end, we denote

$$\Xi_k^1 := hb(X_k^{f,f}, U_k^f) - hb(X_k^{f,c-}, U_k^f) - hb(X_k^{f,c+}, U_{k+1}^f)$$
$$\Xi_k^2 := hb(X_k^{c-,f}, U_k^{f,1}) - hb(X_k^{c-,c-}, U_k^{f,1}) - hb(X_k^{c-,c+}, U_{k+1}^{f,1})$$
$$\Xi_k^3 := hb(X_k^{c+,f}, U_k^{f,2}) - hb(X_k^{c+,c-}, U_k^{f,2}) - hb(X_k^{c+,c+}, U_{k+1}^{f,2})$$
$$\Xi_k := \Xi_k^1 - \frac{1}{2}\left(\Xi_k^2 + \Xi_k^3\right) \text{ and } \Upsilon_k := hb(X_{k+1}^{f,f}, U_{k+1}^f)$$
$$- \frac{1}{2}hb(X_{k+1}^{c-,f}, U_{k+1}^{f,1}) - \frac{1}{2}hb(X_{k+1}^{c+,f}, U_{k+1}^{f,2})$$

and hence, observing that all the noise variables cancel out, we can write $\Psi_{k+2} = \Psi_k + \Xi_k + \Upsilon_k$. Thus we have $\mathbb{E}|\Psi_{k+2}|^2 = \mathbb{E}|\Psi_k|^2 + 2\mathbb{E}\langle\Psi_k, \Xi_k\rangle + 2\mathbb{E}\langle\Psi_k, \Upsilon_k\rangle + \mathbb{E}|\Xi_k + \Upsilon_k|^2$. We will bound $\mathbb{E}|\Xi_k + \Upsilon_k|^2 \leq 2\mathbb{E}|\Xi_k|^2 + 2\mathbb{E}|\Upsilon_k|^2$ and we will first deal with the terms involving Υ_k. We will need an additional auxiliary Markov chain (moving as a fine chain with respect to the discretisation parameter) defined as

$$X_{k+2} = X_{k+1} + ha(X_{k+1}) + \beta\sqrt{h}Z_{k+2},$$

$$X_{k+1} = X_{kh} + ha(X_k) + \beta\sqrt{h}Z_{k+1}.$$

Using $b(x, U) = \frac{1}{2}b(x, U^1) + \frac{1}{2}b(x, U^2)$, we have

$$\Upsilon_k = hb(X_{k+1}^{f,f}, U_{k+1}^f) - hb(X_{k+1}, U_{k+1}^f)$$
$$- \frac{1}{2}h\left(b(X_{k+1}^{c-,f}, U_{k+1}^{f,1}) - b(X_{k+1}, U_{k+1}^{f,1})\right)$$
$$- \frac{1}{2}h\left(b(X_{k+1}^{c+,f}, U_{k+1}^{f,2}) - b(X_{k+1}, U_{k+1}^{f,2})\right).$$

First, in order to deal with $\mathbb{E}\langle\Psi_k, \Upsilon_k\rangle$, we use Taylor's formula to write

$$\Upsilon_k = h\Bigg(\sum_{|\alpha|=1} D^\alpha b(X_{k+1}, U_{k+1}^f)\left(X_{k+1}^{f,f} - X_{k+1}\right)^\alpha$$
$$+ \sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha b\left(X_{k+1} + t\left(X_{k+1}^{f,f} - X_{k+1}\right), U_{k+1}^f\right)dt$$
$$\left(X_{k+1}^{f,f} - X_{k+1}\right)^\alpha\Bigg)$$
$$- \frac{1}{2}h\Bigg(\sum_{|\alpha|=1} D^\alpha b(X_{k+1}, U_{k+1}^{f,1})\left(X_{k+1}^{c-,f} - X_{k+1}\right)^\alpha$$
$$+ \sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha b\left(X_{k+1} + t\left(X_{k+1}^{c-,f} - X_{k+1}\right), U_{k+1}^{f,1}\right)dt$$
$$\left(X_{k+1}^{c-,f} - X_{k+1}\right)^\alpha\Bigg)$$
$$- \frac{1}{2}h\Bigg(\sum_{|\alpha|=1} D^\alpha b(X_{k+1}, U_{k+1}^{f,2})\left(X_{k+1}^{c+,f} - X_{k+1}\right)^\alpha$$
$$+ \sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha b\left(X_{k+1} + t\left(X_{k+1}^{c+,f} - X_{k+1}\right), U_{k+1}^{f,2}\right)dt$$
$$\left(X_{k+1}^{c+,f} - X_{k+1}\right)^\alpha\Bigg).$$

Hence, using Assumptions 4.7 and 4.8, we have

$$\mathbb{E}\langle\Psi_k, \Upsilon_k\rangle = h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_{k+1})$$
$$\left(X_{k+1}^{f,f} - \frac{1}{2}X_{k+1}^{c-,f} - \frac{1}{2}X_{k+1}^{c+,f}\right)^\alpha\rangle$$
$$+ h\langle\Psi_k, \sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha a\left(X_{k+1} + t\left(X_{k+1}^{f,f} - X_{k+1}\right)\right)dt$$
$$\left(X_{k+1}^{f,f} - X_{k+1}\right)^\alpha\rangle$$
$$+ h\langle\Psi_k, \sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha a\left(X_{k+1} + t\left(X_{k+1}^{c-,f} - X_{k+1}\right)\right)dt$$
$$\left(X_{k+1}^{c-,f} - X_{k+1}\right)^\alpha\rangle$$
$$+ h\langle\Psi_k, \sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha a\left(X_{k+1} + t\left(X_{k+1}^{c+,f} - X_{k+1}\right)\right)dt$$
$$\left(X_{k+1}^{c+,f} - X_{k+1}\right)^\alpha\rangle.$$

Now we can use Young's inequality for each term above with some ε_1, ε_2, ε_3, $\varepsilon_4 > 0$ respectively and, using Assumption 4.3 (condition (4.14)), we get

$$
\begin{aligned}
\mathbb{E}\langle \Psi_k, \Upsilon_k \rangle &\leq \frac{1}{2} h \varepsilon_1 C_{a^{(1)}} \mathbb{E}\,|\Psi_k|^2 \\
&+ \frac{1}{2} h \frac{1}{\varepsilon_1} C_{a^{(1)}} \mathbb{E} \left| X_{k+1}^{f,f} - \frac{1}{2} X_{k+1}^{c-,f} - \frac{1}{2} X_{k+1}^{c+,f} \right|^2 \\
&+ \frac{1}{2} h \varepsilon_2 C_{a^{(2)}} \mathbb{E}\,|\Psi_k|^2 + \frac{1}{2} h \frac{1}{\varepsilon_2} C_{a^{(2)}} \mathbb{E} \left| X_{k+1}^{f,f} - X_{k+1} \right|^4 \\
&+ \frac{1}{2} h \varepsilon_3 C_{a^{(2)}} \mathbb{E}\,|\Psi_k|^2 + \frac{1}{2} h \frac{1}{\varepsilon_3} C_{a^{(2)}} \mathbb{E} \left| X_{k+1}^{c-,f} - X_{k+1} \right|^4 \\
&+ \frac{1}{2} h \varepsilon_4 C_{a^{(2)}} \mathbb{E}\,|\Psi_k|^2 + \frac{1}{2} h \frac{1}{\varepsilon_4} C_{a^{(2)}} \mathbb{E} \left| X_{k+1}^{c+,f} - X_{k+1} \right|^4 .
\end{aligned}
$$

Now note that the second term on the right hand side above is of order $\mathcal{O}(h^2/s^2)$ due to Lemma 5.3. For the other three terms, we need the following auxiliary result.

Lemma 5.4 *Let Assumptions 4.3, 4.5, 4.6 and 4.9 hold. Assuming $X_0^{f,f} = X_0$, there exists a constant $C > 0$ such that for all $k \geq 1$,*

$$
\mathbb{E} \left| X_k^{f,f} - X_k \right|^4 \leq C \frac{1}{s^2} h^2 .
$$

The proof is similar to the proof of Lemma 5.2 and can be found in Appendix 5.

The reasoning in the proof of Lemma 5.4 applies also to $X_k^{c-,f}$ and $X_k^{c+,f}$ in place of $X_k^{f,f}$. Hence we see that $\mathbb{E}\langle \Psi_k, \Upsilon_k \rangle$ is bounded from above by an expression of the form $C_1(\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4) h \mathbb{E}|\Psi_k|^2 + C_2 h$, where ε_i for $i \in \{1, \ldots 4\}$ can be chosen as small as necessary and the constant C_2 is of the correct order in s and h, i.e., of order $\mathcal{O}(h^2/s^2)$. We will explain later how to handle the other terms and how to choose the values of ε_i for $i \in \{1, \ldots 4\}$. Now in order to deal with $\mathbb{E}|\Upsilon_k|^2$ we use a different decomposition for Υ_k, namely we write

$$
\begin{aligned}
\Upsilon_k &= \frac{1}{2} h b(X_{k+1}^{f,f}, U_{k+1}^{f,1}) + \frac{1}{2} h b(X_{k+1}^{f,f}, U_{k+1}^{f,2}) \\
&\quad - \frac{1}{2} h b(X_{k+1}^{c-,f}, U_{k+1}^{f,1}) - \frac{1}{2} h b(X_{k+1}^{c+,f}, U_{k+1}^{f,2}) \\
&= \frac{1}{2} h b(X_{k+1}^{f,f}, U_{k+1}^{f,1}) - \frac{1}{2} h b(\bar{X}_{k+1}^{c,f}, U_{k+1}^{f,1}) \\
&\quad + \frac{1}{2} h b(\bar{X}_{k+1}^{c,f}, U_{k+1}^{f,1}) - \frac{1}{2} h b(X_{k+1}^{c-,f}, U_{k+1}^{f,1}) \\
&\quad + \frac{1}{2} h b(X_{k+1}^{f,f}, U_{k+1}^{f,2}) - \frac{1}{2} h b(\bar{X}_{k+1}^{c,f}, U_{k+1}^{f,2}) \\
&\quad + \frac{1}{2} h b(\bar{X}_{k+1}^{c,f}, U_{k+1}^{f,2}) - \frac{1}{2} h b(X_{k+1}^{c+,f}, U_{k+1}^{f,2}),
\end{aligned}
$$

where $\bar{X}_{k+1}^{c,f} := \frac{1}{2} \left(X_{k+1}^{c-,f} + X_{k+1}^{c+,f} \right)$. Hence, using (4.15) we obtain

$$
\begin{aligned}
\mathbb{E}\,|\Upsilon_k|^2 &\leq \frac{3}{4} h^2 \bar{L}^2 \mathbb{E} \left| X_{k+1}^{f,f} - \bar{X}_{k+1}^{c,f} \right|^2 + \frac{3}{4} h^2 \bar{L}^2 \mathbb{E} \left| X_{k+1}^{f,f} - \bar{X}_{k+1}^{c,f} \right|^2 \\
&+ \frac{3}{4} h^2 \mathbb{E} \left| b(\bar{X}_{k+1}^{c,f}, U_{k+1}^{f,1}) - b(X_{k+1}^{c-,f}, U_{k+1}^{f,1}) \right. \\
&\left. + b(\bar{X}_{k+1}^{c,f}, U_{k+1}^{f,2}) - b(X_{k+1}^{c+,f}, U_{k+1}^{f,2}) \right|^2 .
\end{aligned}
$$

The first two terms on the right hand side above are identical and have the correct order in s and h due to Lemma 5.3. On the other hand, the third term can be dealt with by one more application of Taylor's formula (cf. the calculation for the term J_{33} in the proof of Lemma 5.3 in Appendix 4 for more details).

The terms involving Ξ_k can be handled using similar ideas as above. In particular, for $\mathbb{E}|\Xi_k|^2$ we have the following result.

Lemma 5.5 *Let Assumptions 4.3, 4.5, 4.6, 4.7 and 4.9 hold. Assuming all the auxiliary chains introduced above are initiated at the same point, there exists a constant $C_{1,\Xi} > 0$ such that for all $k \geq 1$,*

$$
\mathbb{E}|\Xi_k|^2 \leq C_{1,\Xi} \frac{1}{s^2} h^2 .
$$

The proof of Lemma 5.5 can be found in Appendix 5. The last term to deal with is $\mathbb{E}\langle \Psi_k, \Xi_k \rangle$. We have the following result.

Lemma 5.6 *Let Assumptions 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9 and 4.10 hold. Assuming all the auxiliary chains introduced above are initiated at the same point, there exist constants $C_{2,\Xi}$, $C_{3,\Xi}$ and $C_{4,\Xi} > 0$ such that for all $k \geq 1$,*

$$
\mathbb{E}\langle \Psi_k, \Xi_k \rangle \leq (C_{2,\Xi}\varepsilon - C_{3,\Xi}) h \mathbb{E}|\Psi_k|^2 + C_{4,\Xi} \frac{1}{s^2} h^2 ,
$$

where $\varepsilon > 0$ can be chosen arbitrarily small.

The crucial insight about the bound in Lemma 5.6 is that, thanks to the special structure of the term $\Xi_k = \Xi_k^1 - \frac{1}{2}(\Xi_k^2 + \Xi_k^3)$, we can extract from $\mathbb{E}\langle \Psi_k, \Xi_k \rangle$, after a few applications of Taylor's formula, a term of the form $h \mathbb{E}\langle \Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)(\Psi_k)^\alpha \rangle$, which, due to Assumptions 4.3, can be bounded from above by $-Kh\mathbb{E}|\Psi_k|^2$. This gives us the term with the constant $C_{3,\Xi}$ in Lemma 5.6. Then, after combining all our estimates and choosing all the $\varepsilon_i > 0$ small enough, we can indeed conclude that for any $k \geq 1$ we have $\mathbb{E}|\Psi_{k+2}|^2 \leq (1 - ch)\mathbb{E}|\Psi_k|^2 + Ch^3$ with C of order $\mathcal{O}(s^{-2})$, which, as explained above, finishes the proof.

The proof of Lemma 5.6 is lengthy and tedious but elementary and hence is moved to Appendix 5.

Appendix: Bounds on moments of subsampling estimators

In this section we present bounds both for subsampling with and without replacement (Welling and Teh 2011; Shamir 2016). We fix $s, m \in \mathbb{N}$ such that $s < m$. Let $\theta_i \in \mathbb{R}^d$, for $i = 1, \ldots, m$. Moreover, let $U = (U_i)_{i=1\ldots,s}$ be a collection of s independent random variables, uniformly distributed over the set $\{1, \ldots, m\}$. We define

$$a(x) = \frac{1}{m} \sum_{i=1}^{m} \hat{b}(x, \theta_i) \quad \text{and} \quad b(x, U) = \frac{1}{s} \sum_{i=1}^{s} \hat{b}(x, \theta_{U_i}). \tag{6.1}$$

Here $\hat{b} : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^d$ is a kernel such that for any x, $y \in \mathbb{R}^d$ and any $\theta \in \mathbb{R}^k$ we have

$$\begin{aligned} |\hat{b}(x, \theta) - \hat{b}(y, \theta)| &\leq L|x - y| \qquad \text{and} \\ \langle x - y, \hat{b}(x, \theta) - \hat{b}(y, \theta) \rangle &\leq -K|x - y|^2. \end{aligned} \tag{6.2}$$

for some $L, K > 0$. Hence b is an unbiased estimator of a that corresponds to sampling with replacement s terms from the sum of m terms defining a, cf. Example 2.15 in Majka et al. (2020). Moreover, Assumptions 4.3 and 4.4 are satisfied with constants L, K and $\bar{L} = L$. We will now verify Assumptions 4.5 and 4.6.

Note that related calculations for second moments of subsampling estimators were carried out in Majka et al. (2020) (see Example 2.15 therein) for the drift a and its estimator b in (6.1) rescaled by m, that is, for $a(x) = \sum_{i=1}^{m} \hat{b}(x, \theta_i)$ and $b(x, U_k) = \frac{m}{s} \sum_{i=1}^{s} \hat{b}(x, \theta_{U_i})$. Hence, obviously, all the upper bounds on second moments obtained in Majka et al. (2020) still hold for a and b given by (6.1), after rescaling by $1/m^2$.

Based on the calculations in Majka et al. (2020), we know that if we assume that for all θ and x we have

$|\hat{b}(x, \theta)|^2 \leq C(1 + |x|^2)$ with some constant $C > 0$, then $\mathbb{E}|b(x, U) - a(x)|^2 \leq \frac{1}{s} C(1 + |x|^2)$, which verifies Assumption 4.5 for the subsampling with replacement scheme. Let us now define a new estimator $b^{wor}(x, U) := \frac{1}{s} \sum_{j=1}^{m} \hat{b}(x, \theta_j) Z_j$, where $(Z_j)_{j=1}^{m}$ are correlated random variables such that $\mathbb{P}(Z_j = 1) = \frac{s}{m}, \mathbb{P}(Z_j = 0) = 1 - \frac{s}{m}$ and $\mathbb{P}(Z_i = 1, Z_j = 1) = \binom{m-2}{s-2} / \binom{m}{s}$ for any $i, j \in \{1, \ldots, m\}$ such that $i \neq j$. Note that this definition of b^{wor} corresponds to sampling s terms from the sum of m terms defining a in (6.1) without replacement, see Example 2.15 in Majka et al. (2020). It is immediate to check that this estimator of a is indeed unbiased. In order to bound the variance, we can first check that $\text{Cov}(Z_i, Z_j) = \frac{s(s-1)}{m(m-1)} - \frac{s^2}{m^2} = -\frac{s(1-\frac{s}{m})}{m(m-1)}$. We have

$$\begin{aligned} &\mathbb{E}|b^{wor}(x, U) - a(x)|^2 \\ &= \mathbb{E} \left| \frac{1}{s} \sum_{j=1}^{m} \hat{b}(x, \theta_j) Z_j - \frac{1}{s} \sum_{j=1}^{m} \frac{s}{m} \hat{b}(x, \theta_j) \right|^2 \\ &= \frac{1}{s^2} \mathbb{E} \left| \sum_{j=1}^{m} \hat{b}(x, \theta_j)(Z_j - \frac{s}{m}) \right|^2 \\ &= \frac{1}{s^2} \left(\mathbb{E} \sum_{j=1}^{m} \hat{b}(x, \theta_j)^2 \left(Z_j - \frac{s}{m} \right)^2 \right. \\ &\quad \left. + \mathbb{E} \sum_{i,j=1,i\neq j}^{m} \hat{b}(x, \theta_j) \left(Z_j - \frac{s}{m} \right) \hat{b}(x, \theta_i)(Z_i - \frac{s}{m}) \right). \end{aligned}$$

Note now that we have

$$\begin{aligned} &\mathbb{E} \left(Z_i - \frac{s}{m} \right) \left(Z_j - \frac{s}{m} \right) \\ &= \mathbb{E} \left(Z_i Z_j - \frac{s}{m} Z_i - \frac{s}{m} Z_j + \frac{s^2}{m^2} \right) \\ &= \frac{\binom{m-2}{s-2}}{\binom{m}{s}} - 2 \frac{s^2}{m^2} + \frac{s^2}{m^2} \\ &= \frac{s(s-1)}{m(m-1)} - \frac{s^2}{m^2} = -\frac{s(m-s)}{m^2(m-1)}. \end{aligned}$$

Hence we can easily check that $\mathbb{E}|b^{wor}(x, U) - a(x)|^2 \leq \frac{1}{s}(1 - \frac{s}{m})C(1 + |x|^2)$. Thus we see that the upper bound on the variance of the estimator b^{wor} that we obtained is equal to the upper bound on the variance of b multiplied by $(1 - \frac{s}{m})$. In particular, this confirms that Assumption 4.5 holds also for the subsampling without replacement scheme.

Let us now explain how to estimate the fourth centered moments required for Assumption 4.6, based on the assump-

tion that for all θ and x we have $|\hat{b}(x,\theta)|^4 \leq C(1+|x|^4)$. We have

$$\mathbb{E}|b(x,U) - a(x)|^4 = \frac{1}{s^4}\mathbb{E}\left|\sum_{i=1}^{s}(\hat{b}(x,\theta_{U_i}) - a(x))\right|^4$$

$$= \frac{1}{s^4}\sum_{i=1}^{s}\mathbb{E}\left|\hat{b}(x,\theta_{U_i}) - a(x)\right|^4$$

$$+ 4\frac{1}{s^4}\sum_{i,j=1,i<j}^{s}\mathbb{E}\left(\hat{b}(x,\theta_{U_i}) - a(x)\right)^3\left(\hat{b}(x,\theta_{U_j}) - a(x)\right)$$

$$+ 6\frac{1}{s^4}\sum_{i,j=1,i<j}^{s}\mathbb{E}\left(\hat{b}(x,\theta_{U_i}) - a(x)\right)^2\left(\hat{b}(x,\theta_{U_j}) - a(x)\right)^2$$

$$+ 24\frac{1}{s^4}\sum_{i,j,k,l=1,i<j<k<l}^{s}\mathbb{E}\left(\hat{b}(x,\theta_{U_i}) - a(x)\right)$$

$$\left(\hat{b}(x,\theta_{U_j}) - a(x)\right)\left(\hat{b}(x,\theta_{U_k}) - a(x)\right)$$

$$\times\left(\hat{b}(x,\theta_{U_l}) - a(x)\right)$$

$$+ 12\frac{1}{s^4}\sum_{i,j,k=1,i<j<k}^{s}\mathbb{E}\left(\hat{b}(x,\theta_{U_i}) - a(x)\right)^2$$

$$\left(\hat{b}(x,\theta_{U_j}) - a(x)\right)\left(\hat{b}(x,\theta_{U_k}) - a(x)\right).$$

Since $(\hat{b}(x,\theta_{U_i}) - a(x))$ are mutually independent, centered random variables, we see that

$$\mathbb{E}|b(x,U) - a(x)|^4 = \frac{1}{s^4}\sum_{i=1}^{s}\mathbb{E}\left|\hat{b}(x,\theta_{U_i}) - a(x)\right|^4$$

$$+ 6\frac{1}{s^4}\sum_{i,j=1,i<j}^{s}\mathbb{E}\left(\hat{b}(x,\theta_{U_i}) - a(x)\right)^2\left(\hat{b}(x,\theta_{U_j}) - a(x)\right)^2.$$

(6.3)

We can now compute for any $i = 1, \ldots, m$

$$\mathbb{E}\left|\hat{b}(x,\theta_{U_i}) - a(x)\right|^4 = \sum_{i=1}^{m}\left(\hat{b}(x,\theta_i) - a(x)\right)^4\frac{1}{m}$$

$$\leq \frac{1}{m}\sum_{i=1}^{m}\left(\hat{b}(x,\theta_i)^4 - 4\hat{b}(x,\theta_i)^3 a(x) + 6\hat{b}(x,\theta_i)^2 a(x)^2\right)$$

$$\leq C(1+|x|^4),$$

where we used the linear growth conditions for $a(x)$ and for $\hat{b}(x,\theta)$. Hence we see that the first term on the right hand side of (6.3) can be bounded by $\frac{1}{s^3}C(1+|x|^4)$. On the other hand, using $\mathbb{E}\left(\hat{b}(x,\theta_{U_i}) - a(x)\right)^2 \leq C(1+|x|^2)$ we see that the second term on the right hand side of (6.3) can be bounded by $\frac{1}{s^4}\binom{s}{2}C(1+|x|^4)$ and hence we obtain

$\mathbb{E}|b(x,U) - a(x)|^4 \leq \frac{1}{s^2}C(1+|x|^4)$. By analogy, for the estimator without replacement b^{wor} we have

$$\mathbb{E}|b^{wor}(x,U) - a(x)|^4 = \frac{1}{s^4}\mathbb{E}\left|\sum_{i=1}^{m}\hat{b}(x,\theta_i)\left(Z_i - \frac{s}{m}\right)\right|^4$$

$$= \frac{1}{s^4}\sum_{i=1}^{m}\mathbb{E}\left|\hat{b}(x,\theta_i)\left(Z_i - \frac{s}{m}\right)\right|^4$$

$$+ 4\frac{1}{s^4}\sum_{i,j=1,i<j}^{m}\mathbb{E}\left(\hat{b}(x,\theta_i)\left(Z_i - \frac{s}{m}\right)\right)^3\left(\hat{b}(x,\theta_j)\left(Z_j - \frac{s}{m}\right)\right)$$

$$+ 6\frac{1}{s^4}\sum_{i,j=1,i<j}^{m}\mathbb{E}\left(\hat{b}(x,\theta_i)\left(Z_i - \frac{s}{m}\right)\right)^2\left(\hat{b}(x,\theta_j)\left(Z_j - \frac{s}{m}\right)\right)^2$$

$$+ 24\frac{1}{s^4}\sum_{i,j,k,l=1,i<j<k<l}^{m}\mathbb{E}\left(\hat{b}(x,\theta_i)\left(Z_i - \frac{s}{m}\right)\right)$$

$$\left(\hat{b}(x,\theta_j)\left(Z_j - \frac{s}{m}\right)\right)\left(\hat{b}(x,\theta_k)\left(Z_k - \frac{s}{m}\right)\right)$$

$$\times\left(\hat{b}(x,\theta_l)\left(Z_l - \frac{s}{m}\right)\right)$$

$$+ 12\frac{1}{s^4}\sum_{i,j,k=1,i<j<k}^{m}\mathbb{E}\left(\hat{b}(x,\theta_i)\left(Z_i - \frac{s}{m}\right)\right)^2$$

$$\left(\hat{b}(x,\theta_j)\left(Z_j - \frac{s}{m}\right)\right)\left(\hat{b}(x,\theta_k)\left(Z_k - \frac{s}{m}\right)\right).$$

(6.4)

Recall that the random variables Z_i are not independent and hence we need to compute all the terms in the sum above. We have

$$\frac{1}{s^4}\sum_{i=1}^{m}\mathbb{E}\left|\hat{b}(x,\theta_i)\left(Z_i - \frac{s}{m}\right)\right|^4$$

$$= \frac{1}{s^4}\sum_{i=1}^{m}|\hat{b}(x,\theta_i)|^4\left[\left(\frac{s}{m}\right)^4\left(1 - \frac{s}{m}\right) + \left(1 - \frac{s}{m}\right)^4\frac{s}{m}\right]$$

$$= \frac{1}{s^4}\sum_{i=1}^{m}|\hat{b}(x,\theta_i)|^4\frac{s(m-s)}{m^4}(3s^2 - 3ms + m^2)$$

$$\leq \frac{1}{s^3}\left(1 - \frac{s}{m}\right)C(1+|x|^4).$$

Let us now focus on the fourth term on the right hand side of (6.4). We have

$$\left(Z_i - \frac{s}{m}\right)\left(Z_j - \frac{s}{m}\right)\left(Z_k - \frac{s}{m}\right)\left(Z_l - \frac{s}{m}\right) = Z_i Z_j Z_k Z_l$$

$$- \frac{s}{m}\left(Z_i Z_j Z_k + Z_i Z_k Z_l + Z_j Z_k Z_l + Z_i Z_j Z_l\right)$$

$$+ \frac{s^2}{m^2}\left(Z_i Z_k + Z_j Z_k + Z_j Z_l + Z_i Z_j + Z_i Z_l + Z_k Z_l\right)$$

$$- \frac{s^3}{m^3}\left(Z_i + Z_j + Z_k + Z_l\right) + \frac{s^4}{m^4}$$

and hence, using the definition of the random variables Z_i,

$$
\begin{aligned}
&\mathbb{E}\left(Z_i - \frac{s}{m}\right)\left(Z_j - \frac{s}{m}\right)\left(Z_k - \frac{s}{m}\right)\left(Z_l - \frac{s}{m}\right) \\
&= \binom{m-4}{s-4}/\binom{m}{s} - 4\frac{s}{m}\binom{m-3}{s-3}/\binom{m}{s} \\
&\quad + 6\frac{s^2}{m^2}\binom{m-2}{s-2}/\binom{m}{s} - 4\frac{s^3}{m^3}\frac{s}{m} + \frac{s^4}{m^4} \\
&= \frac{(s-2)(s-1)s}{(m-2)(m-1)m}\left(\frac{s-3}{m-3} - \frac{s}{m}\right) \\
&\quad + 3\frac{(s-1)s^2}{(m-1)m^2}\left(\frac{s}{m} - \frac{s-2}{m-2}\right) + 3\frac{s^3}{m^3}\left(\frac{s-1}{m-1} - \frac{s}{m}\right).
\end{aligned}
$$

Some straightforward computations allow us then to conclude that the fourth term on the right hand side of (6.4) is bounded by $\frac{1}{s^2}(1 - \frac{s}{m})C(1 + |x|^4)$. Dealing in a similar way with the remaining terms, by tedious by otherwise simple computations we can conclude that

$$
\mathbb{E}|b^{wor}(x, U) - a(x)|^4 \le \frac{1}{s^2}(1 - \frac{s}{m})C(1 + |x|^4).
$$

Appendix: On the advisability of subsampling: a simple example

In this section we illustrate the issues that arise in the analysis of the dependence of the cost of SGAs on the parameters m and s. Let us begin by discussing the MSE estimates (1.3) in more detail.

In order to disentangle various approximation errors in our analysis, it is useful to consider the SDE

$$
dY_t = -\frac{1}{2m}\nabla V(Y_t, \nu^m)dt + \frac{1}{\sqrt{m}}dW_t, \tag{7.1}
$$

where $(W_t)_{t\ge0}$ is the standard Brownian motion in $\mathbb{R}^d$. We remark that (7.1) is the time-changed SDE $d\bar{Y}_t = -\nabla V(\bar{Y}_t, \nu^m)dt + \sqrt{2}dW_t$ and they both have the same limiting stationary distribution π, cf. (Durmus et al. 2017; Xifara et al. 2014). In the analysis of the mean square error, for $t = kh$, we can estimate

$$
\begin{aligned}
\mathrm{MSE}(\mathcal{A}^{f,k,N}) \le\ &|(f, \pi) - (f, \mathrm{Law}(Y_t))| \\
&+ |(f, \mathrm{Law}(Y_t)) - (f, \mathrm{Law}(X_k))| + \left(N^{-1}\mathbb{V}[f(X_k)]\right)^{1/2}, \tag{7.2}
\end{aligned}
$$

where $X_{k+1} = X_k - \frac{h}{2m}\nabla V(X_k, \nu^s) + \sqrt{h/m}Z_{k+1}$ with i.i.d. $(Z_k)_{k=1}^\infty$ with the standard normal distribution. The three terms above are, in order, bias (due to the simulation up to a finite time $t > 0$), weak time discretisation error and the Monte Carlo variance. We choose to work with the SDE (7.1),

to mitigate the effect of m on the Lipschitz and convexity constants that play the key role in the first two errors in (7.2), see the discussion in Nagapetyan et al. (2017). Consequently, we focus on the last term in (7.2), i.e., the variance of $\mathcal{A}^{f,k,N}$, in our analysis.

For convenience we assume that $h = 1/n$ for some $n \ge 1$, which corresponds to taking n steps in each unit time interval. There are numerous results in the stochastic analysis literature for bounding the first term on the right hand side of (7.2) by a quantity of order $\mathcal{O}(e^{-t})$, under fairly general assumptions on ∇V, see e.g. (Eberle 2016; Eberle et al. 2019). Moreover, in our previous paper (Majka et al. 2020) we carried out the weak error analysis (see Theorem 1.5 therein) that provides an upper bound for the second term in (7.2) of order $\mathcal{O}(h)$. Hence, for any algorithm $\mathcal{A}^{f,k,N}$ based on the chain $(X_k)_{k=0}^\infty$ given above, we have

$$
\mathrm{MSE}(\mathcal{A}^{f,k,N}) \lesssim e^{-\lambda t} + \frac{1}{n} + \frac{1}{\sqrt{\Lambda(s,m)N}}, \tag{7.3}
$$

for some $\lambda > 0$ (which is the exponential rate of convergence in the L^1-Wasserstein distance of the SDE (7.1) to π). Here $\Lambda(s, m)$ is a quantity whose exact value depends on the properties of V and the function f, cf. the discussion below. Fix $\varepsilon > 0$ and set $\mathrm{MSE}(\mathcal{A}^{f,k,N}) \lesssim \varepsilon$. This enforces the following choice of the parameters $t \approx \lambda^{-1}\log(\varepsilon^{-1})$, $\Lambda(s, m)N \approx \varepsilon^{-2}$, $n \approx \varepsilon^{-1}$. The cost of simulation of our algorithm is defined as the product of the number of paths N, the number of iterations k of each path and the number of data points s in each iteration. Since $t = kh$ and $h = 1/n$, we have

$$
\mathrm{cost}(\mathcal{A}^{f,k,N}) = tnNs \approx \frac{s}{\Lambda(s,m)}\log(\varepsilon^{-1})\varepsilon^{-3}. \tag{7.4}
$$

The main difficulty in quantifying the cost of such Monte Carlo algorithms stems from the fact that the value of $\Lambda(s, m)$ is problem-specific and depends substantially on the interplay between parameters m, s and h. Hence, one may obtain different costs (and thus different answers to the question of profitability of using mini-batching) for different models and different data regimes (Nagapetyan et al. 2017).

In order to gain some insight into possible values of $\Lambda(s, m)$, we consider a simple example of an Ornstein-Uhlenbeck Markov chain $(X_k)_{k=0}^\infty$ given by

$$
X_{k+1} = X_k - \alpha X_k h + \left(\frac{1}{m}\sum_{i=1}^m \xi_i\right)h + \sqrt{h/m}Z_{k+1}, \tag{7.5}
$$

where $\alpha > 0$ and $(\xi_i)_{i=1}^m$ are data points in $\mathbb{R}^k$, and its stochastic gradient counterpart $(\bar{X}_k)_{k=0}^\infty$ given by

$$
\bar{X}_{k+1} = \bar{X}_k - \alpha\bar{X}_k h + \left(\frac{1}{s}\sum_{i=1}^s \xi_{U_i^k}\right)h + \sqrt{h/m}Z_{k+1}, \tag{7.6}
$$

where for all $i \in \{1, \dots, s\}$ and for all $k \geq 0$ we have $U_i^k \sim \text{Unif}(\{1, \dots, m\})$ and all the random variables U_i^k are mutually independent. Denoting $b := \frac{1}{m} \sum_{i=1}^m \xi_i$ and $\bar{b}^k := \frac{1}{s} \sum_{i=1}^s \xi_{U_i^k}$, we easily see that

$$X_k = (1 - \alpha h)^k X_0 + \sum_{j=1}^k (1 - \alpha h)^{k-j} (bh + \sqrt{h/m} Z_j)$$

and $\bar{X}_k = (1 - \alpha h)^k \bar{X}_0 + \sum_{j=1}^k (1 - \alpha h)^{k-j} (\bar{b}^{j-1} h + \sqrt{h/m} Z_j)$. Since $\mathbb{V}[Z_j] = 1$ for all $j \geq 1$, we observe that

$$\mathbb{V}[X_k] = (1 - \alpha h)^{2k} \mathbb{V}[X_0] + \frac{h}{m} \sum_{j=1}^k (1 - \alpha h)^{2(k-j)}. \quad (7.7)$$

Moreover, $\mathbb{V}[\bar{X}_k] = (1 - \alpha h)^{2k} \mathbb{V}[\bar{X}_0] + \sum_{j=1}^k (1 - \alpha h)^{2(k-j)} \left(h^2 \mathbb{V}[\bar{b}^{j-1}] + \frac{h}{m} \right)$. Following the calculations in Example 2.15 in Majka et al. (2020), we see that for any $j \geq 1$

$$\mathbb{V}[\bar{b}^j] = \frac{1}{s} \left[\frac{1}{m} \sum_{j=1}^m \xi_j^2 - \left(\frac{1}{m} \sum_{j=1}^m \xi_j \right)^2 \right],$$

which, assuming $\bar{X}_0 = X_0$, shows

$$\mathbb{V}[\bar{X}_k] = \mathbb{V}[X_k]$$
$$+ \frac{h^2}{s} \left[\frac{1}{m} \sum_{j=1}^m \xi_j^2 - \left(\frac{1}{m} \sum_{j=1}^m \xi_j \right)^2 \right] \sum_{j=1}^k (1 - \alpha h)^{2(k-j)}. \quad (7.8)$$

Since the sum $\sum_{j=1}^k (1 - \alpha h)^{2(k-j)}$ is of order $1/h$, we infer that $\mathbb{V}[X_k]$ is of order $1/m$, whereas $\mathbb{V}[\bar{X}_k]$ is of order $1/m + h/s$. Note that this corresponds to taking $f(x) = x$ in $\mathcal{A}^{f,k,N}$ and demonstrates that even in this simple case it is not clear whether the algorithm $\mathcal{A}^{f,k,N}$ based on $\bar{X}_k$ is more efficient than the one based on X_k, since the exact values of their costs (7.4) depend on the relation between m, s and h. Note that our analysis in this case is exact, since we used equalities everywhere.

Let us also consider the case of $f(x) = x^2$, which turns out to be more cumbersome. We first assume that $\bar{X}_0 = X_0$ and observe that then $\mathbb{E}[\bar{X}_k] = \mathbb{E}[X_k]$ for all $k \geq 0$ and hence it is sufficient to compare the variances of the centered versions of $\bar{X}_k$ and X_k. More precisely, in our analysis of the algorithm $\mathcal{A}^{f,k,N}$ we want to look at $\mathbb{V}[f(X_k - \mathbb{E}[X_k])]$ and $\mathbb{V}[f(\bar{X}_k - \mathbb{E}[\bar{X}_k])]$ with $f(x) = x^2$ and hence we will compare their respective upper bounds $\mathbb{E}|X_k - \mathbb{E}[X_k]|^4$ and

$\mathbb{E}|\bar{X}_k - \mathbb{E}[\bar{X}_k]|^4$. First we observe that

$$\mathbb{E}|X_k - \mathbb{E}[X_k]|^4 = \mathbb{E} \left| (1 - \alpha h)^k (X_0 - \mathbb{E}[X_0]) \right.$$
$$\left. + \sum_{j=1}^k (1 - \alpha h)^{k-j} \frac{\sqrt{h}}{\sqrt{m}} Z_j \right|^4. \quad (7.9)$$

Hence we can expand the fourth power of the sum as in Sect. 1 and, after taking into account all the cross-terms, we see that $\mathbb{E}|X_k - \mathbb{E}[X_k]|^4$ is of order h/m^2. On the other hand, using $(a + b)^4 \leq 8a^4 + 8b^4$ we get

$$\mathbb{E}|\bar{X}_k - \mathbb{E}[\bar{X}_k]|^4 \leq 8\mathbb{E}|X_k - \mathbb{E}[X_k]|^4$$
$$+ 8\mathbb{E} \left| \sum_{j=1}^k (1 - \alpha h)^{k-j} (\bar{b}^{j-1} - \mathbb{E}[\bar{b}^{j-1}]) h \right|^4. \quad (7.10)$$

Now the analysis follows again Sect. 1 in expanding the fourth power of the sum. Note that similarly as in (6.3) the terms involving $\mathbb{E}(\bar{b}^{j-1} - \mathbb{E}[\bar{b}^{j-1}])$ and $\mathbb{E}(\bar{b}^{j-1} - \mathbb{E}[\bar{b}^{j-1}])^3$ vanish and hence we are left with the terms involving $\mathbb{E}(\bar{b}^{j-1} - \mathbb{E}[\bar{b}^{j-1}])^2$ and $\mathbb{E}(\bar{b}^{j-1} - \mathbb{E}[\bar{b}^{j-1}])^4$, for which we can use the bounds obtained in Example 2.15 in Majka et al. (2020) and in Sect. 1, to conclude that the second term on the right hand side of (7.10) is of order h^3/s^2. Hence we conclude that for the case $f(x) = x^2$ the algorithm based on X_k has the variance of order h/m^2, while the algorithm based on $\bar{X}_k$ has the variance of order $h/m^2 + h^3/s^2$.

Finally, let us analyse $f(x) = \sqrt{x}$. To this end, we again use centered versions and compare $\mathbb{E}|X_k - \mathbb{E}[X_k]|$ with $\mathbb{E}|\bar{X}_k - \mathbb{E}[\bar{X}_k]|$. Similarly as in (7.9) we observe that $\mathbb{E}|X_k - \mathbb{E}[X_k]|$ is of order $1/\sqrt{hm}$ (remembering that $\sum_{j=1}^k (1 - \alpha h)^{k-j}$ is of order $1/h$). Moreover, similarly as in (7.10) we have

$$\mathbb{E}|\bar{X}_k - \mathbb{E}[\bar{X}_k]| \leq \mathbb{E}|X_k - \mathbb{E}[X_k]|$$
$$+ \mathbb{E} \left| \sum_{j=1}^k (1 - \alpha h)^{k-j} (\bar{b}^{j-1} - \mathbb{E}[\bar{b}^{j-1}]) h \right|. \quad (7.11)$$

Hence, recalling once again from Example 2.15 in Majka et al. (2020) that $\mathbb{E}(\bar{b}^j - \mathbb{E}[\bar{b}^j])^2$ is of order $1/s$ for any $j \geq 1$, we can use Jensen's inequality to conclude that $\mathbb{E}(\bar{b}^j - \mathbb{E}[\bar{b}^j])$ is of order $1/\sqrt{s}$ and, consequently, that the second term on the right hand side of (7.11) is also of order $1/\sqrt{s}$ (after cancellation of h with the sum $\sum_{j=1}^k (1 - \alpha h)^{k-j}$ which, as we already pointed out, is of order $1/h$). Hence for $f(x) = \sqrt{x}$ we observe that X_k leads to the variance of order $1/\sqrt{hm}$, whereas $\bar{X}_k$ leads to the variance of order $1/\sqrt{hm} + 1/\sqrt{s}$. Again, in all these cases, determining which term is the leading one depends on the interplay between m, s and h.

Appendix: Uniform bounds on fourth moments of SGLD

Lemma 7.1 *Let Assumptions 4.3, 4.6 and 4.9 hold. Then there exists a constant $C_{IEul}^{(4)} > 0$ such that for the Markov chain $(X_k)_{k=0}^{\infty}$ given by*

$$X_{k+1} = X_k + hb(X_k, U_k) + \beta\sqrt{h}Z_{k+1}$$

with pairwise independent $(U_k)_{k=0}^{\infty}$ satisfying (4.3) and $(Z_k)_{k=0}^{\infty}$ i.i.d., $Z_k \sim N(0, I)$ and independent of $(U_k)_{k=0}^{\infty}$, we have

$$\mathbb{E}|X_k|^4 \le C_{IEul}^{(4)}$$

for all $k \ge 1$ and $h \in (0, h_0)$, where $C_{IEul}^{(4)} := \mathbb{E}|X_0|^4 + \frac{C_{ult}^{(4)}}{c}$ with $C_{ult}^{(4)} > 0$ given by (7.13) and $c, h_0 > 0$ determined by (7.12).

Proof By a standard computation, we have

$$
\begin{aligned}
\mathbb{E}|X_{k+1}|^4 &\le \mathbb{E}|X_k|^4 + h^4\mathbb{E}|b(X_k, U_k)|^4 \\
&\quad + \beta^4 h^2 \mathbb{E}|Z_{k+1}|^4 + 6h^2\mathbb{E}|X_k|^2|b(X_k, U_k)|^2 \\
&\quad + 6\beta^2 h\mathbb{E}|X_k|^2|Z_{k+1}|^2 + 6\beta^2 h^3\mathbb{E}|b(X_k, U_k)|^2|Z_{k+1}|^2 \\
&\quad + 4h\mathbb{E}|X_k|^2\langle X_k, b(X_k, U_k)\rangle \\
&\quad + 4\beta\sqrt{h}\mathbb{E}|X_k|^2\langle X_k, Z_{k+1}\rangle \\
&\quad + 4h^3\mathbb{E}|b(X_k, U_k)|^2\langle b(X_k, U_k), X_k\rangle \\
&\quad + 4\beta h^3\mathbb{E}|b(X_k, U_k)|^2\langle b(X_k, U_k), Z_{k+1}\rangle \\
&\quad + 4\beta^3 h^{3/2}\mathbb{E}|Z_{k+1}|^2\langle Z_{k+1}, X_k\rangle \\
&\quad + 4\beta^3 h^{3/2}\mathbb{E}|Z_{k+1}|^2\langle Z_{k+1}, b(X_k, U_k)\rangle \\
&\quad + 6\beta h^{3/2}\mathbb{E}|X_k|^2\langle b(X_k, U_k), Z_{k+1}\rangle \\
&\quad + 6\beta h^{5/2}\mathbb{E}|b(X_k, U_k)|^2\langle X_k, Z_{k+1}\rangle \\
&\quad + 6\beta^2 h^2\mathbb{E}|Z_{k+1}|^2\langle X_k, b(X_k, U_k)\rangle =: \sum_{j=1}^{15} \tilde{I}_j.
\end{aligned}
$$

By conditioning on X_k and U_k and using properties of the multivariate normal distribution, we see that $\tilde{I}_8 = \tilde{I}_{10} = \tilde{I}_{11} = \tilde{I}_{12} = \tilde{I}_{13} = \tilde{I}_{14} = 0$. Hence we have

$$
\begin{aligned}
\mathbb{E}|X_{k+1}|^4 &\le \mathbb{E}|X_k|^4 + h^4\mathbb{E}|b(X_k, U_k)|^4 + \beta^4 h^2\mathbb{E}|Z_{k+1}|^4 \\
&\quad + 6h^2\mathbb{E}|X_k|^2|b(X_k, U_k)|^2 \\
&\quad + 6\beta^2 h\mathbb{E}|X_k|^2|Z_{k+1}|^2 + 6\beta^2 h^3\mathbb{E}|b(X_k, U_k)|^2|Z_{k+1}|^2 \\
&\quad + 4h\mathbb{E}|X_k|^2\langle X_k, b(X_k, U_k)\rangle \\
&\quad + 4h^3\mathbb{E}|b(X_k, U_k)|^2\langle b(X_k, U_k), X_k\rangle \\
&\quad + 6\beta^2 h^2\mathbb{E}|Z_{k+1}|^2\langle X_k, b(X_k, U_k)\rangle =: \sum_{j=1}^9 I_j.
\end{aligned}
$$

Now observe that due to Assumptions 4.9 and 4.6 we have $\mathbb{E}|b(x, U)|^4 \le \bar{L}_0^{(4)}(1 + |x|^4)$, where

$$\bar{L}_0^{(4)} := 8(\sigma^{(4)} + L_0^{(4)}).$$

Indeed, we have

$$
\begin{aligned}
\mathbb{E}|b(x, U)|^4 &= \mathbb{E}|b(x, U) - a(x) + a(x)|^4 \\
&\le 8\mathbb{E}|b(x, U) - a(x)|^4 + 8\mathbb{E}|a(x)|^4.
\end{aligned}
$$

Moreover, we have $\left(\mathbb{E}|b(x, U)|^3\right)^{4/3} \le \mathbb{E}\left(|b(x, U)|^3\right)^{4/3} = \mathbb{E}|b(x, U)|^4 \le \bar{L}_0^{(4)}(1 + |x|^4)$ and hence

$$
\begin{aligned}
\mathbb{E}|b(x, U)|^3 &\le \left(\bar{L}_0^{(4)}\right)^{3/4}\left(1 + |x|^4\right)^{3/4} \\
&\le \left(\bar{L}_0^{(4)}\right)^{3/4}\left(1 + |x|^3\right).
\end{aligned}
$$

These auxiliary estimates allow us to bound $I_2 \le h^4\bar{L}_0^{(4)}\left(1 + \mathbb{E}|X_k|^4\right)$ and

$$
\begin{aligned}
I_8 &\le 4h^3\mathbb{E}|b(X_k, U_k)|^3|X_k| \\
&\le 4h^3\left(\bar{L}_0^{(4)}\right)^{3/4}\left(C_{IEul}^{1/2} + \mathbb{E}|X_k|^4\right).
\end{aligned}
$$

Moreover, by conditioning, we get

$$
\begin{aligned}
I_3 &= \beta^4 h^2(d^2 + 2d) \\
I_4 &\le 6h^2\bar{L}_0\left(\mathbb{E}|X_k|^2 + \mathbb{E}|X_k|^4\right) \le 6h^2\bar{L}_0 C_{IEul} + 6h^2\bar{L}_0\mathbb{E}|X_k|^4 \\
I_5 &= 6\beta^2 hd\mathbb{E}|X_k|^2 \le 6\beta^2 hdC_{IEul} \\
I_6 &= 6\beta^2 h^3 d\mathbb{E}|b(X_k, U_k)|^2 \le 6\beta^2 h^3 d\bar{L}_0(1 + \mathbb{E}|X_k|^2) \\
&\le 6\beta^2 h^3 d\bar{L}_0(1 + C_{IEul}) \\
I_7 &= 4h\mathbb{E}|X_k|^2\langle X_k, a(X_k)\rangle \le 4hM_2\mathbb{E}|X_k|^2 - 4hM_1\mathbb{E}|X_k|^4 \\
I_9 &\le 6\beta^2 h^2 dM_2 - 6\beta^2 h^2 dM_1\mathbb{E}|X_k|^2,
\end{aligned}
$$

where in I_4 and I_6 we used (5.5) and in I_7 and I_9 we used (5.4). Hence we obtain

$$
\begin{aligned}
\mathbb{E}|X_{k+1}|^4 &\le -4hM_1\mathbb{E}|X_k|^4 \\
&\quad + \left(1 + h^4\bar{L}_0^{(4)} + 6h^2\bar{L}_0 + 4h^3\left(\bar{L}_0^{(4)}\right)^{3/4}\right)\mathbb{E}|X_k|^4 \\
&\quad + \beta^4 h^2(d^2 + 2d) + 6\beta^2 h^2 dM_2 \\
&\quad + 4h^3\left(\bar{L}_0^{(4)}\right)^{3/4}C_{IEul}^{1/2} + 4hM_2 C_{IEul} \\
&\quad + 6\beta^2 h^3 d\bar{L}_0(1 + C_{IEul}) \\
&\quad + 6\beta^2 hdC_{IEul} + 6h^2\bar{L}_0 C_{IEul} + h^4\bar{L}_0^{(4)}.
\end{aligned}
$$

We can now choose constants $c, h_0 > 0$ such that for all $h \in (0, h_0)$ we have

$$h^4\bar{L}_0^{(4)} + 6h^2\bar{L}_0 + 4h^3\left(\bar{L}_0^{(4)}\right)^{3/4} - 4hM_1 \le -ch \quad (7.12)$$

holds for all $h \in (0, h_0)$. Then, putting

$$
\begin{aligned}
C_{ult}^{(4)} := {} & \beta^4 h_0 (d^2 + 2d) + 6\beta^2 h_0 d M_2 \\
& + 4h_0^2 \left(\bar{L}_0^{(4)}\right)^{3/4} C_{IEul}^{1/2} + 4M_2 C_{IEul} \\
& + 6\beta^2 h_0^2 d \bar{L}_0 (1 + C_{IEul}) + 6\beta^2 d C_{IEul} \\
& + 6h_0 \bar{L}_0 C_{IEul} + h_0^3 \bar{L}_0^{(4)}
\end{aligned} \tag{7.13}
$$

we get $\mathbb{E}|X_{k+1}|^4 \leq (1 - ch)\mathbb{E}|X_k|^4 + C_{ult}^{(4)} h$ for all $h \in (0, h_0)$, and hence

$$
\begin{aligned}
\mathbb{E}|X_{k+1}|^4 & \leq (1 - ch)^{k+1} \mathbb{E}|X_0|^4 + \sum_{j=0}^{k} C_{ult}^{(4)} (1 - ch)^j h \\
& \leq (1 - ch)^{k+1} \mathbb{E}|X_0|^4 + \frac{C_{ult}^{(4)}}{c}.
\end{aligned}
$$

$\square$

Appendix: Proofs for AMLMC via subsampling

Proof of Lemma 5.2

Proof We have

$$
\begin{aligned}
\mathbb{E}\left|X_{k+1}^{c+} - X_{k+1}^{c-}\right|^4 & = \mathbb{E}\left|X_k^{c+} + h b(X_k^{c+}, U_k^{f,2}) \right. \\
& \quad \left. - X_k^{c-} - h b(X_k^{c-}, U_k^{f,1})\right|^4 \\
& \leq \mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4 \\
& \quad + 4\mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^2 \left\langle X_k^{c+} - X_k^{c-},\right. \\
& \quad \left. h b(X_k^{c+}, U_k^{f,2}) - h b(X_k^{c-}, U_k^{f,1})\right\rangle \\
& \quad + 6\mathbb{E}\left[\left|X_k^{c+} - X_k^{c-}\right|^2 \left|h b(X_k^{c+}, U_k^{f,2}) - h b(X_k^{c-}, U_k^{f,1})\right|^2\right] \\
& \quad + 4\mathbb{E}\left\langle X_k^{c+} - X_k^{c-}, h b(X_k^{c+}, U_k^{f,2}) - h b(X_k^{c-}, U_k^{f,1})\right\rangle \\
& \quad \left|h b(X_k^{c+}, U_k^{f,2}) - h b(X_k^{c-}, U_k^{f,1})\right|^2 \\
& \quad + \mathbb{E}\left|h b(X_k^{c+}, U_k^{f,2}) - h b(X_k^{c-}, U_k^{f,1})\right|^4 \\
& =: B_1 + B_2 + B_3 + B_4 + B_5.
\end{aligned}
$$

We obtain $B_2 \leq -4hK\mathbb{E}|X_k^{c+} - X_k^{c-}|^4$ by conditioning on X_k^{c+} and X_k^{c-} and using Assumption 4.3(ii). In order to deal with B_5 we write

$$
\begin{aligned}
b(X_k^{c+}, & U_k^{f,2}) - b(X_k^{c-}, U_k^{f,1}) \\
& = b(X_k^{c+}, U_k^{f,2}) - a(X_k^{c+}) + a(X_k^{c-}) - a(X_k^{c+}) \quad (7.14) \\
& \quad + a(X_k^{c-}) - b(X_k^{c-}, U_k^{f,1}).
\end{aligned}
$$

We will now use the inequality $\left(\sum_{j=1}^{n} a_j\right)^k \leq n^{k-1} \left(\sum_{j=1}^{n} a_j^k\right)$, which holds for all $a_j \geq 0$ and all integers $k \geq 2$ and $n \geq 1$ due to the Hölder inequality for sums. Hence we have $(a + b + c)^4 \leq 27(a^4 + b^4 + c^4)$ and we obtain

$$
\begin{aligned}
B_5 \leq 27h^4 \Big(& \mathbb{E}\left|b(X_k^{c+}, U_k^{f,2}) - a(X_k^{c+})\right|^4 \\
& + \mathbb{E}\left|a(X_k^{c+}) - a(X_k^{c-})\right|^4 \\
& + \mathbb{E}\left|b(X_k^{c-}, U_k^{f,1}) - a(X_k^{c-})\right|^4 \Big).
\end{aligned}
$$

Hence, due to Assumption 4.6, we get $B_5 \leq 27h^4 \left(2\sigma^{(4)}(1 + C_{IEul}^{(4),(2h)}) + L^4 \mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4\right)$. Using the Cauchy-Schwarz inequality, we now have

$$
\begin{aligned}
B_4 & \leq 4h^3 \left(\mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^2 \left|b(X_k^{c+}, U_k^{f,2}) - b(X_k^{c-}, U_k^{f,1})\right|^2\right)^{1/2} \\
& \quad \times \left(\mathbb{E}\left|b(X_k^{c+}, U_k^{f,2}) - b(X_k^{c-}, U_k^{f,1})\right|^4\right)^{1/2} \\
& \leq 4h^3 \left(\left(\mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4\right)^{1/2} \right. \\
& \quad \left(\mathbb{E}\left|b(X_k^{c+}, U_k^{f,2}) - b(X_k^{c-}, U_k^{f,1})\right|^4\right)^{1/2}\Big)^{1/2} \\
& \quad \times \left(27 \left(2\sigma^{(4)}(1 + C_{IEul}^{(4),(2h)}) + L^4 \mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4\right)\right)^{1/2}.
\end{aligned}
$$

Hence, after applying the inequalities $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ and $(a + b)^{1/2} \leq a^{1/2} + b^{1/2}$ several times, we obtain

$$
\begin{aligned}
B_4 & \leq 4h^3 \left(\frac{1}{2}\mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4 + \frac{1}{2}\mathbb{E}\left|b(X_k^{c+}, U_k^{f,2}) - b(X_k^{c-}, U_k^{f,1})\right|^4\right)^{1/2} \\
& \quad \times \left(\left(54\sigma^{(4)}(1 + C_{IEul}^{(4),(2h)})\right)^{1/2} + \left(27L^4 \mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4\right)^{1/2}\right) \\
& \leq 4h^3 \left(\frac{1}{2}\mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4 + \frac{27}{2}L^4 \mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4 \right. \\
& \quad \left. + \frac{54}{2}\sigma^{(4)}(1 + C_{IEul}^{(4),(2h)})\right)^{1/2} \\
& \quad \times \left(\left(54\sigma^{(4)}(1 + C_{IEul}^{(4),(2h)})\right)^{1/2} + \left(27L^4 \mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4\right)^{1/2}\right) \\
& \leq 4h^3 \left(\left(\frac{1 + 27L^4}{2}27L^4\right)^{1/2} \mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4\right. \\
& \quad + \left(\frac{1 + 27L^4}{2}\mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4\right)^{1/2} \left(54\sigma^{(4)}(1 + C_{IEul}^{(4),(2h)})\right)^{1/2} \\
& \quad + 27\sqrt{2}\sigma^{(4)}(1 + C_{IEul}^{(4),(2h)}) + \left(27\sigma^{(4)}(1 + C_{IEul}^{(4),(2h)})\right)^{1/2} \\
& \quad \left(27L^4 \mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4\right)^{1/2}\Big) \\
& \leq 4h^3 \left(\left(\frac{1 + 27L^4}{2}27L^4\right)^{1/2} + \frac{1 + 81L^4}{4}\right) \mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4 \\
& \quad + 4h^3 \left(27\sqrt{2} + \frac{81}{2}\right)\sigma^{(4)}(1 + C_{IEul}^{(4),(2h)}).
\end{aligned}
$$

In order to deal with B_3, we again use the decomposition (7.14). Then, conditioning on X_k^{c+} and X_k^{c-}, and using Assumption 4.5, we get

$$
B_3 \le 6h^2 \Big(3\mathbb{E}\left(|X_k^{c+} - X_k^{c-}|^2 \sigma^2 (1 + |X_k^{c+}|^2)\right)
$$
$$
+ 3L^2 \mathbb{E}|X_k^{c+} - X_k^{c-}|^4
$$
$$
+ 3\mathbb{E}\left(|X_k^{c+} - X_k^{c-}|^2 \sigma^2 (1 + |X_k^{c-}|^2)\right) \Big).
$$

Now, using the Cauchy-Schwarz inequality and $ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2$, we have

$$
\mathbb{E}\left(|X_k^{c+} - X_k^{c-}|^2 \sigma^2 (1 + |X_k^{c+}|^2) h^2\right)
$$
$$
\le \left(\mathbb{E}|X_k^{c+} - X_k^{c-}|^4 h\varepsilon\right)^{1/2} \left(\sigma^4 \mathbb{E}(1 + |X_k^{c+}|^2)^2 h^3 \frac{1}{\varepsilon}\right)^{1/2}
$$
$$
\le \frac{1}{2}\mathbb{E}|X_k^{c+} - X_k^{c-}|^4 h\varepsilon + \frac{1}{2\varepsilon}h^3 \mathbb{E}\left(1 + |X_k^{c+}|^2\right)^2 \sigma^4
$$
$$
\le \frac{1}{2}\mathbb{E}|X_k^{c+} - X_k^{c-}|^4 h\varepsilon + \frac{1}{2\varepsilon}h^3 \left(1 + 2C_{IEul}^{(2h)} + C_{IEul}^{(4),(2h)}\right) \sigma^4
$$

for some $\varepsilon > 0$ to be specified later. Hence

$$
B_3 \le 18\mathbb{E}|X_k^{c+} - X_k^{c-}|^4 h\varepsilon + 18\frac{1}{\varepsilon}h^3
$$
$$
\left(1 + 2C_{IEul}^{(2h)} + C_{IEul}^{(4),(2h)}\right)
$$
$$
\sigma^4 + 18h^2 L^2 \mathbb{E}|X_k^{c+} - X_k^{c-}|^4.
$$

Note that here $\sigma^4 = (\sigma^2)^2$, where σ^2 is given in Assumption 4.5, whereas $\sigma^{(4)}$ appearing in the bounds on B_4 and B_5 is possibly a different quantity, given in Assumption 4.6. Combining all our estimates together, we obtain

$$
\mathbb{E}|X_{k+1}^{c+} - X_{k+1}^{c-}|^4 \le C_1 h^3
$$
$$
+ \Big(1 - 4hK + 18h\varepsilon + 18h^2 L^2 + 27h^4 L^4
$$
$$
+ 4h^3 \left(\left(\frac{1 + 27L^4}{2}27L^4\right)^{1/2} + \frac{1 + 81L^4}{4}\right)\Big)
$$
$$
\times \mathbb{E}|X_k^{c+} - X_k^{c-}|^4,
$$

where

$$
C_1 := 18\frac{1}{\varepsilon}\left(1 + 2C_{IEul}^{(2h)} + C_{IEul}^{(4),(2h)}\right)\sigma^4
$$
$$
+ 4\left(27\sqrt{2} + 54h_0\right)\sigma^{(4)}(1 + C_{IEul}^{(4),(2h)}). \tag{7.15}
$$

Hence, choosing h, c_1, ε such that

$$
-4hK + 18h\varepsilon + 18h^2 L^2 + 27h^4 L^4
$$
$$
+ 4h^3 \left(\left(\frac{1 + 27L^4}{2}27L^4\right)^{1/2}\right.
$$
$$
\left. + \frac{1 + 81L^4}{4}\right) \le -c_1 h, \tag{7.16}
$$

we obtain, for any $k \ge 1$, $\mathbb{E}|X_{k+1}^{c+} - X_{k+1}^{c-}|^4 \le (1 - c_1 h)^k \mathbb{E}|X_0^{c+} - X_0^{c-}|^4 + \sum_{j=0}^k C_1(1 - c_1 h)^j h^3$. Taking $X_0^{c+} = X_0^{c-}$ and bounding the sum above by an infinite sum gives $\mathbb{E}|X_k^{c+} - X_k^{c-}|^4 \le (C_1/c_1)h^2$ for all $k \ge 1$ and finishes the proof. □

Proof of Lemma 5.3

Proof Using $b(X_k^f, U_k^f) = \frac{1}{2}b(X_k^f, U_k^{f,1}) + \frac{1}{2}b(X_k^f, U_k^{f,2})$, we have

$$
\mathbb{E}\left|X_{k+1}^f - \bar{X}_{k+1}^c\right|^2 = \mathbb{E}\left|X_k^f - \bar{X}_k^c\right|^2
$$
$$
+ 2h\mathbb{E}\left\langle X_k^f - \bar{X}_k^c, \frac{1}{2}b(X_k^f, U_k^{f,1}) + \frac{1}{2}b(X_k^f, U_k^{f,2})\right.
$$
$$
\left. - \frac{1}{2}b(X_k^{c-}, U_k^{f,1}) - \frac{1}{2}b(X_k^{c+}, U_k^{f,2})\right\rangle
$$
$$
+ h^2 \mathbb{E}\left|\frac{1}{2}b(X_k^f, U_k^{f,1}) + \frac{1}{2}b(X_k^f, U_k^{f,2}) - \frac{1}{2}b(X_k^{c-}, U_k^{f,1})\right.
$$
$$
\left. - \frac{1}{2}b(X_k^{c+}, U_k^{f,2})\right|^2 =: J_1 + J_2 + J_3.
$$
$$
\tag{7.17}
$$

We begin by bounding J_2. We have

$$
J_2 = h\mathbb{E}\left\langle X_k^f - \bar{X}_k^c, b(X_k^f, U_k^{f,1}) - b(\bar{X}_k^c, U_k^{f,1})\right\rangle
$$
$$
+ h\mathbb{E}\left\langle X_k^f - \bar{X}_k^c, b(\bar{X}_k^c, U_k^{f,1}) - b(X_k^{c-}, U_k^{f,1})\right\rangle
$$
$$
+ h\mathbb{E}\left\langle X_k^f - \bar{X}_k^c, b(X_k^f, U_k^{f,2}) - b(\bar{X}_k^c, U_k^{f,2})\right\rangle
$$
$$
+ h\mathbb{E}\left\langle X_k^f - \bar{X}_k^c, b(\bar{X}_k^c, U_k^{f,2}) - b(X_k^{c+}, U_k^{f,2})\right\rangle
$$
$$
=: I_1 + I_2 + I_3 + I_4.
$$

By conditioning on X_k^f, X_k^{c+} and X_k^{c-} and using Assumption 4.3 (specifically condition (4.13)) we get $I_1 = I_3 = h\mathbb{E}\left\langle X_k^f - \bar{X}_k^c, a(X_k^f) - a(\bar{X}_k^c)\right\rangle \le -hK\mathbb{E}\left|X_k^f - \bar{X}_k^c\right|^2$, while for the other terms we have $I_2 = h\mathbb{E}\left\langle X_k^f - \bar{X}_k^c, a(\bar{X}_k^c)\right.$

$-a(X_k^{c-})\rangle$ and $I_4 = h\mathbb{E}\left\langle X_k^f - \bar{X}_k^c, a(\bar{X}_k^c) - a(X_k^{c+})\right\rangle$. We now use the Taylor formula for a and (5.6) to write

$$
\begin{aligned}
I_2 + I_4 &= h\mathbb{E}\Big\langle X_k^f - \bar{X}_k^c, -\sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha a\left(\bar{X}_k^c\right.\\
&\quad + t\left(\bar{X}_k^c - X_k^{c-}\right)\big) dt \left(\bar{X}_k^c - X_k^{c-}\right)^\alpha \\
&\quad - \sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha a\left(\bar{X}_k^c + t\left(\bar{X}_k^c - X_k^{c+}\right)\right) dt \left(\bar{X}_k^c - X_k^{c+}\right)^\alpha\Big\rangle \\
&\le hC_{a^{(2)}}\mathbb{E}|X_k^f - \bar{X}_k^c| \cdot |\bar{X}_k^c - X_k^{c-}|^2 \\
&\quad + hC_{a^{(2)}}\mathbb{E}|X_k^f - \bar{X}_k^c| \cdot |\bar{X}_k^c - X_k^{c+}|^2 \\
&= \frac{1}{2}C_{a^{(2)}}\mathbb{E}\left[|X_k^f - \bar{X}_k^c|h^{1/2} \cdot |X_k^{c+} - X_k^{c-}|^2 h^{1/2}\right] \\
&\le \frac{1}{2}C_{a^{(2)}}\left(\mathbb{E}|X_k^f - \bar{X}_k^c|^2\varepsilon_1 h\right)^{1/2} \cdot \left(\mathbb{E}|X_k^{c+} - X_k^{c-}|^4 \frac{1}{\varepsilon_1}h\right)^{1/2} \\
&\le \frac{1}{4}C_{a^{(2)}}\mathbb{E}|X_k^f - \bar{X}_k^c|^2\varepsilon_1 h + \frac{1}{4}C_{a^{(2)}}\mathbb{E}|X_k^{c+} - X_k^{c-}|^4 \frac{1}{\varepsilon_1}h,
\end{aligned}
$$

for some $\varepsilon_1 > 0$ whose exact value will be specified later, where we used the Cauchy-Schwarz inequality and the elementary inequality $ab \le \frac{1}{2}(a^2 + b^2)$.

We now come back to (7.17) and deal with J_3. We have

$$
\begin{aligned}
J_3 &= \frac{1}{4}h^2\mathbb{E}\Big|b(X_k^f, U_k^{f,1}) - b(\bar{X}_k^c, U_k^{f,1}) + b(\bar{X}_k^c, U_k^{f,1}) \\
&\quad - b(X_k^{c-}, U_k^{f,1}) \\
&\quad + b(X_k^f, U_k^{f,2}) - b(\bar{X}_k^c, U_k^{f,2}) + b(\bar{X}_k^c, U_k^{f,2}) \\
&\quad - b(X_k^{c+}, U_k^{f,2})\Big|^2 \\
&\le \frac{3}{4}h^2\bar{L}^2\mathbb{E}\left|X_k^f - \bar{X}_k^c\right|^2 + \frac{3}{4}h^2\bar{L}^2\mathbb{E}\left|X_k^f - \bar{X}_k^c\right|^2 \\
&\quad + \frac{3}{4}h^2\mathbb{E}\Big|b(\bar{X}_k^c, U_k^{f,1}) - b(X_k^{c-}, U_k^{f,1}) + b(\bar{X}_k^c, U_k^{f,2}) \\
&\quad - b(X_k^{c+}, U_k^{f,2})\Big|^2 =: J_{31} + J_{32} + J_{33},
\end{aligned}
$$

where we used the Lipschitz condition (4.15). Note that we have $J_{31} + J_{32} = \frac{3}{2}h^2\bar{L}^2\mathbb{E}\left|X_k^f - \bar{X}_k^c\right|^2$. On the other hand, in order to deal with J_{33}, we use the Taylor theorem to write

$$
\begin{aligned}
&b(\bar{X}_k^c, U_k^{f,1}) - b(X_k^{c-}, U_k^{f,1}) \\
&= -\Big[\sum_{|\alpha|=1} D^\alpha b(\bar{X}_k^c, U_k^{f,1})\left(\bar{X}_k^c - X_k^{c-}\right)^\alpha \\
&\quad + \sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha b\left(\bar{X}_k^c + t\left(\bar{X}_k^c - X_k^{c-}\right),\right. \\
&\quad \left.U_k^{f,1}\right) dt \left(\bar{X}_k^c - X_k^{c-}\right)^\alpha\Big].
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
&b(\bar{X}_k^c, U_k^{f,1}) - b(X_k^{c-}, U_k^{f,1}) + b(\bar{X}_k^c, U_k^{f,2}) - b(X_k^{c+}, U_k^{f,2}) \\
&= \frac{1}{2}\sum_{|\alpha|=1}\Big(D^\alpha b(\bar{X}_k^c, U_k^{f,1}) - D^\alpha a(\bar{X}_k^c) + D^\alpha a(\bar{X}_k^c) \\
&\quad - D^\alpha b(\bar{X}_k^c, U_k^{f,2})\Big)\left(X_k^{c+} - X_k^{c-}\right)^\alpha \\
&\quad + \frac{1}{4}\sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha b\left(\bar{X}_k^c + t\left(\bar{X}_k^c - X_k^{c-}\right), U_k^{f,1}\right) dt \\
&\quad \left(X_k^{c+} - X_k^{c-}\right)^\alpha \\
&\quad - \frac{1}{4}\sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha b\left(\bar{X}_k^c + t\left(\bar{X}_k^c - X_k^{c+}\right), U_k^{f,2}\right) dt \\
&\quad \left(X_k^{c+} - X_k^{c-}\right)^\alpha.
\end{aligned}
$$

Recall that we assume (in Assumption 4.7) that $\mathbb{E}\,|\nabla b(x, U) - \nabla a(x)|^4 \le \sigma^{(4)}(1 + |x|^4)$ and that $|D^\alpha b(x, U)| \le C_{b^{(2)}}$ for all multiindices α with $|\alpha| = 2$. Hence we have

$$
\begin{aligned}
J_{33} &\le \frac{3}{4}h^2\mathbb{E}\left[\left|\nabla b(\bar{X}_k^c, U_k^{f,1}) - \nabla a(\bar{X}_k^c)\right|^2 \left|X_k^{c+} - X_k^{c-}\right|^2\right] \\
&\quad + \frac{3}{4}h^2\mathbb{E}\left[\left|\nabla a(\bar{X}_k^c) - \nabla b(\bar{X}_k^c, U_k^{f,2})\right|^2 \left|X_k^{c+} - X_k^{c-}\right|^2\right] \\
&\quad + \frac{3}{8}h^2C_{b^{(2)}}^2\mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4 \\
&\le \frac{3}{4}h^3\sigma^{(4)}(1 + \mathbb{E}|\bar{X}_k^c|^4) + \frac{3}{4}h\mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4 \\
&\quad + \frac{3}{8}h^2C_{b^{(2)}}^2\mathbb{E}\left|X_k^{c+} - X_k^{c-}\right|^4,
\end{aligned}
$$

where in the second inequality we used Young's inequality. Combining all our estimates together, we see that if we choose h, c_2 and $\varepsilon_1 > 0$ such that

$$
-hK + \frac{1}{4}dC_{a^{(2)}}\varepsilon_1 h + \frac{3}{2}h^2\bar{L}^2 \le -c_2 h, \tag{7.18}
$$

then we obtain $\mathbb{E}\left|X_{k+1}^f - \bar{X}_{k+1}^c\right|^2 \le (1 - c_2 h)\mathbb{E}\left|X_k^f - \bar{X}_k^c\right|^2 + C_2 h^3$, where

$$
\begin{aligned}
C_2 &:= \frac{3}{4}\sigma^{(4)}(1 + C_{IEul}^{(4)}) \\
&\quad + \frac{C_1}{c_1}\left(\frac{1}{4}dC_{a^{(2)}}\frac{1}{\varepsilon_1} + \frac{3}{4} + \frac{3}{8}h_0 C_{b^{(2)}}^2\right). \tag{7.19}
\end{aligned}
$$

We can now finish the proof exactly as we did in Lemma 5.2. $\qquad\square$

Appendix: Proofs for MASGA

Proof of Lemma 5.4 The argument is very similar to the proof of Lemma 5.2 and in fact even simpler as here we have only

one inaccurate drift. For completeness, we give here an outline of the proof anyway. We have

$$
\mathbb{E}\left|X_{k+1}^{f,f} - X_{k+1}\right|^4
$$

$$
= \mathbb{E}\left|X_k^{f,f} - X_k + hb(X_k^{f,f}, U_k^f) - ha(X_k)\right|^4
$$

$$
\leq \mathbb{E}\left|X_k^{f,f} - X_k\right|^4 + 4\mathbb{E}\left|X_k^{f,f} - X_k\right|^2
$$

$$
\left\langle X_k^{f,f} - X_k, hb(X_k^{f,f}, U_k^f) - ha(X_k)\right\rangle
$$

$$
+ 6\mathbb{E}\left|X_k^{f,f} - X_k\right|^2 \left|hb(X_k^{f,f}, U_k^f) - ha(X_k)\right|^2
$$

$$
+ 4\mathbb{E}\left\langle X_k^{f,f} - X_k, hb(X_k^{f,f}, U_k^f) - ha(X_k)\right\rangle
$$

$$
\left|hb(X_k^{f,f}, U_k^f) - ha(X_k)\right|^2
$$

$$
+ \mathbb{E}\left|hb(X_k^{f,f}, U_k^f) - ha(X_k)\right|^4
$$

$$
=: B_1 + B_2 + B_3 + B_4 + B_5.
$$

By conditioning and Assumption 4.3, we have $B_2 \leq -4hK\mathbb{E}\left|X_k^{f,f} - X_k\right|^4$. Furthermore,

$$
B_5 \leq h^4 \mathbb{E}\left|b(X_k^{f,f}, U_k^f) - a(X_k^{f,f}) + a(X_k^{f,f}) - a(X_k)\right|^4
$$

$$
\leq 8h^4 \mathbb{E}\left|b(X_k^{f,f}, U_k^f) - a(X_k^{f,f})\right|^4
$$

$$
+ 8h^4 \mathbb{E}\left|a(X_k^{f,f}) - a(X_k^{f,f})\right|^4
$$

$$
\leq 8h^4 \sigma^{(4)}(1 + C_{IEul}^{(4)}) + 8h^4 L^4 \mathbb{E}\left|X_k^{f,f} - X_k\right|^4,
$$

where we used Assumptions 4.6, 4.3 and Lemma 7.1. It is now clear that the terms B_3 and B_4 can be dealt with exactly as the corresponding terms in the proof of Lemma 5.2 and we obtain essentially the same estimates with sligthly different constants, which are, however, of the same order in s and h. □

Proof of Lemma 5.5 We denote

$$
\Xi_k^A := h\left(b(X_k^{f,f}, U_k^f) - \frac{1}{2}\left(b(X_k^{c-,f}, U_k^{f,1}) + b(X_k^{c+,f}, U_k^{f,2})\right)\right)
$$

$$
\Xi_k^B := h\left(b(X_k^{f,c-}, U_k^f) - \frac{1}{2}\left(b(X_k^{c-,c-}, U_k^{f,1}) + b(X_k^{c+,c-}, U_k^{f,2})\right)\right)
$$

$$
\Xi_k^C := h\left(b(X_k^{f,c+}, U_{k+1}^f) - \frac{1}{2}\left(b(X_k^{c-,c+}, U_{k+1}^{f,1}) + b(X_k^{c+,c+}, U_{k+1}^{f,2})\right)\right)
$$

and we have $\Xi_k = \Xi_k^A - \Xi_k^B - \Xi_k^C$. Then, using $b(x, U) = \frac{1}{2}b(x, U^1) + \frac{1}{2}b(x, U^2)$, we have

$$
\Xi_k^A = \frac{1}{2}hb(X_k^{f,f}, U_k^{f,1}) - \frac{1}{2}hb(\bar{X}_k^{c,f}, U_k^{f,1})
$$

$$
+ \frac{1}{2}hb(\bar{X}_k^{c,f}, U_k^{f,1})
$$

$$
- \frac{1}{2}hb(X_k^{c-,f}, U_k^{f,1})
$$

$$
+ \frac{1}{2}hb(X_k^{f,f}, U_k^{f,2}) - \frac{1}{2}hb(\bar{X}_k^{c,f}, U_k^{f,2})
$$

$$
+ \frac{1}{2}hb(\bar{X}_k^{c,f}, U_k^{f,2})
$$

$$
- \frac{1}{2}hb(X_k^{c+,f}, U_k^{f,2})
$$

and hence

$$
\mathbb{E}|\Xi_k^A|^2 \leq \frac{3}{4}h\bar{L}\mathbb{E}\left|X_k^{f,f} - \bar{X}_k^{c,f}\right|^2 + \frac{3}{4}h\bar{L}\mathbb{E}\left|X_k^{f,f} - \bar{X}_k^{c,f}\right|^2
$$

$$
+ \frac{3}{4}h\mathbb{E}\left|b(\bar{X}_k^{c,f}, U_k^{f,1}) - b(X_k^{c-,f}, U_k^{f,1})\right|
$$

$$
+ b(\bar{X}_k^{c,f}, U_k^{f,2}) - b(X_k^{c+,f}, U_k^{f,2})\right|^2.
$$

Note that the first two terms on the right hand side above are identical and have the correct order in s and h due to Lemma 5.3. Furthermore, the last term can be dealt with by applying Taylor's formula twice in $\bar{X}_k^{c,f}$ and using the argument from the proof of Lemma 5.3 for the term J_{33} therein. Bounds for $\mathbb{E}|\Xi_k^B|^2$ and $\mathbb{E}|\Xi_k^C|^2$ can be obtained in exactly the same way. □

Proof of Lemma 5.6 We need to introduce an auxiliary chain

$$
X_{k+2}^c = X_k^c + 2ha(X_k^c) + \beta\sqrt{2h}\hat{Z}_{k+2}.
$$

Let us begin with bounding $\mathbb{E}\langle\Psi_k, \Xi_k^1\rangle$. Recall that $\bar{X}_k^{f,c} = \frac{1}{2}\left(X_k^{f,c-} + X_k^{f,c+}\right)$. We denote

$$
\Xi_k^{1,1} := hb(X_k^{f,c-}, U_k^f) - hb(\bar{X}_k^{f,c}, U_k^f) + hb(X_k^{f,c+}, U_k^f)
$$

$$
- hb(\bar{X}_k^{f,c}, U_k^f)
$$

$$
\Xi_k^{1,2} := hb(X_k^{f,f}, U_k^f) - hb(X_k, U_k^f),
$$

$$
\Xi_k^{1,3} := hb(\bar{X}_k^{f,c}, U_k^f) - hb(X_k^c, U_k^f)
$$

$$
\Xi_k^{1,4} := hb(\bar{X}_k^{f,c}, U_{k+1}^f) - hb(X_k^c, U_{k+1}^f)
$$

and we see that $\Xi_k^1 = -\Xi_k^{1,1} + \Xi_k^{1,2} - \Xi_k^{1,3} - \Xi_k^{1,4} + hb(X_k, U_k^f) - hb(X_k^c, U_k^f) - hb(X_k^c, U_{k+1}^f)$. By analogy, we define

$$
\Xi_k^{2,1} := hb(X_k^{c-,c-}, U_k^{f,1}) - hb(\bar{X}_k^{c-,c}, U_k^{f,1})
$$

$$
+ hb(X_k^{c-,c+}, U_k^{f,1}) - hb(\bar{X}_k^{c-,c}, U_k^{f,1})
$$

$$
\Xi_k^{2,2} := hb(X_k^{c-,f}, U_k^{f,1}) - hb(X_k, U_k^{f,1}),
$$

$$
\Xi_k^{2,3} := hb(\bar{X}_k^{c-,c}, U_k^{f,1}) - hb(X_k^c, U_k^{f,1})
$$

$$
\Xi_k^{2,4} := hb(\bar{X}_k^{c-,c}, U_{k+1}^{f,1}) - hb(X_k^c, U_{k+1}^{f,1})
$$

$$
\Xi_k^{3,1} := hb(X_k^{c+,c-}, U_k^{f,2})
$$

$$
- hb(\bar{X}_k^{c+,c}, U_k^{f,2}) + hb(X_k^{c+,c+}, U_k^{f,2}) - hb(\bar{X}_k^{c+,c}, U_k^{f,2})
$$

$$
\Xi_k^{3,2} := hb(X_k^{c+,f}, U_k^{f,2}) - hb(X_k, U_k^{f,2}),
$$

$$\Xi_k^{3,3} := hb(\bar{X}_k^{c+,c}, U_k^{f,2}) - hb(X_k^c, U_k^{f,2})$$
$$\Xi_k^{3,4} := hb(\bar{X}_k^{c+,c}, U_{k+1}^{f,2}) - hb(X_k^c, U_{k+1}^{f,2})$$

and hence, since $\Xi_k = \Xi_k^1 - \frac{1}{2}\left(\Xi_k^2 + \Xi_k^3\right)$, we see that

$$\Xi_k = -\left(\Xi_k^{1,1} - \frac{1}{2}\left(\Xi_k^{2,1} + \Xi_k^{3,1}\right)\right)$$
$$+ \left(\Xi_k^{1,2} - \frac{1}{2}\left(\Xi_k^{2,2} + \Xi_k^{3,2}\right)\right)$$
$$- \left(\Xi_k^{1,3} - \frac{1}{2}\left(\Xi_k^{2,3} + \Xi_k^{3,3}\right)\right)$$
$$- \left(\Xi_k^{1,4} - \frac{1}{2}\left(\Xi_k^{2,4} + \Xi_k^{3,4}\right)\right) + Rh,$$

$$R = b(X_k, U_k^f) - \frac{1}{2}\left(b(X_k, U_k^{f,1}) + b(X_k, U_k^{f,2})\right)$$
$$- b(X_k^c, U_k^f)$$
$$+ \frac{1}{2}\left(b(X_k^c, U_k^{f,1}) + b(X_k^c, U_k^{f,2})\right)$$
$$- b(X_k^c, U_{k+1}^f) + \frac{1}{2}\left(b(X_k^c, U_{k+1}^{f,1}) + b(X_k^c, U_{k+1}^{f,2})\right) = 0,$$

since $b(x, U) = \frac{1}{2}b(x, U^1) + \frac{1}{2}b(x, U^2)$. We now write

$$\Xi_k^{1,1} = h \sum_{|\alpha|=1} D^\alpha b(\bar{X}_k^{f,c}, U_k^f)\left(X_k^{f,c-} - \bar{X}_k^{f,c}\right)^\alpha$$
$$+ h \sum_{|\alpha|=2} \int_0^1 (1-t)D^\alpha b\left(\bar{X}_k^{f,c} + t\left(X_k^{f,c-} - \bar{X}_k^{f,c}\right), U_k^f\right) dt$$
$$\left(X_k^{f,c-} - \bar{X}_k^{f,c}\right)^\alpha$$
$$+ h \sum_{|\alpha|=1} D^\alpha b(\bar{X}_k^{f,c}, U_{k+1}^f)\left(X_k^{f,c+} - \bar{X}_k^{f,c}\right)^\alpha$$
$$+ h \sum_{|\alpha|=2} \int_0^1 (1-t)D^\alpha b\left(\bar{X}_k^{f,c} + t\left(X_k^{f,c+} - \bar{X}_k^{f,c}\right), U_{k+1}^f\right) dt$$
$$\left(X_k^{f,c+} - \bar{X}_k^{f,c}\right)^\alpha.$$

Note that $X_k^{f,c-} - \bar{X}_k^{f,c} = -\left(X_k^{f,c+} - \bar{X}_k^{f,c}\right) = \frac{1}{2}\left(X_k^{f,c+} - X_k^{f,c-}\right)$. Hence

$$\mathbb{E}\langle\Psi_k, \Xi_k^{1,1}\rangle = h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(\bar{X}_k^{f,c})\left(X_k^{f,c-} - \bar{X}_k^{f,c}\right)^\alpha$$
$$+ \sum_{|\alpha|=2} \int_0^1 (1-t)D^\alpha a\left(\bar{X}_k^{f,c} + t\left(X_k^{f,c-} - \bar{X}_k^{f,c}\right)\right)$$
$$dt\left(X_k^{f,c-} - \bar{X}_k^{f,c}\right)^\alpha + \sum_{|\alpha|=1} D^\alpha a(\bar{X}_k^{f,c})\left(X_k^{f,c+} - \bar{X}_k^{f,c}\right)^\alpha$$
$$+ \sum_{|\alpha|=2} \int_0^1 (1-t)D^\alpha a\left(\bar{X}_k^{f,c} + t\left(X_k^{f,c+} - \bar{X}_k^{f,c}\right)\right)$$
$$dt\left(X_k^{f,c+} - \bar{X}_k^{f,c}\right)^\alpha\rangle = h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=2} \int_0^1 (1-t)D^\alpha a$$

$$\left(\bar{X}_k^{f,c} + t\left(X_k^{f,c-} - \bar{X}_k^{f,c}\right)\right)$$
$$dt\left(X_k^{f,c-} - \bar{X}_k^{f,c}\right)^\alpha + \sum_{|\alpha|=2} \int_0^1 (1-t)D^\alpha a$$
$$\left(\bar{X}_k^{f,c} + t\left(X_k^{f,c+} - \bar{X}_k^{f,c}\right)\right)$$
$$dt\left(X_k^{f,c+} - \bar{X}_k^{f,c}\right)^\alpha\rangle \le 2hC_{a^{(2)}}\varepsilon_5\mathbb{E}|\Psi_k|^2$$
$$+ \frac{1}{8}hC_{a^{(2)}}\frac{1}{\varepsilon_5}\mathbb{E}\left|X_k^{f,c+} - X_k^{f,c-}\right|^4,$$

where we used Young's inequality with some $\varepsilon_5 > 0$ to be specified later. From Lemma 5.1 we know that $\mathbb{E}\left|X_k^{f,c+} - X_k^{f,c-}\right|^4$ has the correct order in s and h. Similarly, we can show

$$\mathbb{E}\langle\Psi_k, \Xi_k^{2,1}\rangle \le 2hC_{a^{(2)}}\varepsilon_6\mathbb{E}|\Psi_k|^2$$
$$+ \frac{1}{8}hC_{a^{(2)}}\frac{1}{\varepsilon_6}\mathbb{E}\left|X_k^{c-,c+} - X_k^{c-,c-}\right|^4$$
$$\mathbb{E}\langle\Psi_k, \Xi_k^{3,1}\rangle \le 2hC_{a^{(2)}}\varepsilon_7\mathbb{E}|\Psi_k|^2$$
$$+ \frac{1}{8}hC_{a^{(2)}}\frac{1}{\varepsilon_7}\mathbb{E}\left|X_k^{c+,c+} - X_k^{c+,c-}\right|^4$$

for some $\varepsilon_6, \varepsilon_7 > 0$ and we also conclude that the second terms on the right hand side above have the correct order in s and h. Note that in order to deal with the terms $\Xi_k^{1,1}$, $\Xi_k^{2,1}$ and $\Xi_k^{3,1}$ we did not use the structure of our estimator and we just dealt with each of them separately. This will be different in the case of the expression $\left(\Xi_k^{1,2} - \frac{1}{2}\left(\Xi_k^{2,2} + \Xi_k^{3,2}\right)\right)$, where we will use its structure in order to produce an additional antithetic term $X_k^{f,f} - \frac{1}{2}\left(X_k^{c-,f} + X_k^{c+,f}\right)$ on the right hand side of $\mathbb{E}\langle\Psi_k, \Xi_k^{1,2} - \frac{1}{2}\left(\Xi_k^{2,2} + \Xi_k^{3,2}\right)\rangle$ below. Indeed, we first write

$$\Xi_k^{1,2} = h \sum_{|\alpha|=1} D^\alpha b(X_k, U_k^f)\left(X_k^{f,f} - X_k\right)^\alpha$$
$$+ h \sum_{|\alpha|=2} \int_0^1 (1-t)D^\alpha b\left(X_k + t\left(X_k^{f,f} - X_k\right), U_k^f\right) dt\left(X_k^{f,f} - X_k\right)^\alpha.$$

Then, expanding $\Xi_k^{2,2}$ and $\Xi_k^{3,2}$ in an analogous way, we see that

$$\mathbb{E}\langle\Psi_k, \Xi_k^{1,2} - \frac{1}{2}\left(\Xi_k^{2,2} + \Xi_k^{3,2}\right)\rangle$$
$$= h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(X_k^{f,f} - \frac{1}{2}\left(X_k^{c-,f} + X_k^{c+,f}\right)\right)^\alpha$$
$$+ \sum_{|\alpha|=2} \int_0^1 (1-t)D^\alpha a\left(X_k + t\left(X_k^{f,f} - X_k\right)\right) dt\left(X_k^{f,f} - X_k\right)^\alpha$$
$$- \frac{1}{2}\sum_{|\alpha|=2} \int_0^1 (1-t)D^\alpha a\left(X_k + t\left(X_k^{c-,f} - X_k\right)\right) dt\left(X_k^{c-,f} - X_k\right)^\alpha$$
$$- \frac{1}{2}\sum_{|\alpha|=2} \int_0^1 (1-t)D^\alpha a\left(X_k + t\left(X_k^{c+,f} - X_k\right)\right) dt\left(X_k^{c+,f} - X_k\right)^\alpha\rangle$$

$$=: h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(X_k^{f,f} - \frac{1}{2}\left(X_k^{c-,f} + X_k^{c+,f}\right)\right)^\alpha\rangle + h\mathbb{E}\langle\Psi_k, \hat{\Xi}_k^2\rangle.$$

Using Young's inequality with some $\varepsilon_8 > 0$, we can now bound

$$\mathbb{E}\langle\Psi_k, \hat{\Xi}_k^2\rangle \le 3\varepsilon_8 C_{a^{(2)}}\mathbb{E}|\Psi_k|^2 + \frac{1}{\varepsilon_8}C_{a^{(2)}}\left[\mathbb{E}\left|X_k^{f,f} - X_k\right|^4\right.$$
$$\left. + \mathbb{E}\left|X_k^{c-,f} - X_k\right|^4 + \mathbb{E}\left|X_k^{c+,f} - X_k\right|^4\right],$$

whereas the remaining antithetic term will be used later. Note that all the fourth moments above have the correct order in s and h due to Lemma 5.4.

Now we turn our attention to $\left(\Xi_k^{1,3} - \frac{1}{2}\left(\Xi_k^{2,3} + \Xi_k^{3,3}\right)\right)$. We start with $\Xi_k^{1,3}$ by writing

$$\Xi_k^{1,3} = h\sum_{|\alpha|=1} D^\alpha b(X_k^c, U_k^f)\left(\bar{X}_k^{f,c} - X_k^c\right)^\alpha$$
$$+ h\sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha b\left(X_k^c + t\left(\bar{X}_k^{f,c} - X_k^c\right), U_k^f\right)dt$$
$$\left(\bar{X}_k^{f,c} - X_k^c\right)^\alpha =: \Xi_k^{1,3,1} + \Xi_k^{1,3,2}.$$

Note that $\mathbb{E}\langle\Psi_k, \Xi_k^{1,3,2}\rangle \le h\varepsilon_9 C_{a^{(2)}}\mathbb{E}|\Psi_k|^2 + h\varepsilon_9^{-1}C_{a^{(2)}}\mathbb{E}\left|\bar{X}_k^{f,c} - X_k^c\right|^4$ for some $\varepsilon_9 > 0$. We have

Lemma 7.2 *Under the assumptions of Lemma 5.4, there is a constant $C > 0$ such that for all $k \ge 1$,*

$$\mathbb{E}\left|\bar{X}_k^{f,c} - X_k^c\right|^4 \le C\frac{1}{s^2}h^2.$$

Proof We notice that

$$\mathbb{E}\left|\bar{X}_k^{f,c} - X_k^c\right|^4 = \mathbb{E}\left|\frac{1}{2}\left(X_k^{f,c-} - X_k^c + X_k^{f,c+} - X_k^c\right)\right|^4$$
$$\le \frac{1}{2}\mathbb{E}\left|X_k^{f,c-} - X_k^c\right|^4 + \frac{1}{2}\mathbb{E}\left|X_k^{f,c+} - X_k^c\right|^4$$

and then use an analogue of Lemma 5.4 for the coarse chain. □

On the other hand,

$$\mathbb{E}\langle\Psi_k, \Xi_k^{1,3,1}\rangle = h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k^c)\left(\bar{X}_k^{f,c} - X_k^c\right)^\alpha\rangle$$
$$= h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(\bar{X}_k^{f,c} - X_k\right)^\alpha\rangle$$
$$+ h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1}\left(D^\alpha a(X_k^c)\left(\bar{X}_k^{f,c} - X_k^c\right)^\alpha\right.$$
$$\left. - D^\alpha a(X_k)\left(\bar{X}_k^{f,c} - X_k\right)^\alpha\right)\rangle$$
$$=: \mathbb{E}\langle\Psi_k, \Xi_k^{1,3,1,1}\rangle + \mathbb{E}\langle\Psi_k, \Xi_k^{1,3,1,2}\rangle.$$

Now observe that

$$\mathbb{E}\langle\Psi_k, \Xi_k^{1,3,1,2}\rangle = h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1}\left(D^\alpha a(X_k^c)\left(\bar{X}_k^{f,c} - X_k^c\right)^\alpha\right.$$
$$\left. - D^\alpha a(X_k)\left(\bar{X}_k^{f,c} - X_k\right)^\alpha\right)\rangle$$
$$+ h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1}\left(D^\alpha a(X_k)\left(\bar{X}_k^{f,c} - X_k^c\right)^\alpha\right.$$
$$\left. - D^\alpha a(X_k)\left(\bar{X}_k^{f,c} - X_k\right)^\alpha\right)\rangle$$
$$= h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1}\left(D^\alpha a(X_k^c)\right.$$
$$\left. - D^\alpha a(X_k)\right)\left(\bar{X}_k^{f,c} - X_k^c\right)^\alpha\rangle$$
$$+ h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(X_k - X_k^c\right)^\alpha\rangle.$$

$$(7.20)$$

Recall that we are dealing now with the group $-\left(\Xi_k^{1,3} - \frac{1}{2}\left(\Xi_k^{2,3} + \Xi_k^{3,3}\right)\right)$. Similarly as above, for $\Xi_k^{2,3}$ and $\Xi_k^{3,3}$, we have the terms

$$\Xi_k^{2,3,2} = h\sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha$$
$$b\left(X_k^c + t\left(\bar{X}_k^{c-,c} - X_k^c\right), U_k^{f,1}\right)dt\left(\bar{X}_k^{c-,c} - X_k^c\right)^\alpha$$

$$\Xi_k^{3,3,2} = h\sum_{|\alpha|=2}\int_0^1 (1-t)D^\alpha$$
$$b\left(X_k^c + t\left(\bar{X}_k^{c+,c} - X_k^c\right), U_k^{f,2}\right)dt\left(\bar{X}_k^{c+,c} - X_k^c\right)^\alpha$$

for which

$$\mathbb{E}\langle\Psi_k, \Xi_k^{2,3,2}\rangle \le h\varepsilon_{10}C_{a^{(2)}}\mathbb{E}|\Psi_k|^2 + h\frac{1}{\varepsilon_{10}}C_{a^{(2)}}\mathbb{E}\left|\bar{X}_k^{c-,c} - X_k^c\right|^4$$

$$\mathbb{E}\langle\Psi_k, \Xi_k^{3,3,2}\rangle \le h\varepsilon_{11}C_{a^{(2)}}\mathbb{E}|\Psi_k|^2 + h\frac{1}{\varepsilon_{11}}C_{a^{(2)}}\mathbb{E}\left|\bar{X}_k^{c+,c} - X_k^c\right|^4$$

for some $\varepsilon_{10}, \varepsilon_{11} > 0$, and we can again apply Lemma 7.2 to conclude that the fourth moments above have the correct order in s and h. On the other hand, repeating the analysis for $\Xi_k^{1,3,1}$ above, we see that

$$\mathbb{E}\langle\Psi_k, \Xi_k^{2,3,1,1}\rangle = h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(\bar{X}_k^{c-,c} - X_k\right)^\alpha\rangle$$

$$\mathbb{E}\langle\Psi_k, \Xi_k^{3,3,1,1}\rangle = h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(\bar{X}_k^{c+,c} - X_k\right)^\alpha\rangle,$$

whereas

$$\mathbb{E}\langle\Psi_k, \Xi_k^{2,3,1,2}\rangle = h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1}\left(D^\alpha a(X_k^c) - D^\alpha a(X_k)\right)$$
$$\left(\bar{X}_k^{c-,c} - X_k^c\right)^\alpha\rangle$$
$$+ h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(X_k - X_k^c\right)^\alpha\rangle$$

$$\mathbb{E}\langle\Psi_k, \Xi_k^{3,3,1,2}\rangle = h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1}\left(D^\alpha a(X_k^c) - D^\alpha a(X_k)\right)$$
$$\left(\bar{X}_k^{c+,c} - X_k^c\right)^\alpha\rangle$$
$$+ h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(X_k - X_k^c\right)^\alpha\rangle.$$

$$(7.21)$$

Recall that

$$\mathbb{E}\langle\Psi_k, -\left(\Xi_k^{1,3} - \frac{1}{2}\left(\Xi_k^{2,3} + \Xi_k^{3,3}\right)\right)\rangle$$

$$= \mathbb{E}\langle\Psi_k, -\left(\Xi_k^{1,3,1} - \frac{1}{2}\left(\Xi_k^{2,3,1} + \Xi_k^{3,3,1}\right)\right)\rangle$$

$$+ \mathbb{E}\langle\Psi_k, -\left(\Xi_k^{1,3,2} - \frac{1}{2}\left(\Xi_k^{2,3,2} + \Xi_k^{3,3,2}\right)\right)\rangle$$

and, due to our discussion above, $\mathbb{E}\langle\Psi_k, -\left(\Xi_k^{1,3,2} - \frac{1}{2}\left(\Xi_k^{2,3,2} + \Xi_k^{3,3,2}\right)\right)\rangle$ is of the correct order in s and h. Hence we focus on

$$\mathbb{E}\langle\Psi_k, -\left(\Xi_k^{1,3,1} - \frac{1}{2}\left(\Xi_k^{2,3,1} + \Xi_k^{3,3,1}\right)\right)\rangle$$

$$= -h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(\bar{X}_k^{f,c} - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right)^\alpha\rangle$$

$$- h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1}\left(D^\alpha a(X_k^c) - D^\alpha a(X_k)\right)\left(\bar{X}_k^{f,c}\right.$$

$$\left. - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right)^\alpha\rangle,$$

where the first term on the right hand side above comes from the expression $\mathbb{E}\langle\Psi_k, \Xi_k^{1,3,1,1}\rangle - \frac{1}{2}\left(\mathbb{E}\langle\Psi_k, \Xi_k^{2,3,1,1}\rangle\right.$

$\left. + \mathbb{E}\langle\Psi_k, \Xi_k^{3,3,1,1}\rangle\right)$ and the second term comes from

$$\mathbb{E}\langle\Psi_k, \Xi_k^{1,3,1,2}\rangle - \frac{1}{2}\left(\mathbb{E}\langle\Psi_k, \Xi_k^{2,3,1,2}\rangle + \mathbb{E}\langle\Psi_k, \Xi_k^{3,3,1,2}\rangle\right).$$

$$(7.22)$$

Note that each term in (7.22) was a sum of two terms, however, all the second terms in (7.22) cancelled out, since they were all of the same form, cf. (7.20) and (7.21). Now we will combine the first term on the right hand side of $\mathbb{E}\langle\Psi_k, -\left(\Xi_k^{1,3,1} - \frac{1}{2}\left(\Xi_k^{2,3,1} + \Xi_k^{3,3,1}\right)\right)\rangle$ with a term from a previous group. To this end, recall that

$$\mathbb{E}\langle\Psi_k, \Xi_k^{1,2} - \frac{1}{2}\left(\Xi_k^{2,2} + \Xi_k^{3,2}\right)\rangle$$

$$= h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(X_k^{f,f} - \frac{1}{2}\left(X_k^{c-,f} + X_k^{c+,f}\right)\right)^\alpha\rangle$$

$$+ h\mathbb{E}\langle\Psi_k, \hat{\Xi}_k^2\rangle,$$

where $h\mathbb{E}\langle\Psi_k, \hat{\Xi}_k^2\rangle$ is of the correct order in s and h, and notice that we have

$$h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(X_k^{f,f} - \frac{1}{2}\left(X_k^{c-,f} + X_k^{c+,f}\right)\right)^\alpha\rangle$$

$$- h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(\bar{X}_k^{f,c} - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right)^\alpha\rangle$$

$$= h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(X_k^{f,f} - \frac{1}{2}\left(X_k^{c-,f} + X_k^{c+,f}\right)\right.$$

$$\left. - \left(\bar{X}_k^{f,c} - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right)\right)^\alpha\rangle$$

$$= h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k)\left(\Psi_k\right)^\alpha\rangle \le -Kh\mathbb{E}|\Psi_k|^2.$$

In the inequality above we used the fact that Assumption 4.3 implies that for all $x, y \in \mathbb{R}^d$ we have $\langle y, \sum_{|\alpha|=1} D^\alpha a(x)y^\alpha\rangle \le -K|y|^2$. On the other hand, the first terms in (7.22) give

$$h\mathbb{E}\langle\Psi_k, \sum_{|\alpha|=1}\left(D^\alpha a(X_k^c) - D^\alpha a(X_k)\right)\left(\bar{X}_k^{f,c} - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right)^\alpha\rangle$$

$$\le 2h\varepsilon_{12}C_{a^{(1)}}\mathbb{E}|\Psi_k|^2 + 2h\frac{1}{\varepsilon_{12}}C_{a^{(1)}}\mathbb{E}\left|\bar{X}_k^{f,c} - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right|^2$$

for some $\varepsilon_{12} > 0$ to be chosen later. Now we use

Lemma 7.3 *Under the assumptions of Lemma 5.3, there is a $C > 0$ such that for all $k \ge 1$ we have $\mathbb{E}\left|\bar{X}_k^{f,c}\right.$*
$\left. - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c}\right)\right|^2 \le Ch^2/s^2$.

Proof Notice that

$$
\mathbb{E}\left| \bar{X}_k^{f,c} - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c} \right) \right|^2
$$

$$
\leq \frac{1}{2}\mathbb{E}\left| X_k^{f,c-} - \frac{1}{2}\left(X_k^{c-,c-} + X_k^{c+,c-} \right) \right|^2
$$

$$
+ \frac{1}{2}\mathbb{E}\left| X_k^{f,c+} - \frac{1}{2}\left(X_k^{c-,c+} + X_k^{c+,c+} \right) \right|^2
$$

and that both terms above correspond to antithetic estimators with respect to subsampling for coarse chains, hence Lemma 5.3 applies. □

Hence it only remains to deal with $\mathbb{E}\langle \Psi_k, -\left(\Xi_k^{1,4} - \frac{1}{2}\left(\Xi_k^{2,4} + \Xi_k^{3,4} \right) \right)\rangle$. We have

$$
\Xi_k^{1,4} = h\sum_{|\alpha|=1} D^\alpha b(X_k^c, U_{k+1}^f)\left(\bar{X}_k^{f,c} - X_k^c \right)^\alpha
$$

$$
+ h\sum_{|\alpha|=2} \int_0^1 (1-t) D^\alpha b\left(X_k^c + t\left(\bar{X}_k^{f,c} - X_k^c \right), U_{k+1}^f \right) dt
$$

$$
\left(\bar{X}_k^{f,c} - X_k^c \right)^\alpha
$$

$$
\Xi_k^{2,4} = h\sum_{|\alpha|=1} D^\alpha b(X_k^c, U_{k+1}^{f,1})\left(\bar{X}_k^{c-,c} - X_k^c \right)^\alpha
$$

$$
+ h\sum_{|\alpha|=2} \int_0^1 (1-t) D^\alpha b\left(X_k^c + t\left(\bar{X}_k^{c-,c} - X_k^c \right), U_{k+1}^{f,1} \right) dt
$$

$$
\left(\bar{X}_k^{c-,c} - X_k^c \right)^\alpha
$$

$$
\Xi_k^{3,4} = h\sum_{|\alpha|=1} D^\alpha b(X_k^c, U_{k+1}^{f,2})\left(\bar{X}_k^{c+,c} - X_k^c \right)^\alpha + h\sum_{|\alpha|=2} \int_0^1 (1-t) D^\alpha b
$$

$$
\times \left(X_k^c + t\left(\bar{X}_k^{c+,c} - X_k^c \right), U_{k+1}^{f,2} \right) dt \left(\bar{X}_k^{c+,c} - X_k^c \right)^\alpha
$$

and hence

$$
\mathbb{E}\langle \Psi_k, -\left(\Xi_k^{1,4} - \frac{1}{2}\left(\Xi_k^{2,4} + \Xi_k^{3,4} \right) \right)\rangle
$$

$$
= -h\mathbb{E}\langle \Psi_k, \sum_{|\alpha|=1} D^\alpha a(X_k^c)\left(\bar{X}_k^{f,c} - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c} \right) \right)^\alpha \rangle
$$

$$
- h\mathbb{E}\langle \Psi_k, \sum_{|\alpha|=2} \int_0^1 (1-t) D^\alpha a\left(X_k^c + t\left(\bar{X}_k^{f,c} - X_k^c \right) \right) dt
$$

$$
\left(\bar{X}_k^{f,c} - X_k^c \right)^\alpha \rangle
$$

$$
- h\mathbb{E}\langle \Psi_k, \sum_{|\alpha|=2} \int_0^1 (1-t) D^\alpha a\left(X_k^c + t\left(\bar{X}_k^{c-,c} - X_k^c \right) \right) dt
$$

$$
\left(\bar{X}_k^{c-,c} - X_k^c \right)^\alpha \rangle
$$

$$
- h\mathbb{E}\langle \Psi_k, \sum_{|\alpha|=2} \int_0^1 (1-t) D^\alpha a\left(X_k^c + t\left(\bar{X}_k^{c+,c} - X_k^c \right) \right) dt
$$

$$
\left(\bar{X}_k^{c+,c} - X_k^c \right)^\alpha \rangle
$$

$$
\leq h\varepsilon_{13} C_{a^{(1)}}\mathbb{E}|\Psi_k|^2 + h\frac{1}{\varepsilon_{13}} C_{a^{(1)}}\mathbb{E}\left| \bar{X}_k^{f,c} - \frac{1}{2}\left(\bar{X}_k^{c-,c} + \bar{X}_k^{c+,c} \right) \right|^2
$$

$$
+ h\varepsilon_{14} C_{a^{(2)}}\mathbb{E}|\Psi_k|^2 + h\frac{1}{\varepsilon_{14}} C_{a^{(2)}}\mathbb{E}\left| \bar{X}_k^{f,c} - X_k^c \right|^4
$$

$$
+ h\varepsilon_{15} C_{a^{(2)}}\mathbb{E}|\Psi_k|^2 + h\frac{1}{\varepsilon_{15}} C_{a^{(2)}}\mathbb{E}\left| \bar{X}_k^{c-,c} - X_k^c \right|^4
$$

$$
+ h\varepsilon_{16} C_{a^{(2)}}\mathbb{E}|\Psi_k|^2 + h\frac{1}{\varepsilon_{16}} C_{a^{(2)}}\mathbb{E}\left| \bar{X}_k^{c+,c} - X_k^c \right|^4,
$$

for some $\varepsilon_{13}, \varepsilon_{14}, \varepsilon_{15}, \varepsilon_{16} > 0$. Using Lemmas 7.2 and 7.3, we see that all the fourth moments above have the correct order in s and h. This concludes our estimates for $\mathbb{E}\langle \Psi_k, \Xi_k \rangle$. □

References

Aicher, C., Ma, Y.-A., Foti, N.J., Fox, E.B.: Stochastic gradient MCMC for state space models. SIAM J. Math. Data Sci. **1**(3), 555–587 (2019)

Baker, J., Fearnhead, P., Fox, E.B., Nemeth, C.: Control variates for stochastic gradient MCMC. Stat. Comput. **29**(3), 599–615 (2019)

Barkhagen, M., Chau, N.H., Moulines, É., Rásonyi, M., Sabanis, S., Zhang, Y.: On stochastic gradient Langevin dynamics with dependent data streams in the logconcave case. Bernoulli (2020)

Ben Alaya, M., Kebaier, A.: Central limit theorem for the multilevel Monte Carlo Euler method. Ann. Appl. Probab. **25**(1), 211–234 (2015)

Ben Alaya, M., Kebaier, A., Tram Ngo, T.B.: Central Limit Theorem for the σ-antithetic multilevel Monte Carlo method. arXiv e-prints, page arXiv:2002.08834 (2020)

Blackard, J.A., Dean, D.J.: Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. Comput. Electron. Agric. **24**(3), 131–151 (1999)

Brosse, N., Durmus, A., Moulines, E.: The promises and pitfalls of stochastic gradient langevin dynamics. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31, pp. 8268–8278. Curran Associates Inc (2018)

Chatterji, N., Flammarion, N., Ma, Y., Bartlett, P., Jordan, M.: On the theory of variance reduction for stochastic gradient Monte Carlo. In: Dy, J., Krause, A. (eds.), Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pp. 764–773, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR

Chau, H.N., Rasonyi, M.: Stochastic Gradient Hamiltonian Monte Carlo for Non-Convex Learning in the Big Data Regime. arXiv e-prints. arXiv: 1903.10328 (2019)

Chau, N.H., Moulines, É., Rásonyi, M., Sabanis, S., Zhang, Y.: On stochastic gradient Langevin dynamics with dependent data streams: the fully non-convex case. arXiv e-prints arXiv:1905.13142 (2019)

Cheng, X., Chatterji, N.S., Abbasi-Yadkori, Y., Bartlett, P.L., Jordan, M.I.: Sharp convergence rates for Langevin dynamics in the non-convex setting. arXiv e-prints, page arXiv:1805.01648 (2018)

Cornish, R., Vanetti, P., Bouchard-Cote, A., Deligiannidis, G., Doucet, A.: Scalable Metropolis-Hastings for exact Bayesian inference with large datasets. In K. Chaudhuri and R. Salakhutdinov, editors, Proceedings of the 36th International Conference on Machine Learning, volume 97 of Proceedings of Machine Learning Research, pp. 1351–1360, Long Beach, California, USA, 09–15 Jun 2019. PMLR

Crisan, D., Del Moral, P., Houssineau, J., Jasra, A.: Unbiased multi-index Monte Carlo. Stoch. Anal. Appl. **36**(2), 257–273 (2018)

Dalalyan, A.S.: Theoretical guarantees for approximate sampling from smooth and log-concave densities. J. R. Stat. Soc. Ser. B. Stat. Methodol. **79**(3), 651–676 (2017)

Dalalyan, A.S., Karagulyan, A.: User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. Stochastic Process. Appl. **129**(12), 5278–5311 (2019)

Deniz Akyildiz, Ö., Sabanis, S.: Nonasymptotic analysis of Stochastic Gradient Hamiltonian Monte Carlo under local conditions for nonconvex optimization. arXiv e-prints arXiv:2002.05465, (Feb. 2020)

Dereich, S.: General multilevel adaptations for stochastic approximation algorithms II: CLTs. Stochastic Process. Appl. **132**, 226–260 (2021)

Dereich, S., Müller-Gronbach, T.: General multilevel adaptations for stochastic approximation algorithms of Robbins–Monro and Polyak-Ruppert type. Numer. Math. **142**(2), 279–328 (2019)

Dubey, K.A., Reddi, S.J., Williamson, S.A., Poczos, B., Smola, A.J., Xing, E.P.: Variance reduction in Stochastic Gradient Langevin Dynamics. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29, pp. 1154–1162. Curran Associates Inc (2016)

Durmus, A., Eberle, A.: Asymptotic bias of inexact Markov Chain Monte Carlo methods in high dimension. arXiv e-prints, page arXiv:2108.00682 (2021)

Durmus, A., Moulines, E.: Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. Ann. Appl. Probab. **27**(3), 1551–1587 (2017)

Durmus, A., Roberts, G.O., Vilmart, G., Zygalakis, K.C.: Fast Langevin based algorithm for MCMC in high dimensions. Ann. Appl. Probab. **27**(4), 2195–2237 (2017)

Eberle, A.: Reflection couplings and contraction rates for diffusions. Probab. Theory Rel. Fields **166**(3–4), 851–886 (2016)

Eberle, A., Guillin, A., Zimmer, R.: Quantitative Harris-type theorems for diffusions and McKean–Vlasov processes. Trans. Am. Math. Soc. **371**(10), 7135–7173 (2019)

Eberle, A., Majka, M.B.: Quantitative contraction rates for Markov chains on general state spaces. Electron. J. Probab. **24**, 36 (2019)

Frikha, N.: Multi-level stochastic approximation algorithms. Ann. Appl. Probab. **26**(2), 933–985 (2016)

Gao, X., Gürbüzbalaban, M., Zhu, L.: Global Convergence of Stochastic Gradient Hamiltonian Monte Carlo for Non-Convex Stochastic Optimization: Non-Asymptotic Performance Bounds and Momentum-Based Acceleration. arXiv e-prints arXiv:1809.04618, (Sep 2018)

Giles, M.B.: Multilevel Monte Carlo path simulation. Oper. Res. **56**(3), 607–617 (2008)

Giles, M.B.: Multilevel Monte Carlo methods. Acta Numer. **24**, 259–328 (2015)

Giles, M.B., Haji-Ali, A.-L.: Multilevel nested simulation for efficient risk estimation. SIAM/ASA J. Uncertain. Quantif. **7**(2), 497–525 (2019)

Giles, M.B., Majka, M.B., Szpruch, L., Vollmer, S.J., Zygalakis, K.C.: Multi-level Monte Carlo methods for the approximation of invariant measures of stochastic differential equations. Stat. Comput. **30**(3), 507–524 (2020)

Giles, M.B., Szpruch, L.: Antithetic multilevel Monte Carlo estimation for multidimensional SDEs. In: Monte Carlo and quasi-Monte Carlo methods 2012, volume 65 of Springer Proc. Math. Stat., pp. 367–384. Springer, Heidelberg (2013)

Giles, M.B., Szpruch, L.: Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without Lévy area simulation. Ann. Appl. Probab. **24**(4), 1585–1620 (2014)

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates Inc (2014)

Haji-Ali, A.-L., Nobile, F., Tempone, R.: Multi-index Monte Carlo: when sparsity meets sampling. Numer. Math. **132**(4), 767–806 (2016)

Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference. J. Mach. Learn. Res. (2013)

Hu, K., Ren, Z., Siska, D., Szpruch, L.: Mean-Field Langevin Dynamics and Energy Landscape of Neural Networks. arXiv e-prints arXiv:1905.07769 (2019)

Hwang, C.-R.: Laplace's method revisited: weak convergence of probability measures. Ann. Probab. **8**(6), 1177–1182 (1980)

Jabir, J.-F., Šiška, D., Szpruch, Ł.: Mean-Field Neural ODEs via Relaxed Optimal Control. arXiv e-prints arXiv:1912.05475 (2019)

Jacob, P.E., O'Leary, J., Atchadé, Y.F.: Unbiased Markov chain Monte Carlo methods with couplings. J. R. Stat. Soc. Ser. B. Stat. Methodol. **82**(3), 543–600 (2020)

Jourdain, B., Kebaier, A.: Non-asymptotic error bounds for the multilevel Monte Carlo Euler method applied to SDEs with constant diffusion coefficient. Electron. J. Probab. **24**, 12–34 (2019)

Kebaier, A.: Statistical Romberg extrapolation: a new variance reduction method and applications to option pricing. Ann. Appl. Probab. **15**(4), 2681–2705 (2005)

Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

Ma, Y.-A., Chen, T., Fox, E.: A complete recipe for stochastic gradient mcmc. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28, pp. 2917–2925. Curran Associates Inc, UK (2015)

Ma, Y.-A., Chen, Y., Jin, C., Flammarion, N., Jordan, M.I.: Sampling can be faster than optimization. Proc. Natl. Acad. Sci. USA **116**(42), 20881–20885 (2019)

Majka, M.B., Mijatović, A., Szpruch, L.: Non-asymptotic bounds for sampling algorithms without log-concavity. Ann. Appl. Probab. **30**(4), 1534–1581 (2020)

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)

Moulines, E., Bach, F.R.: Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 24, pp. 451–459. Curran Associates Inc (2011)

Nagapetyan, T., Duncan, A.B., Hasenclever, L., Vollmer, S.J., Szpruch, L., Zygalakis, K.: The True Cost of Stochastic Gradient Langevin Dynamics. arXiv e-prints arXiv:1706.02692 (2017)

Nemeth, C., Fearnhead, P.: Stochastic gradient Markov chain Monte Carlo. arXiv e-prints, page arXiv:1907.06986 (2019)

Raginsky, M., Rakhlin, A., Telgarsky, M.: Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In: Kale, S., Shamir, O. (eds.), Proceedings of the 2017 Conference on Learning Theory, volume 65 of Proceedings of Machine Learning Research, pp. 1674–1703. Amsterdam, Netherlands, 07–10 Jul 2017. PMLR

Rhee, C.-H., Glynn, P.W.: Unbiased estimation with square root convergence for SDE models. Oper. Res. **63**(5), 1026–1043 (2015)

Shamir, O.: Without-replacement sampling for stochastic gradient methods. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems 29, pp. 46–54. Curran Associates Inc (2016)

Szpruch, Ł., Tse, A.: Antithetic multilevel particle system sampling method for McKean-Vlasov SDEs. arXiv e-prints arXiv:1903.07063 (2019)

Teh, Y.W., Thiery, A.H., Vollmer, S.J.: Consistency and fluctuations for stochastic gradient Langevin dynamics. J. Mach. Learn. Res. **17**, 7–33 (2016)

Villani, C.: Optimal Transport: Old and New, vol. 338. Springer (2009)

S. J. Vollmer, K. C. Zygalakis, and Y. W. Teh. Exploration of the (non-)asymptotic bias and variance of stochastic gradient Langevin dynamics. J. Mach. Learn. Res., 17:159 2016

Welling, M., Teh, Y.W.: Bayesian learning via Stochastic Gradient Langevin Dynamics. In L. Getoor and T. Scheffer, editors, Proceedings of the 28th International Conference on Machine Learning (ICML-11), ICML '11, pp. 681–688, New York, NY, USA, (2011). ACM

Xifara, T., Sherlock, C., Livingstone, S., Byrne, S., Girolami, M.: Langevin diffusions and the Metropolis-adjusted Langevin algorithm. Statist. Probab. Lett. **91**, 14–19 (2014)

Zhang, Y., Deniz Akyildiz, Ö., Damoulas, T., Sabanis, S.: Nonasymptotic estimates for Stochastic Gradient Langevin Dynamics under local conditions in nonconvex optimization. arXiv e-prints arXiv:1910.02008 (2019)

Zou, D., Xu, P., Gu, Q.: Stochastic gradient hamiltonian Monte Carlo methods with recursive variance reduction. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 3835–3846. Curran Associates Inc (2019)