Submitted to Bernoulli

Variance reduction for Markov chains via martingale representations

D. BELOMESTNY[1,3] and E. MOULINES[2,3] and S. SAMSONOV[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 additive functionals of Markov chains relying on a novel discrete time martingale representation. Our approach is fully non-asymptotic and does not require the knowledge of the stationary distribution (and even any type of ergodicity) or specific structure of the underlying density. By rigorously analyzing the convergence properties of the proposed algorithm, we show that its cost-to-variance relation is indeed smaller than one of the naive algorithm. The numerical performance of the new method is illustrated for Langevin-type Markov Chain Monte Carlo (MCMC) methods.

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

1. Introduction

Markov chains and Markov Chain Monte Carlo algorithms play crucial role in modern numerical analysis, finding various applications in such research areas as Bayesian inference, reinforcement learning, online learning. As an illustration, suppose that we aim at computing $\pi(f) := \mathsf{E}[f(X)]$, where X is a random vector in $\mathcal{X} \subseteq \mathbb{R}^d$ admitting a density π and f is a square-integrable function $f : \mathcal{X} \to \mathbb{R}$, $f \in L^2(\pi)$. Typically it is not possible to compute $\pi(f)$ analytically, leading to the Monte Carlo methods as a popular solution. Given an independent identically distributed observations $X_1, \dots, X_n$ from π, we might estimate $\pi(f)$ by $\hat{\pi}(f) := 1/n \sum_{k=1}^n f(X_k)$. The variance of such estimate equals $\sigma^2(f)/n$ with $\sigma^2(f)$ being the variance of the integrand. The first way to obtain tighter estimate $\hat{\pi}(f)$ is simply to increase the sample size n. Unfortunately, this solution might be prohibitively costly, especially when the dimension d is large enough or sampling from π is complicated. An alternative approach is to decrease $\sigma^2(f)$ by constructing a new Monte-Carlo experiment with the same expectation as the original one, but with a lower variance. Methods to achieve this are known as variance reduction techniques. Introduction to many of them can be found in [24], [13] and [12]. Recently one witnessed a revival of interest in efficient variance reduction methods for Markov chains, mostly with applications to MCMC algorithms; see for example [8], [20], [21], [25], [6] and references therein.

One of the popular approaches to variance reduction is the control variates method. We aim at constructing cheaply computable random variable ζ (control variate) 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 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 [9, Section 3.2]. 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 [1], [20], and Stein operator in [6]. 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.

Unfortunately it is not always possible to generate independent observations distributed according to π. To overcome this issue one might consider MCMC algorithms, where the exact samples from π are replaced by (X_p), $p = 0, 1, 2, \ldots$, forming a Markov chain with a marginal distribution of X_n converging to π in a suitable metrics as n grows. It is still possible to apply the control variates method in the similar manner to a simple Monte-Carlo case, yet the choice of the optimal control variate becomes much more difficult. Due to significant correlations between the elements of the Markov chain it might be not enough to minimize the marginal variances of $(X_p)_{p\geq 0}$ as it was in independent case. Instead one may choose the control variate by minimizing the corresponding asymptotic variance of the chain as it is suggested in [2]. At the same time it is possible to express the optimal control variate in terms of the solution of the Poisson equation for the corresponding Markov chain (X_p). As it was observed in [16, 17], for a time-homogeneous Markov chain X_p with a stationary distribution π, the function $U_G(x) := G(x) - \mathsf{E}[G(X_1)|X_0 = x]$ has zero mean with respect to π for an arbitrary real-valued function $G : \mathcal{X} \to \mathbb{R}$, such that $G \in L^1(\pi)$. Hence, $U_G(x)$ is a valid control functional for a suitable choice of G, with the best G given by a solution of the Poisson equation $\mathsf{E}[G(X_1)|X_0 = x] - G(x) = -f(x) + \pi(f)$. For such G we obtain $f(x) + U_G(x) = f(x) - f(x) + \pi(f) = \pi(f)$ leading to an ideal estimator with zero variance. Despite the fact that 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, [16] 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 [6] 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 [6] proposed another approach tailored to diffusion setting which does not 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, 11, 18, 23, 22] 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 its cost-to-variance relation.

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 5 we analyze the performance of the proposed variance reduction algorithm in the case of Unadjusted Langevin Algorithm (ULA). Finally, numerical examples are presented in Section 6.

Notations. We use the notations $\mathbb{N} = \{1, 2, \ldots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. We denote $\boldsymbol{\varphi}(z) = (2\pi)^{-d/2} \exp(-|z|^2/2)$, $z \in \mathbb{R}^d$ probability density function of the $d-$dimensional standard normal distribution. For $x \in \mathbb{R}^d$ and $r > 0$ let $B_r(x) = \{y \in \mathbb{R}^d | |y - x| < r\}$ where $|\cdot|$ is a standard Euclidean norm. For the twice differentiable function $g : \mathbb{R}^d \to \mathbb{R}$ we denote by $D^2 g(x)$ its Hessian at point x. For $m \in \mathbb{N}$, a smooth function $h \colon \mathbb{R}^{d \times m} \to \mathbb{R}$ with arguments being denoted $(y_1, \ldots, y_m)$, $y_i \in \mathbb{R}^d$, $i = 1, \ldots, m$, a multi-index $\mathbf{k} = (k_i) \in \mathbb{N}_0^d$, and $j \in \{1, \ldots, m\}$, we use the notation $\partial_{y_j}^{\mathbf{k}} h$ for the multiple derivative of h with respect to the components of y_j:

$$\partial_{y_j}^{\mathbf{k}} h(y_1, \ldots, y_m) := \partial_{y_j^d}^{k_d} \ldots \partial_{y_j^1}^{k_1} h(y_1, \ldots, y_m), \quad y_j = (y_j^1, \ldots, y_j^d).$$

In the particular case $m = 1$ we drop the subscript y_1 in that notation. For probability measures μ and ν on $\mathbb{R}^d$ denote by $\|\mu - \nu\|_{\mathrm{TV}}$ the total variation distance between μ and ν, that is,

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

where $\mathcal{B}(\mathbb{R}^d)$ is a Borel σ-algebra of $\mathbb{R}^d$. For a bounded Borel function $f : \mathbb{R}^d \to \mathbb{R}$ denote $\mathrm{osc}(f) := \sup_{x \in \mathbb{R}^d} f(x) - \inf_{x \in \mathbb{R}^d} f(x)$. Given a function $V : \mathbb{R}^d \to \mathbb{R}$, for a function $f : \mathbb{R}^d \to \mathbb{R}$ we define $\|f\|_V = \sup_{x \in \mathbb{R}^d} \{|f(x)|/V(x)\}$ and the corresponding $V-$norm between probability measures μ and ν on $\mathcal{B}(\mathbb{R}^d)$ as

$$d_V(\mu, \nu) = \|\mu - \nu\|_V = \sup_{\|f\|_V \le 1} \left[\int_{\mathbb{R}^d} f(x)\, d\mu(x) - \int_{\mathbb{R}^d} f(x)\, d\nu(x) \right] .$$

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 $(X_p)_{p\in\mathbb{N}_0}$ such that π is its unique invariant measure. In this paper we study a class of MCMC algorithms with $(X_p)_{p\in\mathbb{N}_0}$ satisfying the the following recurrence relation:

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

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 [10, Theorem 1.3.6]) and many well-known MCMC algorithms can be represented in the 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 $\mu = \nabla U$ for some continuously differentiable function U, then Markov chain (2) can be used to approximately sample from the density

$$\pi(x) = Z^{-1}\mathrm{e}^{-U(x)/2}, \quad Z = \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)/2}\,\mathrm{d}x, \tag{3}$$

provided that $Z < \infty$. This method is usually referred to as Unadjusted Langevin Algorithm (ULA). Denote by W the standard $\mathbb{R}^m$-valued Brownian motion. 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 [11].

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*

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

where

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

This transition is reversible with respect to π and therefore preserves the stationary density π; see [10, Chapter 2]. 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 [19]. The Metropolis-Adjusted Langevin algorithm (MALA) takes (2) as proposal, that is,

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

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

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

where $(U_p)_{p\ge1}$ is an i.i.d. sequence of uniformly distributed on $[0,1]$ random variables independent of $(Z_p)_{p\ge1}$. Thus we recover (1) with $\xi_p = (U_p, Z_p) \in \mathbb{R}^{d+1}$ and

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

Example 3 *Let $(X_t)_{t\ge0}$ be the unique strong solution to a SDE of the form:*

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \quad t \ge 0, \tag{4}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ 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 $(X_t)_{t\ge0}$ is a Markov process and let L denote its infinitesimal generator defined by

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

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. [23])

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

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. [20]) impossible in this case.

3. Martingale representation

In this section we provide a general discrete-time martingale representation for Markov chains of the type (1) which is used later 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 $[0, 1]$.

Let $(\xi_p)_{p\in\mathbb{N}}$ be i.i.d. $m-$dimensional random vectors. We denote by $(\mathcal{G}_p)_{p\in\mathbb{N}_0}$ the filtration generated by $(\xi_p)_{p\in\mathbb{N}}$ with the convention $\mathcal{G}_0 = \text{triv}$. For $x \in \mathbb{R}^d, y \in \mathbb{R}^m$ and $k \in \mathbb{N}$, let $\Phi_k(x, y)$ be a function mapping $\mathbb{R}^{d+m}$ to $\mathbb{R}^d$. Then we set for $l \leq p$

$$X_{l,p}^x := G_{l,p}(x, \xi_l, \ldots, \xi_p), \tag{5}$$

with the functions $G_{l,p} : \mathbb{R}^{d+m\times(p-l+1)} \to \mathbb{R}^d$ defined as

$$G_{l,p}(x, y_l, \ldots, y_p) := \Phi_p(\cdot, y_p) \circ \Phi_{p-1}(\cdot, y_{p-1}) \circ \cdots \circ \Phi_l(x, y_l). \tag{6}$$

Note that $\left(X_{0,p}^x\right)_{p\in\mathbb{N}_0}$ is a Markov chain with values in $\mathbb{R}^d$ of the form (1), starting at $X_0 = x$. In the sequel we write X_p^x and G_p as a shorthand notation for $X_{0,p}^x$ and $G_{0,p}$, respectively.

Theorem 4 *For all $p \in \mathbb{N}$, $q \le p$, $j < q \le p$, any Borel bounded functions $f : \mathbb{R}^d \to \mathbb{R}$ and all $x \in \mathbb{R}^d$ the following representation holds in $L^2(\mathsf{P})$*

$$f(X_q^x) = \mathsf{E}\left[f(X_q^x)\middle|\mathcal{G}_j\right] + \sum_{k=1}^{\infty}\sum_{l=j+1}^{q} a_{q,l,k}(X_{l-1}^x)\phi_k(\xi_l), \tag{7}$$

where $(X_p^x)_{p\ge 0}$ is given in (5) and for all $y \in \mathbb{R}^d$

$$a_{q,l,k}(y) = \mathsf{E}\left[f(X_{l-1,q}^y)\phi_k(\xi_l)\right], \quad q \ge l, \quad k \in \mathbb{N}. \tag{8}$$

Remark 5 *Denote by $\mathcal{F}_2^p$ class of Borel functions $f : \mathbb{R}^d \to \mathbb{R}$ such that for any $n \le p$, any indices $1 \le i_1 < \ldots < i_n \le p$ and $y \in \mathbb{R}^d$,*

$$\mathsf{E}\left[\left|f(\Phi_{i_n}(\cdot,\xi_{i_n}) \circ \Phi_{i_{n-1}}(\cdot,\xi_{i_{n-1}}) \circ \cdots \circ \Phi_{i_1}(y,\xi_{i_1}))\right|^2\right] < \infty \tag{9}$$

Then the statement of Theorem 4 remains valid for $f \in \mathcal{F}_2^p$.

If all the functions Φ_l, $l \ge 1$, in (5) are equal, then the condition (9) reduces to $\mathsf{E}\left[f^2(X_q^y)\right] < \infty$ for all $q \le p$ and $y \in \mathbb{R}^d$. Let us denote this class of functions by $\mathcal{F}_{2,\mathrm{hom}}^p$.

Corollary 6 *Let $(X_p^x)_{p\ge 0}$ be a homogeneous Markov chain of the form (5) with $\Phi_l = \Phi$, $l \ge 1$. Then for all $p \in \mathbb{N}$, $q \le p$, $j < q \le p$, $f \in \mathcal{F}_{2,hom}^p$ and $x \in \mathbb{R}^d$ it holds in $L^2(\mathsf{P})$,*

$$f(X_q^x) = \mathsf{E}\left[f(X_q^x)\middle|\mathcal{G}_j\right] + \sum_{k=1}^{\infty}\sum_{l=j+1}^{q} \bar{a}_{q-l,k}(X_{l-1}^x)\phi_k(\xi_l)$$

where for all $y \in \mathbb{R}^d$,

$$\bar{a}_{r,k}(y) = \mathsf{E}\left[f(X_r^y)\phi_k(\xi_1)\right] \quad r,k \in \mathbb{N} \tag{10}$$

Another equivalent representation of the coefficients $a_{p,l,k}$ turns out to be more useful to construct estimators:

Proposition 7 *Let $q \ge l, k \in \mathbb{N}$. Then the coefficients $a_{q,l,k}$ in (8) can be alternatively represented as*

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

with $Q_{q,l}(x) = \mathsf{E}\left[f(X_{l,q}^x)\right]$, $q \ge l$. In the homogeneous case $\Phi_l = \Phi$, the coefficients $\bar{a}_{r,k}$ in (10) are given respectively for all $l \ge 1$, by

$$\bar{a}_{r,k}(x) = \mathsf{E}\left[\phi_k(\xi)\, Q_r(\Phi(x,\xi))\right] \quad \text{with } Q_r(x) = \mathsf{E}\left[f(X_r^x)\right], \ r \in \mathbb{N}. \tag{11}$$

4. Variance reduction

Next we show how the representation (7) can be used to reduce the variance of MCMC algorithms. For the sake of clarity, in the sequel we consider only the time homogeneous case ($\Phi_l = \Phi$ for all $l \in \mathbb{N}$). Define

$$\pi_n^N(f) = \frac{1}{n} \sum_{p=N+1}^{N+n} f(X_p^x), \tag{12}$$

where $N \in \mathbb{N}_0$ is the length of the burn-in period and $n \in \mathbb{N}$ is the number of effective samples. Fix some $K \in \mathbb{N}$ and denote

$$\begin{aligned} M_{K,n}^N(f) &= \frac{1}{n} \sum_{p=N+1}^{N+n} \left[\sum_{k=1}^{K} \sum_{l=N+1}^{p} \bar{a}_{p-l,k}(X_{l-1}^x)\phi_k(\xi_l) \right] \\ &= \frac{1}{n} \sum_{k=1}^{K} \sum_{l=N+1}^{N+n} A_{N+n-l,k}(X_{l-1}^x)\phi_k(\xi_l), \end{aligned} \tag{13}$$

where

$$A_{q,k}(y) = \sum_{r=0}^{q} \bar{a}_{r,k}(y), \quad q = 0, \dots, n-1. \tag{14}$$

Since X_{l-1}^x is independent of ξ_l and $\mathsf{E}[\phi_k(\xi_l)] = 0$, $k \neq 0$, we obtain

$$\mathsf{E}[g(X_{l-1}^x)\phi_k(\xi_l)] = \mathsf{E}[g(X_{l-1}^x)\mathsf{E}[\phi_k(\xi_l) \,|\, \mathcal{G}_{l-1}]] = 0$$

for any function $g \in L^2(\pi)$. Hence the r.v. $M_{K,n}^N(f)$ has zero mean and can be viewed as a control variate. The representation (8) suggests also that the variance of the estimator

$$\pi_{K,n}^N(f) = \pi_n^N(f) - M_{K,n}^N(f) \tag{15}$$

should be small for K large enough. Indeed, since $\mathsf{E}[\phi_k(\xi_l)\phi_{k'}(\xi_l)] = 0$ if $k \neq k'$, we obtain

$$\mathsf{Var}[\pi_{K,n}^N(f)] = \frac{1}{n^2} \sum_{k=K+1}^{\infty} \sum_{l=1}^{n} \mathsf{E}[A_{n-l,k}^2(X_{N+l-1}^x)], \tag{16}$$

Hence $\mathsf{Var}[\pi_{K,n}^N(f)]$ is small provided that the coefficients $A_{s,k}$ decay fast enough as $k \to \infty$. In Section 5 we provide a detailed theoretical analysis of this decay for ULA (see Example 1).

The coefficients $(\bar{a}_{l,k})$ need to be estimated before one can apply the proposed variance reduction approach. One way to estimate them is to use nonparametric regression. We present a generic regression algorithm and then in Section 6 give further implementation

details. Our algorithm starts with estimating the functions Q_r for $r = 0, \dots, n-1$, defined in (11). We first generate T paths of the chain X (the so-called "training paths"):

$$\mathcal{D} = \left\{ (X_1^{(s)}, \dots, X_{N+n}^{(s)}), \quad s = 1, \dots, T \right\}$$

with $X_0^{(s)} = x$, $s = 1, \dots, T$. Then we solve the least squares optimization problems

$$\widehat{Q}_r \in \arg\min_{\psi \in \Psi} \sum_{s=1}^{T} \left| f(X_{r+N}^{(s)}) - \psi(X_N^{(s)}) \right|^2, \quad r = 0, \dots, n-1, \tag{17}$$

where Ψ is a class of real-valued functions on $\mathbb{R}^d$. We then set

$$\widehat{a}_{r,k}(x) = \int \widehat{Q}_r(\Phi(x,z))\phi_k(z) P_\xi(\mathrm{d}z), \quad \widehat{A}_{q,k}(x) = \min\left\{ W_k, \sum_{r=0}^{q} \widehat{a}_{r,k}(x) \right\}, \tag{18}$$

where (W_k) is a sequence of truncation levels. Finally, we construct the empirical estimate of $M_{K,n}^N(f)$ in the form

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

Obviously, $\mathsf{E}[\widehat{M}_{K,n}^N(f)|\mathcal{D}] = 0$ and $\widehat{M}_{K,n}^N(f)$ is a valid control variate in that it does not introduce any bias. By the Jensen inequality and orthonormality of $(\phi_k)_{k\geq 0}$,

$$\begin{aligned} &\mathsf{E}\left[\left| \widehat{M}_{K,n}^N(f) - M_{K,n}^N(f) \right|^2 \middle| \mathcal{D} \right] \\ &\qquad \leq \frac{1}{n^2} \sum_{k=1}^{K} \sum_{l=1}^{n} \mathsf{E}\left[|A_{n-l,k}(X_{N+l-1}^x) - \widehat{A}_{n-l,k}(X_{N+l-1}^x)|^2 \middle| \mathcal{D} \right]. \end{aligned}$$

Set

$$\widehat{\pi}_{K,n}^N(f) := \pi_n^N(f) - \widehat{M}_{K,n}^N(f).$$

Combining this with (16) results in

$$\begin{aligned} \mathsf{Var}[\widehat{\pi}_{K,n}^N(f)|\mathcal{D}] = \frac{1}{n^2} &\sum_{k=K+1}^{\infty} \sum_{l=1}^{n} \mathsf{E}[A_{n-l,k}^2(X_{N+l-1}^x)] \\ &+ \frac{1}{n^2} \sum_{k=1}^{K} \sum_{l=1}^{n} \mathsf{E}\left[|A_{n-l,k}(X_{N+l-1}^x) - \widehat{A}_{n-l,k}(X_{N+l-1}^x)|^2 \middle| \mathcal{D} \right]. \end{aligned} \tag{19}$$

In the next section we analyze the case of ULA algorithm and show that under some mild conditions (Lipschitz continuity and convexity outside a ball of the potential U),

$$\mathsf{E}[A_{l,k}^2(X_{N+l-1}^x)] \leq R_k^2, \quad k \in \mathbb{N}, \quad l = 0, \dots, n-1,$$

for a sequence (R_k) satisfying $\sum_{k=1}^{\infty} R_k^2 \le C$ with constant C not depending on γ and n (see (29)). Hence if we take $W_k = R_k$ as truncation parameters in (18), then $\widehat{A}_{n-l,k}$ converges to

$$\bar{A}_{q,k} := \arg\min_{\psi\in\Psi} \mathsf{E}[|A_{q,k}(X_N) - \psi(X_N)|^2]$$

as $T \to \infty$, see Section 11 in [14]. Moreover, it follows from (19) that

$$\mathsf{Var}[\widehat{\pi}^N_{K,n}(f)|\mathcal{D}] \le \frac{2C}{n}.$$

If the class of functions Ψ is a linear one of dimension $\mathtt{D}_\Psi$, then the cost of computing the coefficients $\widehat{A}_{l,k}$ for all $l = 1,\dots,n$, and $k = 1,\dots,K$, is of order $\mathtt{D}_\Psi K T^2 n$. Given $(\widehat{A}_{l,k})$ the cost of computing $\widehat{\pi}^N_{K,n}$ is proportional to $\mathtt{D}_\Psi K n$. At the same time the variance of the standard estimate $\pi^N_n(f)$ is of order $1/(n\gamma)$ and this bound can not be improved, see Lemma 8 and Remark 9. Thus, the ratio of the corresponding cost-to-variance characteristics is of order

$$\frac{\mathrm{cost}(\widehat{\pi}^N_{K,n})\mathsf{Var}[\widehat{\pi}^N_{K,n}(f)|\mathcal{D}]}{\mathrm{cost}(\pi^N_n)\mathsf{Var}[\pi^N_n(f)]} \le CK\gamma T^2 \mathtt{D}_\Psi. \tag{20}$$

Thus, for any fixed $K \ge 1$, our variance reduction method is advantageous if $\gamma \ll \min\{1/(\mathtt{D}_\Psi T^2), 1/\sqrt{n}\}$. Note that in order to achieve convergence of the corresponding MSE to zero, we need to let $\gamma \to 0$ as the invariant measure of the ULA with a constant time step γ is not equal to π.

5. Analysis of variance reduced ULA

In this section we perform the convergence analysis of the ULA algorithm. We use the notations of Example 1. For the sake of clarity and notational simplicity we restrict our attention to the constant time step, that is, we take $\gamma_k = \gamma$ for any $k \in \mathbb{N}$. 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 $\mathcal{G}_p = \sigma(Z_1,\dots,Z_p)$, $p \in \mathbb{N}$, and $\mathcal{G}_0 = \mathrm{triv}$. Given N and n as above, for $K \in \mathbb{N}$, denote

$$M^N_{K,n}(f) := \frac{1}{n} \sum_{0<\|\mathbf{k}\|\le K} \sum_{l=1}^{n} A_{n-l,\mathbf{k}}(X_{N+l-1}) \mathbf{H_k}(Z_{N+l})$$

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

$$A_{q,\mathbf{k}}(y) := \sum_{r=0}^{q} \bar{a}_{r,\mathbf{k}}(y). \tag{21}$$

For an estimator $\rho_n^N(f) \in \{\pi_n^N(f), \pi_{K,n}^N(f)\}$ of $\pi(f)$ (see (12) and (15)), we shall be interested in the Mean Squared Error (MSE), which can be decomposed as the sum of the squared bias and the variance:

$$\mathrm{MSE}\left[\rho_n^N(f)\right] = \mathsf{E}\left[\left\{\rho_n^N(f) - \pi(f)\right\}^2\right] = \left\{\mathsf{E}[\rho_n^N(f)] - \pi(f)\right\}^2 + \mathsf{Var}[\rho_n^N(f)]. \tag{22}$$

Our analysis is carried out under the following two assumptions.

(H1) [**Lipschitz continuity**] The potential U is differentiable and ∇U is Lipschitz continuous, that is, there exists $L < \infty$ such that

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

(H2) [**Convexity**] The potential U is of the form $U(x) = U_0(x) + \Delta(x)$, where Δ has compact support and $U_0 \in C^2(\mathbb{R}^d)$ with

$$\mathfrak{m}|x|^2 \le \langle D^2 U_0(x), x\rangle \le \mathfrak{M}|x|^2, \quad x \in \mathbb{R}$$

for some constants $\mathfrak{m}, \mathfrak{M} > 0$.

Let π be the probability measure on $\mathbb{R}^d$ with density $\pi(x)$ of the form (3); for $\gamma > 0$, define the Markov kernel Q_γ associated to one step of the ULA algorithm by

$$\mathrm{Q}_\gamma(x, A) = \int_A \frac{1}{(2\pi\gamma)^{d/2}} \exp\left\{-\frac{1}{2\gamma}|y - x + \gamma\nabla U(x)|^2\right\} \mathrm{d}y \tag{23}$$

for any $A \in \mathcal{B}(\mathbb{R}^d)$. Note that for a homogeneous Markov chain $\left(X_p^x\right)_{p\in\mathbb{N}_0}$ of the form (2) with constant step size γ, $\mathsf{P}(X_n^x \in A) = \mathrm{Q}_\gamma^n(x, A)$. Due to the martingale transform structure of $M_{K,n}^N(f)$, we have

$$\mathsf{E}\left[M_{K,n}^N(f)\right] = 0.$$

Hence both estimators $\pi_n^N(f)$ and $\pi_{K,n}^N(f)$ have the same bias. Under assumptions **(H1)** and **(H2)**, the corresponding Markov chain has a unique stationary distribution π_γ which is different from π. From [11, Theorem 10], it follows that there exist $\gamma_0 > 0$ and $C < \infty$ such that for all $\gamma \in (0, \gamma_0]$, we get

$$\|\pi - \pi_\gamma\|_{\mathrm{TV}} \le C\sqrt{\gamma} \tag{24}$$

Let us now derive an upper bound for the variance of the classical estimator (12) for ULA-based chain. Define

$$V(x) = 1 + |x|^2. \tag{25}$$

The following lemma follows from Lemma 63 in Appendix.

Lemma 8 *Assume* **(H1)** *and* **(H2)**. *Let f be a bounded Borel function and $(X_p^x)_{p\in\mathbb{N}_0}$ be a Markov chain generated by ULA with constant step size γ. Then, there exist $C<\infty$ and $\gamma_0>0$, such that for all $n\in\mathbb{N}$, $\gamma\in(0,\gamma_0]$ and $x\in\mathbb{R}^d$,*

$$\mathsf{Var}\left[\pi_n^N(f)\right] \le C\left(\frac{1}{n\gamma}+\left(\frac{V(x)+1}{n\gamma}\right)^2\right). \tag{26}$$

Remark 9 *The bound in Lemma 8 is sharp and cannot be improved even in the case of a normal distribution π. Indeed, ULA for the standard normal distribution takes form*

$$X_{k+1}=(1-2\gamma)X_k+\sqrt{\gamma}\xi_{k+1},\quad \xi_{k+1}\sim\mathcal{N}(0,1).$$

Thus, setting $X_0=0$, we obtain

$$\mathsf{Var}_0[\pi_n^0(f)] \;=\; \mathsf{Var}_0\left[\frac{1}{n}\sum_{i=1}^n X_i\right]=\frac{1}{n^2}\mathsf{Var}\left[\sum_{i=1}^n\sum_{j=1}^i\sqrt{\gamma}(1-2\gamma)^{i-j}\xi_j\right]\ge\frac{1}{n\gamma}-\frac{2(1-2\gamma)}{n^2\gamma^2}.$$

One of the main results of this paper is the following upper bound for the functions $(A_{q,\mathbf{k}})$ in (21).

Theorem 10 *Fix some $K\ge 1$ and suppose that a bounded function f and $\mu=\nabla U$ are $K\times d$ times continuously differentiable and for all $x\in\mathbb{R}^d$ and $\mathbf{k}$ satisfying $0<\|\mathbf{k}\|\le K$,*

$$|\partial^{\mathbf{k}}f(x)|\le B_f,\quad |\partial^{\mathbf{k}}\mu(x)|\le B_\mu\,. \tag{27}$$

Moreover assume that

$$\max_{j\ge N}\sum_{p=j+1}^{\infty}\gamma\left\|\prod_{k=j+1}^{p}(I-\gamma\mathsf{D}^2U(X_{k-1}^x))\right\|_r\le\Xi,\quad r=2,4, \tag{28}$$

for some $\Xi=\Xi(x)>0$ not depending on γ, where for any random matrix A we define $\|A\|_r=(\mathsf{E}|A|^r)^{1/r}$. Then, there exists a constant $C_K<\infty$ such that

$$A_{q,\mathbf{k}}^2(x)\le C_K\,\Xi(x)\,,\quad 1\le|\mathbf{k}|\le K,\quad q=0,\dots,n-1, \tag{29}$$

and

$$\sum_{\|\mathbf{k}\|\ge K+1}A_{q,\mathbf{k}}^2(x)\le C_K\sum_{I\subseteq\{1,\dots,d\},\,I\neq\emptyset}\left(\frac{\gamma}{2}\right)^{|I|K-1}\Xi(x)\,, \tag{30}$$

where the sum in (29) runs over all nonempty subsets I of the set $\{1,\dots,d\}$.

Corollary 11 *Suppose that* **(H2)** *holds with $\Delta\in C^2(\mathbb{R}^d)$ satisfying $|\mathsf{D}^2\Delta(x)|\le D$ for all $x\in\mathbb{R}^d$ and some constant $D>0$. Then for $\gamma<1/\mathfrak{M}$ we have*

$$\sum_{p=j+1}^{\infty}\gamma\left\|\prod_{k=j+1}^{p}(I-\gamma\mathsf{D}^2U(X_{k-1}^x))\right\|_r \;\le\; \frac{1}{\mathfrak{m}}+\frac{D}{\mathfrak{m}^2(1-\gamma\mathfrak{M})}.$$

Hence under assumptions of Theorem 10, there exists a constant $\widetilde{C}_K < \infty$ such that for all $n \in \mathbb{N}$ and $x \in \mathbb{R}^d$,

$$\mathsf{Var}\left[\pi^N_{K,n}(f)\right] \le \widetilde{C}_K n^{-1}\gamma^{K-1}\,. \tag{31}$$

Proof We have

$$\prod_{k=j+1}^{p}\left(I-\gamma \mathsf{D}^2 U(X_{k-1})\right) = \prod_{k=j+1}^{p}\left(I-\gamma \mathsf{D}^2 U_0(X_{k-1})\right) + \gamma\sum_{l=j+1}^{p}\prod_{k=j+1}^{l-1}\left(I-\gamma \mathsf{D}^2 U_0(X_{k-1})\right)\mathsf{D}^2\Delta(X_{l-1})\prod_{k=l+1}^{p}\left(I-\gamma \mathsf{D}^2 U_0(X_{k-1})\right),$$

where empty products are equal 1 by definition. Hence

$$\begin{aligned}\left|\prod_{k=j+1}^{p}\left(I-\gamma \mathsf{D}^2 U(X_{k-1})\right)\right| &\le \left(1+\frac{\gamma(p-j)D}{1-\gamma\mathfrak{M}}\right)\prod_{k=j+1}^{p}\left|I-\gamma \mathsf{D}^2 U_0(X_{k-1})\right| \\ &\le \left(1+\frac{\gamma(p-j)D}{1-\gamma\mathfrak{M}}\right)(1-\gamma\mathfrak{m})^{p-j}.\end{aligned}$$

□

Let us sketch the main steps of the proof. First using integration by parts, we prove that

$$A_{q,\mathbf{k}}(x) = \gamma^{|\mathbf{k}'|/2}\frac{\sqrt{(\mathbf{k}-\mathbf{k}')!}}{\sqrt{\mathbf{k}!}}\mathsf{E}\left[\partial^{\mathbf{k}'}_{Z_1}F(x,Z_1,\ldots,Z_q)\mathbf{H}_{\mathbf{k}-\mathbf{k}'}(Z_1)\right], \tag{32}$$

where $F_q(x,Z_1,\ldots,Z_q) := \sum_{r=0}^{q} f(X^x_r)$ and $\partial^{\mathbf{k}'}_{Z_1}$ stands for a weak partial derivative of the functional F_s that also can be viewed as discretised version of Malliavin derivative. Now by taking $\mathbf{k}' = \mathbf{k}$, we get

$$A^2_{q,\mathbf{k}}(x) \le \gamma^{|\mathbf{k}|}\mathsf{Var}\left(\sum_{p=1}^{q}\partial^{\mathbf{K}_I}_{Z_1} f\left(X^x_p\right)\right), \quad q=0,\ldots,n-1.$$

Also from (32) we can derive that

$$\sum_{\|\mathbf{k}\|\ge K+1} A^2_{q,\mathbf{k}}(x) \le \sum_{I\subseteq\{1,\ldots,d\},\, I\neq\emptyset}\left(\frac{\gamma}{2}\right)^{|I|K}\mathsf{Var}\left(\sum_{p=1}^{q}\partial^{\mathbf{K}_I}_{Z_1} f\left(X^x_p\right)\right),$$

where

$$\mathbf{K}_I = K(\mathbb{1}_I(1),\ldots,\mathbb{1}_I(d)).$$

Finally, using the Gaussian Poincare inequality, we show that under assumptions (27) and (28) it holds

$$\mathsf{Var}\left(\sum_{p=1}^{q}\partial_{Z_1}^{\mathbf{K}_I} f\left(X_p^x\right)\right) \le C_K\gamma^{-1}\Xi(x), \quad q=0,\dots,n-1, \tag{33}$$

for some positive constants $\varkappa, C_K$ not depending on n and γ.

6. Numerical analysis

In this section we illustrate the performance of the proposed variance reduction method for ULA. First we construct a polynomial approximation for each $Q_r(x)$ in the form:

$$\widehat{Q}_r(x) = \sum_{\|\mathbf{s}\|\le m}\widehat{\beta}_{\mathbf{s}}x^{\mathbf{s}}, \quad s=(s_1,\dots,s_d)\in\mathbb{N}_0^d.$$

The coefficients $\widehat{\beta}_{\mathbf{s}}\in\mathbb{R}$ are obtained using a modified least-squares criteria based on T independent training trajectories $\left\{\left(X_1^{(s)},\dots,X_{N+n}^{(s)}\right)\right\}_{s=1}^{T}$. More precisely we define $\widehat{Q}_0(x)=f(x)$ and for $1\le r\le n-1$, we set

$$\widehat{Q}_r = \operatorname*{arg\,min}_{\psi\in\Psi_m}\sum_{s=1}^{T}\left|f(X_{N+r}^{(s)})-\psi(X_N^{(s)})\right|^2, \tag{34}$$

where Ψ_m is a class of polynomials $\psi(x)=\sum_{\|\mathbf{s}\|\le m}\alpha_{\mathbf{s}}x^{\mathbf{s}}$ with $\alpha_{\mathbf{s}}\in\mathbb{R}$. 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 obtain closed-form expression for the estimates $\widehat{a}_{r,k}(x)$ of functions $\bar{a}_{r,k}(x)$ in (11). Namely, for all $x=(x^1,\dots,x^d)\in\mathbb{R}^d$,

$$\begin{aligned}\widehat{a}_{r,\mathbf{k}}(x) &= \int \mathbf{H}_{\mathbf{k}}(z)\widehat{Q}_r(x-\gamma\mu(x)+\sqrt{\gamma}z)P_\xi(\mathrm{d}z)\\ &= \sum_{\|\mathbf{s}\|\le m}\beta_{\mathbf{s}}\prod_{i=1}^{d}P_{k_i,s_i}(x^i-\gamma\mu_i(x)),\end{aligned} \tag{35}$$

where for any integer k,s $P_{k,s}$ is a one-dimensional polynomial of degree at most s with analytically known coefficients. We estimate Q_r only for $r<n_{\mathrm{trunc}}$ where the truncation level n_{trunc} may depend on d and γ. It allows us to use a smaller amount of training trajectories to approximate $Q_r(x)$. Finally we construct a truncated version of the estimator (15):

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

where

$$\widehat{M}^N_{K,n,n_{\text{trunc}}}(f) \quad = \quad \frac{1}{n}\sum_{p=N+1}^{N+n}\left[\sum_{0<\|\mathbf{k}\|\le K}\sum_{l=N+1}^{p}\widehat{a}_{p-l,\mathbf{k}}(X_{l-1})\mathbf{H}_{\mathbf{k}}(\xi_l)\mathbb{1}\{|p-l|<n_{\text{trunc}}\}\right].$$

6.1. Comparison with vanilla ULA

In this subsection we aim at comparing cost of variance reduction, achieved by the proposed algorithm, to the cost of vanilla ULA algorithm. We consider samples, generated by ULA with π being either the standard normal distribution in dimension d or the mixture of two $d-$dimensional standard Gaussians of the form

$$\pi(x) = \frac{1}{2\sqrt{(2\pi)^d}}\left(\mathrm{e}^{-(1/2)\|x-\mu\|^2} + \mathrm{e}^{-(1/2)\|x+\mu\|^2}\right)$$

where $d = 2$ and $\mu = (0.5, 0.5)$. For both examples, we aimed at estimating $\pi(f)$ with $f(x) = \sum_{i=1}^d x_i$ and $f(x) = \sum_{i=1}^d x_i^2$. We used constant step size $\gamma = 0.1$ and sampled $T = 5\times 10^4$ independent training trajectories, each one with the burn-in period $n_{\text{burn}} = 100$. Then we solve the least squares problems (34) using the first order polynomial approximations for $f(x) = \sum_{i=1}^d x_i$ and second order polynomial approximations for $f(x) = \sum_{i=1}^d x_i^2$ as described in the previous section. We set $K = 1$ or $K = 2$ for $f(x) = \sum_{i=1}^d x_i$ or $f(x) = \sum_{i=1}^d x_i^2$ respectively. Then we estimate the cost-variance ratio

$$\mathcal{R}(K,N,n,n_{\text{trunc}}) = \frac{\text{cost}(\pi_n^N)\mathsf{Var}[\pi_n^N(f)]}{\text{cost}(\widehat{\pi}^N_{K,n})\mathsf{Var}[\widehat{\pi}^N_{K,n,n_{\text{trunc}}}(f)|\mathcal{D}]} \tag{36}$$

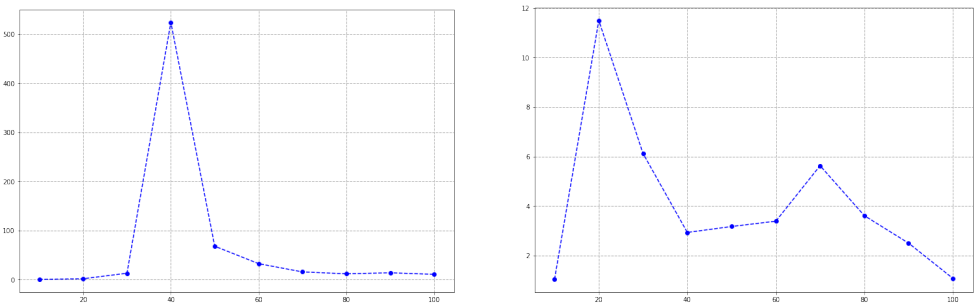
by its empirical counterpart, computed over 24 independent trajectories with $n = 2\times 10^3$ and $N = 2\times 10^3$. Since for fixed K the cost of computing $\pi_n^N(f)$ is proportional to computing function f, we set for $K = 1$

$$\text{cost}(\widehat{\pi}^N_{K,n}) = \text{cost}(\pi_n^N)\times n_{\text{trunc}}\times 2$$

since there are only 2 non-zero functions among $a_{r,\mathbf{k}}(x)$ for fixed r. Note that in this case each $a_{r,\mathbf{k}}(x)$ is a linear function, which can be computed at the same cost as f. Similarly, for $K = 2$ we set

$$\text{cost}(\widehat{\pi}^N_{K,n}) = \text{cost}(\pi_n^N)\times n_{\text{trunc}}\times 8$$

since there are at most 8 non-zero functions among $a_{r,\mathbf{k}}(x)$ for fixed r. Variance reduction costs for Gaussian potential and different n_{trunc} are summarized in Figure 1, and for the Gaussian mixture - in Figure 2. Note that for both examples the proposed algorithm allows us to obtain a sufficient gain in variance reduction cost.



(a) $f(x) = \sum_{i=1}^{d} x_i$ (b) $f(x) = \sum_{i=1}^{d} x_i^2$

Figure 1: Cost-variance ratios (36) as functions of the truncation level n_{trunc} for two-dimensional standard Gaussian distribution and different test functions.

6.2. Gaussian mixtures

We consider ULA with π given by the mixture of two equally-weighted d-dimensional Gaussian distributions of the following form

$$\pi(x) = \frac{1}{2\sqrt{(2\pi)^d |\Sigma|}} \left(\mathrm{e}^{-(1/2)(x-\mu)^T \Sigma^{-1}(x-\mu)} + \mathrm{e}^{-(1/2)(x+\mu)^T \Sigma^{-1}(x+\mu)} \right) \tag{37}$$

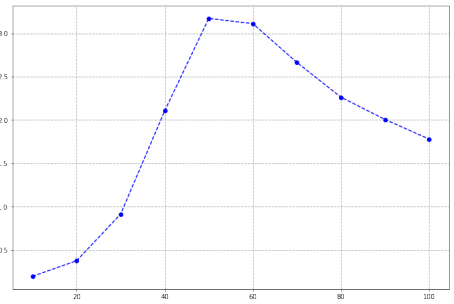
where $\mu \in \mathbb{R}^d$, Σ is a positive-definite $d \times d$ matrix and $|\Sigma|$ is its determinant. The function $U(x)$ and its gradient are given by

$$U(x) = \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) - \ln\left(1 + \mathrm{e}^{-2\mu^\top \Sigma^{-1} x}\right)$$

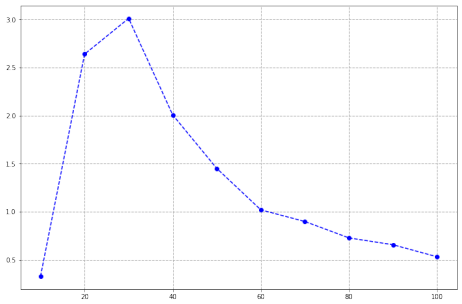
and

$$\nabla U(x) = \Sigma^{-1}(x-\mu) + 2\left(1 + \mathrm{e}^{2\mu^\top \Sigma^{-1} x}\right)^{-1} \Sigma^{-1}\mu,$$

respectively. In our experiments we considered $d = 2$, $\mu = (0.5, 0.5)$ and randomly generated positive-definite matrix Σ (heterogeneous structure). In order to approximate the expectation $\pi(f)$ with $f(x) = \sum_{i=1}^{d} x_i$ or $f(x) = \sum_{i=1}^{d} x_i^2$ we used constant step size $\gamma = 0.1$ and sampled $T = 5 \times 10^4$ independent training trajectories, each one of size $n = 50$ with the burn-in period $n_{\text{burn}} = 100$. Then we solve the least squares problems (34) using the first order polynomial approximations for $f(x) = \sum_{i=1}^{d} x_i$ and second order approximations for $f(x) = \sum_{i=1}^{d} x_i^2$ as described in the previous section. Hence we set $K = 1$ for $f(x) = \sum_{i=1}^{d} x_i$ or $K = 2$ for $f(x) = \sum_{i=1}^{d} x_i^2$. We set the truncation level $n_{\text{trunc}} = 50$. To test our variance reduction algorithm, we generated 100 independent trajectories of length $n = 2 \times 10^3$ and plot boxplots of the corresponding ergodic averages in Figure 3. Our approach (MDCV, Martingale decomposition control variates) is compared to other variance reduction methods of [20] and [2]. In the baselines we use first order polynomials in case of $K = 1$ and second order polynomials for $K = 2$.



(a) $f(x) = \sum_{i=1}^{d} x_i$



(b) $f(x) = \sum_{i=1}^{d} x_i^2$

Figure 2: Cost-variance ratios (36) as functions of the truncation level n_{trunc} for a mixture of two-dimensional Gaussian distributions and different test functions.

6.3. Banana shape distribution

The "Banana-shape" distribution, proposed by [15], can be obtained from a d-dimensional Gaussian vector with zero mean and covariance $\text{diag}(p, 1, \ldots, 1)$ by applying transformation $\varphi_b(x) : \mathbb{R}^d \to \mathbb{R}^d$ of the form

$$\varphi_b(x_1, \ldots, x_n) = (x_1, x_2 + bx_1^2 - pb, x_3, \ldots, x_d)$$

where $p > 0$ and $b > 0$ are parameters; b accounts for the curvature of density's level sets. The potential U is given by

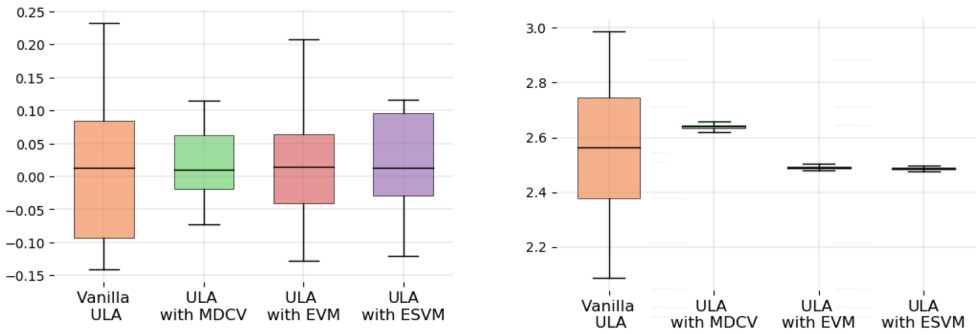
$$U(x_1, \ldots, x_d) = \frac{x_1^2}{2} + \frac{(x_2 + bx_1^2 - pb)^2}{2} + \frac{1}{2}\sum_{k=3}^{d} x_k^2 .$$

The quantity of interest is the expectation of $f(x) = x_2$. We set $p = 100, b = 0.1$ and consider $d = 8$. We solve the least squares problems (34) using the second-order polynomial approximations for the coefficients $\hat{a}_{p,\mathbf{k}}$ as described in the previous section, hence we set $K = 2$. We use $T = 5 \times 10^4$ independent training trajectories, each of size $n = 50$ with the burn-in $n_{\text{burn}} = 100$. We set the truncation level $n_{\text{trunc}} = 50$. To test our variance reduction algorithm, we generated 100 independent trajectories of length $n = 2 \times 10^3$ and display boxplots of the ergodic averages in Figure 4.

6.4. Binary Logistic Regression

In this section, we consider logistic regression. Let $\mathsf{Y} = (\mathsf{Y}_1, \ldots, \mathsf{Y}_n) \in \{0, 1\}^m$ be binary response variables, $\mathsf{X} \in \mathbb{R}^{m \times d}$ be a feature matrix and $\theta \in \mathbb{R}^d$ - vector of regression parameters. We define log-likelihood of i-th observation as

$$\ell(\mathsf{Y}_i|\theta, \mathsf{X}_i) = \mathsf{Y}_i \mathsf{X}_i^T \theta - \ln\left(1 + \mathrm{e}^{\mathsf{X}_i^T \theta}\right)$$



(a) $f(x) = \sum_{i=1}^{d} x_i$ (b) $f(x) = \sum_{i=1}^{d} x_i^2$

Figure 3: Boxplots of ergodic averages from the variance reduced ULA algorithms for the Gaussian mixture model. The compared estimators are the ordinary empirical average (Vanilla), our estimator of control variates (MDCV), zero variance estimator (ZV) and control variates obtained with empirical spectral variance minimisation (ESVM)

In order to estimate θ according to given data, the Bayesian approach introduces prior distribution $\pi_0(\theta)$ and consider the posterior density $\pi(\theta|\mathsf{Y},\mathsf{X})$ using Bayes' rule.

In the case of Gaussian prior $\pi_0(\theta) \sim \mathcal{N}(0, \sigma^2 I_d)$, the unnormalized posterior density takes the form:

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

Thus we obtain

$$\begin{aligned} U(\theta) &= -\mathsf{Y}^T\mathsf{X}\theta + \sum_{i=1}^{m} \ln\left(1+\mathrm{e}^{\theta^T\mathsf{X}_i}\right) + \frac{1}{2\sigma^2}\|\theta\|_2^2, \\ \nabla U(\theta) &= -\mathsf{Y}^T\mathsf{X} + \sum_{i=1}^{m} \frac{\mathsf{X}_i}{1+\mathrm{e}^{-\theta^T\mathsf{X}_i}} + \frac{1}{\sigma^2}\theta. \end{aligned}$$

To demonstrate the performance of the proposed control variates approach in the above Bayesian logistic regression model, we take a simple dataset from [20], 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 $n = 1 \times 10^4$, we take step size $\gamma = 0.1$ 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 $\hat{a}_{p-l,\mathbf{k}}$, based on $T = 10$ training trajectories. We put $n_{trunc} = 50$ and $K = 1$. 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 generate 100 independent test trajectories of length $n = 10^3$. In

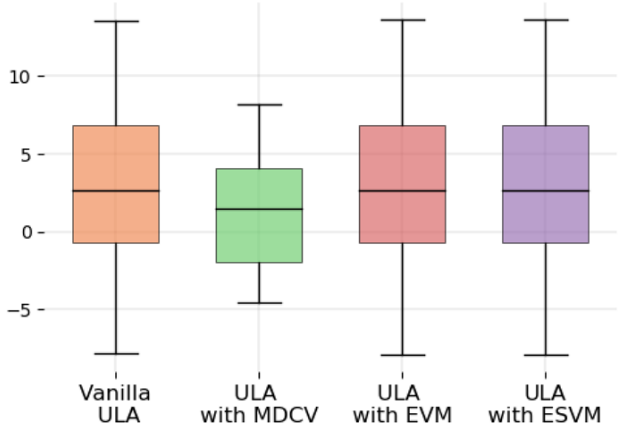


Figure 4: Boxplots of ergodic averages from the variance reduced ULA algorithms for the Banana shape density. The compared estimators are the ordinary empirical average (Vanilla), our estimator of control variates (MDCV), zero variance estimator (ZV) control variates obtained with empirical spectral variance minimisation (ESVM).

Figure 5 we compare our approach to the variance reduction methods of [20] and [2]. In the baselines we use first order polynomials for the same reasons as in the Gaussian mixtures example.

7. Proofs

7.1. Proof of Theorem 4

The expansion obviously holds for $p = q = 1$ and $j = 0$. Indeed, since $(\phi_k)_{k\geq 0}$ is a complete orthonormal system in $L^2(\mathbb{R}^d, P_\xi)$, it holds in $L^2(\mathsf{P})$ that

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

for any bounded f with $a_{1,1,k}(x) = \mathsf{E}[f(X_1^x)\phi_k(\xi_1)]$. Assume now that (7) holds for some $m < p$, any $q \leq m$, $j < q \leq m$ and all bounded f. Let us prove that the induction assumption holds for $q = m + 1$ and all $j < q$. The orthonormality and completeness of the system $(\phi_k)_{k=0}^{\infty}$ implies that for any bounded f and $y \in \mathbb{R}^d$, $\lim_{n\to\infty} \Psi_{n,m+1,1}(y) = 0$ with $\Psi_{n,m+1,1}(y) = \mathsf{E}[|f(X_{m,m+1}^y) - f_{n,m+1,1}(y)|^2]$ and

$$f_{n,m+1,1}(y) = \mathsf{E}[f(X_{m,m+1}^y)] + \sum_{k=1}^{n} \mathsf{E}[f(X_{m,m+1}^y)\phi_k(\xi_{m+1})]\phi_k(\xi_{m+1}).$$

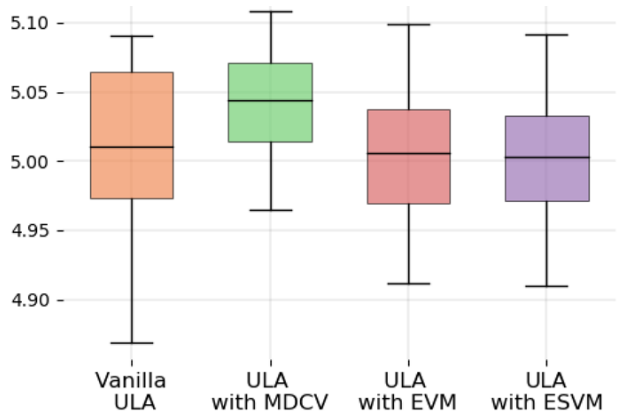


Figure 5: Boxplots of ergodic averages from the variance reduced ULA algorithms for the Logistic Regression. The compared estimators are the ordinary empirical average (Vanilla), our estimator of control variates (MDCV), zero variance estimator (ZV) control variates obtained with empirical spectral variance minimisation (ESVM).

Note that for any $y \in \mathbb{R}^d$ and $n \in \mathbb{N}$ it holds $\Psi_{n,m+1,1}(y) \leq \Psi_{0,m+1,1}(y)$ and

$$\begin{aligned} \mathsf{E}\left[\psi_{0,m+1,1}(X_m^x)\right] &= \mathsf{E}\left[\mathsf{E}\left[\psi_{0,m+1,1}(X_m^x)\middle|\,\mathcal{G}_m\right]\right] = \mathsf{E}\left[\mathsf{E}\left[\left|f(X_{m,m+1}^{X_m^x}) - \mathsf{E}f(X_{m,m+1}^{X_m^x})\right|^2\middle|\,X_m^x\right]\right] \\ &= \mathsf{E}\left[\left|f(X_{m+1}^x) - \mathsf{E}f(X_{m+1}^x)\right|^2\right] = \mathsf{Var}f(X_{m+1}^x) < \infty \end{aligned}$$

Hence, by Lebesgue dominated convergence theorem, $\lim_{n\to\infty} \mathsf{E}\left[\psi_{n,m+1,1}(X_m^x)\right] = 0$ and since for all $y \in \mathbb{R}^d$ the expectation $\mathsf{E}[f(X_{m,m+1}^y)]$ is a version of $\mathsf{E}\left[f(X_{m,m+1}^y)|\mathcal{G}_m\right]$, it holds in $L^2(\mathsf{P})$ that

$$f(X_{m+1}^x) = \mathsf{E}[f(X_{m+1}^x)|\mathcal{G}_m] + \sum_{k=1}^{\infty} a_{m+1,m+1,k}(X_m^x)\phi_k(\xi_{m+1}) \tag{38}$$

where

$$a_{m+1,m+1,k}(y) = \mathsf{E}\left[f(X_{m,m+1}^y)\phi_k\left(\xi_{m+1}\right)\right]$$

which is the required statement in the case $q = m+1$ and $j = m$. Now assume that $j < m$. Set $g(y) = \mathsf{E}\left[f(X_{m,m+1}^y)\right]$. Note that P-a.s. it holds $g(X_m^x) = \mathsf{E}\left[f(X_m^x)\middle|\,\mathcal{G}_m\right]$ and g is bounded by construction. Hence the induction hypothesis applies, and we get

$$\mathsf{E}\left[f(X_{m+1}^x)\middle|\,\mathcal{G}_m\right] = \mathsf{E}\left[f(X_{m+1}^x)\middle|\,\mathcal{G}_j\right] + \sum_{k\geq 1}\sum_{l=j+1}^{m} a_{m+1,l,k}(X_{l-1}^x)\phi_k(\xi_l) \tag{39}$$

with

$$a_{m+1,l,k}(X_{l-1}^x) = \mathsf{E}\left[\mathsf{E}\left[f(X_{m+1}^x)\middle|\,\mathcal{G}_m\right]\phi_k(\xi_l)\middle|\,\mathcal{G}_{l-1}\right] = \mathsf{E}\left[f(X_{m+1}^x)\phi_k(\xi_l)\middle|\,X_{l-1}^x\right].$$

where for $y \in \mathbb{R}^d$,

$$a_{m+1,l,k}(y) = \mathsf{E}\left[f(X^y_{l-1,m+1})\phi_k(\xi_l)\right]$$

Formulas (38) and (39) conclude the induction step for $q = m+1$ and all $j < q$ and hence the proof.

7.2. Proof of Lemma 8

In the sequel, the notation $A(\gamma, n, x) \lesssim B(\gamma, n, x)$ means that there exist $\gamma_0 > 0$, and $C < \infty$ such that for all $\gamma \in (0, \gamma_0]$, $x \in \mathbb{R}^d$, $n \in \mathbb{N}$, $A(\gamma, n, x) \le CB(\gamma, n, x)$.

Under assumptions **(H1)** and **(H2)** the Markov chain $(X^x_p)_{p\in\mathbb{N}_0}$ is $V-$geometrically ergodic in the sense of (57) with $V(x) = 1 + |x|^2$ and $\rho \in (0, 1)$ specified in Lemma 19. Due to Lemma 22,

$$\mathsf{Var}[\pi^N_n(f)] \lesssim \frac{1}{n(1-\rho^{1/2})} + \left(\frac{\rho^N V(x)+1}{n(1-\rho^{1/2})}\right)^2$$

It follows from (60) that $1 - \rho^{1/2} \ge C\gamma$ with a constant C not depending on γ, N and n, thus yielding an estimate

$$\mathsf{Var}[\pi^N_n(f)] \lesssim \frac{1}{n\gamma} + \left(\frac{\rho^N V(x)+1}{n\gamma}\right)^2.$$

7.3. Proof of Theorem 10

For $l \le p$ and $x \in \mathbb{R}^d$, we have the representation

$$X^x_p = G_p(x, \sqrt{\gamma}Z_1, \ldots, \sqrt{\gamma}Z_p),$$

where the function $G_p : \mathbb{R}^{d\times(p+1)} \to \mathbb{R}^d$ is defined as

$$G_p(x, y_1, \ldots, y_p) := \Phi(\cdot, y_p) \circ \Phi(\cdot, y_{p-1}) \circ \ldots \circ \Phi(x, y_1) \tag{40}$$

with, for $x, y \in \mathbb{R}^d$, $\Phi(x, y) = x - \gamma\mu(x) + y$. As a consequence, for a function $f\colon \mathbb{R}^d \to \mathbb{R}$ as in Section 2, we have

$$f(X_p) = f \circ G_p(X_0, \sqrt{\gamma}Z_1, \ldots, \sqrt{\gamma}Z_p).$$

In what follows, for $\mathbf{k} \in \mathbb{N}^d_0$, we use the shorthand notation

$$\partial^{\mathbf{k}}_1 f(X_p) := \partial^{\mathbf{k}}_1[f \circ G_p](X_0, \sqrt{\gamma}Z_1, \ldots, \sqrt{\gamma}Z_p) \tag{41}$$

whenever the function $f \circ G_p$ is smooth enough (that is, f and μ need to be smooth enough). Finally, for a multi-index $\mathbf{k} = (k_i) \in \mathbb{N}^d_0$, we use the notation $\mathbf{k}! := k_1! \cdot \ldots \cdot k_d!$

Lemma 12 *For any* $\mathbf{k}, \mathbf{k}' \in \mathbb{N}_0^d$ *such that* $\mathbf{k}' \le \mathbf{k}$ *componentwise and* $\|\mathbf{k}'\| \le K$, *the following representation holds*

$$\bar{a}_{p,\mathbf{k}}(x) = \left(\gamma^{|\mathbf{k}'|}\frac{(\mathbf{k}-\mathbf{k}')!}{\mathbf{k}!}\right)^{1/2} \mathsf{E}\left[\partial_1^{\mathbf{k}'} f(X_p^x)\mathbf{H}_{\mathbf{k}-\mathbf{k}'}(Z_1)\right],$$

where $\bar{a}_{p,\mathbf{k}}$ *is defined in* (11).

Proof Note 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 $l = p$)

$$\begin{aligned}
\bar{a}_{p,\mathbf{k}}(x) &= \int_{\mathbb{R}^d}\dots\int_{\mathbb{R}^d} f\circ G_p(x,\sqrt{\gamma}z_1,\dots,\sqrt{\gamma}z_p)\mathbf{H}_{\mathbf{k}}(z_1)\boldsymbol{\varphi}(z_1)\prod_{j=2}^{p}\boldsymbol{\varphi}(z_j)\,\mathrm{d}z_1\dots\mathrm{d}z_p\\
&= \frac{1}{\sqrt{\mathbf{k}!}}\int_{\mathbb{R}^d}\dots\int_{\mathbb{R}^d} f\circ G_p(x,\sqrt{\gamma}z_1,\dots,\sqrt{\gamma}z_p)(-1)^{|\mathbf{k}|}\partial^{\mathbf{k}}\boldsymbol{\varphi}(z_1)\prod_{j=2}^{p}\boldsymbol{\varphi}(z_j)\,\mathrm{d}z_1\dots\mathrm{d}z_p\\
&= \frac{\gamma^{|\mathbf{k}'|/2}}{\sqrt{\mathbf{k}!}}\int_{\mathbb{R}^d}\dots\int_{\mathbb{R}^d} \partial_1^{\mathbf{k}'}[f\circ G_p](x,\sqrt{\gamma}z_1,\dots,\sqrt{\gamma}z_p)(-1)^{|\mathbf{k}-\mathbf{k}'|}\partial^{\mathbf{k}-\mathbf{k}'}\boldsymbol{\varphi}(z_1)\prod_{j=2}^{p}\boldsymbol{\varphi}(z_j)\,\mathrm{d}z_1\dots\mathrm{d}z_p\\
&= \gamma^{|\mathbf{k}'|/2}\frac{\sqrt{(\mathbf{k}-\mathbf{k}')!}}{\sqrt{\mathbf{k}!}}\mathsf{E}\left[\partial_{y_1}^{\mathbf{k}'}[f\circ G_p](x,\sqrt{\gamma}Z_1,\dots,\sqrt{\gamma}Z_p)\mathbf{H}_{\mathbf{k}-\mathbf{k}'}(Z_1)\right].
\end{aligned}$$

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 $\mathbf{k}' \ne \mathbf{k}$, $\|k'\| \le K$, we get applying first Lemma 12,

$$A_{s,\mathbf{k}}(x) = \left(\gamma^{|\mathbf{k}'|}\frac{(\mathbf{k}-\mathbf{k}')!}{\mathbf{k}!}\right)^{1/2} \mathsf{E}\left[\sum\nolimits_{r=1}^{s}\{\partial_1^{\mathbf{k}'} f(X_r^x) - \mathsf{E}[\partial_1^{\mathbf{k}'} f(X_r^x)]\}\mathbf{H}_{\mathbf{k}-\mathbf{k}'}(Z_1)\right]$$

where $A_{s,\mathbf{k}}$ is defined in (14). Assume that μ and f are $K \times d$ times continuously differentiable. Then, given $\mathbf{k} \in \mathbb{N}_0^d$, by taking $\mathbf{k}' = \mathbf{k}'(\mathbf{k}) = K(\mathbb{1}_{\{k_1>K\}}\dots,\mathbb{1}_{\{k_d>K\}})$, we get

$$\begin{aligned}
\sum_{\mathbf{k}:\,\|\mathbf{k}\|\ge K+1} A_{s,\mathbf{k}}^2(x) &= \sum_{\mathbf{k}:\,\|\mathbf{k}\|\ge K+1}\left(\gamma^{|\mathbf{k}'|}\frac{(\mathbf{k}-\mathbf{k}')!}{\mathbf{k}!}\right)Q_s(\mathbf{k}',\mathbf{k}-\mathbf{k}')\\
&= \left\{\sum_{I\subseteq\{1,\dots,d\},\,I\ne\emptyset}\gamma^{|I|K}\sum_{\mathbf{m}_I\in\mathbb{N}_I^d}\frac{\mathbf{m}_I!}{(\mathbf{m}_I+\mathbf{K}_I)!}\right\}\left\{\sum_{\mathbf{m}_{I^c}\in\mathbb{N}_{0,I^c}^d,\,\|\mathbf{m}_{I^c}\|\le K} Q_s(\mathbf{K}_I,\mathbf{m}_I+\mathbf{m}_{I^c})\right\},
\end{aligned} \tag{42}$$

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

$$Q_s(\mathbf{r},\mathbf{q}) = \left\{ \mathsf{E}\left[\sum_{p=1}^{s} \left\{ \partial_1^{\mathbf{r}} f\left(X_p^x\right) - \mathsf{E}\left[\partial_1^{\mathbf{r}} f\left(X_p^x\right)\right] \right\} \mathbf{H}_{\mathbf{q}}(Z_1) \right] \right\}^2 .$$

In (42) 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

$$\begin{aligned} \sum_{\mathbf{k}:\ \|\mathbf{k}\|\ge K+1} A_{s,\mathbf{k}}^2(x) &\le \sum_{I\subseteq\{1,\dots,d\},\, I\neq\emptyset} (\gamma/2)^{|I|K} \\ &\quad \times \sum_{\mathbf{m}_I\in\mathbb{N}_I^d} \sum_{\mathbf{m}_{I^c}\in\mathbb{N}_{0,I^c}^d,\, \|\mathbf{m}_{I^c}\|\le K} Q(\mathbf{K}_I, \mathbf{m}_I+\mathbf{m}_{I^c}) \\ &\le \sum_{I\subseteq\{1,\dots,d\},\, I\neq\emptyset} (\gamma/2)^{|I|K} \sum_{\mathbf{m}\in\mathbb{N}_0^d} Q(\mathbf{K}_I,\mathbf{m}). \end{aligned} \tag{43}$$

The Parseval identity implies that for any function $\varphi : \mathbb{R}^d \to \mathbb{R}$ satisfying $\mathsf{E}[\varphi^2(Z_1)] < \infty$,

$$\sum_{\mathbf{m}\in\mathbb{N}_0^d} \{\mathsf{E}[\varphi(Z_1)\mathbf{H}_{\mathbf{m}}(Z_1)]\}^2 \le \mathsf{E}[\{\varphi(Z_1)\}^2]$$

Using this identity in (43) implies

$$\sum_{\mathbf{k}:\ \|\mathbf{k}\|\ge K+1} A_{s,\mathbf{k}}^2(x) \quad \le \quad \sum_{I\subseteq\{1,\dots,d\},\, I\neq\emptyset} \left(\frac{\gamma}{2}\right)^{|I|K} \mathsf{Var}\left(\sum_{p=1}^{s} \partial_1^{\mathbf{K}_I} f\left(X_p^x\right) \right)$$

Next we show that under the conditions of Theorem 10

$$\mathsf{Var}\left(\sum_{p=1}^{q} \partial_1^{\mathbf{K}_I} f\left(X_p^x\right) \right) \le C\gamma^{-1}\Xi(x), \quad q = 1,\dots,n,$$

for all x and some constants $C, \varkappa > 0$ not depending on n and γ. To keep the notational burden at a reasonable level, we present the proof only in one-dimensional case. Multi-dimensional extension is straightforward but requires involved notations. First, we need to prove several auxiliary results.

Lemma 13 *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*

$$0 \le x_p \le \alpha_p x_{p-1} + \gamma\epsilon_p, \quad p \in \mathbb{N}, \tag{44}$$

$$0 \le \epsilon_p \le \overline{C}_1 \prod_{k=1}^{p} \alpha_k^2, \quad p \in \mathbb{N}, \tag{45}$$

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

$$\gamma \sum_{r=1}^{\infty} \prod_{k=1}^{r} \alpha_k \le \overline{C}_2 \tag{46}$$

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 (44) recursively, we get

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

where we use the convention $\prod_{k=p+1}^{p} := 1$. Substituting estimate (45) into the right-hand side, we obtain

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

which, together with (46), completes the proof. □

In what follows, we use the notation

$$\alpha_k = 1 - \gamma\mu'(X_{k-1}^x), \quad k \in \mathbb{N}. \tag{47}$$

The assumption (27) implies that $|\mu'(x)| \le B_\mu$ for some constant $B_\mu > 0$ and all $x \in \mathbb{R}^d$. Without loss of generality we suppose that $\gamma B_\mu < 1$.

Lemma 14 *Under assumptions of Theorem 10, for all natural $r \le K$ and $l \le p$, there exist constants C_r (not depending on l and p) such that*

$$\left|\partial_{y_l}^r X_p^x\right| \le C_r \prod_{k=l+1}^{p} (1 - \gamma\mu'(X_{k-1}^x)) \quad a.s. \tag{48}$$

where $\partial_{y_l}^r X_p^x$ is defined in (41). Moreover, we can choose $C_1 = 1$.

Lemma 15 *Under assumptions of Theorem 10, for all natural $r \le K$, $j \ge l$ and $p > j$, we have*

$$\left|\partial_{y_j}\partial^r_{y_l} X^x_p\right| \le c_r \prod_{k=l+1}^{p} (1 - \gamma\mu'(X^x_{k-1})), \quad a.s. \tag{49}$$

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

Proof The last statement is straightforward. We fix natural numbers $j \ge l$ and prove (49) for all $p > j$ by induction in r. First, for $p > j$, we write

$$\partial_{y_l} X^x_p = \left[1 - \gamma\mu'(X^x_{p-1})\right] \partial_{y_l} X^x_{p-1}$$

and differentiate this identity with respect to y_j

$$\partial_{y_j}\partial_{y_l} X^x_p = \left[1 - \gamma\mu'(X^x_{p-1})\right] \partial_{y_j}\partial_{y_l} X^x_{p-1} - \gamma\mu''(X^x_{p-1})\partial_{y_j} X^x_{p-1}\partial_{y_l} X^x_{p-1}.$$

By Lemma 14, we have

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

with a suitable constant. By Lemma 13 applied to bound $|\partial_{y_j}\partial_{y_l} X^x_p|$ for $p \ge j+1$ (notice that $\partial_{y_j}\partial_{y_l} X^x_j = 0$, that is, $\overline{C}_0$ in Lemma 13 is zero, while $\overline{C}_1$ in Lemma 13 has the form const $\prod_{k=l+1}^{j} \alpha_k$), we obtain (49) for $r = 1$. The induction hypothesis is now that the inequality

$$\left|\partial_{y_j}\partial^k_{y_l} X^x_p\right| \le c_k \prod_{s=l+1}^{p} \alpha_s \tag{50}$$

holds for all natural $k < r$ ($\le K$) and $p > j$. We need to show (50) for $k = r$. Faà di Bruno's formula implies for $2 \le r \le K$ and $p > l$

$$\begin{aligned}
\partial^r_{y_l} X^x_p = &\left[1 - \gamma\mu'(X^x_{p-1})\right] \partial^r_{y_l} X^x_{p-1} \\
&- \gamma \sum \frac{r!}{m_1! \dots m_{r-1}!} \mu^{(m_1+\dots+m_{r-1})}(X^x_{p-1}) \prod_{k=1}^{r-1} \left(\frac{\partial^k_{y_l} X^x_{p-1}}{k!}\right)^{m_k},
\end{aligned} \tag{51}$$

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

$$1 \cdot m_1 + 2 \cdot m_2 + \dots + (r-1) \cdot m_{r-1} = r. \tag{52}$$

Notice that we work with $(r-1)$-tuples rather than with r-tuples because the term containing $\partial^r_{y_l} X^x_{p-1}$ on the right-hand side of (51) is listed separately. For $p > j$, we then have

$$\begin{aligned}\partial_{y_j}\partial^r_{y_l} X^x_p &= \left[1-\gamma_p\mu'(X^x_{p-1})\right]\partial_{y_j}\partial^r_{y_l} X^x_{p-1} - \gamma\mu''(X^x_{p-1})\partial^r_{y_l} X^x_{p-1}\partial_{y_j} X^x_{p-1} \\ &- \gamma\sum \frac{r!}{m_1!\ldots m_{r-1}!}\mu^{(m_1+\ldots+m_{r-1}+1)}(X^x_{p-1})\partial_{y_j} X^x_{p-1}\prod_{k=1}^{r-1}\left(\frac{\partial^k_{y_l} X^x_{p-1}}{k!}\right)^{m_k} \\ &- \gamma\sum \frac{r!}{m_1!\ldots m_{r-1}!}\mu^{(m_1+\ldots+m_{r-1})}(X^x_{p-1})\partial_{y_j}\left[\prod_{k=1}^{r-1}\left(\frac{\partial^k_{y_l} X^x_{p-1}}{k!}\right)^{m_k}\right] \\ &= \left[1-\gamma\mu'(X^x_{p-1})\right]\partial_{y_j}\partial^r_{y_l} X^x_{p-1} + \gamma\epsilon_{l,j,p},\end{aligned} \tag{53}$$

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

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

Using Lemma 14, induction hypothesis (50) and the fact that $m_1+\ldots+m_{r-1}\ge 2$ for $(r-1)$-tuples of nonnegative integers satisfying (52), we can bound $|\epsilon_{l,j,p}|$ as follows

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

with some constant "const" depending on $B_\mu, r, C_1,\ldots,C_r, c_1,\ldots,c_{r-1}$. Thus, (53) now implies

$$|\partial_{y_j}\partial^r_{y_l} X^x_p| \le \alpha_p|\partial_{y_j}\partial^r_{y_l} X^x_{p-1}| + \gamma\,\text{const}\prod_{k=l+1}^{j}\alpha_k\prod_{k=j+1}^{p}\alpha_k^2,\quad p\ge j+1.$$

We can again apply Lemma 13 to bound $|\partial_{y_j}\partial^r_{y_l} X^x_p|$ for $p\ge j+1$ (notice that $\partial_{y_j}\partial^r_{y_l} X^x_j = 0$, that is, $\overline{C}_0$ in Lemma 13 is zero, while $\overline{C}_1$ in Lemma 13 has the form const $\prod_{k=l+1}^{j}\alpha_k$), and we obtain (50) for $k=r$. This concludes the proof. □

Lemma 16 *Under assumptions of Theorem 10, it holds*

$$\mathsf{Var}\left[\sum_{p=1}^{q}\partial_{y_1}^K f\left(X_p^x\right)\right] \le C_K\gamma^{-1}\Xi(x), \quad q=1,\dots,n, \tag{54}$$

where C_K is a constant not depending on x, n and γ.

Proof The expression $\sum_{p=1}^{q}\partial_{y_1}^K f(X_p^x)$ can be viewed as a deterministic function of $x, Z_1, Z_2, \dots, Z_q$

$$\sum_{p=1}^{q}\partial_{y_1}^K f(X_p^x) = F(x, Z_1, Z_2, \dots, Z_q).$$

By the (conditional) Gaussian Poincaré inequality, we have

$$\mathsf{Var}\left[\sum_{p=1}^{q}\partial_{y_1}^K f\left(X_p^x\right)\right] \le \mathsf{E}_x\left[\|\nabla_Z F(x, Z_1, Z_2, \dots, Z_q)\|^2\right],$$

where $\nabla_Z F = (\partial_{Z_1}F, \dots, \partial_{Z_q}F)$, and $\|\cdot\|$ denotes the Euclidean norm. Notice that $\partial_{Z_j}F = \sqrt{\gamma}\,\partial_{y_j}F$ and hence

$$\mathsf{Var}\left[\sum_{p=1}^{q}\partial_{y_1}^K f\left(X_p^x\right)\right] \le \gamma^2\sum_{j=1}^{q}\mathsf{E}\left[\left(\sum_{p=1}^{q}\partial_{y_j}\partial_{y_1}^K f\left(X_p^x\right)\right)^2\right].$$

It is straightforward to check that $\partial_{y_j}\partial_{y_1}^K f(X_p^x) = 0$ whenever $p < j$. Therefore, we get

$$\mathsf{Var}\left[\sum_{p=1}^{q}\partial_{y_1}^K f\left(X_p^x\right)\right] \le \gamma^2\sum_{j=1}^{q}\mathsf{E}\left[\left(\sum_{p=j}^{q}\partial_{y_j}\partial_{y_1}^K f\left(X_p^x\right)\right)^2\right]. \tag{55}$$

Now fix p and j, $p \ge j$, in $\{1,\dots,q\}$. By Faà di Bruno's formula

$$\partial_{y_1}^K f\left(X_p^x\right) = \sum \frac{K!}{m_1!\dots m_K!} f^{(m_1+\dots+m_K)}(X_p^x)\prod_{k=1}^{K}\left(\frac{\partial_{y_1}^k X_p^x}{k!}\right)^{m_k},$$

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

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

Then

$$\begin{aligned}
\partial_{y_j}\partial_{y_1}^K f\left(X_p^x\right) &= \sum \frac{K!}{m_1!\dots m_K!} f^{(m_1+\dots+m_K+1)}(X_p^x)\left[\partial_{y_j}X_p^x\right]\prod_{k=1}^{K}\left(\frac{\partial_{y_1}^k X_p^x}{k!}\right)^{m_k} \\
&\quad + \sum \frac{K!}{m_1!\dots m_K!} f^{(m_1+\dots+m_K)}(X_p^x)\sum_{s=1}^{K}\frac{m_s}{s!}\left(\frac{\partial_{y_1}^s X_p^x}{s!}\right)^{m_s-1} \\
&\quad \times\left[\partial_{y_j}\partial_{y_1}^s X_p^x\right]\prod_{k\le K,\,k\ne s}\left(\frac{\partial_{y_1}^k X_p^x}{k!}\right)^{m_k}.
\end{aligned}$$

Using the bounds of Lemmas 14 and 15, we obtain

$$\left|\partial_{y_j}\partial_{y_1}^K f\left(X_p^x\right)\right| \le A_K \prod_{k=2}^{p} \alpha_k \tag{56}$$

with a suitable constant A_K. Substituting this in (55), we proceed as follows

$$\begin{aligned}
\mathsf{Var}_x\left[\sum_{p=1}^{q} \partial_{y_1}^K f\left(X_p^x\right)\right] &\le \gamma^2 A_K^2 \sum_{j=1}^{q} \mathsf{E}\left(\sum_{p=j}^{q}\prod_{k=2}^{p}\alpha_k\right)^2 \\
&\le \frac{\gamma^2 A_K^2}{(1-\gamma B_\mu)^2}\mathsf{E}\sum_{j=1}^{q}\left(\sum_{p=j+1}^{q+1}\prod_{k=2}^{p}\alpha_k\right)^2 \\
&\le \frac{\gamma^2 A_K^2}{(1-\gamma B_\mu)^3}\mathsf{E}\sum_{j=1}^{q}\prod_{k=1}^{j}\alpha_k\left(\sum_{p=j+1}^{q+1}\prod_{k=j+1}^{p}\alpha_k\right)^2
\end{aligned}$$

Now, from the Hölder inequality, we obtain (with $\|X\|_p = (\mathsf{E}X^p)^{\frac{1}{p}}$)

$$\begin{aligned}
\mathsf{E}\left[\sum_{j=1}^{q}\prod_{k=l}^{j}\alpha_k\left(\sum_{p=j+1}^{q+1}\prod_{k=j+1}^{p}\alpha_k\right)^2\right] &\le \sum_{j=1}^{q}\left\|\prod_{k=1}^{j}\alpha_k\right\|_2\left\|\sum_{p=j+1}^{q+1}\prod_{k=j+1}^{p}\alpha_k\right\|_4^2 \\
&\le \sum_{j=1}^{q}\left\|\prod_{k=1}^{j}\alpha_k\right\|_2\left(\sum_{p=j+1}^{q+1}\left\|\prod_{k=j+1}^{p}\alpha_k\right\|_4\right)^2 .
\end{aligned}$$

Now assumption (28) implies (54). □

Appendix A: Geometric ergodicity of the Unadjusted Langevin Algorithm

Definition 1 *Let $(X_p)_{p\in\mathbb{N}_0}$ be a Markov chain taking values in some space X with the Markov kernel Q and stationary distribution π. We say that $(X_p)_{p\in\mathbb{N}_0}$ is $V-$geometrically ergodic for a given function $V : X \to [1;+\infty)$ if there exist real numbers $C > 0$ and $0 < \rho < 1$ such that for any $n \in \mathbb{N}$,*

$$d_V(\delta_x Q^n, \pi) \le C\rho^n V(x) \tag{57}$$

In the sequel, without loss of generality, we assume that $(X_p)_{p\in\mathbb{N}_0}$ follows ULA dynamics (2) with $\nabla U(0) = 0$ and set

$$V(x) = 1 + |x|^2 . \tag{58}$$

Lemma 17 *Assume* **(H1)** *and* **(H2)**. *Then there exists* $K_2 \geq 0$ *such that for any* $|x| \geq K_2$ *it holds* $\langle \nabla U(x), x\rangle \geq (\mathfrak{m}/2)|x|^2$.

Proof Suppose that the function Δ in **(H2)** is supported in the ball $\{x : |x| \leq K_1\}$, then we have

$$\begin{aligned}\langle \nabla U(x), x\rangle &= \int_0^{K_1/|x|} D^2U(tx)[x^{\otimes 2}]\mathrm{d}t + \int_{K_1/|x|}^1 D^2U(tx)[x^{\otimes 2}]\mathrm{d}t \\ &\geq \mathfrak{m}|x|^2\{1 - K_1(1+L/\mathfrak{m})/|x|\}\,, \quad \text{for all} \quad |x| \geq K_1,\end{aligned}$$

which proves the statement with $K_2 = 2K_1(1 + L/\mathfrak{m})$. □

Lemma 18 *Assume* **(H1)** *and* **(H2)**. *Then, for any* $\gamma \in (0, \bar{\gamma}]$ *with* $\bar{\gamma} = \mathfrak{m}/2L^2$ *the kernel* Q_γ *from* (23) *satisfies drift condition*

$$Q_\gamma V(x) \leq \rho^\gamma V(x) + \gamma C \tag{59}$$

with $\rho = \exp\left(-\frac{\mathfrak{m}}{2}\right)$, $C = 2K_2^2(L+\mathfrak{m}) + d + \mathfrak{m}$ *with* K_2 *from Lemma 17, and* $\mathfrak{m}$ *from* **(H2)**.

Proof Note first that

$$Q_\gamma V(x) = 1 + |x - \gamma\nabla U(x)|^2 + \gamma d$$

Let us first consider the case $x \notin B(0, K_2)$. It remains to notice that, by Lemma 17, we get

$$|x - \gamma\nabla U(x)|^2 = |x|^2 - 2\gamma\langle \nabla U(x), x\rangle + \gamma^2|\nabla U(x)|^2 \leq \left(1 - \gamma\mathfrak{m} + \gamma^2L^2\right)|x|^2$$

Since $\gamma < \frac{\mathfrak{m}}{2L^2}$, we obtain

$$Q_\gamma V(x) \leq (1 - \gamma\mathfrak{m} + \gamma^2L^2)V(x) + \gamma(d + \mathfrak{m} - \gamma L^2) \leq \exp\Big(-\frac{\gamma\mathfrak{m}}{2}\Big)V(x) + \gamma(d+\mathfrak{m})$$

Consider now the case $x \in B(0, K_2)$. Then simply using $|x - \gamma\nabla U(x)|^2 \leq (1 + L\gamma)^2|x|^2$, we obtain

$$\begin{aligned}Q_\gamma V(x) &\leq (1 - \gamma\mathfrak{m} + \gamma^2L^2)V(x) + \gamma\left((\mathfrak{m} - \gamma L^2)(1 + |x|^2) + d + L(2 + L\gamma)|x|^2\right) \leq \\ &\leq \exp\Big(-\frac{\gamma\mathfrak{m}}{2}\Big)V(x) + \gamma(2K_2^2(L+\mathfrak{m}) + d + \mathfrak{m})\end{aligned}$$

□

It is known that under assumption **(H1)** the Markov chain generated by ULA with constant step size γ would have unique stationary distribution π_γ, which is different from π. Yet this chain will be $V-$geometrically ergodic due to [10, Theorem 19.4.1]. Namely, the following lemma holds:

Lemma 19 *Assume* **(H1)** *and* **(H2)**. *Then for* $0 < \gamma < \overline{\gamma} = \min(\mathfrak{m}/4L^2, 1/L)$, *and for any* $x \in \mathbb{R}^d$ *it holds*

$$d_V(\delta_x Q_\gamma^n, \pi_\gamma) \le C\rho^n \left(V(x) + \pi_\gamma(V)\right)$$

with $V(x) = 1 + |x|^2$ *and constants*

$$C = \left(1 + \exp\left(-\frac{\mathfrak{m}\gamma}{2}\right)\right)\left(1 + \frac{\overline{b}}{(1-\varepsilon)(1 - \exp\left(-\frac{\mathfrak{m}\gamma}{2}\right) - \frac{2b}{2+K^2})}\right),$$

$$b = \gamma(2K_2^2(L+\mathfrak{m}) + d + \mathfrak{m}), \quad \overline{b} = b\exp\left(-\frac{\mathfrak{m}\gamma}{2}\right) + K^2 + 1,$$

$$\varepsilon = 2\Phi\left(-K\sqrt{\left(1+\frac{1}{L}\right)(1+2L)}\right),$$

$$\rho = \exp\left\{-\frac{\gamma m}{4}\frac{\log(1-\varepsilon)}{\log(1-\varepsilon) + \log\left(\exp\left(-\frac{\mathfrak{m}\gamma}{2}\right) + \frac{2b}{K^2+2}\right)}\right\}, \tag{60}$$

where $K_2 > 0$ *is from Lemma 17 and* $K > 0$ *is a positive parameter such that*

$$K^2 > 4\frac{2K_2^2(L+m) + d + m}{m}.$$

Proof Note that the condition **(H1)** implies that for any $x, y \in \mathbb{R}^d$,

$$|x - y - \gamma(\nabla U(x) - \nabla U(y))|^2 \le (1 + 2L\gamma + \gamma^2)|x-y|^2$$

Hence, [5, Corollary 5] implies that

$$\|\delta_x Q_\gamma^{[1/\gamma]} - \delta_y Q_\gamma^{[1/\gamma]}\|_{\mathrm{TV}} \le 2(1-\varepsilon)$$

with $\varepsilon = 2\Phi\left(-K\sqrt{\left(1+\frac{1}{L}\right)(1+2L)}\right)$ for all x, y such that $|x|, |y| \le K$, which means that the kernel Q_γ satisfies $([1/\gamma], \varepsilon)$-Doeblin condition on the set $\{x : V(x) < 1 + K^2\}$. Together with drift condition (59) it allows to apply [10, Theorem 19.4.1] with appropriate constants. □

Appendix B: Covariance estimation for $V-$geometrically ergodic Markov chains

In this section we assume that $(X_p)_{p\in\mathbb{N}_0}$ is a $V-$geometrically ergodic Markov chain and prove bounds on the variance of ergodic average $\pi_n^N(f)$ of the form (12). We use the same technique as in [2] to control autocovariances for a given Markov chain. We start from some auxiliary lemmas:

Lemma 20 *Let $(X_p)_{p\geq 0}$ be a $V-$geometrically ergodic Markov chain with a stationary distribution π, and let f be a function with $\|f\|_{V^{\frac{1}{2}}} < \infty$. Set $\tilde{f}(x) = f(x) - \pi(f)$, then it holds for some constant $C > 0$,*

$$|\mathsf{E}_x[\tilde{f}(X_0)\tilde{f}(X_s)]| \leq 2C\rho^{s/2}\|\tilde{f}\|^2_{V^{\frac{1}{2}}} V(x), \tag{61}$$

$$|\mathsf{E}_\pi[\tilde{f}(X_0)\tilde{f}(X_s)]| \leq 2C\pi(V)\rho^{s/2}\|\tilde{f}\|^2_{V^{\frac{1}{2}}}. \tag{62}$$

Lemma 21 *Let $(X_p)_{p\in\mathbb{N}_0}$ be a $V-$geometrically ergodic Markov chain with a stationary distribution π, and let f be a function with $\|f\|_{V^{\frac{1}{2}}} < \infty$. Set $\tilde{f}(x) = f(x) - \pi(f)$, then*

$$\mathsf{cov}_x\left[f(X_k), f(X_{k+s})\right] \leq C^2V^2(x)\rho^{2k+s} + C^2\rho^{k+s/2}\|\tilde{f}\|_{V^{\frac{1}{2}}} V(x) + C\pi(V)\rho^{s/2}\|\tilde{f}\|_{V^{\frac{1}{2}}} \tag{63}$$

Proof Note that

$$\begin{aligned}\mathsf{cov}_x\left[f(X_k), f(X_{k+s})\right] &= \mathsf{E}_x\left[f(X_k) - \pi(f)\right]\left[f(X_{k+s}) - \pi(f)\right] \\ &\quad + \left[\pi(f) - \mathsf{E}_x f(X_{k+s})\right]\left[\mathsf{E}_x f(X_k) - \pi(f)\right].\end{aligned}$$

Due to $V-$ergodicity (57),

$$\left|\left[\pi(f) - \mathsf{E}_x f(X_{k+s})\right]\left[\mathsf{E}_x f(X_k) - \pi(f)\right]\right| \leq C^2V^2(x)\rho^{2k+s}.$$

To bound the first term note that

$$\begin{aligned}&\left|\mathsf{E}_x\left[f(X_k) - \pi(f)\right]\left[f(X_{k+s}) - \pi(f)\right] - \mathsf{cov}_\pi\left[f(X_k), f(X_{k+s})\right]\right| \\ &\qquad \leq \int_{\mathbb{R}^d} \left|\mathsf{E}_y\left[\tilde{f}(X_0)\tilde{f}(X_s)\right]\right| \left|\mathsf{P}^k(x, dy) - \pi(dy)\right| \\ &\qquad \leq 2C^2\rho^{s/2}\|\tilde{f}\|^2_{V^{\frac{1}{2}}} \|\mathsf{P}^k(x, dy) - \pi(dy)\|_V \leq 2C^2\rho^{k+s/2}\|\tilde{f}\|^2_{V^{\frac{1}{2}}} V(x),\end{aligned}$$

where we have used Lemma 20 and ergodicity of $(X_k)_{k\geq 0}$. This implies (63). □

Now we state and prove the main result of this section on the variance bound for the estimator $\pi^n_N(f)$ in the case of a $V-$geometrically ergodic Markov chain.

Lemma 22 *Let $(X_p)_{p\in\mathbb{N}_0}$ be a $V-$geometrically ergodic Markov chain with a stationary distribution π, and let f be a function with $\|f\|_{V^{\frac{1}{2}}} < \infty$. Then*

$$\mathsf{Var}_x\left[\pi^n_N(f)\right] \leq \frac{C_1\pi(V)\|\tilde{f}\|^2_{V^{1/2}}}{n(1-\rho^{1/2})} + C_2\left(\frac{V(x)+1}{n(1-\rho)}\right)^2 \tag{64}$$

with constants $C_1, C_2 > 0$ and $\rho \in (0, 1)$ from (57).

Proof Note that

$$\mathsf{Var}_x\left[\frac{1}{n}\sum_{k=N+1}^{N+n} f(X_k)\right] = \underbrace{\frac{1}{n^2}\sum_{k=N+1}^{N+n}\mathsf{Var}_x\left[f(X_k)\right]}_{S_1} + \underbrace{\frac{2}{n^2}\sum_{k=N+1}^{N+n-1}\sum_{s=1}^{n-k-1}\mathsf{cov}_x\left[f(X_k), f(X_{k+s})\right]}_{S_2}.$$

Now we bound each summand using Lemma 21:

$$S_1 \le \frac{1}{n}C\pi(V)\|\tilde{f}\|_{V^{1/2}} + \frac{\rho^N C^2\left(V^2(x)+V(x)\|\tilde{f}\|_{V^{1/2}}\right)}{n^2(1-\rho)},$$

$$\begin{aligned} S_2 &\le \frac{2}{n^2}\sum_{k=N+1}^{N+n-1}\left[\frac{C^2V^2(x)\rho^{2k+1}}{1-\rho} + \frac{2C^2V(x)\|\tilde{f}\|^2_{V^{1/2}}\rho^{k+1/2}}{1-\rho^{1/2}} + \frac{2C\|\tilde{f}\|^2_{V^{1/2}}\pi(V)}{1-\rho^{1/2}}\right] \le \\ &\le \frac{2C^2\rho^{2N}V^2(x)}{n^2(1-\rho)^2} + \frac{4C^2\rho^{N+1/2}\|\tilde{f}\|^2_{V^{1/2}}V(x)}{n^2(1-\rho)(1-\rho^{1/2})} + \frac{4C\|\tilde{f}\|^2_{V^{1/2}}\pi(V)}{n(1-\rho^{1/2})} \end{aligned}$$

Hence (64) follows. □

References

[1] Roland Assaraf and Michel Caffarel. Zero-variance principle for Monte Carlo algorithms. *Physical review letters*, 83(23):4682, 1999.

[2] D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, and S. Samsonov. Variance reduction for Markov chains with application to MCMC. *arXiv preprint arXiv:1910.03643*, 2019.

[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] Valentin De Bortoli and Alain Durmus. Convergence of diffusions and their discretizations: from continuous to discrete processes and back. *arXiv preprint arXiv:1904.09808*, 2019.

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

[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] A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.
[12] Paul Glasserman. *Monte Carlo methods in financial engineering*, volume 53. Springer Science & Business Media, 2013.
[13] Emmanuel Gobet. *Monte-Carlo methods and stochastic processes*. CRC Press, Boca Raton, FL, 2016. From linear to non-linear.
[14] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.
[15] Heikki Haario, Eero Saksman, and Johanna Tamminen. Adaptive proposal distribution for random walk metropolis algorithm. *Computational Statistics*, 14(3):375–396, 1999.
[16] Shane G Henderson. *Variance reduction via an approximating Markov process*. PhD thesis, Stanford University, 1997.
[17] Shane G. Henderson and Burt Simon. Adaptive simulation using perfect control variates. *J. Appl. Probab.*, 41(3):859–876, 09 2004.
[18] Vincent Lemaire. An adaptive scheme for the approximation of dissipative systems. *Stochastic Process. Appl.*, 117(10):1491–1518, 2007.
[19] K. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24:101–121, 1996.
[20] Antonietta Mira, Reza Solgi, and Daniele Imparato. Zero variance markov chain Monte Carlo for bayesian estimators. *Statistics and Computing*, 23(5):653–662, 2013.
[21] Chris J. Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages n/a–n/a, 2016.
[22] 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.
[23] Gilles Pagès and Fabien Panloup. Weighted multilevel Langevin simulation of invariant measures. *Ann. Appl. Probab.*, 28(6):3358–3417, 2018.
[24] Reuven Y Rubinstein and Dirk P Kroese. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons, 2016.
[25] L. F. South, C. J. Oates, A. Mira, and C. Drovandi. Regularised zero-variance control variates. *arXiv preprint arXiv:1811.05073*, 2018.