Submitted to Bernoulli

Variance reduction for MCMC methods via martingale representations

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

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

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

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

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

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

1. Introduction

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

Suppose that we wish to compute $\pi(f) := \mathsf{E}\left[f(X)\right]$, 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 [?]. 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 [?]. 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 [?] and Stein operator in [?]. 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 [?]. 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, $p = 0, 1, 2, \ldots$, 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 [?] 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 [?] and [?]. 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, [?] 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 [?] and [?] series-type control variates are introduced and studied for reversible Markov chains. It is assumed in [?] that the one-step conditional expectations can be computed explicitly for a set of basis functions. The authors in [?] 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 [?, ?, ?, ?, ?] and references therein. We propose a generic variance reduction method for these and other types algorithms, which is purely non-asymptotic and does not require that the conditional expectations of the corresponding Markov chain can be computed or that the generator is known analytically. Moreover, we do not need to assume stationarity or/and sampling under the invariant distribution π. We rigorously analyse the convergence of the method and study its complexity. It is shown that our variance-reduced Langevin algorithm outperforms the standard Langevin algorithms in terms of complexity.

The paper is organized as follows. In Section 2 we set up the problem and introduce some notations. Section 3 contains a novel martingale representation and shows how this representation can be used for variance reduction. In Section 4 we analyze the performance of the proposed variance reduction algorithm in the case of Unadjusted Langevin Algorithm (ULA). Section **??** studies the complexity of the variance reduced ULA. Finally, numerical examples are presented in Section 5.

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 $D^2 g$ 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 $\partial^{\mathbf{k}}_{y_j} h$ for the multiple derivative of h with respect to the components of y_j:

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

In the particular case $m = 1$ we drop the subscript y_1 in that notation.

2. Setup

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

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

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

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

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 [**?**]) and many well-known MCMC algorithms can be represented in form (1). Let us consider two popular examples.

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

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

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

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

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 [**?**] and [**?**].*

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

Given $X_p = x$;

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

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

where

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

*This transition is reversible with respect to π and therefore preserves the stationary density π; see [**?**, 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 [**?**]. 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. [**?**])*

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

where $\Gamma_n = \gamma_1 + \ldots + \gamma_n$ *and* $(\gamma_n)_{n\ge1}$ *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. [**?**]) impossible in this case. On the contrary, our method can be directly used to reduce the variance of the ergodic estimator* π_n^γ *without explicit knowledge of* π*.*

3. Martingale representation and variance reduction

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

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

with $\xi \sim P_\xi$. Notice that this implies that the random variables $\phi_k(\xi)$, $k \geq 1$, are centered. As an example, we can take multivariate Hermite polynomials for the ULA algorithm and a tensor product of Shifted Legendre polynomials for "uniform part" and Hermite polynomials for "Gaussian part" of the random variable $\xi = (u, z)^T$ in MALA, as the Shifted Legendre polynomials are orthogonal with respect to the Lebesgue measure on $[0, 1]$. Let $(\Omega, \mathcal{F}, \mathsf{P})$ be a probability space, $(\xi_p)_{p\in\mathbb{N}}$ be i.i.d. $d_\xi-$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 $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^{d_\xi}$ let $\Phi_k(x, y)$ be a function mapping $\mathbb{R}^{d+d_\xi}$ to $\mathbb{R}^d$. Then we can define for $l \leq p$

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

with the function $G_{l,p} : \mathbb{R}^{d+d_\xi\times(p-l+1)} \to \mathbb{R}^d$ 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 $\left(X^x_{0,p}\right)_{p\in\mathbb{N}_0}$ is a Markov chain with values in $\mathbb{R}^d$ of the form (1), starting at $X_0 = x$. We write X^x_p and G_p as a shorthand notation for $X^x_{0,p}$ and $G_{0,p}$, respectively.

Theorem 1 *Let $(X^x_p)_{p\geq 0}$ be a Markov chain of the form (5). Then, for $p > j$ and for all Borel functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathsf{E}\left[\left|f(X^x_p)\right|^2\right] < \infty$, the following representation holds in $L^2(\mathsf{P})$*

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

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

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

Proof The expansion obviously holds for $p = 1$ 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

$$f(X^x_1) = \mathsf{E}[f(X^x_1)] + \sum_{k\geq 1} a_{1,1,k}(x)\phi_k(\xi_1)$$

with $a_{1,1,k}(x) = \mathsf{E}[f(X^x_1)\phi_k(\xi_1)]$. Assume now that (7) holds for $p = q$, all $j < q$, and Borel-measurable functions f with $\mathsf{E}\left[|f(X^x_q)|^2\right] < \infty$. Let us prove that the induction assumption holds for $p = q + 1$. Given f with $\mathsf{E}\left[|f(X^x_{q+1})|^2\right] < \infty$, due to the orthonormality and completeness of the system (ϕ_k), we get that for any $y \in \mathbb{R}^d$ it holds in $L^2(\mathsf{P})$

$$f(X^y_{q,q+1}) = \mathsf{E}f(X^y_{q,q+1}) + \sum_{k=1}^{\infty} \mathsf{E}[f(X^y_{q,q+1})\phi_k(\xi_{q+1})]\phi_k(\xi_{q+1})$$

which means that

$$\lim_{n\to\infty} \mathsf{E}[|f(X^y_{q,q+1}) - \mathsf{E}f(X^y_{q,q+1}) - \sum_{k=1}^{n} \mathsf{E}[f(X^y_{q,q+1})\phi_k(\xi_{q+1})]\phi_k(\xi_{q+1})|^2] = 0 \tag{9}$$

Let us define sequence $\psi_{q+1,n}(y) = \mathsf{E}[|f(X^y_{q,q+1}) - \mathsf{E}f(X^y_{q,q+1})|^2] - \sum_{k=1}^{n} \mathsf{E}^2[f(X^y_{q,q+1})\phi_k(\xi_{q+1})]$. Note that $\psi_{q+1,n}(y)$ coincides with the left side of 9, and due to Parseval identity $\Psi_{q+1,n}(y) \to 0, n \to \infty$ for any $y \in \mathbb{R}^d$. Note that $\psi_n(y) < \mathsf{E}[|f(X^y_{q,q+1}) - \mathsf{E}f(X^y_{q,q+1})|^2] \le \frac{\|f\|^2_\infty}{4}$. Hence, by Lebesgue dominated convergence theorem, $\mathsf{E}\psi_n(X^x_q)^2 \to 0$ and since $\mathsf{E}[f(X^y_{q,q+1})\phi_k(\xi_{q+1})]$ is a version of conditional expectation $\mathsf{E}[f(X^x_{q+1}|X^x_q = y]$, it holds in $L^2(\mathsf{P})$

$$f(X^x_{q+1}) = \mathsf{E}[f(X^x_{q+1})|\mathcal{G}_k] + \sum_{k=1}^{\infty} a_{q+1,q+1,k}(X^x_q)\phi_k(\xi_{q+1}) \tag{10}$$

where

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

We thus arrive at

$$f(X^x_{q+1}) = \mathsf{E}\left[f(X^x_{q+1})\middle| X^x_q\right] + \sum_{k\ge1} a_{q+1,q+1,k}(X^x_q)\phi_k(\xi_{q+1}), \tag{11}$$

which is the required statement in the case $j = q$. Now assume $j < q$. The random variable $\mathsf{E}\left[f(X^x_{q+1})\middle| X^x_q\right]$ is square integrable and has the form $g(X^x_q)$, hence the induction hypothesis applies, and we get

$$\mathsf{E}\left[f(X^x_{q+1})\middle| X^x_q\right] = \mathsf{E}\left[f(X^x_{q+1})\middle| X^x_j\right] + \sum_{k\ge1}\sum_{l=j+1}^{q} a_{q+1,l,k}(X^x_{l-1})\phi_k(\xi_l) \tag{12}$$

with

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

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

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

Formulas (11) and (12) conclude the induction step for $p = q + 1$ and hence the proof. □

Corollary 1 *If all the kernels Φ_l, $l \geq 1$, are equal, then the representation* (7) *takes the form*

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

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

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

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

Proposition 2 *The coefficients $a_{p,l,k}$ in* (8) *can be alternatively represented as*

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

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

$$Q_{p,l}(x) = \mathsf{E}\left[Q_{p,l+1}(X_{l+1}^x)\right]. \tag{13}$$

In the case $\Phi_l = \Phi$ for all $l \geq 1$, we have

$$\bar{a}_{r,k}(x) = \mathsf{E}\left[\phi_k\left(\xi\right) Q_r\left(\Phi(x, \xi)\right)\right] \tag{14}$$

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

Next we show how the representation (7) can be used to reduce the variance of MCMC algorithms. Consider the case of time homogeneous transition kernels and introduce a weighted average estimator $\pi_n^N(f)$ of the form

$$\pi_n^N(f) = \frac{1}{n} \sum_{p=N+1}^{N+n} f(X_p^x), \tag{15}$$

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

$$\begin{aligned} M_{K,n}^N(f) &= \frac{1}{n} \sum_{p=N+1}^{N+n} \left[\sum_{k=1}^{K} \sum_{l=N+1}^{p} \bar{a}_{p-l,k}(X_{l-1}) \phi_k(\xi_l) \right] \\ &= \sum_{k=1}^{K} \sum_{l=N+1}^{N+n} \left(1 + \frac{N-l}{n}\right) A_{N+n-l,k}(X_{l-1}) \phi_k\left(\xi_l\right), \end{aligned} \tag{16}$$

where

$$A_{s,k}(x) = \frac{1}{s} \sum_{r=1}^{s} \bar{a}_{r,k}(x). \tag{17}$$

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

$$\mathsf{E}[A_{N+n-l,k}(X_{l-1})\phi_k(\xi_l)] = \mathsf{E}[A_{N+n-l,k}(X_{l-1})\mathsf{E}\left[\phi_k(\xi_l)\,|\,\mathcal{G}_{l-1}\right]] = 0$$

and hence the r.v. $M^N_{K,n}(f)$ has zero mean and can be viewed as a control variate. The representation (8) suggests that the variance of the variance-reduced estimator

$$\pi^N_{K,n}(f) = \pi^N_n(f) - M^N_{K,n}(f) \tag{18}$$

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

$$\mathsf{Var}[\pi^N_{K,n}(f)] \leq \sum_{k=K+1}^{\infty} \sum_{l=1}^{n} \mathsf{E}[A^2_{n-l,k}(X_{N+l-1})] \tag{19}$$

and $\mathsf{Var}[\pi^N_{K,n}(f)]$ is small provided that the coefficients $A_{s,k}$ decay fast enough as $k \to \infty$. In the next section we provide a detailed analysis of $\mathsf{Var}[\pi^N_{K,n}(f)]$ for the ULA algorithm (see Example 1).

4. Analysis of variance reduced ULA

In this section we perform the convergence analysis of the ULA algorithm. 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,

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

Sergey

N and N0 were defined here before

Eric

these notations are used before; it is best to put all the notations in a short paragraph just after the introduction

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

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

with $\|\mathbf{k}\| = \max_i k_i$. For an estimator $\rho(f) \in \{\pi^N_n(f), \pi^N_{K,n}(f)\}$ of $\pi(f)$ (see (15) and (18)), we shall be interested in the Mean Squared Error (MSE), which can be decomposed as the sum of the squared bias and the variance:

$$\mathrm{MSE}\left[\rho(f)\right] = \mathsf{E}\left[\left\{\rho(f) - \pi(f)\right\}^2\right] = \left\{\mathsf{E}[\rho(f)] - \pi(f)\right\}^2 + \mathsf{Var}[\rho(f)]. \tag{20}$$

Our analysis is carried out under the following two assumptions:

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

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

(H2) [**Convexity outside a ball**] There exist $K_U > 0$, $M_U > 0$ and $m_U > 0$ such that for any $x \notin B_{K_U}(0)$ it holds

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

First we analyse the biases of the estimators $\pi_n^N(f)$ and $\pi_{K,n}^N(f)$.

Squared bias: 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 conditional bias. Notice that this remains true also when we substitute the coefficients $a_{p,l,\mathbf{k}}$ in (16) with some independent approximations $\widehat{a}_{p,l,\mathbf{k}}$.

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

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

For a bounded Borel function f, we can estimate the conditional bias similarly to [**?**, Section 4]:

Sergey

Changed notation here, put variance of the noise equal to γ

$$\left\{\mathsf{E}[\pi_{K,n}^N(f)] - \pi(f)\right\}^2 = \left(\mathsf{E}[\pi_n^N(f)] - \pi(f)\right)^2 \leq \frac{\operatorname{osc}(f)^2}{n} \sum_{p=N+1}^{N+n} \|\delta_x \mathrm{Q}_\gamma^p - \pi(\cdot)\|_{\mathrm{TV}}^2, \quad (21)$$

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

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

Different specific upper bounds for the squared bias can be deduced from (21) using results of Section 3 in [**?**] on bounds in the total variation distance. It is known that under assumptions **(H1)** and **(H2)**, the corresponding Markov chain has a unique stationary distribution π_γ, which is different from π. Moreover, as shown in Proposition, (Appendix)

Eric

put a precise reference here

$$\|\delta_x \mathrm{Q}_\gamma^n - \pi_\gamma\|_{\mathrm{TV}} \leq C_1 \rho^{\gamma n}\left(V(x) + \pi_\gamma(V)\right), \quad \|\pi - \pi_\gamma\|_{\mathrm{TV}} \leq C_2\sqrt{\gamma} \quad (22)$$

for $V(x) = 1 + \|x\|^2$, some $\rho \in (0, 1)$ and constants C_1, $C_2 > 0$ independent of γ and n.

Variance: An upper bound for the variance of the classical estimator (15) under **(H1)** and **(H2)** is derived in Section. In particular, for any bounded function f,

$$\mathsf{Var}[\pi_n^N(f)] \lesssim (n\gamma)^{-1}\,, \tag{23}$$

where $\lesssim$ stands for inequality up to a constant depending on constants L_U, m_U, M_U and $\|f\|_\infty$. One of the main results of this paper is the following upper bound for the variance of $\pi_{K,n}^N(f)$.

Theorem 3 *Assume* **(H1)** *and* **(H2)**. *Suppose additionally that a bounded function f and $\mu = \nabla U$ are $K \times d \geq 2$ times continuously differentiable and for all $x \in \mathbb{R}^d$ and $\mathbf{k}$ satisfying $0 < \|\mathbf{k}\| \leq K$,*

$$|\partial^{\mathbf{k}} f(x)| \leq B_f, \quad |\partial^{\mathbf{k}} \mu(x)| \leq B_\mu\,. \tag{24}$$

Then it holds

$$\mathsf{Var}\left(\pi_{K,n}^N(f)\right) \lesssim n^{-1}\gamma^{K-2}\,, \tag{25}$$

Sergey

removed repeating explanation of $\lesssim$

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

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

where $F_s(X_0, Z_1, \ldots, Z_s) = s^{-1}\sum_{r=0}^{s} f(X_r)$ and $\partial_{Z_1}^{\mathbf{k}'}$ stands for a weak partial derivative of the functional F_s that also can be viewed as discretised version of Malliavin derivative. From (26), we derive

$$\sum_{\mathbf{k}\colon\, \|\mathbf{k}\|\geq K+1} A_{s,\mathbf{k}}^2(x) \leq \sum_{I\subseteq\{1,\ldots,d\},\, I\neq\emptyset}\left(\frac{\gamma}{2}\right)^{|I|K}\mathsf{Var}_x\left(s^{-1}\sum\nolimits_{p=1}^{s}\partial_{y_1}^{\mathbf{K}_I}f\left(X_p\right)\right),$$

where the sum runs over all nonempty subsets I of the set $\{1, \ldots, d\}$ and

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

Finally we show via the Poincare inequality that under smoothness assumption (24) it holds

$$\mathsf{Var}_x\left(s^{-1}\sum\nolimits_{p=1}^{s}\partial_{y_1}^{\mathbf{K}_I}f\left(X_p\right)\right) \leq W_s(x)$$

for all x and some family of functions $(W_s)_{s\geq 1}$ such that the sum $\sum_{s=1}^{n}\mathsf{E}[W_s(X_{N+n-s-1})]$ is bounded uniformly in $n \in \mathbb{N}$.

Eric

Since we do note state was W_s is, it is a bit difficult to guess why we are using a Poincare inequality there

5. Numerical results

We will perform numerical experiments based on ULA with constant step size γ. Suppose that each function $Q_r(x)$ from 14 can be well approximated by polynomials, and we constructed a polynomial approximation for each $Q_r(x)$ in the form:

$$\widehat{Q}_r(x) = \sum_{\|\mathbf{s}\| \le m} \beta_{\mathbf{s}} x^{\mathbf{s}}, \quad s = (s_1, \ldots, s_d)$$

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

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

we obtain closed-form expression for the estimates $\widehat{a}_{r,k}(x)$ of functions $\bar{a}_{r,k}(x)$ from 14. Namely, for all $x \in \mathbb{R}^d$,

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

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

$$P_{r,k,s}(x) = \mathsf{E}\left[H_k(\xi_l)(x - \gamma\mu(x) + \sqrt{\gamma}\xi)^s\right]$$

is a one-dimensional polynomial (in x) of degree at most s with analytically known coefficients. In consequence, it is enough to estimate functions $Q_r(x)$. It can be done using a modified least-squares criteria based on T training trajectories $\left(X_1^{(s)}, \ldots, X_{N+n}^{(s)}\right)$, $s = 1, \ldots, T$ with N being the size of burn-in period:

$$\widehat{Q}_r = \underset{\psi \in \Psi}{\arg\min} \sum_{s=1}^{T} \sum_{l=N+1}^{N+n-r} \left| f(X_{l+r}^{(s)}) - \psi(X_l^{(s)}) \right|^2 \tag{27}$$

for $1 \le r \le n-1$, Ψ being the class of polynomials $\Psi = \{\psi(x) | \psi(x) = \sum_{\|\mathbf{s}\| \le m} \alpha_{\mathbf{s}} x^{\mathbf{s}}\}$ and $\widehat{Q}_0(x) = f(x)$ by definition. Certainly one may use another functional class Ψ in , but using polynomials allows to compute closed-form representation of $\widehat{a}_{r,\mathbf{k}}(x)$.

Due to Lemma **??**, it is enough to estimate Q_r for $r < n_{\text{trunc}}$ for some truncation level n_{trunc} depending on d and γ. It allows us to use a smaller amount of training trajectories to approximate $Q_r(x)$.

Sergey

we need to put some lemma about the truncation point here, I will do it

Finally we construct a truncated version of the estimator 18:

$$\pi_{K,n,n_{\text{trunc}}}^N(f) = \pi_n^N(f) - \widehat{M}_{K,n,n_{\text{trunc}}}^N(f),$$

where

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

5.1. Gaussian mixtures

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

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

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

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

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

6. Proofs

6.1. Proof of Theorem 3

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

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

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

$$G_p(x, y_1, \ldots, y_p) := \Phi(\cdot, y_p) \circ \Phi(\cdot, y_{p-1}) \circ \ldots \circ \Phi(x, y_1) \tag{28}$$

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

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

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

$$\partial_1^{\mathbf{k}} f\left(X_p\right) := \partial_1^{\mathbf{k}} [f \circ G_p](X_0, \sqrt{\gamma} Z_1, \ldots, \sqrt{\gamma} Z_p) \tag{29}$$

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 4 *For any $\mathbf{k},\mathbf{k}' \in \mathbb{N}_0^d$ such that $\mathbf{k}' \leq \mathbf{k}$ componentwise and $\|\mathbf{k}'\| \leq K$, 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 (14).

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

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

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

$$\bar{a}_{p,\mathbf{k}}(x) = \int_{\mathbb{R}^d}\ldots\int_{\mathbb{R}^d} f\circ G_p(x,\sqrt{\gamma}z_1,\ldots,\sqrt{\gamma}z_p)\mathbf{H}_\mathbf{k}(z_1)\varphi(z_1)\prod_{j=2}^p\varphi(z_j)\,\mathrm{d}z_1\ldots\mathrm{d}z_p$$

$$= \frac{1}{\sqrt{\mathbf{k}!}}\int_{\mathbb{R}^d}\ldots\int_{\mathbb{R}^d} f\circ G_p(x,\sqrt{\gamma}z_1,\ldots,\sqrt{\gamma}z_p)(-1)^{|\mathbf{k}|}\partial^\mathbf{k}\varphi(z_1)\prod_{j=2}^p\varphi(z_j)\,\mathrm{d}z_1\ldots\mathrm{d}z_p$$

$$= \frac{\gamma^{|\mathbf{k}'|/2}}{\sqrt{\mathbf{k}!}}\int_{\mathbb{R}^d}\ldots\int_{\mathbb{R}^d} \partial_1^{\mathbf{k}'}[f\circ G_p](x,\sqrt{\gamma}z_1,\ldots,\sqrt{\gamma}z_p)(-1)^{|\mathbf{k}-\mathbf{k}'|}\partial^{\mathbf{k}-\mathbf{k}'}\varphi(z_1)\prod_{j=2}^p\varphi(z_j)\,\mathrm{d}z_1\ldots\mathrm{d}z_p$$

$$= \gamma^{|\mathbf{k}'|/2}\frac{\sqrt{(\mathbf{k}-\mathbf{k}')!}}{\sqrt{\mathbf{k}!}}\mathsf{E}\left[\partial_{y_1}^{\mathbf{k}'}[f\circ G_p](x,\sqrt{\gamma}Z_1,\ldots,\sqrt{\gamma}Z_p)\mathbf{H}_{\mathbf{k}-\mathbf{k}'}(Z_1)\right].$$

The last expression yields the result. □

For multi-indices $\mathbf{k},\mathbf{k}' \in \mathbb{N}_0^d$ with $\mathbf{k}' \leq \mathbf{k}$ componentwise and $\mathbf{k}' \neq \mathbf{k}$, $\|k'\| \leq K$, we get applying first Lemma 4,

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

where $A_{s,\mathbf{k}}$ is defined in (17). 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

$$\sum_{\mathbf{k}:\,\|\mathbf{k}\|\geq K+1} A^2_{s,\mathbf{k}}(x) = \sum_{\mathbf{k}:\,\|\mathbf{k}\|\geq K+1} \left(\gamma^{|\mathbf{k}'|}\frac{(\mathbf{k}-\mathbf{k}')!}{\mathbf{k}!}\right) Q_s(\mathbf{k}',\mathbf{k}-\mathbf{k}')$$
$$= \left\{\sum_{I\subseteq\{1,\ldots,d\},\, I\neq\emptyset}\gamma^{|I|K}\sum_{\mathbf{m}_I\in\mathbb{N}^d_I}\frac{\mathbf{m}_I!}{(\mathbf{m}_I+\mathbf{K}_I)!}\right\}\left\{\sum_{\mathbf{m}_{I^c}\in\mathbb{N}^d_{0,I^c},\,\|\mathbf{m}_{I^c}\|\leq K} Q_s(\mathbf{K}_I,\mathbf{m}_I+\mathbf{m}_{I^c})\right\}, \tag{30}$$

where for any two multi-indices $\mathbf{r}$, $\mathbf{q}$ from $\mathbb{N}^d_0$

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

In (30) the first sum runs over all nonempty subsets I of the set $\{1,\ldots,d\}$. For any subset I, $\mathbb{N}^d_I$ 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 $\mathbb{N}^d_{0,I^c}$ 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

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

we get

$$\begin{aligned}\sum_{\mathbf{k}:\,\|\mathbf{k}\|\geq K+1} A^2_{s,\mathbf{k}}(x) \leq & \sum_{I\subseteq\{1,\ldots,d\},\, I\neq\emptyset} (\gamma/2)^{|I|K} \\ & \times \sum_{\mathbf{m}_I\in\mathbb{N}^d_I}\ \sum_{\mathbf{m}_{I^c}\in\mathbb{N}^d_{0,I^c},\,\|\mathbf{m}_{I^c}\|\leq K} Q(\mathbf{K}_I,\mathbf{m}_I+\mathbf{m}_{I^c}) \\ \leq & \sum_{I\subseteq\{1,\ldots,d\},\, I\neq\emptyset} (\gamma/2)^{|I|K} \sum_{\mathbf{m}\in\mathbb{N}^d_0} Q(\mathbf{K}_I,\mathbf{m}).\end{aligned} \tag{31}$$

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

$$\sum_{\mathbf{m}\in\mathbb{N}^d_0}\{\mathsf{E}[\varphi(Z_1)\mathbf{H}_{\mathbf{m}}(Z_1)]\}^2 \leq \mathsf{E}[\{\varphi(Z_1)\}^2]$$

Using this identity in (31) implies

$$\sum_{\mathbf{k}:\,\|\mathbf{k}\|\geq K+1} A^2_{s,\mathbf{k}}(x) \quad\leq\quad \sum_{I\subseteq\{1,\ldots,d\},\, I\neq\emptyset}\left(\frac{\gamma}{2}\right)^{|I|K}\mathsf{Var}\left(s^{-1}\sum\nolimits_{p=1}^{s}\partial_1^{\mathbf{K}_I} f\left(X_p^x\right)\right)$$

Next we show that under the conditions of Theorem 3

$$\mathsf{Var}\left(s^{-1}\sum_{p=1}^{s}\partial_1^{\mathbf{K}_I} f\left(X_p^x\right)\right) \leq W_s(x)$$

for all x and some family of functions $(W_s)_{s\geq 1}$ such that the sum $\sum_{s=1}^{n}\mathsf{E}[W_s(X_{N+n-s-1})]$ is bounded uniformly in $n \in \mathbb{N}$.

To keep the notational burden at a reasonable level, we present the proof only in one-dimensional case. Multidimensional extension is straightforward but requires involved notations. First we need to prove several auxiliary results.

Lemma 5 *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*

$$0 \leq x_p \leq \alpha_p x_{p-1} + \gamma\epsilon_p, \quad p \in \mathbb{N}, \tag{32}$$

$$0 \leq \epsilon_p \leq \overline{C}_1 \prod_{k=1}^{p}\alpha_k^2, \quad p \in \mathbb{N}, \tag{33}$$

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

$$\gamma\sum_{r=1}^{\infty}\prod_{k=1}^{r}\alpha_k \leq \overline{C}_2 \tag{34}$$

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

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

Proof Applying (32) recursively, we get

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

where we use the convention $\prod_{k=p+1}^{p} := 1$. Substituting estimate (33) into the right-hand side, we obtain

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

which, together with (34), completes the proof. □

In what follows, we use the notation

$$\alpha_k = 1 - \gamma\mu'(X_{k-1}), \quad k \in \mathbb{N}. \tag{35}$$

The assumption (24) implies that $|\mu'(x)| \leq B_\mu$ 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 6 *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*

$$\left|\partial^r_{y_l} X_p\right| \le C_r \prod_{k=l+1}^{p} (1 - \gamma\mu'(X_{k-1})) \quad a.s. \tag{36}$$

where $\partial^r_{y_l} X_p$ is defined in (29). Moreover, we can choose $C_1 = 1$.

Lemma 7 *Under assumptions of Theorem 3, for all natural $r \le K$, $j \ge l$ and $p > j$, we have*

$$\left|\partial_{y_j}\partial^r_{y_l} X_p\right| \le c_r \prod_{k=l+1}^{p} (1 - \gamma\mu'(X_{k-1})), \quad a.s. \tag{37}$$

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

Proof The last statement is straightforward. We fix natural numbers $j \ge l$ and prove (37) for all $p > j$ by induction in r. First, for $p > j$, we write

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

and differentiate this identity with respect to y_j

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

By Lemma 6, we have

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

with a suitable constant. By Lemma 5 applied to bound $|\partial_{y_j}\partial_{y_l} X_p|$ for $p \ge j+1$ (notice that $\partial_{y_j}\partial_{y_l} X_j = 0$, that is, $\overline{C}_0$ in Lemma 5 is zero, while $\overline{C}_1$ in Lemma 5 has the form const $\prod_{k=l+1}^{j} \alpha_k$), we obtain (37) for $r = 1$. The induction hypothesis is now that the inequality

$$\left|\partial_{y_j}\partial^k_{y_l} X_p\right| \le c_k \prod_{s=l+1}^{p} \alpha_s \tag{38}$$

holds for all natural $k < r$ ($\le K$) and $p > j$. We need to show (38) for $k = r$. Faà di Bruno's formula implies for $2 \le r \le K$ and $p > l$

$$\begin{aligned}
\partial^r_{y_l} X_p &= [1 - \gamma\mu'(X_{p-1})]\, \partial^r_{y_l} X_{p-1} \\
&\quad - \gamma \sum \frac{r!}{m_1! \ldots m_{r-1}!} \mu^{(m_1 + \ldots + m_{r-1})}(X_{p-1}) \prod_{k=1}^{r-1} \left(\frac{\partial^k_{y_l} X_{p-1}}{k!}\right)^{m_k},
\end{aligned} \tag{39}$$

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

$$1\cdot m_1+2\cdot m_2+\ldots+(r-1)\cdot m_{r-1}=r. \tag{40}$$

Notice that we work with $(r-1)$-tuples rather than with r-tuples because the term containing $\partial^r_{y_l}X_{p-1}$ on the right-hand side of (39) is listed separately. For $p>j$, we then have

$$\begin{aligned}\partial_{y_j}\partial^r_{y_l}X_p&=[1-\gamma_p\mu'(X_{p-1})]\,\partial_{y_j}\partial^r_{y_l}X_{p-1}-\gamma\mu''(X_{p-1})\partial^r_{y_l}X_{p-1}\partial_{y_j}X_{p-1} \qquad (41)\\&-\gamma\sum\frac{r!}{m_1!\ldots m_{r-1}!}\mu^{(m_1+\ldots+m_{r-1}+1)}(X_{p-1})\partial_{y_j}X_{p-1}\prod_{k=1}^{r-1}\left(\frac{\partial^k_{y_l}X_{p-1}}{k!}\right)^{m_k}\\&-\gamma\sum\frac{r!}{m_1!\ldots m_{r-1}!}\mu^{(m_1+\ldots+m_{r-1})}(X_{p-1})\partial_{y_j}\left[\prod_{k=1}^{r-1}\left(\frac{\partial^k_{y_l}X_{p-1}}{k!}\right)^{m_k}\right]\\&=[1-\gamma\mu'(X_{p-1})]\,\partial_{y_j}\partial^r_{y_l}X_{p-1}+\gamma\epsilon_{l,j,p},\end{aligned}$$

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

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

Using Lemma 6, induction hypothesis (38) and the fact that $m_1+\ldots+m_{r-1}\ge 2$ for $(r-1)$-tuples of nonnegative integers satisfying (40), we can bound $|\epsilon_{l,j,p}|$ as follows

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

with some constant "const" depending on $B_\mu,r,C_1,\ldots,C_r,c_1,\ldots,c_{r-1}$. Thus, (41) now implies

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

We can again apply Lemma 5 to bound $|\partial_{y_j}\partial_{y_l}^r X_p|$ for $p \geq j+1$ (notice that $\partial_{y_j}\partial_{y_l}^r X_j = 0$, that is, $\overline{C}_0$ in Lemma 5 is zero, while $\overline{C}_1$ in Lemma 5 has the form const $\prod_{k=l+1}^{j}\alpha_k$), and we obtain (38) for $k = r$. This concludes the proof. □

Lemma 8 *Under assumptions of Theorem 3, it holds*

$$\mathsf{Var}_x\left[\sum_{p=1}^{s}\partial_{y_1}^K f(X_p)\right] \leq B_K \quad a.s.,$$

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

Proof The expression $\sum_{p=1}^{q}\partial_{y_1}^K f(X_p)$ can be viewed as a deterministic function of $X_0, Z_1, Z_2, \ldots, Z_q$

$$\sum_{p=1}^{q}\partial_{y_1}^K f(X_p) = F(X_0, Z_1, Z_2, \ldots, Z_q).$$

By the (conditional) Gaussian Poincaré inequality, we have

$$\mathsf{Var}_x\left[\sum_{p=1}^{q}\partial_{y_1}^K f(X_p)\right] \leq \mathsf{E}_x\left[\|\nabla_Z F(X_0, Z_1, Z_2, \ldots, Z_q)\|^2\right],$$

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

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

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

$$\mathsf{Var}_x\left[\sum_{p=1}^{q}\partial_{y_1}^K f(X_p)\right] \leq \gamma\sum_{j=1}^{q}\mathsf{E}_x\left[\left(\sum_{p=j}^{q}\gamma\partial_{y_j}\partial_{y_1}^K f(X_p)\right)^2\right]. \tag{42}$$

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

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

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

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

Then

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

Using the bounds of Lemmas 6 and 7, we obtain

$$
\left|\partial_{y_j}\partial_{y_1}^K f\left(X_p\right)\right| \le A_K\prod_{k=2}^{p}\alpha_k \tag{43}
$$

with a suitable constant A_K. Substituting this in (42), we proceed as follows

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

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 $\prod_{k=j+1}^{p}\alpha_k \le \exp\left(-\sum_{k=j+1}^{p}\gamma\mu'(X_{k-1})\right)$, 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

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

From Theorem 11 in Appendix A, it follows that

$$\mathsf{E}\left[e^{-u\gamma\left(\sum_{k=l}^{j}[\mu'(X_{k-1})-\pi_\gamma(\mu')]\right)}\right] \leq e^{u^2\varkappa(j-l-1)\gamma^2}$$

with constant $\varkappa = ...$, where we used the fact that $\mu'(x)$ is bounded. Then we have

$$\mathsf{E}\left[e^{-s\sum_{k=l}^{j}\gamma_k\mu'(X_{k-1})}\right] \leq e^{s^2\varkappa\gamma^2(j-l-1)}e^{-s\gamma(j-l-1)\pi_\gamma(\mu')}.$$

Furthermore, since $\mu\pi', \mu'\pi \in L^1(\mathbb{R})$, we have

Denis

Please finish

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

yielding the final bound

$$\zeta_{l,j}(s) \leq e^{s^2\kappa(j-l-1)\gamma^2}e^{-2s\gamma(j-l-1)\alpha}$$

where $\alpha = \int \mu^2(x)\pi(x)\,dx$. Finally

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

Plugging into bound for conditional variance,

$$\mathsf{Var}\left[\sum_{p=l}^{q}\gamma_{p+1}\partial_{y_l}^{K}f\left(X_p\right)\middle|X_{l-1}\right] \leq \frac{A_K^2 e^{2\kappa\Gamma^*}}{4\alpha^2(1-\gamma_1 B_\mu)^3}$$

The proof is completed. □

Appendix A: Bounds for moments of ULA

We shall show that conditions **(H1)** and **(H2)** are sufficient to show that ULA kernel R_γ satisfies the so-called drift condition:

Definition 1 *We say that the Markov kernel R_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 $x \in X$,*

$$RV(x) \leq \lambda^\gamma V(x) + \gamma C$$

Lemma 9 *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 R_γ from* **??** *satisfies drift condition 1 for any $0 < \gamma < \overline{\gamma} = \frac{m}{4L^2}$ with drift function $V(x) = 1 + \|x\|^2$, constants $\lambda = \exp\left(-\frac{m}{2}\right)$, $C = \frac{25K_1^2}{8} + 2d + m$ with K_1 from* **??**, *m from* **(H2)**.

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

$$R_\gamma V(x) = \int_{\mathbb{R}^d} V(y) P_\gamma(x, dy) = \int_{\mathbb{R}^d} (1 + \|y\|^2) \frac{1}{(4\pi\gamma)^{\frac{d}{2}}} \exp\left(-\frac{\|y - x + \gamma\nabla U(x)\|^2}{4\gamma}\right) dy =$$

$$= 1 + \frac{1}{(4\pi\gamma)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \|z + x - \gamma\nabla U(x)\|^2 \exp\left(-\frac{\|z\|^2}{4\gamma}\right) dz$$

Let us first consider the case $x \notin B(0, K_1)$. Note that

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

and the linear term vanishes after integration, moreover, for $Z \sim \mathcal{N}(0, 2\gamma I_d)$, it holds $\mathbb{E}\|Z\|^2 = 2\gamma d$. It remains to notice that due to **H1** and **H2** (which implies **??**),

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

$$\leq (1 - \gamma m)\|x\|^2 + 2\gamma^2 L^2\|x\|^2$$

Thus, plugging everything into expression for $R_\gamma V$ and using $\gamma < \frac{m}{4L}$, we obtain

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

Now let $x \in B(0, K_1)$. Then simply using $\|x - \gamma\nabla U(x)\|^2 \leq 2(1 + L\gamma)^2\|x\|^2$, we obtain

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

$$\leq \exp^{-\frac{\gamma m}{2}} V(x) + \gamma(\frac{25K_1^2}{8} + 2d + m)$$

□

For now let us assume for simplicity that we run ULA with fixed step size γ. Then it is known that under assumption **(H1)** corresponding Markov chain would have unique stationary distribution π_γ, which is different from π. Yet this chain will be $V-$geometrically ergodic due to [**?**, Theorem 19.4.1]. Namely, the following lemma holds:

Lemma 10 *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*

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

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

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

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

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

Proof Note that the condition **(H1)** implies that the Markov kernel $Q_{\boldsymbol{\gamma}}^{1,n}$ satisfies $(1, \varepsilon)$-Doeblin condition with $\varepsilon = 2\Phi\left(-\frac{\sqrt{d}(1+L\gamma)}{2\sqrt{\gamma}}\right)$. Together with drift condition **??** it allows to apply [**?**, Theorem 19.4.1] with appropriate constants. □

Aforementioned lemmas allow us to bound the exponential moment of the additive functional of ULA. Namely, the following theorem holds

Theorem 11 *Suppose that ULA kernel* **??** *satisfies assumptions* **(H1)** *and* **(H2)***, and let $X_1, \ldots, X_n, \ldots$ be generated by ULA with constant step size γ. Let $X_0 = x$ be fixed. Then for any bounded function $g(x) : \mathbb{R}^d \to \mathbb{R}, |g(x)| \leq M$, it holds*

$$\mathsf{E}\left[\exp\left(-\rho\gamma\sum_{k=l}^{p}(g(X_k) - \pi_\gamma(g))\right)\right] \leq C_1 \exp\left(s^2\kappa\gamma^2(p-l+1)\right)$$

for some absolute constant C_1 not depending on l, p.

Proof Conditions **(H1)** and **(H2)** imply that the Markov kernel R_γ is $V-$ergodic (lemma 10) and satisfies drift condition 9. Due to [**?**, Theorem 3, Fact 3], we obtain the following bound

$$\mathsf{E}_x\left[\exp\left(-s\gamma\sum_{k=l}^{p}(g(X_k) - \mathsf{E}_x g(X_k))\right)\right] \leq \exp\left(s^2\kappa\gamma^2(p-l+1)\right)$$

with constant $\kappa = ...$ Note that

Sergey

I will put precise constant later

$$\mathsf{E}\left[\exp\left(-s\gamma\sum_{k=l}^{p}(g(X_k)-\pi_\gamma(g))\right)\right]=$$

$$=\mathsf{E}_x\left[\exp\left(-s\gamma\sum_{k=l}^{p}(g(X_k)-\mathsf{E}_x g(X_k))\right)\right]\exp\left(s\gamma\sum_{k=l}^{p}(\pi_\gamma(g)-\mathsf{E}_x g(X_k))\right)$$

Using lemma 10,

$$|\mathsf{E}g(X_k)-\pi_\gamma(g)|\leq Md_V(\pi_\gamma,\delta_x Q_{\boldsymbol{\gamma}}^{1,n})\leq CM\rho^k\left(V(x)+\pi_\gamma(V)\right)$$

Hence,

$$\exp\left(s\gamma\sum_{k=l}^{p}(\pi_\gamma(g)-\mathsf{E}_x g(X_k))\right)\leq\exp\left(s\gamma CM(V(x)+\pi_\gamma(V))\frac{\rho^l-\rho^{p+1}}{1-\rho}\right)$$

and the statement follows. □

Let us formulate and prove prove analogue of the previous theorem for inhomogeneous Markov chain with step size $\gamma_1,\ldots,\gamma_n,\ldots$. Note that under assumptions $\sum_{p=1}^{\infty}\gamma_p=\infty$, $\sum_{p=1}^{\infty}\gamma_p^2<\infty$ and **(H1)** the stationary distribution of ULA-based chain will be equal to π. We will need additional assumption

(H3) There exist such constants $\rho>0, K_3>0$ that

$$d_V(\delta_x Q_{\boldsymbol{\gamma}}^{1,n},\pi)\leq K_3V(x)\rho^n$$

Theorem 12 *Suppose that*

$$\sum_{p=1}^{\infty}\gamma_p=\infty,\quad\sum_{p=1}^{\infty}\gamma_p^2<\infty.$$

Under assumptions **(H1)**, **(H2)** *and* **(H3)**, *for any bounded function g on $\mathbb{R}^d$ we have*

$$\mathsf{E}\left[\exp\left(-s\left(\sum_{k=l}^{p}\gamma_{k+1}g(X_k)-\left(\sum_{k=l}^{p}\gamma_{k+1}\right)\pi(g)\right)\right)\right]\leq C \tag{44}$$

for all natural $0<l\leq p<\infty$ with constant C not depending on l,p.. Moreover,

$$\mathsf{Var}_x\left(\sum_{k=l}^{l+n-1}\frac{\gamma_{k+1}}{\Gamma_{l+1,l+n}}g(X_k)-\mathsf{E}_x\sum_{k=l}^{l+n-1}\frac{\gamma_{k+1}}{\Gamma_{l+1,l+n}}g(X_k)\right)\leq\frac{C_1}{\Gamma_{l+1,l+n}} \tag{45}$$

Proof

Note that under assumptions **(H1)**, **(H2)** and **(H3)** it holds due to [**?**, Theorem 4] that

$$\mathsf{E}_x\left[\exp\left(-s\left(\sum_{k=l}^{p}\gamma_{k+1}g(X_k)-\mathsf{E}_x\left(\sum_{k=l}^{p}\gamma_{k+1}g(X_k)\right)\right)\right)\right]\le\exp\left(s^2\kappa\sum_{k=l}^{p}\gamma_{k+1}^2\right)\le\exp\left(s^2\kappa\Gamma^*\right) \tag{46}$$

with $\Gamma^*=\sum_{k=1}^{\infty}\gamma_k^2<\infty$ and $\kappa=\ldots$. Using the same transformation as in theorem 11, and using the bound

Sergey

This theorem is not yet written in the second paper

$$|\mathsf{E}_x f(X_k)-\pi(f)|\le K_3 M\rho^{\Gamma_{1,k}}V(x)$$

we obtain

$$\mathsf{E}\left[\exp\left(-s\left(\sum_{k=l}^{p}\gamma_{k+1}g(X_k)-\left(\sum_{k=l}^{p}\gamma_{k+1}\right)\pi(g)\right)\right)\right]\le\exp\left(s^2\kappa\Gamma^*\right)\exp\left(K_3MV(x)\sum_{k=l}^{p}\rho^{\Gamma_{1,k}}\gamma_{k+1}\right)\le$$

$$\le\exp\left(s^2\kappa\Gamma^*\right)\exp\left(K_3MV(x)\rho^{\Gamma_{1,k}}\int\limits_0^{+\infty}\rho^y\,dy\right)=\exp\left(s^2\kappa\Gamma^*\right)\exp\left(K_3MV(x)\frac{\rho^{\Gamma_{1,k}}}{\ln\rho}\right)$$

□

which proves 44. Note that from 46 by Taylor expansion for small s we obtain

$$1+\frac{s^2}{2}\mathsf{Var}_x\left(\sum_{k=l}^{l+n-1}\gamma_{k+1}g(X_k)-\mathsf{E}_x\sum_{k=l}^{l+n-1}\gamma_{k+1}g(X_k)\right)+O(s^3)\le1+s^2\kappa\sum_{k=l+1}^{l+n}\gamma_k^2+O(s^4)$$

Since it holds for arbitarily small s, the variance is bounded by

$$\mathsf{Var}_x\left(\sum_{k=l}^{l+n-1}\gamma_{k+1}g(X_k)-\mathsf{E}_x\sum_{k=l}^{l+n-1}\gamma_{k+1}g(X_k)\right)\le2\kappa\sum_{k=l+1}^{l+n}\gamma_k^2\le2\kappa\Gamma_{l+1,l+n}$$

Now we need to divide both parts by $\Gamma_{l+1,l+n}^2$ to get 45. The proof is completed.