Submitted to Bernoulli

Variance reduction for MCMC methods via martingale representations

D. BELOMESTNY[1,3] and E. MOULINES[2,3] and S. SAMSONOV[3] and N. SHAGADATOV[3]

[1] *Duisburg-Essen University, Faculty of Mathematics, D-45127 Essen Germany*

[2] *Centre de Mathématiques Appliquées, UMR 7641, Ecole Polytechnique, France*

[3] *National University Higher School of Economics, Moscow, Russia*

In this paper we propose an efficient variance reduction approach for MCMC algorithms relying on a novel discrete time martingale representation for Markov chains. Our approach is fully non-asymptotic and does not require any type of ergodicity or special product structure of the underlying density. By rigorously analyzing the convergence properties of the proposed algorithm, we show that it's complexity is indeed asymptotically smaller than one of the original MCMC algorithm. The numerical performance of the new method is illustrated in the case of Gaussian mixtures and Bayesian regression models.

MSC 2010 subject classifications: Primary 60G40, 60G40; secondary 91G80.
Keywords: MCMC, variance reduction, martingale representation.

1. Introduction

Monte Carlo integration typically has an error variance of the form σ^2/n, where n is a sample size and σ^2 is the variance of integrand. We can make the variance smaller by using a larger value of n. Alternatively, we can reduce σ^2 instead of increasing the sample size n. To this end, one can try to construct a new Monte Carlo experiment with the same expectation as the original one but with a lower variance σ^2. Methods to achieve this are known as variance reduction techniques. Variance reduction plays an important role in Monte Carlo and Markov Chain Monte Carlo methods. Introduction to many of the variance reduction techniques can be found in [6], [22], [14] and [13]. Recently one witnessed a revival of interest in efficient variance reduction methods for MCMC algorithms, see for example [8], [19], [5] and references therein.

Suppose that we wish to compute $\pi(f) := \mathsf{E}[f(X)]$, where X is a random vector-valued in $\mathcal{X} \subseteq \mathbb{R}^d$ with a density π and $f : \mathcal{X} \to \mathbb{R}$ with $f \in L^2(\pi)$. The idea of the so-called control variates variance reduction method is to find a cheaply computable random variable ζ with $\mathsf{E}[\zeta] = 0$ and $\mathsf{E}[\zeta^2] < \infty$, such that the variance of the r.v. $f(X) - \zeta$ is small. The complexity of the problem of constructing classes Z of control variates ζ satisfying $\mathsf{E}[\zeta] = 0$ essentially depends on the degree of our knowledge on π. For example, if π is analytically known and satisfies some regularity conditions, one can

apply the well-known technique of polynomial interpolation to construct control variates enjoying some optimality properties, see, for example, Section 3.2 in [9]. Alternatively, if an orthonormal system in $L^2(\pi)$ is analytically available, one can build control variates ζ as a linear combination of the corresponding basis functions, see [4]. Furthermore, if π is known only up to a normalizing constant (which is often the case in Bayesian statistics), one can apply the recent approach of control variates depending only on the gradient $\nabla \log \pi$ using Schrödinger-type Hamiltonian operator in [19] and so-called Stein operator in [5]. In some situations π is not known analytically, but X can be represented as a function of simple random variables with known distribution. Such situation arises, for example, in the case of functionals of discretized diffusion processes. In this case a Wiener chaos-type decomposition can be used to construct control variates with nice theoretical properties, see [3]. Note that in order to compare different variance reduction approaches, one has to analyze their complexity, that is, the number of numerical operations required to achieve a prescribed magnitude of the resulting variance.

The situation becomes much more difficult in the case of MCMC algorithms, where one has to work with a Markov chain X_p, $p = 0, 1, 2, \dots$, whose marginal distribution converges to π as time grows. One important class of the variance reduction methods in this case is based on the so-called Poisson equation for the corresponding Markov chain. It was observed in Henderson [15] that if a time-homogeneous Markov chain (X_p) is stationary with stationary distribution π, then for any real-valued function $G \in L^1(\pi)$ defined on the state space of the Markov chain (X_p), the function $U(x) := G(x) - \mathsf{E}[G(X_1)|X_0 = x]$ has zero mean with respect to π. The best choice for the function G corresponds to a solution of the so-called Poisson equation $\mathsf{E}[G(X_1)|X_0 = x] - G(x) = -f(x) + \pi(f)$. Moreover, it is also related to the minimal asymptotic variance in the corresponding central limit theorem, see [11] and [19]. Although the Poisson equation involves the quantity of interest $\pi(f)$ and can not be solved explicitly in most cases, this idea still can be used to construct some approximations for the optimal zero-variance control variates. For example, Henderson [15] proposed to compute approximations for the solution of the Poisson equation for specific Markov chains with particular emphasis on models arising in stochastic network theory. In [8] and [5] series-type control variates are introduced and studied for reversible Markov chains. It is assumed in [8] that the one-step conditional expectations can be computed explicitly for a set of basis functions. The authors in [5] proposed another approach tailored to diffusion setting which doesn't require the computation of integrals of basis functions and only involves applications of the underlying generator.

In this paper we focus on the Langevin type algorithms which got much attention recently, see [7, 12, 16, 21, 20] and references therein. We propose a generic variance reduction method for these and other types algorithms, which is purely non-asymptotic and does not require that the conditional expectations of the corresponding Markov chain can be computed or that the generator is known analytically. Moreover, we do not need to assume stationarity or/and sampling under the invariant distribution π. We rigorously analyse the convergence of the method and study its complexity. It is shown that our variance-reduced Langevin algorithm outperforms the standard Langevin algorithms in terms of complexity.

The paper is organized as follows. In Section 2 we set up the problem and introduce some notations. Section 3 contains a novel martingale representation and shows how this representation can be used for variance reduction. In Section 4 we analyze the performance of the proposed variance reduction algorithm in the case of Unadjusted Langevin Algorithm (ULA). Section 6 studies the complexity of the variance reduced ULA. Finally, numerical examples are presented in Section 7.

2. Setup

Let $\mathcal{X}$ be a domain in $\mathbb{R}^d$. Our aim is to numerically compute expectations of the form

$$\pi(f) = \int_{\mathcal{X}} f(x)\pi(\mathrm{d}x),$$

where $f : \mathcal{X} \longrightarrow \mathbb{R}$ and π is a probability measure supported on $\mathcal{X}$. If the dimension of the space $\mathcal{X}$ is large and $\pi(f)$ can not be computed analytically, one can apply Monte Carlo methods. However, in many practical situations direct sampling from π is impossible and this precludes the use of plain Monte Carlo methods in this case. One popular alternative to Monte Carlo is Markov Chain Monte Carlo (MCMC), where one is looking for a discrete time (possibly non-homogeneous) Markov chain $(X_p)_{p\geq 0}$ such that π is its unique invariant measure. In this paper we study a class of MCMC algorithms with $(X_p)_{p\geq 0}$ satisfying the the following recurrence relation:

$$X_p = \Phi_p(X_{p-1}, \xi_p), \quad p = 1, 2, \dots, \quad X_0 = x_0, \tag{1}$$

for some i.i.d. random vectors $\xi_p \in \mathbb{R}^m$ with distribution P_ξ and some Borel-measurable functions $\Phi_p \colon \mathcal{X} \times \mathbb{R}^m \to \mathcal{X}$. In fact, this is quite general class of Markov chains (see Theorem 1.3.6 in [10]) and many well-known MCMC algorithms can be represented in form (1). Let us consider two popular examples.

Example 1 (Unadjusted Langevin Algorithm) *Fix a sequence of positive time steps $(\gamma_p)_{p\geq 1}$. Given a Borel function $\mu \colon \mathbb{R}^d \to \mathbb{R}^d$, consider a non-homogeneous discrete-time Markov chain $(X_p)_{p\geq 0}$ defined by*

$$X_{p+1} = X_p - \gamma_{p+1}\mu(X_p) + \sqrt{\gamma_{p+1}}Z_{p+1}, \tag{2}$$

where $(Z_p)_{p\geq 1}$ is an i.i.d. sequence of d-dimensional standard Gaussian random vectors. If $\mu = \frac{\nabla U}{2}$ for some continuously differentiable function U, then Markov chain (2) can be used to approximately sample from the density

$$\pi(x) = \mathrm{const}\, \mathrm{e}^{-U(x)}, \quad \mathrm{const} = 1 \Big/ \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)}\, \mathrm{d}x, \tag{3}$$

provided that $\int_{\mathbb{R}^d} \mathrm{e}^{-U(x)}\, \mathrm{d}x$ is finite. This method is usually referred to as Unadjusted Langevin Algorithm (ULA). In fact the Markov chain (2) arises as the Euler-Maruyama discretization of the Langevin diffusion

$$\mathrm{d}Y_t = -\mu(Y_t)\, \mathrm{d}t + \mathrm{d}W_t$$

with nonnegative time steps $(\gamma_p)_{p\geq 1}$, and, under mild technical conditions, the latter Langevin diffusion admits π of (3) as its unique invariant distribution; see [7] and [12].

Example 2 (Metropolis-Adjusted Langevin Algorithm) *The Metropolis-Hastings algorithm associated with a target density π requires the choice of a sequence of conditional densities $(q_p)_{p\geq 1}$ also called proposal or candidate kernels. The transition from the value of the Markov chain X_p at time p and its value at time $p+1$ proceeds via the following transition step:*

Given $X_p = x$;

1. *Generate $Y_p \sim q_p(\cdot|x)$;*
2. *Put*

$$X_{p+1} = \begin{cases} Y_p, & \textit{with probability } \alpha(x, Y_p), \\ x, & \textit{with probability } 1-\alpha(x, Y_p), \end{cases}$$

where

$$\alpha(x,y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\frac{q_p(x|y)}{q_p(y|x)}\right\}.$$

Then, as shown in Metropolis et al. [18], this transition is reversible with respect to π and therefore preserves the stationary density π. If q_p have a wide enough support to eventually reach any region of the state space $\mathcal{X}$ with positive mass under π, then this transition is irreducible and π is a maximal irreducibility measure [17]. The Metropolis-Adjusted Langevin algorithm (MALA) takes (2) as proposal, that is,

$$q_p(y|x) = (\gamma_{p+1})^{-d/2}\boldsymbol{\varphi}\Big([y - x + \gamma_{p+1}\mu(x)]/\sqrt{\gamma_{p+1}}\Big),$$

where $\boldsymbol{\varphi}(z) = (2\pi)^{-d/2}\exp\{-|z|^2/2\}$, $z \in \mathbb{R}^d$, denotes the density of a d-dimensional standard Gaussian random vector. The MALA algorithms usually provide noticeable speed-ups in convergence for most problems. It is not difficult to see that the MALA chain can be compactly represented in the form

$$X_{p+1} = X_p + \mathbb{1}\big(U_{p+1} \leq \alpha(X_p, Y_p)\big)(Y_p - X_p),\ Y_p \quad = X_p - \gamma_{p+1}\mu(X_p) + \sqrt{\gamma_{p+1}}Z_{p+1},$$

where $(U_p)_{p\geq 1}$ is an i.i.d. sequence of uniformly distributed on $[0,1]$ random variables independent of $(Z_p)_{p\geq 1}$. Thus we recover (1) with $\xi_p = (U_p, Z_p) \in \mathbb{R}^{d+1}$ and

$$\Phi_p(x,(u,z)^\top) = x + \mathbb{1}\big(u \leq \alpha(x, x - \gamma_p\mu(x) + \sqrt{\gamma_p}z)\big)(-\gamma_p\mu(x) + \sqrt{\gamma_p}z).$$

Example 3 *Let $(X_t)_{t\in[0,T]}$ be the unique strong solution to a SDE of the form:*

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t, \quad t \geq 0, \tag{4}$$

where W is a standard $\mathbb{R}^m$-valued Brownian motion, $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ are locally Lipschitz continuous functions with at most linear growth. The process $(X_t)_{t\geq 0}$

is a Markov process and let L denote its infinitesimal generator defined by

$$Lg = b^\top \nabla g + \frac{1}{2} Tr(\sigma^\top D^2 g \sigma)$$

for any $g \in C^2(\mathbb{R}^d)$. If there exists a continuously twice differentiable Lyapunov function $V : \mathbb{R}^d \to \mathbb{R}_+$ such that

$$\sup_{x \in \mathbb{R}^d} LV(x) < \infty, \quad \limsup_{|x| \to \infty} LV(x) < 0,$$

then there is an invariant probability measure π for X, that is, $X_t \sim \pi$ for all $t > 0$ if $X_0 \sim \pi$. Invariant measures are crucial in the study of the long term behaviour of stochastic differential systems (4). Under some additional assumptions, the invariant measure π is ergodic and this property can be exploited to compute the integrals $\pi(f)$ for $f \in L^2(\pi)$ by means of ergodic averages. The idea is to replace the diffusion X by a (simulable) discretization scheme of the form (see e.g. [21])

$$\bar{X}_{n+1} = \bar{X}_n + \gamma_{n+1} b(\bar{X}_n) + \sigma(\bar{X}_n)(W_{\Gamma_{n+1}} - W_{\Gamma_n}), \quad n \geq 0, \quad \bar{X}_0 = X_0,$$

where $\Gamma_n = \gamma_1 + \ldots + \gamma_n$ and $(\gamma_n)_{n \geq 1}$ is a non-increasing sequence of time steps. Then for a function $f \in L^2(\pi)$ we can approximate $\pi(f)$ via

$$\pi_n^\gamma(f) = \frac{1}{\Gamma_n} \sum_{i=1}^n \gamma_i f(\bar{X}_{i-1}).$$

Due to typically high correlation between $X_0, X_1, \ldots$ variance reduction is of crucial importance here. As a matter of fact, in many cases there is no explicit formula for the invariant measure and this makes the use of gradient based variance reduction techniques (see e.g. [19]) impossible in this case. On the contrary, our method can be directly used to reduce the variance of the ergodic estimator π_n^γ without explicit knowledge of π.

3. Martingale representation and variance reduction

In this section we give a general discrete-time martingale representation for the Markov chain (1) which is used below to construct an efficient variance reduction algorithm. Let $(\phi_k)_{k \in \mathbb{Z}_+}$ be a complete orthonormal system in $L^2(\mathbb{R}^m, P_\xi)$ with $\phi_0 \equiv 1$. In particular, we have

$$\mathsf{E}[\phi_i(\xi)\phi_j(\xi)] = \delta_{ij}, \quad i, j \in \mathbb{Z}_+$$

with $\xi \sim P_\xi$. Notice that this implies that the random variables $\phi_k(\xi)$, $k \geq 1$, are centered. As an example, we can take multivariate Hermite polynomials for the ULA algorithm and a tensor product of Shifted Legendre polynomials for "uniform part" and Hermite polynomials for "gaussian part" of the random variable $\xi = (u, z)^T$ in MALA, as the Shifted Legendre polynomials are orthogonal with respect to the Lebesgue measure on $[0, 1]$. In the sequel, we denote by $(\mathcal{G}_p)_{p \in \mathbb{Z}_+}$ the filtration generated by generated by $(\xi_p)_{p \in \mathbb{N}^*}$ with the convention $\mathcal{G}_0 = \text{triv}$.

Theorem 1 *Let f be a Borel function $\mathbb{R}^d \to \mathbb{R}$ such that $\mathsf{E}\left[|f(X_p)|^2\right] < \infty$, with a Markov chain $(X_p)_{p\geq 0}$ defined in (1). Then, for $p > j$, the following representation holds in $L^2(P)$*

$$f(X_p) = \mathsf{E}\left[f(X_p)|\,\mathcal{G}_j\right] + \sum_{k=1}^{\infty}\sum_{l=j+1}^{p} a_{p,l,k}(X_{l-1})\phi_k\left(\xi_l\right), \tag{5}$$

where

$$a_{p,l,k}(x) = \mathsf{E}\left[\,f(X_p)\phi_k\left(\xi_l\right)|\,X_{l-1} = x\right], \quad p \geq l, \quad k \in \mathbb{N}. \tag{6}$$

Proof The expansion obviously holds for $p = 1$ and $j = 0$. Indeed, due to the orthonormality and completeness of the system (ϕ_k), we have

$$f(X_1) = \mathsf{E}\left[f(X_1)\right] + \sum_{k\geq 1} a_{1,1,k}(X_0)\phi_k(\xi_1)$$

with

$$a_{1,1,k}(x_0) = \mathsf{E}\left[\,f(X_1)\phi_k\left(\xi_1\right)|\,X_0 = x_0\right],$$

provided $\mathsf{E}\left[|f(X_1)|^2\right] < \infty$. Recall that $\mathcal{G}_l = \sigma(\xi_1, \ldots, \xi_l)$, $l = 1, 2, \ldots$, and $\mathcal{G}_0 = \text{triv}$. Suppose that (5) holds for $p = q$, all $j < q$, and all Borel-measurable functions f with $\mathsf{E}\left[|f(X_q)|^2\right] < \infty$. Let us prove it for $p = q + 1$. Given f with $\mathsf{E}\left[|f(X_p)|^2\right] < \infty$, due to the orthonormality and completeness of the system (ϕ_k), we get by conditioning on $\mathcal{G}_q$,

$$f(X_p) = \mathsf{E}\left[\,f(X_p)|\,\mathcal{G}_q\right] + \sum_{k\geq 1} \alpha_{p,q+1,k}\phi_k(\xi_{q+1}),$$

where

$$\alpha_{p,q+1,k} = \mathsf{E}\left[\,f(X_p)\phi_k(\xi_{q+1})|\,\mathcal{G}_q\right].$$

By the Markov property of (X_l), we have $\mathsf{E}[f(X_p)|\mathcal{G}_q] = \mathsf{E}[f(X_p)|X_q]$. Furthermore, a calculation involving intermediate conditioning on $\mathcal{G}_{q+1}$ and the recurrence relation $X_{q+1} = \Phi_{q+1}(X_q, \xi_{q+1})$ verifies that

$$\alpha_{p,q+1,k} = \mathsf{E}\left[\,f(X_p)\phi_k(\xi_{q+1})|\,X_q\right] = a_{p,q+1,k}(X_q)$$

for suitably chosen Borel-measurable functions $a_{p,q+1,k}$. We thus arrive at

$$f(X_p) = \mathsf{E}\left[\,f(X_p)|\,X_q\right] + \sum_{k\geq 1} a_{p,q+1,k}(X_q)\phi_k(\xi_{q+1}), \tag{7}$$

which is the required statement in the case $j = q$. Now assume $j < q$. The random variable $\mathsf{E}\left[\,f(X_p)|\,X_q\right]$ is square integrable and has the form $g(X_q)$, hence the induction hypothesis applies, and we get

$$\mathsf{E}\left[\,f(X_p)|\,X_q\right] = \mathsf{E}\left[\,f(X_p)|\,X_j\right] + \sum_{k\geq 1}\sum_{l=j+1}^{q} a_{p,l,k}(X_{l-1})\phi_k(\xi_l) \tag{8}$$

with

$$a_{p,l,k}(X_{l-1}) = \mathsf{E}\left[\mathsf{E}\left[f(X_p)\middle|\mathcal{G}_q\right]\phi_k(\xi_l)\middle|\mathcal{G}_{l-1}\right] = \mathsf{E}\left[f(X_p)\phi_k(\xi_l)\middle|\mathcal{G}_{l-1}\right]$$
$$= \mathsf{E}\left[f(X_p)\phi_k(\xi_l)\middle|X_{l-1}\right].$$

Formulas (7) and (8) conclude the proof. □

From numerical point of view another representation of the coefficients $a_{p,l,k}$ turns out to be more useful.

Proposition 2 *The coefficients $a_{p,l,k}$ in (6) can be alternatively represented as*

$$a_{p,l,k}(x) = \mathsf{E}\left[\phi_k(\xi)\, Q_{p,l}\left(\Phi_l(x,\xi)\right)\right]$$

with $Q_{p,l}(x) = \mathsf{E}\left[f(X_p)\middle|X_l = x\right]$, $p \geq l$. The functions $(Q_{p,l})_{l=0}^{p}$ can be computed by the backward recurrence: $Q_{p,p}(x) = f(x)$ and for $l \in \{0, \cdots, p-1\}$

$$Q_{p,l}(x) = \mathsf{E}\left[Q_{p,l+1}(X_{l+1})\middle|X_l = x\right]. \tag{9}$$

Next we show how the representation (5) can be used to efficiently reduce the variance of MCMC algorithms. Let $(\gamma_p)_{p\in\mathbb{N}}$ be a sequence of positive and non-increasing step sizes with $\sum_{p=1}^{\infty}\gamma_p = \infty$ and, for $n, l \in \mathbb{N}$, $n \leq l$, we set

$$\Gamma_{n,l} = \sum_{p=n}^{l}\gamma_p.$$

Consider a weighted average estimator $\pi_n^N(f)$ of the form

$$\pi_n^N(f) = \sum_{p=N+1}^{N+n}\omega_{p,n}^N f(X_p), \quad \omega_{p,n}^N = \gamma_{p+1}\Gamma_{N+2,N+n+1}^{-1}, \tag{10}$$

where $N \in \mathbb{N}_0$ is the length of the burn-in period and $n \in \mathbb{N}$ the number of effective samples. Given N and n as above, for $K \in \mathbb{N}$, denote

$$M_{K,n}^N(f) = \sum_{k=1}^{K}\sum_{l=N+1}^{N+n}\bar{a}_{l,k}(X_{l-1})\phi_k(\xi_l), \tag{11}$$

where

$$\bar{a}_{l,k}(x) = \sum_{p=l}^{N+n}\omega_{p,n}^N a_{p,l,k}(x) = \mathsf{E}\left[\left(\sum_{p=l}^{N+n}\omega_{p,n}^N f(X_p)\right)\phi_k(\xi_l)\middle|X_{l-1} = x\right]. \tag{12}$$

Since X_{l-1} is independent of ξ_l and $\mathsf{E}[\phi_k(\xi_l)] = 0$, $k \geq 1$, the r.v. $M_{K,n}^N(f)$ has zero mean and can be viewed as a control variate.

4. Analysis of variance reduced ULA

The representation (6) suggests that the variance of the variance-reduced estimator

$$\pi^N_{K,n}(f) = \pi^N_n(f) - M^N_{K,n}(f) \tag{13}$$

should be small for K large enough. In this section we provide an analysis of the variance-reduced ULA algorithm (see Example 1). We shall use the notations $\mathbb{N} = \{1, 2, \ldots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. By H_k, $k \in \mathbb{N}_0$, we denote the normalized Hermite polynomial on $\mathbb{R}$, that is,

$$H_k(x) = \frac{(-1)^k}{\sqrt{k!}} \mathrm{e}^{x^2/2} \frac{\partial^k}{\partial x^k} \mathrm{e}^{-x^2/2}, \quad x \in \mathbb{R}.$$

For a multi-index $\mathbf{k} = (k_i) \in \mathbb{N}_0^d$, $\mathbf{H_k}$ denotes the normalized Hermite polynomial on $\mathbb{R}^d$, that is,

$$\mathbf{H_k}(\mathbf{x}) = \prod_{i=1}^d H_{k_i}(x_i), \quad \mathbf{x} = (x_i) \in \mathbb{R}^d.$$

In what follows, we also use the notation $|\mathbf{k}| = \sum_{i=1}^d k_i$ for $\mathbf{k} = (k_i) \in \mathbb{N}_0^d$, and we set $\mathcal{G}_p = \sigma(Z_1, \ldots, Z_p)$, $p \in \mathbb{N}$, and $\mathcal{G}_0 = \mathrm{triv}$. Given N and n as above, for $K \in \mathbb{N}$, denote

$$\begin{aligned} M^N_{K,n}(f) &= \sum_{\mathbf{k}:\, 0<\|\mathbf{k}\|\le K} \sum_{p=N+1}^{N+n} \omega^N_{p,n} \sum_{l=N+1}^{p} a_{p,l,\mathbf{k}}(X_{l-1}) \mathbf{H_k}(Z_l) \\ &= \sum_{\mathbf{k}:\, 0<\|\mathbf{k}\|\le K} \sum_{l=N+1}^{N+n} \bar{a}_{l,\mathbf{k}}(X_{l-1}) \mathbf{H_k}(Z_l) \end{aligned} \tag{14}$$

with $\|\mathbf{k}\| = \max_i k_i$ and

$$\bar{a}_{l,\mathbf{k}}(x) = \sum_{p=l}^{N+n} \omega^N_{p,n} a_{p,l,\mathbf{k}}(x) = \mathsf{E}\left[\left(\sum_{p=l}^{N+n} \omega^N_{p,n} f(X_p) \right) \mathbf{H_k}(Z_l) \,\middle|\, X_{l-1} = x \right].$$

For an estimator $\rho(f) \in \{\pi^N_n(f), \pi^N_{K,n}(f)\}$ of $\pi(f)$ (see (10) and (13)), we are interested in its conditional Mean Squared Error (MSE), which can be decomposed as the sum of the squared conditional bias and the conditional variance:

$$\begin{aligned} \mathrm{MSE}\left[\rho(f)|\mathcal{G}_N\right] &= \mathsf{E}\left[(\rho(f) - \pi(f))^2 | \mathcal{G}_N\right] \\ &= \left(\mathsf{E}[\rho(f)|\mathcal{G}_N] - \pi(f)\right)^2 + \mathsf{Var}\left[\rho(f)|\mathcal{G}_N\right]. \end{aligned} \tag{15}$$

The quantities in (15) are conditioned on $\mathcal{G}_N$, as it reflects the way the estimators are used for MCMC: the path of the Markov chain is simulated only once, and we start to use the realized values of the Markov chain to construct our estimate only after the burn-in period of length N. We also notice that, due to the Markovian structure, the conditioning on $\mathcal{G}_N$ in (15) is equivalent to conditioning on X_N only (this is particularly clear in the case $\rho(f) = \pi^N_n(f)$ but requires some calculation in the remaining case $\rho(f) = \pi^N_{K,n}(f)$).

4.1. Squared conditional bias

Due to the martingale transform structure of $M^N_{K,n}(f)$, we have

$$\mathsf{E}\left[M^N_{K,n}(f)\middle|\,\mathcal{G}_N\right]=0,$$

Hence both estimators $\pi^N_n(f)$ and $\pi^N_{K,n}(f)$ have the same conditional bias. Notice that this remains true also when we substitute the coefficients $a_{p,l,\mathbf{k}}$ in (14) with some independent approximations $\widehat{a}_{p,l,\mathbf{k}}$. For a bounded Borel function f, we can estimate the conditional bias similarly to the beginning of [12, Section 4]:

$$\begin{aligned}\left(\mathsf{E}[\pi^N_{K,n}(f)|\mathcal{G}_N]-\pi(f)\right)^2&=\left(\mathsf{E}[\pi^N_n(f)|\mathcal{G}_N]-\pi(f)\right)^2\\&\le \operatorname{osc}(f)^2\sum_{p=N+1}^{N+n}\omega^N_{p,n}\left\|Q^{N+1,p}_\gamma(X_N,\cdot)-\pi(\cdot)\right\|^2_{TV},\end{aligned}\tag{16}$$

where $\operatorname{osc}(f):=\sup_{x\in\mathbb{R}^d}f(x)-\inf_{x\in\mathbb{R}^d}f(x)$, $\|\mu-\nu\|_{\mathrm{TV}}$ denotes the total variation distance between probability measures μ and ν, that is,

$$\|\mu-\nu\|_{TV}=\sup_{A\in\mathcal{B}(\mathbb{R}^d)}|\mu(A)-\nu(A)|,$$

$\pi(\cdot)$ denotes the probability measure on $\mathbb{R}^d$ with density π of (3); for $\gamma>0$, the Markov kernel R_γ from $(\mathbb{R}^d,\mathcal{B}(\mathbb{R}^d))$ into $(\mathbb{R}^d,\mathcal{B}(\mathbb{R}^d))$ is defined by

$$R_\gamma(x,A)=\int_A\frac{1}{(4\pi\gamma)^{d/2}}\exp\left\{-\frac{1}{4\gamma}\left\|y-x+\gamma\mu(x)\right\|^2\right\}dy,$$

while, for $k,l\in\mathbb{N}$, $k\le l$, the kernel $Q^{k,l}_\gamma$ is $Q^{k,l}_\gamma=R_{\gamma_l}\cdots R_{\gamma_k}$, which, finally, provides the (random) measure $Q^{N+1,p}_\gamma(X_N,\cdot)$ used in the right-hand side of (16). Different specific upper bounds for the squared bias can be deduced from (16) using results of Section 3 in [12] on bounds in the total variation distance.

4.2. Conditional variance

An upper bound for the variance of the classical estimator (10) is provided in in [12, Theorem 17]. As for the estimator (13), we introduce lemma of conditional variance decomposition. It follows from (14), Theorem 1 and independence condition for the sequence $(Z_p)_{p\ge1}$.

Lemma 3 *For variance-reduced estimator* (13)*, defined for the Markov Chain* (1) *the following representation holds*

$$\mathsf{Var}\left[\pi^N_{K,n}(f)\middle|\,\mathcal{G}_N\right]=\sum_{\mathbf{k}:\ \|\mathbf{k}\|\ge K+1}\sum_{l=N+1}^{N+n}\mathsf{E}\left[\left|\bar{a}_{l,\mathbf{k}}(X_{l-1})\right|^2\middle|\,\mathcal{G}_N\right].\tag{17}$$

Now we present an upper bound for the right-hand sides of (17) in case of ULA algorithm.

Theorem 4 *Fix $K \in \mathbb{N}$. Suppose that f and μ are $K+1$ times continuously differentiable and it holds*

$$|\partial^{\mathbf{k}} f(x)| \le B_f, \quad |\partial^{\mathbf{k}} \mu(x)| \le B_\mu, \quad x \in \mathbb{R}^d$$

for all $\mathbf{k}$ with $0 < \|\mathbf{k}\| \le K+1$,

$$\mathrm{J}_\mu(x) \ge b_\mu I, \quad x \in \mathbb{R}^d,$$

for some positive number $b_\mu \in (0, B_\mu]$ where J_μ stands for Jacobian of μ. Let $(\gamma_k)_{k\in\mathbb{N}}$ be a sequence of positive and non-increasing step sizes with $\sum_{k=1}^{\infty} \gamma_k = \infty$. We assume that $\gamma_1 < \frac{1}{B_\mu}$ and that

$$\sum_{r=j}^{\infty} \gamma_r \prod_{k=j}^{r} [1 - \gamma_k b_\mu] \le C, \quad \text{for all } j \in \mathbb{N}, \tag{18}$$

with some constant C. Then it holds

$$\mathsf{Var}\left[\pi^N_{K,n}(f) \middle| \mathcal{G}_N\right] \lesssim \frac{1}{\Gamma^2_{N+2,N+n+1}} \sum_{l=N+1}^{N+n} \sum_{I \subseteq \{1,\dots,d\},\, I\neq\emptyset} \left(\frac{\gamma_l}{2}\right)^{|I|K}, \tag{19}$$

where the sum in (19) is taken over all subsets I of $\{1,\dots,d\}$ and $\lesssim$ stands for inequality up to a constant not depending on n and N.

Remark 1 *Note that in Theorem (4) we don't require that the function μ itself is bounded. Assumption (18) is not restrictive. For instance, a straightforward calculation shows that (18) is satisfied in most interesting case $\gamma_k = \mathrm{const}/k^\alpha$ with $\alpha \in (0,1)$.*

5. Nonparametric regression

The functions $(\bar{a}_{l,k})$ need to be estimated before one can apply the proposed variance reduction approach. This amounts to solve a nonparametric regression problem. We present a generic regression algorithm. Algorithm starts with estimating the functions $\bar{Q}_l$ for $l = N+1,\dots,N+n$, where

$$\bar{Q}_l(x) = \sum_{p=l}^{N+n} \omega^N_{p,n} Q_{p,l}(x) = \omega^N_{l,n} f(x) + \mathsf{E}\left[\sum_{p=l+1}^{N+n} \omega^N_{p,n} f(X_p) \middle| X_l = x\right].$$

We first generate T paths conditionally independent of the σ-algebra generated by the burn-in sequence $X_1,\dots,X_N$:

$$\mathcal{T} = \left\{(X^{(s)}_{N+1},\dots,X^{(s)}_{N+n}), \quad s = 1,\dots,T\right\}$$

of the chain X (the so-called "training paths"). The regression algorithm proceeds with estimating the functions $(\bar{Q}_l)$ by solving the least squares optimization problems

$$\widehat{Q}_l = \arg\min_{\psi\in\Psi} \sum_{s=1}^{T} \left| \sum_{p=l}^{N+n} \omega_{p,n}^N f(X_p^{(s)}) - \psi(X_l^{(s)}) \right|^2 \tag{20}$$

for $l = N+1, \ldots, N+n-1$. where Ψ is a class of functions on $\mathbb{R}^d$. Next we estimate the coefficients $\bar{a}_{l,k}$ using the formula

$$\widehat{a}_{l,k}(x) = \mathsf{E}\left[\phi_k(\xi)\, \widehat{Q}_l(\Phi_l(x,\xi)) \middle| \mathcal{T} \right], \tag{21}$$

where ξ is a random variable independent from $\mathcal{T}$ with law P_ξ, Φ_l is defined in (1) and the expectation is taken according to known distribution P_ξ. Conditioning on $\mathcal{T}$ underlines that estimation of $\widehat{Q}_l$ derived from training trajectories. In many cases integration can be done in closed analytical form.

For example, in the case of ULA algorithm we have P_ξ is the multivariate normal distribution $\mathcal{N}(0, I_d)$ and $\Phi_l(x,\xi) = x - \gamma_l \mu(x) + \sqrt{\gamma_l}\xi$. Therefore if we use polynomials to approximate $(\bar{Q}_l)$, then $(\widehat{a}_{l,k})$ can be even computed in closed form; implementations details are provided in Section 7.

Upon estimating the coefficients $(\widehat{a}_{l,k})$, one can construct an empirical estimate of $M_{K,n}^N(f)$ in the form

$$\widehat{M}_{K,n}^N(f) = \sum_{1\le k\le K} \sum_{l=N+1}^{N+n} \widehat{a}_{l,k}(X_{l-1})\phi_k(\xi_l).$$

Obviously $\mathsf{E}[\widehat{M}_{K,n}^N(f)|\mathcal{T}] = 0$ and $\widehat{M}_{K,n}^N(f)$ is indeed a valid control variate in that it does not introduce any bias.

6. Complexity analysis for ULA

The following result quantifies the error of estimating the functions $(\bar{Q}_l)$ via the algorithm and its proof follows from Theorem 2.2 in [2].

Theorem 5 *Suppose that for any $l \in \{N+1, \ldots, N+n+1\}$,*

$$\mathsf{E}\left[\left(\sum_{p=l}^{N+n} \omega_{p,n}^N f(X_p) - \bar{Q}_l(X_l) \right)^4 \right] \le \sigma_l^4,$$

for some finite positive numbers $\sigma_{N+1}, \ldots, \sigma_{N+n}$. Furthermore, assume that $\Psi = \mathrm{span}\{\psi_1, \ldots, \psi_D\}$, where the functions $\psi_1, \ldots, \psi_D$ are uniformly bounded and satisfy

$$\max_l \sup_{g\in\Psi\setminus\{0\}} \|g\|_\infty^2 / \mathsf{E}[g^2(X_l)|\mathcal{G}_N] \le B < \infty.$$

Then for any values of ε and T such that $2/T \le \varepsilon \le 1$ and

$$T \gtrsim B^2 \left[BD + \log(2/\varepsilon) + \frac{B^2 D^2}{T} \right]$$

it holds with probability at least $1-\varepsilon$,

$$\mathsf{E}\left[\left| \bar{Q}_l(X_l) - \widehat{Q}_l(X_l) \right|^2 \middle| \mathcal{T} \vee \mathcal{G}_N \right] \lesssim \sigma_l^2 B \left(\frac{BD + \log(2/\varepsilon)}{T} + \frac{B^2 D^2}{T^2} \right) + \inf_{\psi \in \Psi} \mathsf{E}\left[\left| \bar{Q}_l(X_l) - \psi(X_l) \right|^2 \middle| \mathcal{G}_N \right], \quad (22)$$

for $l = N+1, \dots, N+n$, with $\lesssim$ standing for inequality up to a universal multiplicative constant.

Now we are able to give a bound for the difference between $M_{K,n}^N$ and $\widehat{M}_{K,n}^N$.

Proposition 6 *Under conditions of Theorem 5, we have with probability at least $1-\varepsilon$,*

$$\begin{aligned} &\mathsf{E}\left[\left| \widehat{M}_{K,n}^N(f) - M_{K,n}^N(f) \right|^2 \middle| \mathcal{T}, \mathcal{G}_N \right] \quad (23) \\ &\quad \lesssim K \left[\sum_{l=N+1}^{N+n} \sigma_l^2 \right] B \left(\frac{BD + \log(2/\varepsilon)}{T} + \frac{B^2 D^2}{T^2} \right) \\ &\quad + K \sum_{l=n+1}^{N+n} \inf_{\psi \in \Psi} \mathsf{E}\left[\left| \bar{Q}_l(X_l) - \psi(X_l) \right|^2 \middle| \mathcal{G}_N \right]. \end{aligned}$$

Proof Using the conditional Cauchy-Schwarz inequality and orthonormality of $(\phi_k)_{k \ge 0}$, we derive

$$\mathsf{E}\left[\left| \widehat{a}_{l,k}(X_{l-1}) - \bar{a}_{l,k}(X_{l-1}) \right|^2 \middle| \mathcal{T} \vee \mathcal{G}_N \right] \le \mathsf{E}\left[\left| \widehat{Q}_l(X_l) - \bar{Q}_l(X_l) \right|^2 \middle| \mathcal{T} \vee \mathcal{G}_N \right]$$

By the Jensen inequality and orthonormality of $(\phi_k)_{k \ge 0}$,

$$\mathsf{E}\left[\left| \widehat{M}_{K,n}^N(f) - M_{K,n}^N(f) \right|^2 \middle| \mathcal{T} \vee \mathcal{G}_N \right] \le \sum_{1 \le k \le K} \sum_{l=N+1}^{N+n} \mathsf{E}\left[\left| \widehat{a}_{l,k}(X_{l-1}) - \bar{a}_{l,k}(X_{l-1}) \right|^2 \middle| \mathcal{T} \vee \mathcal{G}_N \right].$$

□

6.1. Complexity analysis of variance reduction for the ULA algorithm

In the case of ULA one can bound the quantities σ_l^2 using L^p-type Sobolev inequalities (see [1]), see Remark 3 in Section 8. In particular, we can derive that under conditions of Theorem 4 with $K = 1$,

$$\sigma_l^2 \lesssim \frac{1}{\Gamma^2_{N+2,N+n+1}}, \quad l = N+1, \dots, N+n+1,$$

where $\lesssim$ stands for the inequality up to a multiplicative constant not depending on n and N. Using this inequality and combining (23) with Theorem 4, we get for the variance of $\widehat{\pi}^N_{K,n}(f) = \pi^N_n(f) - \widehat{M}^N_{K,n}(f)$, with probability at least $1-\varepsilon$,

$$\begin{aligned}\mathsf{Var}\left[\widehat{\pi}^N_{K,n}(f)\middle|\mathcal{T},\mathcal{G}_N\right] \lesssim\ & \frac{nKB}{\Gamma^2_{N+2,N+n+1}}\left(\frac{BD+\log(2/\varepsilon)}{T}+\frac{B^2D^2}{T^2}\right)\\ &+\frac{1}{\Gamma^2_{N+2,N+n+1}}\sum_{l=N+1}^{N+n}\sum_{I\subseteq\{1,\dots,d\},\, I\neq\emptyset}\left(\frac{\gamma_l}{2}\right)^{|I|K}\\ &+K\sum_{l=n+1}^{N+n}\inf_{\psi\in\Psi}\mathsf{E}\left[\left|\bar{Q}_l(X_l)-\psi(X_l)\right|^2\middle|\mathcal{T}\vee\mathcal{G}_N\right]. \qquad (24)\end{aligned}$$

In order to assess the complexity of the proposed algorithm, we first prove that under some conditions the coefficients $a_{p,l,\mathbf{k}}$ decay exponentially fast as $|p-l|\to\infty$.

Lemma 7 *Suppose that $f, \mu \in C^1(\mathbb{R}^d)$, $0 < b_\mu I \le \mathrm{J}_\mu(x) \le B_\mu I$ and $B_\mu\gamma_l \le 1$, then*

$$\|a_{p,l,\mathbf{k}}\|_\infty \le \sqrt{\gamma_l}\,\|\nabla f\|_\infty \exp\left(-b_\mu\sum_{r=l+1}^{p}\gamma_r\right), \quad \mathbf{k}\in\mathbb{N}^d\setminus\{\mathbf{0}\}.$$

Corollary 1 *Assume that*

$$\gamma_k = \gamma_1 k^{-\alpha}, \quad \alpha\in(0,1), \qquad (25)$$

then

$$\|a_{p,l,\mathbf{k}}\|_\infty \le \sqrt{\gamma_1}\,l^{-\alpha/2}\,\|\nabla f\|_\infty \exp\left(-cb_\mu\gamma_1(p^{1-\alpha}-l^{1-\alpha})\right)$$

for some constant $c > 0$.

Suppose now that the chain is close to stationarity, then $Q_{p,l}$ are functions of $p-l$ only. Furthermore, according to Lemma 7, $a_{p,l,\mathbf{k}}$ are exponentially small for $p-l$ large. As a result we only need to compute a logarithmic (in n) number of functions $Q_{p,l}$. Hence the cost of computing the estimates $\widehat{a}_{p,l,\mathbf{k}}(x)$ for $l = N+1,\dots,N+n$, and $\|\mathbf{k}\| \le K$ using regression on Ψ, is of order

$$\log^{1/(1-\alpha)}(n)K^dTD^2.$$

Suppose for simplicity that all functions $\bar{Q}_l$ are in Ψ for some $D > 0$, that is, the third term in (24) is zero. Then under (25) it is enough to take $K = \lceil 1/\alpha \rceil + 1$ to get

$$\sum_{l=N+1}^{N+n} \sum_{I \subseteq \{1,\dots,d\},\, I \neq \emptyset} \left(\frac{\gamma_l}{2}\right)^{|I|K} \leq C$$

for some constant C not depending on d. Then we have with high probability

$$\mathsf{Var}\left[\widehat{\pi}^N_{K,n}(f) \middle| \mathcal{T} \vee \mathcal{G}_N\right] \lesssim \frac{1}{n^{2(1-\alpha)}} \left[1 + \frac{n}{T}\right],$$

with corresponding cost proportional to $T \log^{1/(1-\alpha)}(n)$, provided $N/n = o(1)$. We should compare this to the standard weighted estimator $\pi^N_n(f)$ with variance

$$\mathsf{Var}\left[\pi^N_n(f) \middle| \mathcal{G}_N\right] \lesssim \frac{1}{n^{1-\alpha}}$$

and cost of order n. Thus while the cost of achieving

$$\mathsf{Var}\left[\pi^N_n(f) \middle| \mathcal{T} \vee \mathcal{G}_N\right] \leq \varepsilon^2$$

is of order $\varepsilon^{-2/(1-\alpha)}$, we get the same bound for $\pi^N_{K,n}(f)$ at a cost of order

$$\varepsilon^{-1/(1-\alpha)} \log^{1/(1-\alpha)}(\varepsilon).$$

In the presence of approximation errors we have to increase D with n to balance the statistical and approximation errors resulting in a smaller complexity reduction. Note that our approach does not require the computation of $\nabla \log(\pi)$ for the construction of control variates. This leads to an additional complexity reduction especially in statistical application where the computation of $\nabla \log(\pi)$ can be very costly.

7. Numerical results

If each function $Q_{p,l}(x)$ can be well approximated by polynomials, then we can use polynomial basis functions in (20) to approximate $\bar{Q}_l$. Suppose that we constructed a polynomial approximation for each function $\bar{Q}_l$ in the form:

$$\widehat{Q}_l(x) = \sum_{\|\mathbf{s}\| \leq m} \beta_{\mathbf{s}} x^{\mathbf{s}}, \quad s = (s_1, \dots, s_d)$$

with some coefficients $\beta_{\mathbf{s}} \in \mathbb{R}$. Then using the identity

$$\xi^j = j! \sum_{r=0}^{\lfloor j/2 \rfloor} \frac{1}{2^r r! \sqrt{(j-2r)!}} H_{j-2r}(\xi), \quad \xi \in \mathbb{R},$$

we derive for all $x \in \mathbb{R}^d$,

$$\widehat{a}_{l,\mathbf{k}}(x) \quad = \quad \mathsf{E}\left[\mathbf{H}_{\mathbf{k}}(x)\widehat{Q}_l(x-\gamma_l\mu(x)+\sqrt{\gamma_l}\xi)\,\middle|\,\mathcal{T}\vee\mathcal{G}_N\right] = \sum_{\|\mathbf{s}\|\le m}\beta_{\mathbf{s}}\prod_{i=1}^{d}P_{i,k_i,s_i}(x),$$

where for all integers i, k, s_i and $x \in \mathbb{R}^d$,

$$P_{l,k,s}(x) = \mathsf{E}\left[H_k(\xi_l)(x_l-\gamma_l\mu_l(x)+\sqrt{\gamma_l}\xi_l)^s\right]$$

is a one-dimensional polynomial (in x) of degree at most s with analytically known coefficients. Alternatively, on practice we can work directly with functions $Q_{p,l}$. Due to (almost) time-homogenous Markov chain generated by ULA, each function $Q_{p,l}$ can be approximated by function, which depends only on the $p-l$, that is,

$$Q_{p,l}(x) = \mathsf{E}\left[f(X_p)|X_l = x\right] \approx Q^{\circ}_{p-l}(x)$$

Thus we have

$$Q^{\circ}_r(x) = \mathsf{E}\left[f(X_{l+r})|X_l = x\right].$$

Consequently, the functions Q°_r can be estimated using a modified least squares criteria

$$\widehat{Q}^{\circ}_r = \arg\min_{\psi\in\Psi}\sum_{s=1}^{T}\sum_{l=N+1}^{N+n-r}\left|f(X^{(s)}_{l+r})-\psi(X^{(s)}_l)\right|^2 \tag{26}$$

for $1 \le r \le n-1$, where $\widehat{Q}^{\circ}_0(x) = f(x)$ by definition. Due to Lemma 7, it is enough to estimate Q°_r for $|r| < n_{\mathrm{trunc}}$ for some truncation level n_{trunc} depending on d and γ. It allows us to use a smaller amount of training trajectories to approximate conditional expectations Q°_r. Finally we construct a truncated version of our estimator:

$$\pi^N_{K,n,n_{\mathrm{trunc}}}(f) = \pi^N_n(f) - M^N_{K,n,n_{\mathrm{trunc}}}(f),$$

where

$$\begin{aligned}\widehat{M}^N_{K,n,n_{\mathrm{trunc}}} \quad = \quad & \sum_{0<\|\mathbf{k}\|\le K}\sum_{p=N+1}^{N+n}\omega^N_{p,n} \\ & \times\sum_{l=N+1}^{p}\widehat{a}_{p-l,\mathbf{k}}(X_{l-1})\mathbf{H}_{\mathbf{k}}(Z_l)\mathbb{1}\{|p-l|<n_{\mathrm{trunc}}\}\end{aligned}$$

and

$$\widehat{a}_{p-l,\mathbf{k}}(x) \quad = \quad \mathsf{E}\left[\mathbf{H}_{\mathbf{k}}(x)\widehat{Q}^{\circ}_{p-l}(x-\gamma_l\mu(x)+\sqrt{\gamma_l}\xi)\,\middle|\,\mathcal{T}\vee\mathcal{G}_N\right].$$

7.1. Gaussian mixtures

We consider a sample generated by ULA with π given by the mixture of two Gaussian distributions with equal weights:

$$\pi(x) = \frac{1}{2(2\pi)^{d/2}} \left(\mathrm{e}^{\frac{-\|x-a\|_2^2}{2}} + \mathrm{e}^{\frac{-\|x+a\|_2^2}{2}} \right), \quad x \in \mathbb{R}^d$$

where $a \in \mathbb{R}^d$ is a given vector. The function $U(x)$ and its gradient are given by

$$U(x) = \frac{1}{2}\|x-a\|_2^2 - \log(1+\mathrm{e}^{-2x^\top a}),\ \nabla U(x) = x - a + 2a(1+\mathrm{e}^{2x^\top a})^{-1},$$

respectively. In our experiments we considered dimensions $d = 2$ and $d = 8$ and take $a = ((2d)^{-1/2}, \ldots, (2d)^{-1/2})^\top$. In order to approximate the expectation $\pi(f)$ with $f(x) = \sum_{i=1}^d x_i$, we have used constant step sizes $\gamma_i = \gamma = 0.2$ and sampled $n = 10^4$ samples for $d = 2$ and $n = 2\times 10^3$ for $d = 8$. We generated $T = 10$ independent "training" trajectories and solved the least squares problems (26) using the first order polynomial approximations for the coefficients $a_{p,\mathbf{k}}$ as described in the previous section. The truncation level n_{trunc} is chosen to be 50. To test our variance reduction algorithm, we generated 100 independent trajectories. In Figure 1 we compare our approach to variance reduction methods of [19] and [5].

7.2. Binary Logistic Regression

Second experiment considers the problem of logistic regression. Suppose we have i.i.d. sample $\{(\mathbf{X}_i, Y_i)\}$ for $i = 1, \ldots, m$ with features $\mathbf{X}_i \in \mathbb{R}^d$ and binary labels $Y_i \in \{0, 1\}$. The binary logistic regression model defines the conditional distribution of Y given X by a logistic function

$$r(\theta, x) = \frac{e^{\theta^T x}}{1 + e^{\theta^T x}},$$

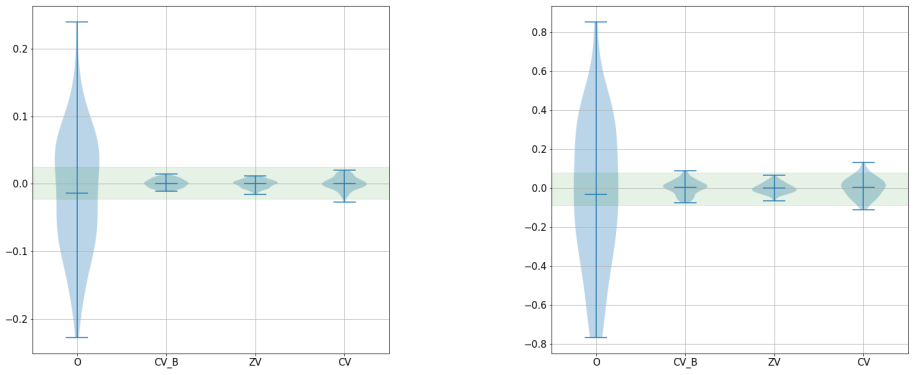
where θ is a parameter of model. In order to estimate θ according to given data, the Bayessian approach introduces prior distribution $\pi_0(\theta)$ and inferences the posterior density $\pi(\theta)$ using Bayes' rule.

In the case of Gaussian prior π_0 with zero mean and covariance matrix $\sigma^2 I_d$, the posterior density takes the form:

$$\pi(\theta) \propto \exp\left\{ -\mathbf{Y}^T\mathbf{X}\theta - \sum_{i=1}^m \log(1 + e^{-\theta^T \mathbf{X}_i}) - \frac{1}{2\sigma^2}\|\theta\|_2^2 \right\},$$

where $\mathbf{Y}$ is defined as $(Y_1, \ldots, Y_m)^T \in \{0, 1\}$ and σ^2 is an additional parameter. We have

$$U(\theta) = \mathbf{Y}^T\mathbf{X}\theta + \sum_{i=1}^m \log(1 + e^{-\theta^T \mathbf{X}_i}) + \frac{1}{2\sigma^2}\|\theta\|_2^2,\ \nabla U(\theta) = \mathbf{X}^T\mathbf{Y} - \sum_{i=1}^m \frac{\mathbf{X}_i}{1 + e^{\theta^T \mathbf{X}_i}} + \frac{1}{\sigma^2}\theta.$$



(a) d = 2 (b) d = 8

Figure 1: Boxplots of ergodic averages from the variance reduced ULA algorithms for the Gaussian mixture model. The compared estimators are the ordinary empirical average (O), our estimator of control variates (CV-B), zero variance estimator (ZV) and the control variates using diffusion approximation (CV), along with the region obtained by the ordinary ULA of the length $10^2 \times n$ (green regions).

To demonstrate the performance of the proposed control variates approach in the above Bayesian logistic regression model, we take a simple dataset from [19], which contains the measurements of four variables on $m = 200$ Swiss banknotes. Prior distribution of the regression parameter $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ assumed to be normal with the covariance matrix $\sigma^2 I_4$, where $\sigma^2 = 100$. To construct trajectories of length $n = 5 \times 10^3$, we take constant step sizes $\gamma_i = 0.1$ for the ULA scheme with $N = 10^3$ burn-in steps. As in the previous experiment we use the first order polynomials approximations to analytically compute the coefficients $\hat{a}_{p-l,\mathbf{k}}$, where $T = 10$ and $n_{trunc} = 50$. The target function is taken to be $f(\theta) = \sum_{i=1}^{i=d} \theta_i$. In order to test our variance reduction algorithm, we generates 100 independent test trajectories. In Figure 2 we compare our approach to variance reduction methods of [19] and [5].

7.3. Binary Probit Regression

Finally we consider a Bayesian probit regression model in its ordinary form, following [19]. More precisely, the log-likelihood of the model looks as follows

$$\mathbf{L}(\mathbf{Y}|\theta, \mathbf{X}) = \sum_{i=1}^{m} \left[\mathbf{Y}_i \log(\Phi(\theta^T \mathbf{X}_i)) + (1 - \mathbf{Y}_i) \log(\Phi(-\theta^T \mathbf{X}_i)) \right],$$

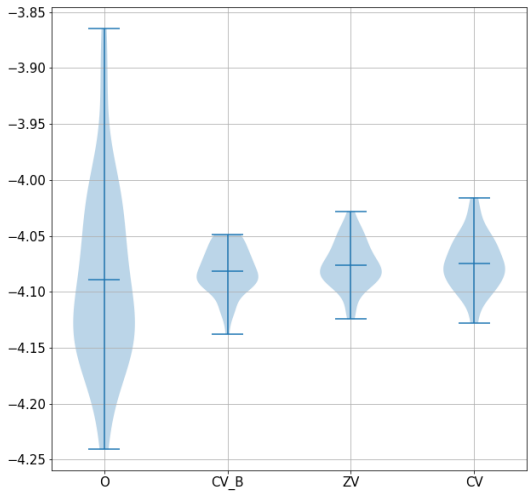


Figure 2: Boxplots of ergodic averages from the variance reduced ULA algorithms for the Logistic Regression. The compared estimators are the ordinary empirical average (O), our estimator of control variates (CV-B), zero variance estimator (ZV) and control variates using diffusion approximation (CV).

where $\theta \in \mathbb{R}^d$ is a parameter of the model and Φ is a cumulative distribution function of the standard normal distribution $\mathcal{N}(0,1)$. Then, if the prior $\pi(\theta) = \mathcal{N}(0, \sigma^2 I_d)$, then the posterior distribution expresses as

$$\pi(\theta|\mathbf{Y},\mathbf{X}) \propto \exp\left\{\sum_{i=1}^{m}\left[\mathbf{Y}_i \log(\Phi(\theta^T\mathbf{X}_i)) + (1-\mathbf{Y}_i)\log(\Phi(-\theta^T\mathbf{X}_i))\right] - \frac{1}{2\sigma^2}\|\theta\|_2^2\right\}$$

Consequently,

$$U(\theta) = -\sum_{i=1}^{m}\left[\mathbf{Y}_i \log(\Phi(\theta^T\mathbf{X}_i)) + (1-\mathbf{Y}_i)\log(\Phi(-\theta^T\mathbf{X}_i))\right] + \frac{1}{2\sigma^2}\|\theta\|_2^2$$

$$\nabla U(\theta) = -\mathbf{X}^T s + \frac{1}{\sigma^2}\theta,$$

where the elements of the m-dimensional column vector s are given by

$$s_i = \mathbf{Y}_i \Omega(\theta^T\mathbf{X}_i) - (1-\mathbf{Y}_i)\Omega(-\theta^T\mathbf{X}_i), \quad i = 1,2,\ldots,m,$$

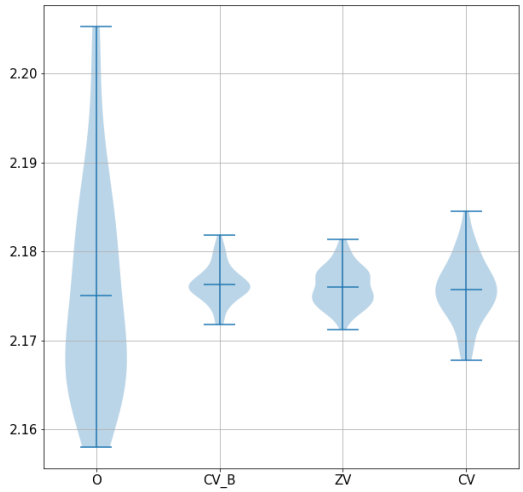


Figure 3: Boxplots of ergodic averages from the variance reduced ULA algorithms for the Probit Regression. The compared estimators are the ordinary empirical average (O), our modified estimator (CV-B), zero variance estimator (ZV) and the estimator with control variates using asymptotic variance minimization (CV).

and $\Omega(\theta^T \mathbf{X}_i) := \phi(\theta^T \mathbf{X}_i)/\Phi(\theta^T \mathbf{X}_i)$, where ϕ stands for the density of the standard normal distribution. As in the previous examples, to show the performance of the proposed variance reduced estimator for the Bayesian probit regression model, banknotes dataset is used and the regularization parameter is taken to be $\sigma^2 = 100$. To generate trajectories of the length $n = 5 \times 10^3$, we take a constant step size $\gamma_i = 0.1$ for the ULA scheme after $N = 10^3$ burn-in steps. We use again the first order polynomials approximations to compute the coefficients $\hat{a}_{p-l,\mathbf{k}}$ and fixed $T = 10$, $n_{trunc} = 50$ and $K = 1$. The target function is now $f(\theta) = \sum_{i=1}^{i=d} \theta_i$. In order to test our variance reduction algorithm, we generate 100 independent test trajectories. In Figure 3 we compare our approach to the variance reduction methods of [19] and [5].

8. Proofs

8.1. Proof of Theorem 4

We start with introducing some notations. For $m \in \mathbb{N}$, a smooth function $h\colon \mathbb{R}^{d\times m} \to \mathbb{R}$ with arguments being denoted $(y_1, \dots, y_m)$, $y_i \in \mathbb{R}^d$, $i = 1, \dots, m$, a multi-index $\mathbf{k} = (k_i) \in \mathbb{N}_0^d$, and $j \in \{1, \dots, m\}$, we use the notation $\partial^{\mathbf{k}}_{y_j} h$ for the multiple derivative of h with respect to the components of y_j:

$$\partial^{\mathbf{k}}_{y_j} h(y_1, \dots, y_m) := \partial^{k_d}_{y_j^d} \dots \partial^{k_1}_{y_j^1} h(y_1, \dots, z_m), \quad y_j = (y_j^1, \dots, y_j^d).$$

In the particular case $m = 1$ we can drop the subscript y_1 in that notation. For $l \le p$, we have the representation

$$X_p = G_{p,l}(X_{l-1}, \sqrt{\gamma_l} Z_l, \dots, \sqrt{\gamma_p} Z_p),$$

where the function $G_{p,l} : \mathbb{R}^{d\times(p-l+2)} \to \mathbb{R}^d$ is defined as

$$G_{p,l}(x, y_l, \dots, y_p) := \Phi_p(\cdot, y_p) \circ \Phi_{p-1}(\cdot, y_{p-1}) \circ \dots \circ \Phi_l(x, y_l) \tag{27}$$

with

$$\Phi_j(x, y) = x - \gamma_j \mu(x) + y, \quad x, y \in \mathbb{R}^d.$$

As a consequence, for a function $f\colon \mathbb{R}^d \to \mathbb{R}$ as in Section 2, we have

$$f(X_p) = f \circ G_{p,l}(X_{l-1}, \sqrt{\gamma_l} Z_l, \dots, \sqrt{\gamma_p} Z_p).$$

In what follows, for $\mathbf{k} \in \mathbb{N}_0^d$, we use the shorthand notation

$$\partial^{\mathbf{k}}_{y_l} f(X_p) := \partial^{\mathbf{k}}_{y_l} [f \circ G_{p,l}](X_{l-1}, \sqrt{\gamma_l} Z_l, \dots, \sqrt{\gamma_p} Z_p) \tag{28}$$

whenever the function $f \circ G_{p,l}$ is smooth enough (that is, f and μ need to be smooth enough). Finally, for a multi-index $\mathbf{k} = (k_i) \in \mathbb{N}_0^d$, we use the notation $\mathbf{k}! := k_1! \cdot \ldots \cdot k_d!$

Lemma 8 *Fix $l \le p$ and some $\mathbf{k}, \mathbf{k}' \in \mathbb{N}_0^d$ with $\mathbf{k}' \le \mathbf{k}$ componentwise. Then the following representation holds*

$$a_{p,l,\mathbf{k}}(X_{l-1}) = \left(\gamma_l^{|\mathbf{k}'|} \frac{(\mathbf{k} - \mathbf{k}')!}{\mathbf{k}!}\right)^{1/2} \mathsf{E}\left[\partial^{\mathbf{k}'}_{y_l} f(X_p) \mathbf{H}_{\mathbf{k}-\mathbf{k}'}(Z_l) \,\middle|\, X_{l-1}\right].$$

Proof Let $\boldsymbol{\varphi}(z) = \frac{1}{(2\pi)^{d/2}} \exp\{-|z|^2/2\}$, $z \in \mathbb{R}^d$, denote the density of a d-dimensional standard Gaussian random vector. We first remark that, for the normalized Hermite polynomial $\mathbf{H}_{\mathbf{k}}$ on $\mathbb{R}^d$, $\mathbf{k} \in \mathbb{N}_0^d$, it holds

$$\mathbf{H}_{\mathbf{k}}(z)\boldsymbol{\varphi}(z) = \frac{(-1)^{|\mathbf{k}|}}{\sqrt{\mathbf{k}!}} \partial^{\mathbf{k}} \boldsymbol{\varphi}(z).$$

This enables to use the integration by parts in vector form as follows (below $\prod_{j=l+1}^{p} := 1$ whenever $l = p$)

$$
\begin{aligned}
a_{p,l,\mathbf{k}}(x) &= \int_{\mathbb{R}^d} \dots \int_{\mathbb{R}^d} f \circ G_{p,l}(x, \sqrt{\gamma_l} z_l, \dots, \sqrt{\gamma_p} z_p) \mathbf{H}_{\mathbf{k}}(z_l) \boldsymbol{\varphi}(z_l) \prod_{j=l+1}^{p} \boldsymbol{\varphi}(z_j)\, dz_l \dots dz_p \\
&= \frac{1}{\sqrt{\mathbf{k}!}} \int_{\mathbb{R}^d} \dots \int_{\mathbb{R}^d} f \circ G_{p,l}(x, \sqrt{\gamma_l} z_l, \dots, \sqrt{\gamma_p} z_p) (-1)^{|\mathbf{k}|} \partial^{\mathbf{k}} \boldsymbol{\varphi}(z_l) \prod_{j=l+1}^{p} \boldsymbol{\varphi}(z_j)\, dz_l \dots dz_p \\
&= \gamma_l^{|\mathbf{k}'|/2} \frac{1}{\sqrt{\mathbf{k}!}} \int_{\mathbb{R}^d} \dots \int_{\mathbb{R}^d} \partial_{y_l}^{\mathbf{k}'} [f \circ G_{p,l}](x, \sqrt{\gamma_l} z_l, \dots, \sqrt{\gamma_p} z_p) \\
&\quad \times (-1)^{|\mathbf{k}-\mathbf{k}'|} \partial^{\mathbf{k}-\mathbf{k}'} \boldsymbol{\varphi}(z_l) \prod_{j=l+1}^{p} \boldsymbol{\varphi}(z_j)\, dz_l \dots dz_p \\
&\quad \times \mathbf{H}_{\mathbf{k}-\mathbf{k}'}(z_l) \boldsymbol{\varphi}(z_l) \prod_{j=l+1}^{p} \boldsymbol{\varphi}(z_j)\, dz_l \dots dz_p \\
&= \gamma_l^{|\mathbf{k}'|/2} \frac{\sqrt{(\mathbf{k}-\mathbf{k}')!}}{\sqrt{\mathbf{k}!}} \mathsf{E}\left[\partial_{y_l}^{\mathbf{k}'} [f \circ G_{p,l}](x, \sqrt{\gamma_l} Z_l, \dots, \sqrt{\gamma_p} Z_p) \mathbf{H}_{\mathbf{k}-\mathbf{k}'}(Z_l) \right].
\end{aligned}
$$

The last expression yields the result. □

For multi-indices $\mathbf{k}, \mathbf{k}' \in \mathbb{N}_0^d$ with $\mathbf{k}' \le \mathbf{k}$ componentwise and $\mathbf{k}' \neq \mathbf{k}$, we get applying first Lemma 8

$$
\begin{aligned}
\bar{a}_{l,\mathbf{k}}(X_{l-1}) &= \left(\gamma_l^{|\mathbf{k}'|} \frac{(\mathbf{k}-\mathbf{k}')!}{\mathbf{k}!} \right)^{1/2} \sum_{p=l}^{N+n} \omega_{p,n}^N \mathsf{E}\left[\partial_{y_l}^{\mathbf{k}'} f(X_p) \mathbf{H}_{\mathbf{k}-\mathbf{k}'}(Z_l) \middle| X_{l-1} \right] \\
&= \left(\gamma_l^{|\mathbf{k}'|} \frac{(\mathbf{k}-\mathbf{k}')!}{\mathbf{k}!} \right)^{1/2} \\
&\quad \times \sum_{p=l}^{N+n} \omega_{p,n}^N \mathsf{E}\left[\left(\partial_{y_l}^{\mathbf{k}'} f(X_p) - \mathsf{E}\left[\partial_{y_l}^{\mathbf{k}'} f(X_p) | X_{l-1} \right] \right) \mathbf{H}_{\mathbf{k}-\mathbf{k}'}(Z_l) \middle| X_{l-1} \right].
\end{aligned}
$$

Assume that μ and f are $K \times d$ times continuously differentiable. Then, given $\mathbf{k} \in \mathbb{N}_0^d$, by taking $\mathbf{k}' = \mathbf{k}'(\mathbf{k}) = (K 1_{\{k_1 > K\}}, \dots, K 1_{\{k_d > K\}})$, for each $l \in \{N+1, \dots, N+n\}$, we get

$$
\begin{aligned}
\sum_{\mathbf{k}:\ \|\mathbf{k}\| \ge K+1} \left[\bar{a}_{l,\mathbf{k}}^2(X_{l-1}) | \mathcal{G}_N \right] &= \sum_{\mathbf{k}:\ \|\mathbf{k}\| \ge K+1} \left(\gamma_l^{|\mathbf{k}'|} \frac{(\mathbf{k}-\mathbf{k}')!}{\mathbf{k}!} \right) Q(\mathbf{k}', \mathbf{k}-\mathbf{k}') \qquad (29) \\
&= \sum_{I \subseteq \{1,\dots,d\},\, I \neq \emptyset} \gamma_l^{|I|K} \sum_{\mathbf{m}_I \in \mathbb{N}_I^d} \frac{\mathbf{m}_I!}{(\mathbf{m}_I + \mathbf{K}_I)!} \\
&\quad \times \sum_{\mathbf{m}_{I^c} \in \mathbb{N}_{0,I^c}^d,\, \|\mathbf{m}_{I^c}\| \le K} Q(\mathbf{K}_I, \mathbf{m}_I + \mathbf{m}_{I^c}),
\end{aligned}
$$

where for any two multi-indices $\mathbf{r}$, $\mathbf{q}$ from $\mathbb{N}_0^d$

$$Q(\mathbf{r},\mathbf{q}) = \mathsf{E}\left\{\left(\mathsf{E}\left[\sum_{p=l}^{N+n}\omega_{p,n}^N\left(\partial_{y_l}^{\mathbf{r}}f\left(X_p\right)-\mathsf{E}\left[\partial_{y_l}^{\mathbf{r}}f\left(X_p\right)|X_{l-1}\right]\right)\mathbf{H}_{\mathbf{q}}(Z_l)\middle|X_{l-1}\right]\right)^2\middle|\mathcal{G}_N\right\}.$$

In (29) the first sum runs over all nonempty subsets I of the set $\{1,\dots,d\}$. For any subset I, $\mathbb{N}_I^d$ stands for a set of multi-indices $\mathbf{m}_I$ with elements $m_i=0$, $i\notin I$, and $m_i\in\mathbb{N}$, $i\in I$. Moreover, $I^c=\{1,\dots,d\}\setminus I$ and $\mathbb{N}_{0,I^c}^d$ stands for a set of multi-indices $\mathbf{m}_{I^c}$ with elements $m_i=0$, $i\in I$, and $m_i\in\mathbb{N}_0$, $i\notin I$. Finally, the multi-index $\mathbf{K}_I$ is defined as $\mathbf{K}_I=(K1_{\{1\in I\}},\dots,K1_{\{d\in I\}})$. Applying the estimate

$$\frac{\mathbf{m}_I!}{(\mathbf{m}_I+\mathbf{K}_I)!}\le(1/2)^{|I|K},$$

we get

$$\begin{aligned}\sum_{\mathbf{k}:\ \|\mathbf{k}\|\ge K+1}\left[\bar a_{l,\mathbf{k}}^2(X_{l-1})|\mathcal{G}_N\right] &\le \sum_{I\subseteq\{1,\dots,d\},\,I\neq\emptyset}(\gamma_l/2)^{|I|K}\\ &\quad\times\sum_{\mathbf{m}_I\in\mathbb{N}_I^d}\sum_{\mathbf{m}_{I^c}\in\mathbb{N}_{0,I^c}^d,\,\|\mathbf{m}_{I^c}\|\le K}Q(\mathbf{K}_I,\mathbf{m}_I+\mathbf{m}_{I^c})\\ &\le \sum_{I\subseteq\{1,\dots,d\},\,I\neq\emptyset}(\gamma_l/2)^{|I|K}\sum_{\mathbf{m}\in\mathbb{N}_0^d}Q(\mathbf{K}_I,\mathbf{m}).\end{aligned}$$

Now using the consequence $\sum_{\mathbf{m}\in\mathbb{N}_0^d}\langle\xi,\mathbf{H}_{\mathbf{m}}(Z_l)\rangle^2\le\langle\xi,\xi\rangle$ of Parseval's identity (the latter is used conditionally on X_{l-1} which is possible because the system $\{\mathbf{H}_{\mathbf{m}}(Z_l)\}_{m\in\mathbb{N}_0^d}$ is orthonormal conditionally on X_{l-1}), we derive

$$\begin{aligned}\sum_{\mathbf{k}:\ \|\mathbf{k}\|\ge K+1}\left[\bar a_{l,\mathbf{k}}^2(X_{l-1})|\mathcal{G}_N\right] &\le \sum_{I\subseteq\{1,\dots,d\},\,I\neq\emptyset}\left(\frac{\gamma_l}{2}\right)^{|I|K}\\ &\quad\times\mathsf{E}\left[\mathsf{Var}\left(\sum_{p=l}^{N+n}\omega_{p,n}^N\partial_{y_l}^{\mathbf{K}_I}f\left(X_p\right)\middle|X_{l-1}\right)\middle|\mathcal{G}_N\right]\\ &= \frac{1}{\Gamma_{N+2,N+n+1}^2}\sum_{I\subseteq\{1,\dots,d\},\,I\neq\emptyset}\left(\frac{\gamma_l}{2}\right)^{|I|K}R_{l,n,N}^{I,K}\end{aligned}$$

with

$$R_{l,n,N}^{I,K}=\mathsf{E}\left[\mathsf{Var}\left(\sum_{p=l}^{N+n}\gamma_{p+1}\partial_{y_l}^{\mathbf{K}_I}f\left(X_p\right)\middle|X_{l-1}\right)\middle|\mathcal{G}_N\right].$$

As a result

$$\mathsf{Var}\left[\pi^N_{K,n}(f)\middle|\mathcal{G}_N\right] \le \frac{1}{\Gamma^2_{N+2,N+n+1}} \sum_{l=N+1}^{N+n} \sum_{I\subseteq\{1,\dots,d\},\, I\neq\emptyset} \left(\frac{\gamma_l}{2}\right)^{|I|K} R^{I,K}_{l,n,N}.$$

Next we show that under the conditions of Theorem 4, the quantity $R^{I,K}_{l,n,N}$ is uniformly bounded in l, n, N, I. For the sake of simplicity we present the proof only in one-dimensional case. First we need to prove several auxiliary results.

Lemma 9 *Let $(x_p)_{p\in\mathbb{N}_0}$ and $(\epsilon_p)_{p\in\mathbb{N}}$ be sequences of nonnegative real numbers satisfying $x_0 = \overline{C}_0$ and*

$$0 \le x_p \le \alpha_p x_{p-1} + \gamma_p \epsilon_p, \quad p \in \mathbb{N}, \tag{30}$$

$$0 \le \epsilon_p \le \overline{C}_1 \prod_{k=1}^{p} \alpha_k^2, \quad p \in \mathbb{N}, \tag{31}$$

where $\alpha_p, \gamma_p \in (0,1)$, $p \in \mathbb{N}$, and $\overline{C}_0, \overline{C}_1$ are some nonnegative constants. Assume

$$\sum_{r=1}^{\infty} \gamma_r \prod_{k=1}^{r} \alpha_k \le \overline{C}_2 \tag{32}$$

for some constant $\overline{C}_2$. Then

$$x_p \le (\overline{C}_0 + \overline{C}_1\overline{C}_2) \prod_{k=1}^{p} \alpha_k, \quad p \in \mathbb{N}.$$

Proof Applying (30) recursively, we get

$$x_p \le \overline{C}_0 \prod_{k=1}^{p} \alpha_k + \sum_{r=1}^{p} \gamma_r \epsilon_r \prod_{k=r+1}^{p} \alpha_k,$$

where we use the convention $\prod_{k=p+1}^{p} := 1$. Substituting estimate (31) into the right-hand side, we obtain

$$x_p \le \left(\overline{C}_0 + \overline{C}_1 \sum_{r=1}^{p} \gamma_r \prod_{k=1}^{r} \alpha_k\right) \prod_{k=1}^{p} \alpha_k,$$

which, together with (32), completes the proof. □

In what follows, we use the notation

$$\alpha_k = 1 - \gamma_k b_\mu, \quad k \in \mathbb{N}. \tag{33}$$

Remark 2 *Notice that, under (18), not only (32) but also*

$$\sum_{r=j}^{\infty} \gamma_r \prod_{k=j}^{r} \alpha_k \le \overline{C}_2 \tag{34}$$

is satisfied with the same constant $\overline{C}_2$ (which is C of (18)) simultaneously for all $j \in \mathbb{N}$. Below this will allow us to apply Lemma 9 to bound double indexed sequences $(x_{j,p})_{j\ge 1, p\ge j}$ satisfying

$$0 \le x_{j,p} \le \alpha_p x_{j,p-1} + \gamma_p \epsilon_{j,p}, \quad p \ge j+1,$$

with suitable $(\epsilon_{j,p})_{j\ge 1, p\ge j+1}$ and the constant $\overline{C}_2$ in (34) being independent of j.

Lemma 10 *Under assumptions of Theorem 4, for all natural $r \le K$ and $l \le p$, there exist constants C_r (not depending on l and p) such that*

$$\left|\partial_{y_l}^r X_p\right| \le C_r \prod_{k=l+1}^{p} \alpha_k \,, \tag{35}$$

where $\partial_{y_l}^r X_p$ is defined in (28). Moreover, we can choose $C_1 = 1$.

Proof The proof is along the same lines as Lemma 11. □

Lemma 11 *Under assumptions of Theorem 4, for all natural $r \le K$, $j \ge l$ and $p > j$, we have*

$$\left|\partial_{y_j}\partial_{y_l}^r X_p\right| \le c_r \prod_{k=l+1}^{p} \alpha_k \tag{36}$$

with some constants c_r not depending on j, l and p, while, for $p \le j$, it holds $\partial_{y_j}\partial_{y_l}^r X_p = 0$.

Proof The last statement is straightforward. We fix natural numbers $j \ge l$ and prove (36) for all $p > j$ by induction in r. First, for $p > j$, we write

$$\partial_{y_l} X_p = [1 - \gamma_p \mu'(X_{p-1})]\, \partial_{y_l} X_{p-1}$$

and differentiate this identity with respect to y_j

$$\partial_{y_j}\partial_{y_l} X_p = [1 - \gamma_p \mu'(X_{p-1})]\, \partial_{y_j}\partial_{y_l} X_{p-1} - \gamma_p \mu''(X_{p-1}) \partial_{y_j} X_{p-1} \partial_{y_l} X_{p-1}.$$

By Lemma 10, we have

$$\begin{aligned} |\partial_{y_j}\partial_{y_l} X_p| &\le \alpha_p |\partial_{y_j}\partial_{y_l} X_{p-1}| + \gamma_p B_\mu \prod_{k=l+1}^{p-1} \alpha_k \prod_{k=j+1}^{p-1} \alpha_k \\ &\le \alpha_p |\partial_{y_j}\partial_{y_l} X_{p-1}| + \gamma_p \mathrm{const} \prod_{k=l+1}^{j} \alpha_k \prod_{k=j+1}^{p} \alpha_k^2, \quad p \ge j+1, \end{aligned}$$

with a suitable constant (we can take, e.g., const $= \frac{B_\mu}{(1-\gamma_1 b_\mu)^2}$). By Lemma 9 together with Remark 2 applied to bound $|\partial_{y_j}\partial_{y_l} X_p|$ for $p \geq j+1$ (notice that $\partial_{y_j}\partial_{y_l} X_j = 0$, that is, $\overline{C}_0$ in Lemma 9 is zero, while $\overline{C}_1$ in Lemma 9 has the form const $\prod_{k=l+1}^{j} \alpha_k$), we obtain (36) for $r = 1$. The induction hypothesis is now that the inequality

$$\left|\partial_{y_j}\partial_{y_l}^k X_p\right| \leq c_k \prod_{s=l+1}^{p} \alpha_s \tag{37}$$

holds for all natural $k < r$ ($\leq K$) and $p > j$. We need to show (37) for $k = r$. Faà di Bruno's formula implies for $2 \leq r \leq K$ and $p > l$

$$\begin{aligned}\partial_{y_l}^r X_p &= \left[1 - \gamma_p \mu'(X_{p-1})\right] \partial_{y_l}^r X_{p-1} \\ &\quad - \gamma_p \sum \frac{r!}{m_1! \ldots m_{r-1}!} \mu^{(m_1+\ldots+m_{r-1})}(X_{p-1}) \prod_{k=1}^{r-1} \left(\frac{\partial_{y_l}^k X_{p-1}}{k!}\right)^{m_k},\end{aligned} \tag{38}$$

where the sum is taken over all $(r-1)$-tuples of nonnegative integers $(m_1, \ldots, m_{r-1})$ satisfying the constraint

$$1 \cdot m_1 + 2 \cdot m_2 + \ldots + (r-1) \cdot m_{r-1} = r. \tag{39}$$

Notice that we work with $(r-1)$-tuples rather than with r-tuples because the term containing $\partial_{y_l}^r X_{p-1}$ on the right-hand side of (38) is listed separately. For $p > j$, we then have

$$\begin{aligned}\partial_{y_j}\partial_{y_l}^r X_p &= \left[1 - \gamma_p \mu'(X_{p-1})\right] \partial_{y_j}\partial_{y_l}^r X_{p-1} - \gamma_p \mu''(X_{p-1}) \partial_{y_l}^r X_{p-1} \partial_{y_j} X_{p-1} \\ &\quad - \gamma_p \sum \frac{r!}{m_1! \ldots m_{r-1}!} \mu^{(m_1+\ldots+m_{r-1}+1)}(X_{p-1}) \partial_{y_j} X_{p-1} \prod_{k=1}^{r-1} \left(\frac{\partial_{y_l}^k X_{p-1}}{k!}\right)^{m_k} \\ &\quad - \gamma_p \sum \frac{r!}{m_1! \ldots m_{r-1}!} \mu^{(m_1+\ldots+m_{r-1})}(X_{p-1}) \partial_{y_j} \left[\prod_{k=1}^{r-1} \left(\frac{\partial_{y_l}^k X_{p-1}}{k!}\right)^{m_k}\right] \\ &= \left[1 - \gamma_p \mu'(X_{p-1})\right] \partial_{y_j}\partial_{y_l}^r X_{p-1} + \gamma_p \epsilon_{l,j,p},\end{aligned} \tag{40}$$

where the last equality defines the quantity $\epsilon_{l,j,p}$. Furthermore,

$$\begin{aligned}\partial_{y_j} \left[\prod_{k=1}^{r-1} \left(\frac{\partial_{y_l}^k X_{p-1}}{k!}\right)^{m_k}\right] &= \sum_{s=1}^{r-1} \frac{m_s}{s!} \left(\frac{\partial_{y_l}^s X_{p-1}}{s!}\right)^{m_s - 1} \partial_{y_j}\partial_{y_l}^s X_{p-1} \\ &\quad \times \prod_{k \leq r-1,\, k \neq s} \left(\frac{\partial_{y_l}^k X_{p-1}}{k!}\right)^{m_k}.\end{aligned}$$

Using Lemma 10, induction hypothesis (37) and the fact that $m_1 + \ldots + m_{r-1} \geq 2$ for

$(r-1)$-tuples of nonnegative integers satisfying (39), we can bound $|\epsilon_{l,j,p}|$ as follows

$$|\epsilon_{l,j,p}| \le B_\mu C_r \prod_{k=l+1}^{p-1} \alpha_k \prod_{k=j+1}^{p-1} \alpha_k + B_\mu \sum \frac{r!}{m_1! \dots m_{r-1}!} \left[\prod_{k=j+1}^{p-1} \alpha_k \right]$$
$$\times \prod_{s=1}^{r-1} \left(\frac{C_s \prod_{k=l+1}^{p-1} \alpha_k}{s!} \right)^{m_s}$$
$$+ B_\mu \sum \frac{r!}{m_1! \dots m_{r-1}!} \sum_{t=1}^{r-1} \frac{m_t}{t!} \left(\frac{C_t \prod_{k=l+1}^{p-1} \alpha_k}{t!} \right)^{m_t - 1} c_t \left[\prod_{k=l+1}^{p-1} \alpha_k \right]$$
$$\times \prod_{s \le r-1,\, s \ne t} \left(\frac{C_s \prod_{k=l+1}^{p-1} \alpha_k}{s!} \right)^{m_s} \le \text{const} \prod_{k=l+1}^{j} \alpha_k \prod_{k=j+1}^{p} \alpha_k^2$$

with some constant "const", which is, in fact, $\frac{1}{(1-\gamma_1 b_\mu)^2}$ times the expression involving $B_\mu, r, C_1, \dots, C_r, c_1, \dots, c_{r-1}$. Thus, (40) now implies

$$|\partial_{y_j} \partial_{y_l}^r X_p| \le \alpha_p |\partial_{y_j} \partial_{y_l}^r X_{p-1}| + \gamma_p \text{const} \prod_{k=l+1}^{j} \alpha_k \prod_{k=j+1}^{p} \alpha_k^2, \quad p \ge j+1.$$

We can again apply Lemma 9 and Remark 2 to bound $|\partial_{y_j} \partial_{y_l}^r X_p|$ for $p \ge j+1$ (notice that $\partial_{y_j} \partial_{y_l}^r X_j = 0$, that is, $\overline{C}_0$ in Lemma 9 is zero, while $\overline{C}_1$ in Lemma 9 has the form const $\prod_{k=l+1}^{j} \alpha_k$), and we obtain (37) for $k = r$. This concludes the proof. □

Lemma 12 *Under assumptions of Theorem 4, for all natural $l \le q$, it holds*

$$\mathsf{Var} \left[\sum_{p=l}^{q} \gamma_{p+1} \partial_{y_l}^K f(X_p) \,\middle|\, X_{l-1} \right] \le B_K \quad a.s.,$$

where B_K is a deterministic bound that does not depend on l and q.

Proof The expression $\sum_{p=l}^{q} \gamma_{p+1} \partial_{y_l}^K f(X_p)$ can be viewed as a deterministic function of $X_{l-1}, Z_l, Z_{l+1}, \dots, Z_q$

$$\sum_{p=l}^{q} \gamma_{p+1} \partial_{y_l}^K f(X_p) = F(X_{l-1}, Z_l, Z_{l+1}, \dots, Z_q).$$

By the (conditional) Gaussian Poincaré inequality, we have

$$\mathsf{Var} \left[\sum_{p=l}^{q} \gamma_{p+1} \partial_{y_l}^K f(X_p) \,\middle|\, X_{l-1} \right] \le \mathsf{E} \left[\|\nabla_Z F(X_{l-1}, Z_l, Z_{l+1}, \dots, Z_q)\|^2 \middle| X_{l-1} \right],$$

where $\nabla_Z F = (\partial_{Z_l} F, \ldots, \partial_{Z_q} F)$, and $\|\cdot\|$ denotes the Euclidean norm. Notice that $\partial_{Z_j} F = \sqrt{\gamma_j}\, \partial_{y_j} F$ and hence

$$\mathsf{Var}\left[\sum_{p=l}^{q} \gamma_{p+1} \partial_{y_l}^K f\left(X_p\right) \Big| X_{l-1}\right] \leq \sum_{j=l}^{q} \gamma_j \mathsf{E}\left[\left(\sum_{p=l}^{q} \gamma_{p+1} \partial_{y_j} \partial_{y_l}^K f\left(X_p\right)\right)^2 \Big| X_{l-1}\right].$$

It is straightforward to check that $\partial_{y_j} \partial_{y_l}^K f(X_p) = 0$ whenever $p < j$. Therefore, we get

$$\mathsf{Var}\left[\sum_{p=l}^{q} \gamma_{p+1} \partial_{y_l}^K f\left(X_p\right) \Big| X_{l-1}\right] \leq \sum_{j=l}^{q} \gamma_j \mathsf{E}\left[\left(\sum_{p=j}^{q} \gamma_{p+1} \partial_{y_j} \partial_{y_l}^K f\left(X_p\right)\right)^2 \Big| X_{l-1}\right]. \quad (41)$$

Now fix p and j, $p \geq j$, in $\{l, \ldots, q\}$. By Faà di Bruno's formula

$$\partial_{y_l}^K f\left(X_p\right) = \sum \frac{K!}{m_1! \ldots m_K!} f^{(m_1+\ldots+m_K)}(X_p) \prod_{k=1}^{K} \left(\frac{\partial_{y_l}^k X_p}{k!}\right)^{m_k},$$

where the sum is taken over all K-tuples of nonnegative integers $(m_1, \ldots, m_K)$ satisfying

$$1 \cdot m_1 + 2 \cdot m_2 + \ldots + K \cdot m_K = K.$$

Then

$$\begin{aligned}
\partial_{y_j} \partial_{y_l}^K f\left(X_p\right) &= \sum \frac{K!}{m_1! \ldots m_K!} f^{(m_1+\ldots+m_K+1)}(X_p) \left[\partial_{y_j} X_p\right] \prod_{k=1}^{K} \left(\frac{\partial_{y_l}^k X_p}{k!}\right)^{m_k} \\
&\quad + \sum \frac{K!}{m_1! \ldots m_K!} f^{(m_1+\ldots+m_K)}(X_p) \sum_{s=1}^{K} \frac{m_s}{s!} \left(\frac{\partial_{y_l}^s X_p}{s!}\right)^{m_s-1} \\
&\quad \times \left[\partial_{y_j} \partial_{y_l}^s X_p\right] \prod_{k \leq K,\, k \neq s} \left(\frac{\partial_{y_l}^k X_p}{k!}\right)^{m_k}.
\end{aligned}$$

Using the bounds of Lemmas 10 and 11, we obtain

$$\left|\partial_{y_j} \partial_{y_l}^K f\left(X_p\right)\right| \leq A_K \prod_{k=l+1}^{p} \alpha_k \quad (42)$$

with a suitable constant A_K. Substituting this in (41), we proceed as follows

$$\begin{aligned}
\mathsf{Var}\left[\sum_{p=l}^{q}\gamma_{p+1}\partial_{y_l}^K f\left(X_p\right)\Big|X_{l-1}\right] &\le A_K^2\sum_{j=l}^{q}\gamma_j\left(\sum_{p=j}^{q}\gamma_{p+1}\prod_{k=l+1}^{p}\alpha_k\right)^2\\
&\le \frac{A_K^2}{(1-\gamma_1 b_\mu)^2}\sum_{j=l}^{q}\gamma_j\left(\sum_{p=j+1}^{q+1}\gamma_p\prod_{k=l+1}^{p}\alpha_k\right)^2\\
&\le \frac{A_K^2}{(1-\gamma_1 b_\mu)^3}\sum_{j=l}^{q}\gamma_j\prod_{k=l}^{j}\alpha_k\left(\sum_{p=j+1}^{q+1}\gamma_p\prod_{k=j+1}^{p}\alpha_k\right)^2\\
&\le \frac{A_K^2}{(1-\gamma_1 b_\mu)^3}C^3 = B_K,
\end{aligned}$$

where C is the bound from (18). The proof is completed. □

Remark 3 *In fact in the proof of Lemma 12 using (42) we have shown that the expectations*

$$\mathsf{E}\left[\|\nabla_Z F(X_{l-1},Z_l,Z_{l+1},\ldots,Z_q)\|^r\big|\,X_{l-1}\right]$$

are uniformly bounded for all natural $l\le q$ and any $r\ge 2$. Indeed it follows from (18) that the quantities

$$\left(\sum_{j=l}^{q}\gamma_j\prod_{k=l}^{j}\alpha_k\right)^{r/2}\left(\sum_{p=j+1}^{q+1}\gamma_p\prod_{k=j+1}^{p}\alpha_k\right)^{r}$$

are bounded. This stronger result in combination with a L^r-Sobolev-type inequalities (see e.g. [1]) implies that also higher order conditional (on X_{l-1}) central moments of the r.v.

$$\sum_{p=l}^{q}\gamma_{p+1}\partial_{y_l}^K f\left(X_p\right)$$

are uniformly bounded in $l\le q$ and p.

8.2. Proof of Lemma 7

We have for any $\mathbf{k}\neq\mathbf{0}$

$$a_{p,l,\mathbf{k}}(x)=\mathsf{E}\left[\mathbf{H}_{\mathbf{k}}\left(Z_l\right)\left[Q_{p,l}\left(\Phi_l(x,Z_l)\right)-Q_{p,l}\left(\Phi_l(x,0)\right)\right]\right].$$

Hence

$$\|a_{p,l,\mathbf{k}}\|_\infty \le \sqrt{\gamma_l}\|\nabla Q_{p,l}\|_\infty.$$

Since $f, \mu \in C^1(\mathbb{R}^d)$, we have

$$\nabla Q_{p,l}(x) = \mathsf{E}\left[(I - \gamma_{l+1}\nabla\mu(x))\nabla Q_{p,l+1}(x - \gamma_{l+1}\mu(x) + \sqrt{\gamma_{l+1}}\xi)\right].$$

Hence

$$\|\nabla Q_{p,l}\|_\infty \le (1 - b_\mu\gamma_{l+1})\|\nabla Q_{p,l+1}\|_\infty$$

and

$$\begin{aligned}\|\nabla Q_{p,l}\|_\infty &\le \|\nabla f\|_\infty \prod_{r=l+1}^{p}(1 - b_\mu\gamma_r) \\ &\le \|\nabla f\|_\infty \exp\left(-b_\mu \sum_{r=l+1}^{p}\gamma_r\right).\end{aligned}$$

As a result

$$\|a_{p,l,\mathbf{k}}\|_\infty \le \sqrt{\gamma_l}\,\|\nabla f\|_\infty \exp\left(-b_\mu \sum_{r=l+1}^{p}\gamma_r\right).$$

Appendix A: Concentration bounds for ULA

For $\gamma > 0$, the Markov kernel R_γ from $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ into $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ is defined by

$$R_\gamma(x, A) = \int_A \frac{1}{(4\pi\gamma)^{d/2}} \exp\left\{-\frac{1}{4\gamma}\|y - x + \gamma\mu(x)\|^2\right\} dy, \quad A \in \mathcal{B}(\mathbb{R}^d),$$

while, for $k, l \in \mathbb{N}$, $k \le l$, the kernel $Q_{\boldsymbol{\gamma}}^{k,l}$ is $Q_{\boldsymbol{\gamma}}^{k,l} = R_{\gamma_l} \cdots R_{\gamma_k}$. We make the following assumptions.

(H1) Assume that ∇U is Lipschitz, that is, there exists $L \ge 0$ such that

$$\|\nabla U(x) - \nabla U(y)\| \le L\|x - y\|, \quad x, y \in \mathbb{R}^d$$

(H2) There exist $K > 0$, $M > 0$ and $m > 0$ such that for any $x \notin B_K(0)$ and any $y \in \mathbb{R}^d$ it holds

$$\langle D^2U(x)y, y\rangle \ge m\|y\|^2, \quad \|D^3U(y)\| \le M.$$

References

[1] Radosł aw Adamczak, Witold Bednorz, and PawełWolff. Moment estimates implied by modified log-Sobolev inequalities. *ESAIM Probab. Stat.*, 21:467–494, 2017.

[2] Jean-Yves Audibert and Olivier Catoni. Robust linear least squares regression. *Ann. Statist.*, 39(5):2766–2794, 2011.

[3] Denis Belomestny, Stefan Häfner, and Mikhail Urusov. Variance reduction for discretised diffusions via regression. *Journal of Mathematical Analysis and Applications*, 458:393–418, 2018.

[4] Tarik Ben Zineb and Emmanuel Gobet. Preliminary control variates to improve empirical regression methods. *Monte Carlo Methods Appl.*, 19(4):331–354, 2013.

[5] Nicolas Brosse, Alain Durmus, Sean Meyn, and Eric Moulines. Diffusion approximations and control variates for mcmc. *arXiv preprint arXiv:1808.01665*, 2018.

[6] P Robert Christian and George Casella. Monte carlo statistical methods, 1999.

[7] Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

[8] Petros Dellaportas and Ioannis Kontoyiannis. Control variates for estimation based on reversible markov chain monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):133–161, 2012.

[9] Ivan T Dimov. *Monte Carlo methods for applied scientists*. World Scientific, 2008.

[10] R Douc, E Moulines, P Priouret, and P Soulier. *Markov Chains*. Springer New York, 2018.

[11] Andrew B Duncan, Tony Lelievre, and GA Pavliotis. Variance reduction using nonreversible Langevin samplers. *Journal of statistical physics*, 163(3):457–491, 2016.

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

[13] Paul Glasserman. *Monte Carlo methods in financial engineering*, volume 53. Springer Science & Business Media, 2013.

[14] Emmanuel Gobet. *Monte-Carlo methods and stochastic processes*. CRC Press, Boca Raton, FL, 2016. From linear to non-linear.

[15] Shane G Henderson. *Variance reduction via an approximating Markov process*. PhD thesis, Stanford University, 1997.

[16] Vincent Lemaire. An adaptive scheme for the approximation of dissipative systems. *Stochastic Process. Appl.*, 117(10):1491–1518, 2007.

[17] K. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24:101–121, 1996.

[18] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[19] Antonietta Mira, Reza Solgi, and Daniele Imparato. Zero variance markov chain Monte Carlo for bayesian estimators. *Statistics and Computing*, 23(5):653–662, 2013.

[20] Gilles Pagès and Fabien Panloup. Ergodic approximation of the distribution of a stationary diffusion: rate of convergence. *Ann. Appl. Probab.*, 22(3):1059–1100, 2012.

[21] Gilles Pagès and Fabien Panloup. Weighted multilevel Langevin simulation of invariant measures. *Ann. Appl. Probab.*, 28(6):3358–3417, 2018.
[22] Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons, 2016.