Submitted to Bernoulli

Variance reduction for MCMC methods 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 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 [22], [13] and [12]. Recently one witnessed a revival of interest in efficient variance reduction methods for MCMC algorithms, see for example [6], [19], [4] 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 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 [8]. 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 [3]. 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 Stein operator in [4]. 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 [2]. 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 Poisson equation for the corresponding Markov chain. It was observed in [16] 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 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 [10] 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, [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 [6] and [4] series-type control variates are introduced and studied for reversible Markov chains. It is assumed in [6] that the one-step conditional expectations can be computed explicitly for a set of basis functions. The authors in [4] 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 [5, 11, 17, 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 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 $\operatorname{osc}(f) := \sup_{x \in \mathbb{R}^d} f(x) - \inf_{x \in \mathbb{R}^d} f(x)$. Given a function $V : \mathbb{R}^d \to \mathbb{R}$, for a function $f : \mathbb{R}^d \to \mathbb{R}$ we define $\|f\|_V = \sup_{x \in \mathbb{R}^d} \frac{|f(x)|}{V(x)}$ 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 Theorem 1.3.6 in [9]) 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

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

provided that $\int_{\mathbb{R}^d} \mathrm{e}^{-\frac{U(x)}{2}}\, \mathrm{d}x$ 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 [5] 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*

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

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 [9, 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 [18]. 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

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

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 $(X_t)_{t\in[0,T]}$ 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 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 $(X_t)_{t\ge0}$ 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. [21])

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

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.*

3. Martingale representation

In this section we give 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 1 *For all* $p \in \mathbb{N}$, $q \leq p$, $j < q \leq p$, *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 $(X_p^x)_{p\geq 0}$ *is given in* (5) *and for all* $y \in \mathbb{R}^d$

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

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

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

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

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

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

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

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

$$\pi_n^N(f) = \frac{1}{n}\sum_{p=N+1}^{N+n} f(X_p^x), \tag{10}$$

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

$$\begin{aligned} M_{K,n}^N(f) &= \frac{1}{n}\sum_{p=N+1}^{N+n}\left[\sum_{k=1}^{K}\sum_{l=N+1}^{p} \bar{a}_{p-l,k}(X_{l-1}^x)\phi_k(\xi_l)\right] \\ &= \frac{1}{n}\sum_{k=1}^{K}\sum_{l=N+1}^{N+n}\sum_{p=0}^{N+n-l} \bar{a}_{p,k}(X_{l-1}^x)\phi_k(\xi_l). \end{aligned} \tag{11}$$

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

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

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

$$\pi^N_{K,n}(f) = \pi^N_n(f) - M^N_{K,n}(f) \tag{12}$$

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

$$\mathsf{Var}[\pi^N_{K,n}(f)] = \frac{1}{n^2} \sum_{k=K+1}^{\infty} \sum_{l=1}^{n} \mathsf{E}[A^2_{n-l,k}(X^x_{N+l-1})], \tag{13}$$

where

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

Hence $\mathsf{Var}[\pi^N_{K,n}(f)]$ 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 first present a generic regression algorithm and then in Section 6 give further implementation details. Our algorithm starts with estimating the functions $A_{q,k}$ for $q = 0, \dots, n-1$, $k = 1, \dots, K$. We first generate T paths of the chain X (the so-called "training paths"):

$$\mathcal{D} = \left\{ (X^{(s)}_1, \dots, X^{(s)}_{N+n}), \quad s = 1, \dots, T \right\}$$

with $X^{(s)}_0 = x$, $s = 1, \dots, T$. Then we solve the least squares optimization problems

$$\widetilde{A}_{q,k} \in \arg\min_{\psi \in \Psi} \sum_{s=1}^{T} \left| \sum_{r=0}^{q} f(X^{(s)}_{r+l_0}) \phi_k(\xi^{(s)}_{l_0+1}) - \psi(X^{(s)}_{l_0}) \right|^2 \tag{15}$$

for $k = 1, \dots, K$, and some l_0 with $l_0 + q \leq N + n$, where Ψ is a class of real-valued functions on $\mathbb{R}^d$ and $\xi^{(s)}_1, \dots, \xi^{(s)}_{N+n}$ are iid random variables with distribution P_ξ used to construct the s-th trajectory of X. As usual in regression analysis (see Section 11 in [14]), we truncate the least squares estimates $(\widetilde{A}_{q,k})$ and set

$$\widehat{A}_{q,k}(y) := \begin{cases} \widetilde{A}_{q,k}(y), & |\widetilde{A}_{q,k}(y)| \leq W_k, \\ W_k \operatorname{sign}(\widetilde{A}_{q,k}(y)), & \text{otherwise.} \end{cases} \tag{16}$$

for a sequence of positive real numbers W_k, $k = 1, \dots, K$. Upon estimating the functions $(A_{q,k})$, one can construct the empirical estimate of $M^N_{K,n}(f)$ in the form

$$\widehat{M}^N_{K,n}(f) := \frac{1}{n} \sum_{k=1}^{K} \sum_{l=N+1}^{N+n} \left[\widehat{A}_{N+n-l,k}(X^x_{l-1}) \phi_k(\xi_l) \right].$$

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

$$\mathsf{E}\left[\left|\widehat{M}^N_{K,n}(f) - M^N_{K,n}(f)\right|^2 \middle| \mathcal{D}\right] \leq \frac{1}{n^2}\sum_{k=1}^{K}\sum_{l=1}^{n}\mathsf{E}\left[|A_{n-l,k}(X^x_{N+l-1}) - \widehat{A}_{n-l,k}(X^x_{N+l-1})|^2 \middle| \mathcal{D}\right].$$

Set

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

Combining this with (13) results in

$$\mathsf{Var}[\widehat{\pi}^N_{K,n}(f)|\mathcal{D}] = \frac{1}{n^2}\sum_{k=K+1}^{\infty}\sum_{l=1}^{n}\mathsf{E}[A^2_{n-l,k}(X^x_{N+l-1})] + \frac{1}{n^2}\sum_{k=1}^{K}\sum_{l=1}^{n}\mathsf{E}\left[|A_{n-l,k}(X^x_{N+l-1}) - \widehat{A}_{n-l,k}(X^x_{N+l-1})|^2 \middle| \mathcal{D}\right]. \quad (17)$$

In the next section we analyze the case of ULA and show that under some rather mild conditions (convexity of the potential U outside a ball around zero)

$$\mathsf{E}[A^2_{l,k}(X^x_{N+l-1})] \leq e^{\varkappa n\gamma^2}R^2_k, \quad k = 1, 2, \ldots, \quad l = 0, \ldots, n-1,$$

for a sequence (R_k) satisfying $\sum_{k=1}^{\infty} R^2_k \leq C$ with constant C not depending on γ and n (see (23)). Hence if we take $W_k = R_k$ as truncation parameters in (16), then $\widehat{A}_{n-l,k}$ converges to

$$\bar{A}_{q,k} := \arg\min_{\psi\in\Psi}\mathsf{E}[|A_{q,k}(X_{l_0}) - \psi(X_{l_0})|]^2$$

as $T \to \infty$, see Section 11 in [14]. Moreover, it follows from (17) that

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

Note that the above estimate is rather rough, as we ignore here the fact that the difference between $\widehat{A}_{n-l,k}$ and $A_{n-l,k}$ is small. 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 $l = 1, \ldots, n$ and $k = 1, \ldots, K$ is of order $\mathtt{D}_\Psi KT^2n$. Given $(\widehat{A}_{l,k})$ the cost of computing $\widehat{\pi}^N_{K,n}$ is proportional to $\mathtt{D}_\Psi Kn$. 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 2 and Remark 2. Thus, the ratio of the corresponding cost-variances is of order

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

Thus, for any fixed $K \geq 1$, 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 $\mathcal{G}_p = \sigma(Z_1, \ldots, Z_p)$, $p \in \mathbb{N}$, and $\mathcal{G}_0 = \mathrm{triv}$. Given N and n as above, for $K \in \mathbb{N}$, denote

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

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

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

For an estimator $\rho(f) \in \{\pi_n^N(f), \pi_{K,n}^N(f)\}$ of $\pi(f)$ (see (10) and (12)), we shall be interested in the Mean Squared Error (MSE), which can be decomposed as the sum of the squared bias and the variance:

$$\mathrm{MSE}\left[\rho(f)\right] = \mathsf{E}\left[\{\rho(f) - \pi(f)\}^2\right] = \{\mathsf{E}[\rho(f)] - \pi(f)\}^2 + \mathsf{Var}[\rho(f)]. \tag{18}$$

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

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

(H2) [**Convexity outside a ball**] There exist $K > 0$ and $m > 0$ such that for any $x \notin B_K(0)$ it holds

$$\langle D^2U(x), x\rangle \geq m\|x\|^2.$$

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

$$\mathrm{Q}_\gamma(x, A) = \int_A \frac{1}{(2\pi\gamma)^{d/2}} \exp\left\{-\frac{1}{2\gamma}\left\|y - x + \gamma\nabla U(x)\right\|^2\right\} \mathrm{d}y \tag{19}$$

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

$$\|\pi - \pi_\gamma\|_{\mathrm{TV}} \leq C\sqrt{\gamma} \tag{20}$$

for some constant $C > 0$. Let us now derive an upper bound for the variance of the classical estimator (10) for ULA-based chain.

Lemma 2 *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, satisfying* **(H1)** *and* **(H2)**. *Assume also that $n\gamma > c_0$ for some fixed $c_0 > 0$, then it holds*

$$\mathsf{Var}[\pi_n^N(f)] \lesssim \frac{V(x)}{n\gamma}, \tag{21}$$

with $V(x) = 1 + \|x\|^2$ where $\lesssim$ stands for inequality up to a constant not depending on γ and n.

Remark 2 *The bound in Lemma 2 is sharp and can not 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

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

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

Theorem 3 *Assume* **(H1)** *and* **(H2)**. *Fix some* $K \geq 1$ *and suppose additionally 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* $0 < \|\mathbf{k}\| \leq K$,

$$|\partial^{\mathbf{k}} f(x)| \leq B_f, \quad |\partial^{\mathbf{k}} \mu(x)| \leq B_\mu \,. \tag{22}$$

Set $V(x) = 1 + \|x\|^2$, *then it holds for* $q = 0, \ldots, n-1$, *and all* x

$$A^2_{q,\mathbf{k}}(x) \leq C_K e^{\varkappa n\gamma^2}(1 + V(x)), \quad 1 \leq |\mathbf{k}| \leq K, \tag{23}$$

and

$$\sum_{\|\mathbf{k}\| \geq K+1} A^2_{q,\mathbf{k}}(x) \leq C_K e^{\varkappa n\gamma^2} \sum_{I \subseteq \{1,\ldots,d\},\, I \neq \emptyset} \left(\frac{\gamma}{2}\right)^{|I|K-1} (1 + V(x)) \tag{24}$$

with some positive constants $\varkappa, C_K$ *not depending on* n *and* γ. *The sum in* (23) *runs over all nonempty subsets* I *of the set* $\{1, \ldots, d\}$

Corollary 2 *Under assumptions of Theorem 3,*

$$\mathsf{Var}\left(\pi^N_{K,n}(f)\right) \lesssim n^{-1}\gamma^{K-1} e^{\varkappa n\gamma^2}\,, \tag{25}$$

where $\lesssim$ *stands for inequality up to a constant not depending on* n, N *and* γ.

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

$$A_{q,\mathbf{k}}(x) = \gamma^{|\mathbf{k}'|/2} \frac{\sqrt{(\mathbf{k}-\mathbf{k}')!}}{\sqrt{\mathbf{k}!}} \mathsf{E}\left[\partial^{\mathbf{k}'}_{Z_1} F(x, Z_1, \ldots, Z_q) \mathbf{H}_{\mathbf{k}-\mathbf{k}'}(Z_1)\right], \tag{26}$$

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

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

Also from (26) we can derive

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

where

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

Finally, we show that under our smoothness assumption (22) it holds

$$\mathsf{Var}\left(\sum_{p=1}^q \partial^{\mathbf{K}_I}_{Z_1} f\left(X^x_p\right)\right) \leq C_K \gamma^{-1} e^{\varkappa n\gamma^2}(1 + V(x)), \quad q = 0, \ldots, n-1, \tag{27}$$

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

Proposition 4 *Let $q \geq l, k \in \mathbb{N}$. Then the coefficients $a_{q,l,k}(x)$ in (8) can be alternatively represented as*

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

with $Q_{q,l}(x) = \mathsf{E}\left[f(X_{l,q}^x)\right]$, $q \geq l$. In the case $\Phi_l = \Phi$ in (5) for all $l \geq 1$, we have

$$\bar{a}_{r,k}(x) = \mathsf{E}\left[\phi_k\left(\xi\right) Q_r\left(\Phi(x,\xi)\right)\right] \tag{28}$$

with $Q_r(x) = \mathsf{E}\left[f(X_r^x)\right]$, $r \in \mathbb{N}$.

We construct a polynomial approximation for each $Q_r(x)$ in (28) in the form:

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

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

$$\widehat{Q}_r = \operatorname*{arg\,min}_{\psi \in \Psi_m} \sum_{s=1}^{T} \left| f(X_{l_0+r}^{(s)}) - \psi(X_{l_0}^{(s)}) \right|^2 \tag{29}$$

for $1 \leq l_0 + r \leq n-1$, where Ψ_m is a class of polynomials

$$\psi(x) = \sum_{\|\mathbf{s}\| \leq m} \alpha_{\mathbf{s}} x^{\mathbf{s}}, \quad \alpha_{\mathbf{s}} \in \mathbb{R}.$$

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 (28). Namely, for all $x \in \mathbb{R}^d$,

$$\begin{aligned} \widehat{a}_{r,\mathbf{k}}(x) &= \mathsf{E}\left[\mathbf{H}_{\mathbf{k}}(\xi)\widehat{Q}_r(x - \gamma\mu(x) + \sqrt{\gamma}\xi) \,\middle|\, \mathcal{D}\right] \\ &= \sum_{\|\mathbf{s}\| \leq m} \beta_{\mathbf{s}} \prod_{i=1}^{d} P_{k_i,s_i}(x_i - \gamma\mu_i(x)), \end{aligned} \tag{30}$$

where for any integer k, s $P_{k,s}$ is a one-dimensional polynomial (in x) of degree at most s with analytically known coefficients. We estimate Q_r only for $r < n_{\text{trunc}}$ 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 (12):

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

where

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

6.1. Gaussian mixtures

We consider ULA with π given by the mixture of two equally-weighted 2-dimensional Gaussian distributions of the following form

$$\pi(x) = \frac{1}{2\sqrt{(2\pi)^d|\Sigma|}}\left(\mathrm{e}^{-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}} + \mathrm{e}^{-\frac{(x+\mu)^T\Sigma^{-1}(x+\mu)}{2}}\right) \tag{31}$$

where $\rho \in (0,1), \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{2\times 2}$ and $|\Sigma|$ is its determinant. The function $U(x)$ and its gradient are given by

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

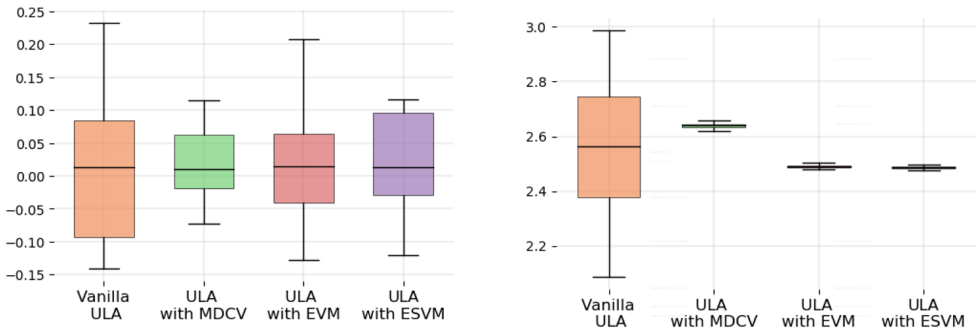
and

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

respectively. In our experiments we take $\mu = (0.5, 0.5)$ and randomly generated positive-definite matrix Σ (heterogeneous structure). In order to approximate the expectation $\pi(f)$ for $f(x) = \sum_{i=1}^d x_i$ or $f(x) = \sum_{i=1}^d x_i^2$, we use ULA with constant step size $\gamma = 0.1$ and sampled $T = 10$ independent training trajectories, each one of size $n = 10^4$ to estimate coefficients $a_{p-l,\mathbf{k}}$. Then we solve the least squares problems (29) using the first or second order polynomial basis depending on f. Respectively, we set $K = 1$ for linear functionals f and $K = 2$ for quadratic functionals. We set the truncation level $n_{\text{trunc}} = 50$. To test our variance reduction algorithm, we generated 100 independent trajectories of length $n = 10^3$ and depict boxplots of the ergodic averages in Figure 1. Our approach (MDCV, Martingale decomposition control variates) is compared to other variance reduction methods of [19] and [1]. In the baselines we use first order polynomials, as they were used as regressors during computations of $\hat{a}_{p-l,\mathbf{k}}$. Next we compute the cost-variance ratios

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

where variances in the above definition are computed using 24 independent trajectories of the length $n = 2 \cdot 10^3$. Note that the cost of $\widehat{\pi}_{K,n,n_{\text{trunc}}}$ is proportional to the number of nonzero coefficients in (30) times n_{trunc}. In Figure 2 cost-variances ratios are shown for for the case of Gaussian distribution with parameters. In particular, in Figure 2a a function $\mathcal{R}(1, 2 \cdot 10^3, 2 \cdot 10^3, n_{\text{trunc}})$ is shown in dependence on n_{trunc} for the case of a linear functional f. In Figure 2b a similar plot of $\mathcal{R}(2, 2 \cdot 10^3, 2 \cdot 10^3, n_{\text{trunc}})$ is shown for the case of a quadratic functional f. Figure 3 presents the corresponding coast-variance ratios for a mixture of two two-dimensional normal distributions as in (31) with $\Sigma = I$ and $\mu = (0.5, 0.5)$.



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

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 (Vanilla), our estimator of control variates (MDCV), zero variance estimator (ZV) and control variates obtained with empirical spectral variance minimisation (ESVM)

6.2. Banana shape distribution

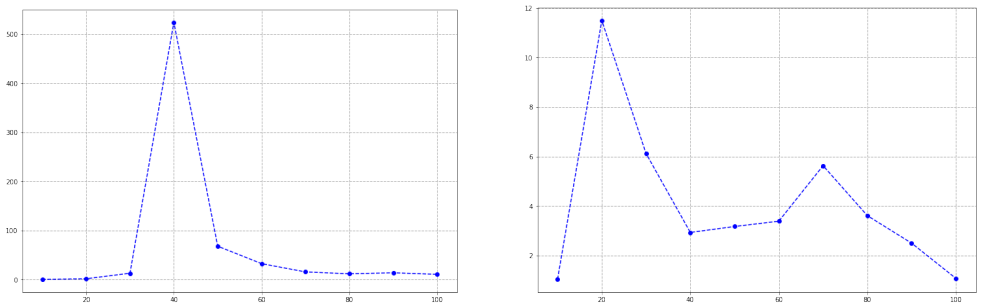
The "Banana-shape" distribution, proposed by [15], can be obtained from a d-dimensional Gaussian vector with zero mean and covariance $\text{diag}(p, 1, \ldots, 1)$ by applying transformation $\varphi_b(x) : \mathbb{R}^d \to \mathbb{R}^d$ of the form

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

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

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

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 (29) using the second-order



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

Figure 2: Cost-variance ratios as a function of the truncation level for two-dimensional standard Gaussian distribution and different test functions.

polynomial approximations for the coefficients $\hat{a}_{p,\mathbf{k}}$ as described in the previous section, hence we set $K = 2$. We use $T = 10$ independent training trajectories, each of size $n = 10^4$. We set the truncation level $n_{\text{trunc}} = 50$. To test our variance reduction algorithm, we generated 100 independent trajectories of length $n = 10^3$ and plot boxplots of the ergodic averages in Figure 4.

6.3. Binary Logistic Regression

Second experiment considers the problem of 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

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

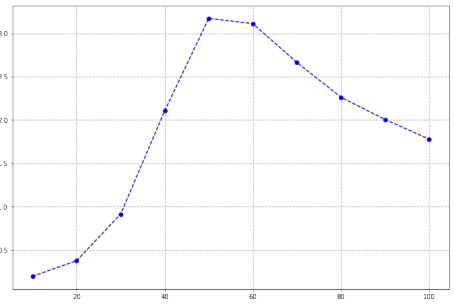
In order to estimate θ according to given data, the Bayesian approach introduces prior distribution $\pi_0(\theta)$ and inferences the posterior density $\pi(\theta|\mathsf{Y}, \mathsf{X})$ using Bayes' rule.

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

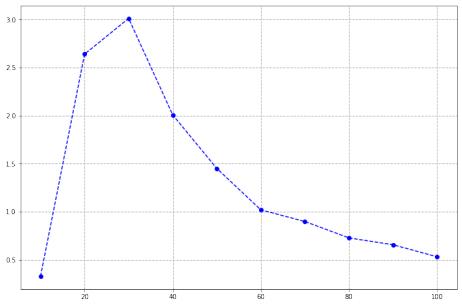
$$\pi(\theta|\mathsf{Y}, \mathsf{X}) \propto \exp\left\{\mathsf{Y}^T \mathsf{X}\theta - \sum_{i=1}^{m} \ln\left(1 + \mathrm{e}^{\mathsf{X}_i \theta}\right) - \frac{1}{2\sigma^2}\|\theta\|_2^2\right\},$$

Thus we obtain

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



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



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

Figure 3: Cost-variance ratios as a function of the truncation level for a mixture of two-dimensional Gaussian distribution and different test functions.

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 $n = 1 \times 10^4$, 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 5 we compare our approach to the variance reduction methods of [19] and [1]. 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 1

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

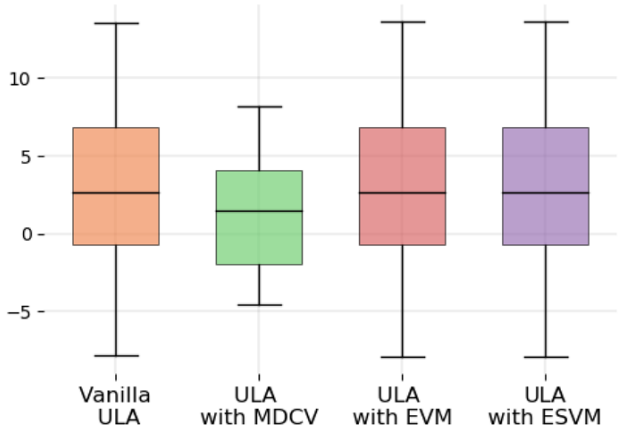


Figure 4: Boxplots of ergodic averages from the variance reduced ULA algorithms for the Banana shape density. The compared estimators are the ordinary empirical average (Vanilla), our estimator of control variates (MDCV), zero variance estimator (ZV) control variates obtained with empirical spectral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

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 $\Psi_{n,m+1,1}(y) = \mathsf{E}[|f(X^y_{m,m+1}) - f_{n,m+1,1}(y)|^2]$ and

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

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

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

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

$$f(X^x_{m+1}) = \mathsf{E}[f(X^x_{m+1})|\mathcal{G}_m] + \sum_{k=1}^{\infty} a_{m+1,m+1,k}(X^x_m)\phi_k(\xi_{m+1}) \tag{32}$$

where

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

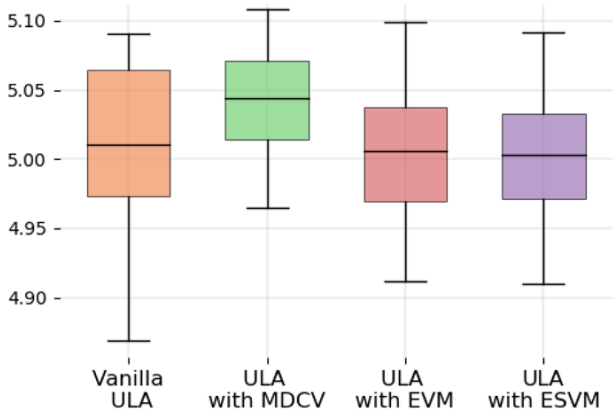


Figure 5: Boxplots of ergodic averages from the variance reduced ULA algorithms for the Logistic Regression. The compared estimators are the ordinary empirical average (Vanilla), our estimator of control variates (MDCV), zero variance estimator (ZV) control variates obtained with empirical spectral variance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

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

$$\mathsf{E}\left[f(X^x_{m+1})\middle|\,\mathcal{G}_m\right] = \mathsf{E}\left[f(X^x_{m+1})\middle|\,\mathcal{G}_j\right] + \sum_{k\geq 1}\sum_{l=j+1}^{m} a_{m+1,l,k}(X^x_{l-1})\phi_k(\xi_l) \tag{33}$$

with

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

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 (32) and (33) conclude the induction step for $q = m + 1$ and all $j < q$ and hence the proof.

7.2. Proof of Lemma 2

Under assumptions **(H1)** and **(H2)** the Markov chain $(X^x_p)_{p\in\mathbb{N}_0}$ is $V-$geometrically ergodic with $V(x) = 1 + \|x\|^2$ and $\rho = \mathrm{e}^{-\kappa\gamma}$ with κ specified in Lemma 11. Due to lemma 13,

$$\mathsf{Var}[\pi^N_n(f)] \lesssim \frac{1}{n(1-\rho^{1/2})}$$

and the result of lemma follows from Taylor expansion of denominator.

7.3. Proof of Theorem 3

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

$$X_p^x = G_p(x, \sqrt{\gamma} Z_1, \dots, \sqrt{\gamma} Z_p),$$

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

$$G_p(x, y_1, \dots, y_p) := \Phi(\cdot, y_p) \circ \Phi(\cdot, y_{p-1}) \circ \dots \circ \Phi(x, y_1) \tag{34}$$

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

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

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

$$\partial_1^{\mathbf{k}} f(X_p) := \partial_1^{\mathbf{k}} [f \circ G_p](X_0, \sqrt{\gamma} Z_1, \dots, \sqrt{\gamma} Z_p) \tag{35}$$

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 $\mathbf{k} = (k_i) \in \mathbb{N}_0^d$, we use the notation $\mathbf{k}! := k_1! \cdot \ldots \cdot k_d!$

Lemma 5 *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 (28).

Proof Let $\boldsymbol{\varphi}(z) = (2\pi)^{-d/2} \exp(-|z|^2/2)$, $z \in \mathbb{R}^d$, 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 $l = p$)

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

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 $\mathbf{k}' \neq \mathbf{k}$, $\|k'\| \le K$, we get applying first Lemma 5,

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

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 $\mathbf{k}' = \mathbf{k}'(\mathbf{k}) = K(\mathbb{1}_{\{k_1>K\}} \ldots, \mathbb{1}_{\{k_d>K\}})$, we get

$$\sum_{\mathbf{k}:\ \|\mathbf{k}\|\ge K+1} A_{s,\mathbf{k}}^2(x) = \sum_{\mathbf{k}:\ \|\mathbf{k}\|\ge K+1}\left(\gamma^{|\mathbf{k}'|}\frac{(\mathbf{k}-\mathbf{k}')!}{\mathbf{k}!}\right) Q_s(\mathbf{k}', \mathbf{k}-\mathbf{k}')$$

$$= \left\{\sum_{I\subseteq\{1,\ldots,d\},\, I\neq\emptyset} \gamma^{|I|K} \sum_{\mathbf{m}_I\in\mathbb{N}_I^d} \frac{\mathbf{m}_I!}{(\mathbf{m}_I+\mathbf{K}_I)!}\right\}\left\{\sum_{\mathbf{m}_{I^c}\in\mathbb{N}_{0,I^c}^d,\, \|\mathbf{m}_{I^c}\|\le K} Q_s(\mathbf{K}_I, \mathbf{m}_I+\mathbf{m}_{I^c})\right\}, \tag{36}$$

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 (36) the first sum runs over all nonempty subsets I of the set $\{1,\ldots,d\}$. 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, $I^c = \{1,\ldots,d\} \setminus I$ 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 $\mathbf{K}_I = (K1_{\{1\in I\}}, \ldots, K1_{\{d\in I\}})$. Applying the estimate

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

we get

$$\begin{aligned}\sum_{\mathbf{k}:\ \|\mathbf{k}\|\ge K+1} A_{s,\mathbf{k}}^2(x) &\le \sum_{I\subseteq\{1,\ldots,d\},\, I\neq\emptyset} (\gamma/2)^{|I|K} \\ &\quad\times \sum_{\mathbf{m}_I\in\mathbb{N}_I^d}\sum_{\mathbf{m}_{I^c}\in\mathbb{N}_{0,I^c}^d,\, \|\mathbf{m}_{I^c}\|\le K} Q(\mathbf{K}_I, \mathbf{m}_I+\mathbf{m}_{I^c}) \\ &\le \sum_{I\subseteq\{1,\ldots,d\},\, I\neq\emptyset} (\gamma/2)^{|I|K} \sum_{\mathbf{m}\in\mathbb{N}_0^d} Q(\mathbf{K}_I, \mathbf{m}).\end{aligned} \tag{37}$$

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 (37) 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 3

$$\mathsf{Var}\left(\sum_{p=1}^{q} \partial_1^{\mathbf{K}_I} f\left(X_p^x\right)\right) \le C\gamma^{-1} e^{\varkappa n\gamma^2}(1+V(x)), \quad q=1,\dots,n,$$

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 6 *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*

$$0 \le x_p \le \alpha_p x_{p-1} + \gamma\epsilon_p, \quad p\in\mathbb{N}, \tag{38}$$

$$0 \le \epsilon_p \le \overline{C}_1 \prod_{k=1}^{p} \alpha_k^2, \quad p\in\mathbb{N}, \tag{39}$$

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

$$\gamma \sum_{r=1}^{\infty} \prod_{k=1}^{r} \alpha_k \le \overline{C}_2 \tag{40}$$

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 (38) 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 (39) 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 (40), completes the proof. □

In what follows, we use the notation

$$\alpha_k = 1 - \gamma\mu'(X^x_{k-1}), \quad k \in \mathbb{N}. \tag{41}$$

The assumption (22) 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 7 *Under assumptions of Theorem 3, for all natural $r \le K$ and $l \le p$, there exist constants C_r (not depending on l and p) such that*

$$\left|\partial^r_{y_l} X^x_p\right| \le C_r \prod_{k=l+1}^{p} (1 - \gamma\mu'(X^x_{k-1})) \quad a.s. \tag{42}$$

where $\partial^r_{y_l} X^x_p$ is defined in (35). Moreover, we can choose $C_1 = 1$.

Lemma 8 *Under assumptions of Theorem 3, for all natural $r \le K$, $j \ge l$ and $p > j$, we have*

$$\left|\partial_{y_j}\partial^r_{y_l} X^x_p\right| \le c_r \prod_{k=l+1}^{p} (1 - \gamma\mu'(X^x_{k-1})), \quad a.s. \tag{43}$$

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 (43) 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 7, 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 6 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 6 is zero, while $\overline{C}_1$ in Lemma 6 has the form

const $\prod_{k=l+1}^{j}\alpha_k$), we obtain (43) for $r=1$. The induction hypothesis is now that the inequality

$$\left|\partial_{y_j}\partial_{y_l}^k X_p^x\right| \le c_k \prod_{s=l+1}^{p}\alpha_s \tag{44}$$

holds for all natural $k<r$ ($\le K$) and $p>j$. We need to show (44) for $k=r$. Faà di Bruno's formula implies for $2\le r\le K$ and $p>l$

$$\begin{aligned}\partial_{y_l}^r X_p^x &= \left[1-\gamma\mu'(X_{p-1}^x)\right]\partial_{y_l}^r X_{p-1}^x \\ &\quad -\gamma\sum\frac{r!}{m_1!\dots m_{r-1}!}\mu^{(m_1+\dots+m_{r-1})}(X_{p-1}^x)\prod_{k=1}^{r-1}\left(\frac{\partial_{y_l}^k X_{p-1}^x}{k!}\right)^{m_k},\end{aligned} \tag{45}$$

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

$$1\cdot m_1+2\cdot m_2+\dots+(r-1)\cdot m_{r-1}=r. \tag{46}$$

Notice that we work with $(r-1)$-tuples rather than with r-tuples because the term containing $\partial_{y_l}^r X_{p-1}^x$ on the right-hand side of (45) is listed separately. For $p>j$, we then have

$$\begin{aligned}\partial_{y_j}\partial_{y_l}^r X_p^x &= \left[1-\gamma_p\mu'(X_{p-1}^x)\right]\partial_{y_j}\partial_{y_l}^r X_{p-1}^x-\gamma\mu''(X_{p-1}^x)\partial_{y_l}^r X_{p-1}^x\partial_{y_j}X_{p-1}^x \\ &\quad -\gamma\sum\frac{r!}{m_1!\dots m_{r-1}!}\mu^{(m_1+\dots+m_{r-1}+1)}(X_{p-1}^x)\partial_{y_j}X_{p-1}^x\prod_{k=1}^{r-1}\left(\frac{\partial_{y_l}^k X_{p-1}^x}{k!}\right)^{m_k} \\ &\quad -\gamma\sum\frac{r!}{m_1!\dots m_{r-1}!}\mu^{(m_1+\dots+m_{r-1})}(X_{p-1}^x)\partial_{y_j}\left[\prod_{k=1}^{r-1}\left(\frac{\partial_{y_l}^k X_{p-1}^x}{k!}\right)^{m_k}\right] \\ &= \left[1-\gamma\mu'(X_{p-1}^x)\right]\partial_{y_j}\partial_{y_l}^r X_{p-1}^x+\gamma\epsilon_{l,j,p},\end{aligned} \tag{47}$$

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

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

Using Lemma 7, induction hypothesis (44) and the fact that $m_1+\dots+m_{r-1}\ge 2$ for

$(r-1)$-tuples of nonnegative integers satisfying (46), 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" depending on $B_\mu, r, C_1, \dots, C_r, c_1, \dots, c_{r-1}$. Thus, (47) now implies

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

We can again apply Lemma 6 to bound $|\partial_{y_j} \partial_{y_l}^r X_p^x|$ for $p \ge j+1$ (notice that $\partial_{y_j} \partial_{y_l}^r X_j^x = 0$, that is, $\overline{C}_0$ in Lemma 6 is zero, while $\overline{C}_1$ in Lemma 6 has the form const $\prod_{k=l+1}^{j} \alpha_k$), and we obtain (44) for $k = r$. This concludes the proof. □

Lemma 9 *Under assumptions of Theorem 3, it holds*

$$\mathsf{Var} \left[\sum_{p=1}^{q} \partial_{y_1}^K f\left(X_p^x\right) \right] \le \gamma^{-1} C e^{\varkappa n \gamma^2} (1 + V(x)), \quad q = 1, \dots, n,$$

where $C, \varkappa > 0$ are constants that do not depend on 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 $\nabla_Z F = (\partial_{Z_1} F, \ldots, \partial_{Z_q} F)$, 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

$$\mathsf{Var}\left[\sum_{p=1}^{q} \partial_{y_1}^K f\left(X_p^x\right)\right] \le \gamma^2 \sum_{j=1}^{q} \mathsf{E}\left[\left(\sum_{p=j}^{q} \partial_{y_j}\partial_{y_1}^K f\left(X_p^x\right)\right)^2\right]. \tag{48}$$

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

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

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

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

Then

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

Using the bounds of Lemmas 7 and 8, we obtain

$$\left|\partial_{y_j}\partial_{y_1}^K f\left(X_p^x\right)\right| \le A_K \prod_{k=2}^{p} \alpha_k \tag{49}$$

with a suitable constant A_K. Substituting this in (48), 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 using the fact that $\prod_{k=j+1}^{p}\alpha_k \le \exp\left(-\sum_{k=j+1}^{p}\gamma\mu'(X_{k-1}^x)\right)$, we get

$$\mathsf{E}\left[\sum_{j=1}^{q}\prod_{k=1}^{j}\alpha_k\left(\sum_{p=j+1}^{q+1}\prod_{k=j+1}^{p}\alpha_k\right)^2\right] \le \sum_{j=1}^{q}\zeta_{1,j}^{1/2}(2)\left(\sum_{p=j+1}^{q+1}\zeta_{j+1,p}^{1/4}(4)\right)^2,$$

where we denote

$$\zeta_{l,j}(u) = \mathsf{E}\left[e^{-u\gamma\sum_{k=l}^{j}\mu'(X_{k-1}^x)}\right], \quad u>0.$$

Note that $\mu'(x)$ is bounded due our assumptions and from Theorem 1 in [7] and Lemma 11 in Appendix A it follows that

$$\mathsf{E}\left[e^{-u\gamma\left(\sum_{k=l}^{j}[\mu'(X_{k-1}^x)-\pi_\gamma(\mu')]\right)}\right] \le C_1 e^{\varkappa n\gamma^2}(1+V(x)), \quad u\gamma\le\gamma_0,$$

for some constants γ_0 and $C_1, \varkappa > 0$. Then we have

$$\zeta_{l,j}(u) = \mathsf{E}\left[e^{-s\gamma\sum_{k=l}^{j}\mu'(X_{k-1}^x)}\right] \le C_1 e^{\varkappa n\gamma^2}(1+V(x))e^{-s\gamma(j-l+1)\pi_\gamma(\mu')}.$$

Furthermore, since $\mu\pi', \mu'\pi \in L^1(\mathbb{R})$ and $\pi'(x) = -\frac{1}{2}\pi(x)\mu(x)$, we have

$$\pi(\mu') = \int \mu'(x)\pi(x)\,dx = -\int \mu(x)\pi'(x)\,dx = \frac{1}{2}\int \mu^2(x)\pi(x)\,dx > 0$$

Note also that $\pi_\gamma(\mu') \geq \pi_\gamma(\mu') - C_2\sqrt{\gamma}$ yielding the bound

$$\zeta_{l,j}(s) \leq C_1 e^{\varkappa n\gamma^2}(1+V(x))e^{-s\gamma(j-l+1)\alpha}$$

where $\alpha = \frac{1}{2}\int \mu^2(x)\pi(x)\,dx - C_1\sqrt{\gamma} > 0$ for sufficiently small γ. Hence we obtain

$$\begin{aligned}\sum_{p=j+1}^{q+1} \zeta_{j+1,p}^{1/4}(4) &\leq (C_1 e^{\varkappa n\gamma^2}(1+V(x)))^{1/4} \sum_{p=j+1}^{q+1} e^{-\gamma(p-j)\alpha} \\ &\leq (C_1 e^{\varkappa n\gamma^2}(1+V(x)))^{1/4} \frac{1}{1-e^{-\gamma\alpha}}\end{aligned}$$

and

$$\begin{aligned}\sum_{j=1}^{q} \zeta_{1,j}^{1/2}(2) &\leq (C_1 e^{\varkappa n\gamma^2}(1+V(x)))^{1/2} \sum_{j=0}^{q} e^{-\gamma j\alpha} \\ &\leq (C_1 e^{\varkappa n\gamma^2}(1+V(x)))^{1/2} \frac{1}{1-e^{-\gamma\alpha}}.\end{aligned}$$

Thus the final bound follows:

$$\mathsf{Var}\left[\sum_{p=1}^{q} \partial_{y_1}^K f\left(X_p^x\right)\right] \leq \gamma^{-1} C_K e^{\varkappa n\gamma^2}(1+V(x))$$

with C_K depending neither on q nor on γ. The proof is completed. □

Appendix A: Bounds for moments of ULA

First let us introduce some definitions

Definition 1 *We say that the Markov kernel Q_γ satisfies Foster-Lyapunov drift condition if there exist a measurable function $V : X \to [1;+\infty)$, real numbers $\lambda \in (0,1)$ and $C > 0$ such that for any $x \in X$,*

$$Q_\gamma V(x) \leq \lambda^\gamma V(x) + \gamma C \tag{50}$$

Definition 2 *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 $n \in \mathbb{N}$,*

$$d_V(\delta_x Q^{1,n}, \pi) \leq C\rho^n V(x) \tag{51}$$

Note that conditions **(H1)** and **(H2)** imply ([4, Lemma 16]) that for some $K_2 > 0$ it holds

$$\langle \nabla U(x), x \rangle \geq \frac{m}{2}\|x\|^2, \quad \|x\| \geq K_2 \tag{52}$$

Now we shall show that conditions **(H1)** and **(H2)** are sufficient to show that ULA kernel Q_γ satisfies the drift condition (50)

Lemma 10 *Assume that the potential $U(x), x \in \mathbb{R}^d$ satisfies conditions* **(H1)** *and* **(H2)** *and without loss of generality consider $\nabla U(0) = 0$. Then the kernel Q_γ from (19) satisfies drift condition (50) for any $0 < \gamma < \overline{\gamma} = \frac{m}{4L}$ with drift function $V(x) = 1 + \|x\|^2$, $\lambda = \exp\left(-\frac{m}{2}\right)$, $C = \frac{25K_2^2}{8} + d + m$ with K_2 from (52), m from* **(H2)**.

Proof Let $V(x) = 1 + \|x\|^2$, then

$$Q_\gamma V(x) = \int\limits_{\mathbb{R}^d} V(y) Q_\gamma(x, dy) = \int\limits_{\mathbb{R}^d} \frac{(1 + \|y\|^2)}{(2\pi\gamma)^{\frac{d}{2}}} \exp\left(-\frac{\|y - x + \gamma\nabla U(x)\|^2}{2\gamma}\right) dy =$$

$$= 1 + \frac{1}{(2\pi\gamma)^{\frac{d}{2}}} \int\limits_{\mathbb{R}^d} \|z + x - \gamma\nabla U(x)\|^2 \exp\left(-\frac{\|z\|^2}{2\gamma}\right) dz$$

Let us first consider the case $x \notin B(0, K_2)$. Note that

$$\|z + x - \gamma\nabla U(x)\|^2 = \|z\|^2 + 2\langle z, x - \gamma\nabla U(x)\rangle + \|x - \gamma\nabla U(x)\|^2$$

and the linear term vanishes after integration, moreover, for $Z \sim \mathcal{N}(0, \gamma I_d)$, it holds $\mathbb{E}\|Z\|^2 = \gamma d$. It remains to notice that due to (52),

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

Since $\gamma < \frac{m}{4L}$, we obtain

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

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

$$Q_\gamma V(x) \leq (1 - \gamma m + 2\gamma^2 L^2)V(x) + \gamma\left((m - 2\gamma L^2)(1 + \|x\|^2) + d + 2(1 + L\gamma)^2\|x\|^2\right) \leq$$

$$\leq \exp^{-\frac{\gamma m}{2}} V(x) + \gamma(\frac{25K_2^2}{8} + d + m)$$

□

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 [9, Theorem 19.4.1]. Namely, the following lemma holds:

Lemma 11 *Assume that the potential $U(x), x \in \mathbb{R}^d$ satisfies conditions* **(H1)** *and* **(H2)**. *Then for $0 < \gamma < \overline{\gamma} = \frac{m}{4L^2}$, for any $x \in X$ it holds*

$$d_V(\delta_x Q_\gamma^{1,n}, \pi_\gamma) \le C\rho^n \left(V(x) + \pi_\gamma(V)\right)$$

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

$$C = \left(1 + \exp\left(-\frac{m\gamma}{2}\right)\right)\left(1 + \frac{\overline{b}}{(1-\varepsilon)(1 - \exp\left(-\frac{m\gamma}{2}\right) - \frac{2b}{1+d})}\right);$$

$$b = \gamma(\frac{25K_1^2}{8} + 2d + m); \quad \overline{b} = b\exp\left(-\frac{m\gamma}{2}\right) + d; \quad \varepsilon = 2\Phi\left(-\frac{\sqrt{d}(1+L\gamma)}{2\sqrt{\gamma}}\right);$$

$$\rho = \exp\left(-\gamma\left(\frac{m}{2} - 2\frac{\frac{25K_1^2}{8} + 2d + m}{d}\right)\frac{\log(1-\varepsilon)}{\log(1-\varepsilon) + \log\left(\exp\left(-\frac{m\gamma}{2}\right) + \frac{2b}{d+1}\right)}\right)$$

Proof Note that the condition **(H1)** implies that the Markov kernel $Q_\gamma^{1,n}$ satisfies $(1,\varepsilon)$-Doeblin condition with $\varepsilon = 2\Phi\left(-\frac{\sqrt{d}(1+L\gamma)}{2\sqrt{\gamma}}\right)$. Together with drift condition (50) it allows to apply [9, 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 (10). We use the same technique as in [1] to control autocovariances for a given Markov chain. We start from auxiliary lemma:

Lemma 12 *Let $(X_p)_{p\in\mathbb{N}_0}$ be a $V-$geometrically ergodic Markov chain with a stationary distribution π, and $f(x)$ be a function with $\|f\|_{V^{\frac{1}{2}}} < \infty$. Let $\tilde{f}(x) = f(x) - \pi(f)$. Then it holds for some constant $C > 0$*

$$|\mathsf{E}_x[\tilde{f}(X_0)\tilde{f}(X_s)]| \le C\rho^{s/2}\|\tilde{f}\|_{V^{\frac{1}{2}}} V(x) \tag{53}$$

For the stationary distribution it holds

$$|\mathsf{E}_\pi[\tilde{f}(X_0)\tilde{f}(X_s)]| \le C\pi(V)\rho^{s/2}\|\tilde{f}\|_{V^{\frac{1}{2}}} \tag{54}$$

Proof Note that

$$|\mathsf{E}_x[\tilde{f}(X_0)\tilde{f}(X_s)]| \le |\tilde{f}(x)|\mathsf{E}_x|\tilde{f}(X_s)| \le \|\tilde{f}\|_{V^{\frac{1}{2}}} V^{\frac{1}{2}}(x) \int\limits_{\mathbb{R}^d} V^{\frac{1}{2}}(y)|\mathsf{P}^s(x,\, dy) - \pi(dy)|$$

By Hoelder inequality,

$$\begin{aligned}\int\limits_{\mathbb{R}^d} V^{\frac{1}{2}}(y)|\mathsf{P}^s(x,\, dy) - \pi(dy)| &\le \left(\int\limits_{\mathbb{R}^d} V(y)|\mathsf{P}^s(x,\, dy) - \pi(dy)|\right)^{1/2} \left(\int\limits_{\mathbb{R}^d} |\mathsf{P}^s(x,\, dy) - \pi(dy)|\right)^{1/2} \le \\ &\le 2\rho^{s/2} V^{1/2}(x)\end{aligned}$$

yielding the first statement of lemma. The second statement can be obtained from the first one by integration with respect to π. □

Lemma 13 *Let $(X_p)_{p\in\mathbb{N}_0}$ be a $V-$geometrically ergodic Markov chain with a stationary distribution π, and $f(x)$ be a function with $\|f\|_{V^{\frac{1}{2}}} < \infty$. Let $\tilde{f}(x) = f(x) - \pi(f)$. Then*

$$\mathsf{cov}_x\left[f(X_k), f(X_{k+s})\right] \le C^2 V^2(x)\rho^{k+s} + C^2\rho^{k+s/2}\|\tilde{f}\|_{V^{\frac{1}{2}}} V(x) + C\pi(V)\rho^{s/2}\|\tilde{f}\|_{V^{\frac{1}{2}}} \quad (55)$$

Proof Note that

$$\mathsf{cov}_x\left[f(X_k), f(X_{k+s})\right] = \mathsf{E}_x\left[f(X_k) - \pi(f)\right]\left[f(X_{k+s}) - \pi(f)\right] + \left[\pi(f) - \mathsf{E}_x f(X_s)\right]\left[\mathsf{E}_x f(X_k) - \pi(f)\right]$$

Due to $V-$ergodicity,

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

To bound the first term note that

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

where the last inequality is due to lemma 12, which implies (55). □

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

Lemma 14 *Let $(X_p)_{p\in\mathbb{N}_0}$ be a $V-$geometrically ergodic Markov chain with a stationary distribution π, and $f(x)$ be a function with $\|f\|_{V^{\frac{1}{2}}} < \infty$. Assume also that $n(1-\rho^{1/2}) > 1$. Then*

$$\mathsf{Var}_x\left[\pi_N^n(f)\right] \le \frac{C\pi(V)\|\tilde{f}\|_{V^{1/2}}}{n(1-\rho^{1/2})} + \mathcal{O}(\frac{1}{n^2(1-\rho^{1/2})}) \quad (56)$$

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 first and second sum using lemma 13:

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

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

Hence (56) follows. □

References

[1] D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, and S. Samsonov. Variance reduction for markov chains with application to mcmc. 2019.

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

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

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

[5] 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.

[6] 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.

[7] Bernard Delyon and A Juditsky. On small perturbations of stable markov operators: unbounded case. *Theory of Probability & Its Applications*, 43(4):577–587, 1999.

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

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

[10] Andrew B Duncan, Tony Lelievre, and GA Pavliotis. Variance reduction using nonreversible Langevin samplers. *Journal of statistical physics*, 163(3):457–491, 2016.
[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] Vincent Lemaire. An adaptive scheme for the approximation of dissipative systems. *Stochastic Process. Appl.*, 117(10):1491–1518, 2007.
[18] K. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, 24:101–121, 1996.
[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.