Martingale representations for Markov chains with application to MCMC ☆

D. Belomestny[a,b,*], E. Moulines[c,b], N. Shagadatov[b], M. Urusov[a]

[a] *Duisburg-Essen University, Essen*
[b] *National University Higher School of Economics, Moscow*
[c] *Centre de Mathématiques Appliquées, UMR 7641, Ecole Polytechnique, France*

Abstract

In this paper we propose an efficient variance reduction approach for MCMC algorithms relying on a novel discrete time martingale representation. Our approach is fully non-asymptotic and does not require any type of ergodicity or special product structure of the underlying density. We rigorously analyze complexity of the proposed algorithm and prove its reduction. The numerical performance is illustrated in the case of Gaussian mixtures and binary regression.

Keywords: Monte Carlo, optimal stopping, regression, boosting
2010 MSC: 65C05, 60F05, 62L10, 65C40, 60J05, 93E35

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 the 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 [1], [2] and [3]. Recently one witnessed a revival of interest in efficient variance reduction methods for MCMC algorithms, see for example [4], [5], [6] and references therein.

Suppose that we wish to compute $\pi(f) := \mathsf{E}[f(X)]$, where X is a random vector-valued in $\mathcal{X} \subseteq \mathbb{R}^d$ with a density π and $f : \mathcal{X} \to \mathbb{R}$ with $f \in L^2(\pi)$. The idea of the so-called control variates variance reduction method is to find a

☆This work has been funded by the Russian Academic Excellence Project '5-100'
*Corresponding author
Email address: denis.belomestny@uni-due.de (D. Belomestny)
URL: www.uni-due.de/~hm0124 (D. Belomestny)

cheaply computable random variable ξ with $\mathsf{E}[\xi] = 0$ and $\mathsf{E}[\xi^2] < \infty$, such that the variance of the r.v. $f(X) - \xi$ is small. The complexity of the problem of constructing classes Ξ of control variates ξ satisfying $\mathsf{E}[\xi] = 0$ 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 [7]. Alternatively, if an orthonormal system in $L^2(\pi)$ is analytically available, one can build control variates ξ as a linear combination of the corresponding basis functions. 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 [5] and further worked out in[8] suggesting control variates depending only on the gradient $\nabla \log \pi$.

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 [9]. 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 so-called Poisson equation for the corresponding Markov chain. It was observed in Henderson [10] 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, provided that $\pi(|G|) < \infty$. The best choice for the function G corresponds to a solution of the so-called Poisson equation $\mathsf{E}[G(X_1)|X_0 = x] - G(x) = -f(x) + \pi(f)$. In fact, the Poisson equation leads to a zero-variance control variate for the empirical mean under π. Moreover, it is also related to the minimal asymptotic variance in the corresponding central limit theorem, see [11] and [5]. Although the Poisson equation involves the quantity of interest $\pi(f)$ and can not be solved explicitly in most cases, this idea still can be used to construct some approximations for the optimal zero-variance control variates. For example, Henderson [10] 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 [4] and [6] series-type control variates are introduced and studied for reversible Markov chains. It is assumed in [4] that the one-step conditional expectations can be computed explicitly for a set of basis functions. The authors in [6] proposed another approach tailored to diffusion setting which doesn't require the computation of integrals of basis functions and only involves applications of the underlying generator.

In this paper we focus on the Langevin type algorithms which got much

attention recently, see [12, 13] 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 explicitely 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 5 studies the complexity of the variance reduced ULA. Finally, numerical examples are presented in Section 6.

2. Setup

Let $\mathcal{X}$ be a domain in $\mathbb{R}^d$. Our aim is to numerically compute the 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, where one is looking for a discrete time (possibly non-homogeneous) Markov chain $(X_p)_{p\geq 0}$ such that π is its unique invariant measure. In this paper we study a class of MCMC algorithms with $(X_p)_{p\geq 0}$ satisfying the the following recurrence relation:

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

for some i.i.d. random vectors $\xi_p \in \mathbb{R}^m$ with distribution P_ξ and some Borel-measurable functions $\Phi_p \colon \mathcal{X} \times \mathbb{R}^m \to \mathcal{X}$. In fact, this is quite general class of Markov chains (see Theorem 1.3.6 in [14]) 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 $\mu = (1/2)\nabla U$ for some continuously differentiable function U, then Markov chain (2) can be used to approximately sample from the density

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

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

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

with nonnegative time steps $(\gamma_p)_{p\geq 1}$, and, under mild technical conditions, the latter Langevin diffusion admits π of (3) as its unique invariant distribution; see [12] and [13].

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 $Y_p \sim q_p(y|x)$;
2. Take

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

where

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

Then, as shown in Metropolis et al. [15], this transition is reversible with respect to π and therefore preserves the stationary density π. If q has 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 [16]. The Metropolis-Adjusted Langevin algorithm (MALA) takes (2) as proposal, that is,

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

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. 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

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

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

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

3. Martingale representation and variance reduction

In this section we give a general discrete-time martingale representation for the chain (1) which later will be used 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

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

Notice that this implies that the random variables $\phi_k(\xi)$, $k \ge 1$, are centered. As an example, we can take multivariate Hermite polynomials for the ULA algorithm and products of Hermite and Legendre polynomials for MALA, as the Legendre polynomials are orthogonal with respect to the Lebesgue measure on $[-1,1]$.

Theorem 1. *Denote by $(\mathcal{G}_p)_{p\in\mathbb{Z}_+}$ a filtration with $\mathcal{G}_0 = \mathrm{triv}$ generated by $(\xi_p)_{p=1,2,\dots,}$. Let f be a Borel function $\mathbb{R}^d \to \mathbb{R}$ such that $\mathsf{E}\left[|f(X_p)|^2\right] < \infty$. Then, for $p > j$, the following representation holds in $L^2(P)$*

$$f(X_p) = \mathsf{E}\left[f(X_p)|\,\mathcal{G}_j\right] + \sum_{k=1}^{\infty}\sum_{l=j+1}^{p} a_{p,l,k}(X_{l-1})\phi_k(\xi_l), \tag{4}$$

where

$$a_{p,l,k}(x) = \mathsf{E}\left[f(X_p)\phi_k(\xi_l)|\,X_{l-1} = x\right], \quad p \ge l, \quad k \in \mathbb{N}. \tag{5}$$

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

$$f(X_1) = \mathsf{E}\left[f(X_1)\right] + \sum_{k\ge1} a_{1,1,k}(X_0)\phi_k(\xi_1)$$

with

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

provided $\mathsf{E}\left[|f(X_1)|^2\right] < \infty$. Recall that $\mathcal{G}_l = \sigma(\xi_1,\dots,\xi_l)$, $l = 1,2,\dots,$ and $\mathcal{G}_0 = \mathrm{triv}$. Suppose that (4) holds for $p = q$, all $j < q$, and all Borel-measurable

functions f with $\mathsf{E}\left[|f(X_q)|^2\right] < \infty$. Let us prove it for $p = q+1$. Given f with $\mathsf{E}\left[|f(X_p)|^2\right] < \infty$, due to the orthonormality and completeness of the system (ϕ_k), we get by conditioning on $\mathcal{G}_q$,

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

where

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

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

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

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

$$f(X_p) = \mathsf{E}\left[\, f(X_p) |\, X_q\right] + \sum_{k\geq 1} a_{p,q+1,k}(X_q)\phi_k(\xi_{q+1}), \tag{6}$$

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

$$\mathsf{E}\left[\, f(X_p) |\, X_q\right] = \mathsf{E}\left[\, f(X_p) |\, X_j\right] + \sum_{k\geq 1}\sum_{l=j+1}^{q} a_{p,l,k}(X_{l-1})\phi_k(\xi_l) \tag{7}$$

with

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

Formulas (6) and (7) conclude the proof.

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

Proposition 2. *The coefficients $a_{p,l,k}$ in (5) 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 $Q_{p,l}(x) = \mathsf{E}\left[\, f(X_p) |\, X_l = x\right]$, $p \geq l$. The functions $Q_{p,l}(x)$ satisfy the recursion

$$Q_{p,l}(x) = \mathsf{E}\left[\, Q_{p,l+1}(X_{l+1}) |\, X_l = x\right], \quad Q_{p,p}(x) = f(x), \tag{8}$$

that is, they can be computed backward via one-step expectations.

Next we show how the representation (4) can be used to efficiently reduce the variance of MCMC algorithms. Let $(\gamma_p)_{p\in\mathbb{N}}$ be a sequence of positive and non-increasing step sizes with $\sum_{p=1}^{\infty}\gamma_p=\infty$ and, for $n,l\in\mathbb{N}$, $n\le l$, we set

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

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

$$\pi_n^N(f)=\sum_{p=N+1}^{N+n}\omega_{p,n}^N f(X_p),\quad \omega_{p,n}^N=\gamma_{p+1}\Gamma_{N+2,N+n+1}^{-1}, \tag{9}$$

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

$$M_{K,n}^N(f)=\sum_{k=1}^{K}\sum_{l=N+1}^{N+n}\bar{a}_{l,k}(X_{l-1})\phi_k(\xi_l), \tag{10}$$

where

$$\bar{a}_{l,n}^{N,k}(x)=\sum_{p=l}^{N+n}\omega_{p,n}^N a_{p,l,k}(x)=\mathsf{E}\left[\left(\sum_{p=l}^{N+n}\omega_{p,n}^N f(X_p)\right)\phi_k(\xi_l)\middle| X_{l-1}=x\right]. \tag{11}$$

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

The coefficients $(\bar{a}_{l,k})$ need to be estimated before one can apply the proposed variance reduction approach. One way of estimating them can be based on nonparametric regression.

Eric: to be discussed should we really present two algorithms

We present two regression algorithms. In both algorithms we first generate T paths conditionally independent of the σ-algebra generated by the burn-in sequence $X_1,\dots,X_N$:

$$\mathcal{T}_{T,n}^N=\left\{(X_{N+1}^{(s)},\dots,X_{N+n}^{(s)}),\quad s=1,\dots,T\right\} \tag{12}$$

of the Markov chain X (the so-called "training paths").

In **Algorithm 1**, we estimate by linear regression the family of functions $\bar{Q}_{l,n}^N$ given for $l=N+1,\dots,N+n$ by

$$\bar{Q}_{l,n}^N(x)=\sum_{p=l}^{N+n}\omega_{p,n}^N Q_{p,l}(x)=\mathsf{E}\left[\sum_{p=l}^{N+n}\omega_{p,n}^N f(X_p)\middle| X_l=x\right]. \tag{13}$$

More precisely, we solve for each $l = N+1, \ldots, N+n-1$. the least squares optimization problems

$$\widehat{Q}_{l,n}^N = \arg\min_{\psi \in \Psi} \sum_{s=1}^{T} \left| \sum_{p=l}^{N+n} \omega_{p,n}^N f(X_p^{(s)}) - \psi(X_l^{(s)}) \right|^2 \tag{14}$$

Next we estimate the coefficients $\bar{a}_{l,n}^{N,k}$ using the formula

$$\hat{a}_{l,n}^{N,k}(x) = \mathsf{E}\left[\phi_k(\xi)\, \widehat{Q}_{l,n}^N\left(\Phi_l(x,\xi)\right) \,\middle|\, \mathcal{T}_{T,n}^N \right], \tag{15}$$

where Φ_l is defined in (1) and Ψ is a class of functions on $\mathbb{R}^d$. and ξ is independent of $\mathcal{T}_{T,n}^N$ with distribution P_ξ. The integration (15) can in many cases be done in closed analytical form. for more details, see Section 6.

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

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

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

4. Analysis of variance reduced ULA

The representation (5) suggests that the variance of the variance-reduced estimator

$$\pi_{K,n}^N(f) = \pi_n^N(f) - M_{K,n}^N(f) \tag{16}$$

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

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

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

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

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 $\mathcal{G}_0 = \mathrm{triv}$ and for $p \in \mathbb{N}$,

$$\mathcal{G}_p = \sigma(Z_1, \ldots, Z_p)\,. \tag{17}$$

Given N and n as above, for $K \in \mathbb{N}$, denote

$$M_{K,n}^N(f) = \sum_{\mathbf{k}:\, 0<\|\mathbf{k}\|\le K} \sum_{p=N+1}^{N+n} \omega_{p,n}^N \sum_{l=N+1}^{p} a_{p,l,\mathbf{k}}(X_{l-1})\mathbf{H}_{\mathbf{k}}(Z_l) \tag{18}$$
$$= \sum_{\mathbf{k}:\, 0<\|\mathbf{k}\|\le K} \sum_{l=N+1}^{N+n} \bar{a}_{l,\mathbf{k}}(X_{l-1})\mathbf{H}_{\mathbf{k}}(Z_l)$$

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

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

For an estimator $\rho(f) \in \{\pi_n^N(f), \pi_{K,n}^N(f)\}$ of $\pi(f)$ (see (9) and (18)), we are interested in its conditional Mean Squared Error (MSE), which can be decomposed as the sum of the squared conditional bias and the conditional variance:

$$\begin{aligned}\mathrm{MSE}\left[\rho(f)|\mathcal{G}_N\right] &= \mathsf{E}\left[(\rho(f)-\pi(f))^2|\mathcal{G}_N\right] \\ &= \left(\mathsf{E}[\rho(f)|\mathcal{G}_N] - \pi(f)\right)^2 + \mathsf{Var}\left[\rho(f)|\mathcal{G}_N\right].\end{aligned} \tag{19}$$

The quantities in (19) are conditioned on $\mathcal{G}_N$, as it reflects the way the estimators are used for MCMC: the path of the Markov chain is simulated only once, and we start to use the realized values of the Markov chain to construct our estimate only after the burn-in period of length N. We also notice that, due to the Markovian structure, the conditioning on $\mathcal{G}_N$ in (19) is equivalent to conditioning on X_N only (this is particularly clear in the case $\rho(f) = \pi_n^N(f)$ but requires some calculation in the remaining case $\rho(f) = \pi_{K,n}^N(f)$).

4.1. Squared conditional bias

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

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

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 (18) with some independent approximations $\widehat{a}_{p,l,\mathbf{k}}$. For a bounded Borel function f, we can estimate the conditional bias similarly to the beginning of [13, Section 4]:

$$\begin{aligned}\left(\mathsf{E}[\pi_{K,n}^N(f)|\mathcal{G}_N] - \pi(f)\right)^2 &= \left(\mathsf{E}[\pi_n^N(f)|\mathcal{G}_N] - \pi(f)\right)^2 \\ &\le \mathrm{osc}(f)^2 \sum_{p=N+1}^{N+n} \omega_{p,n}^N \left\|Q_\gamma^{N+1,p}(X_N,\cdot) - \pi(\cdot)\right\|_{\mathrm{TV}}^2,\end{aligned} \tag{20}$$

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

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

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

$$R_\gamma(x,\cdot) = N\left(x-\gamma\mu(x),\gamma\right),$$

while, for $k,l\in\mathbb{N}$, $k\le l$, the kernel $Q_\gamma^{k,l}$ is

$$Q_\gamma^{k,l} = R_{\gamma_l}\cdots R_{\gamma_k},$$

which, finally, provides the (random) measure $Q_\gamma^{N+1,p}(X_N,\cdot)$ used in the right-hand side of (20).

Different specific upper bounds for the squared bias can be deduced from (20) using results of Section 3 in [13] on bounds in the total variation distance.

4.2. Conditional variance

An upper bound for the variance of the classical estimator (9) is provided in in [13, Theorem 17]. As for estimator (16), it follows from (18) and Proposition 1 applied for $j=N$ that

$$\mathsf{Var}\left[\pi_{K,n}^N(f)\middle|\mathcal{G}_N\right] = \sum_{\mathbf{k}:\ \|\mathbf{k}\|\ge K+1}\ \sum_{l=N+1}^{N+n} \mathsf{E}\left[\left|\bar{a}_{l,\mathbf{k}}(X_{l-1})\right|^2\middle|\mathcal{G}_N\right]. \tag{21}$$

Now we derive an upper bound for the right-hand sides of (21).

Theorem 3. *Fix $K\in\mathbb{N}$. Suppose that f and μ are $K+1$ times continuously differentiable with bounded derivatives*

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

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

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

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

$$\sum_{r=j}^{\infty}\gamma_r\prod_{k=j}^{r}\left[1-\gamma_k b_\mu\right]\le C, \quad \textit{for all } j\in\mathbb{N}, \tag{22}$$

with some constant C. Then it holds

$$\mathsf{Var}\left[\pi^N_{K,n}(f)\middle|\mathcal{G}_N\right] \lesssim \frac{1}{\Gamma^2_{N+2,N+n+1}} \sum_{l=N+1}^{N+n} \sum_{I\subseteq\{1,\dots,d\},\, I\neq\emptyset} \left(\frac{\gamma_l}{2}\right)^{|I|K}, \tag{23}$$

where the sum in (23) *is taken over all subsets I of $\{1,\dots,d\}$ and $\lesssim$ stands for inequality up to a constant not depending on n and N.*

Remark 1. Assumption (22) is not restrictive. For instance, a straightforward calculation shows that (22) is satisfied in most interesting case $\gamma_k = \mathrm{const}/k^\alpha$ with $\alpha \in (0,1)$.

5. Complexity analysis for ULA

We now provide a complexity algorithm of our variance reduced estimator for the ULA algorithm. We focus on our second algorithm where we estimate $\{\bar{Q}^N_{l,n}\}_{l=N}^{N+n}$ by linear regression over K basis functions. We denote by

$$\widetilde{\mathcal{T}}^N_{T,n} = \sigma(\mathcal{T}^N_{T,n}) \vee \mathcal{G}_N\,, \tag{24}$$

where $\mathcal{T}^N_{T,n}$ is the set of training trajectories (see (12)) and $\mathcal{G}_N$ defined in (17).

The first step in the proof is obtained a high-probability bound for the quadratic risk $\mathsf{E}\left[|\bar{Q}^N_{l,n}(X_l) - \widehat{Q}^N_{l,n}(X_l)|^2 \,\middle|\, \widetilde{\mathcal{T}}^N_{T,n}\right]$

Theorem 4. *Suppose that for any $l \in \{N+1,\dots,N+n+1\}$,*

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

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

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

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

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

it holds with probability at least $1-\varepsilon$, for any $l = N+1,\dots,N+n$,

$$\mathsf{E}\left[\left|\bar{Q}^N_{l,n}(X_l) - \widehat{Q}^N_{l,n}(X_l)\right|^2 \middle| \widetilde{\mathcal{T}}^N_{T,n}\right] \lesssim \sigma_l^2 B\left(\frac{BD+\log(2/\varepsilon)}{T} + \frac{B^2D^2}{T^2}\right) + \inf_{\psi\in\Psi} \mathsf{E}\left[\left|\bar{Q}^N_{l,n}(X_l) - \psi(X_l)\right|^2 \middle| \widetilde{\mathcal{T}}^N_{T,n}\right], \tag{25}$$

with $\lesssim$ standing for inequality up to a universal multiplicative constant.

PROOF. The proof follows from [17, Theorem 2.2].

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

Proposition 5. *Under conditions of Theorem 4, we have with probability at least* $1-\varepsilon$,

$$\begin{aligned}
&\mathsf{E}\left[\left|\widehat{M}^N_{K,n}(f) - M^N_{K,n}(f)\right|^2 \middle| \widetilde{\mathcal{T}}^N_{T,n}\right] \qquad (26)\\
&\qquad \lesssim K\left[\sum_{l=N+1}^{N+n}\sigma_l^2\right] B\left(\frac{BD+\log(2/\varepsilon)}{T} + \frac{B^2D^2}{T^2}\right)\\
&\qquad + K\sum_{l=n+1}^{N+n}\inf_{\psi\in\Psi}\mathsf{E}\left[\left|\bar{Q}_l(X_l) - \psi(X_l)\right|^2 \middle| \mathcal{G}_N\right].
\end{aligned}$$

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

$$\mathsf{E}\left[\left|\widehat{a}_{l,k}(X_{l-1}) - \bar{a}_{l,k}(X_{l-1})\right|^2 \middle| \widetilde{\mathcal{T}}^N_{T,n}\right] \leq \mathsf{E}\left[\left|\widehat{Q}^N_{l,n}(X_l) - \bar{Q}^N_{l,n}(X_l)\right|^2 \middle| \widetilde{\mathcal{T}}^N_{T,n}\right]$$

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

$$\begin{aligned}
&\mathsf{E}\left[\left|\widehat{M}^N_{K,n}(f) - M^N_{K,n}(f)\right|^2 \middle| \widetilde{\mathcal{T}}^N_{T,n}\right]\\
&\qquad\qquad \leq \sum_{1\leq k\leq K}\sum_{l=N+1}^{N+n}\mathsf{E}\left[\left|\widehat{a}_{l,k}(X_{l-1}) - \bar{a}_{l,k}(X_{l-1})\right|^2 \middle| \widetilde{\mathcal{T}}^N_{T,n}\right].
\end{aligned}$$

In the case of ULA one can bound the quantities σ_l^2 using Poincaré moment inequalities (see [18]) similar to the proof of Lemma 11. In particular, we can derive that under conditions of Theorem 3 with $K=0$,

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

where $\lesssim$ stands for the inequality up to a multiplicative constant not depending on n and N. Using this inequality and combining (26) with Theorem 3, we get for the variance of $\widehat{\pi}^N_{K,n}(f) = \pi^N_n(f) - \widehat{M}^N_{K,n}(f)$, with probability at least $1-\varepsilon$

$$\begin{aligned}
\mathsf{Var}\left[\widehat{\pi}^N_{K,n}(f)\middle|\, \mathcal{T}^N_{T,n}, \mathcal{G}_N\right] \quad \lesssim \quad & \frac{nKB}{\Gamma^2_{N+2,N+n+1}}\left(\frac{BD+\log(2/\varepsilon)}{T} + \frac{B^2D^2}{T^2}\right)\\
&+\frac{1}{\Gamma^2_{N+2,N+n+1}}\sum_{l=N+1}^{N+n}\sum_{I\subseteq\{1,\ldots,d\},\, I\neq\emptyset}\left(\frac{\gamma_l}{2}\right)^{|I|K}\\
&+K\sum_{l=n+1}^{N+n}\inf_{\psi\in\Psi}\mathsf{E}\left[\left|\bar{Q}_l(X_l) - \psi(X_l)\right|^2\middle|\, \mathcal{G}_N\right]. \qquad (27)
\end{aligned}$$

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

Eric: to be checked: formulation of Lemma 6 in the vector case

Lemma 6. *Suppose that $f, \mu \in C^1(\mathbb{R})$. Assume that there exist $b_\mu > 0$, $B_\mu < \infty$ such that, for all $x \in \mathbb{R}^d$, $0 < b_\mu \, \mathrm{I} \leq \nabla\mu(x)$, $|\nabla\mu(x)| \leq B_\mu$, and $B_\mu \gamma_l \leq 1$. Then*

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

Corollary 1. Assume that $\gamma_k = \gamma_1 k^{-\alpha}$ for some $\alpha \in (0,1)$, then

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

for some constant $c > 0$.

Suppose that the chain is close to stationarity, then $Q_{p,l}$ are functions of $p-l$ only. Furthermore, according to Lemma 6, $a_{p,l,\mathbf{k}}$ are exponentially small for $p-l$ large. As a result we only need to compute a logarithmic (in n) number of functions $Q_{p,l}$. Hence the cost of computing the estimates $\widehat{a}_{p,l,\mathbf{k}}(x)$ for $l = N+1, \ldots, N+n$ and $\|\mathbf{k}\| \leq K$ using regression on Ψ, is of order

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

Suppose for simplicity that all functions $\bar{Q}^N_{l,n}$ are in Ψ for some $D > 0$, that is, the third term in (27) is zero. Then it is enough to take $K = \lceil 1/\alpha \rceil + 1$ to get

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

for some constant not depending on d. As a result we have with high probability

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

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

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

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

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

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

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

6. Numerical results

6.1. Polynomial approximation

In this section we apply polynomial regression to approximate functions $\bar{Q}^N_{l,n}(x)$ appearing in (13). In this case one can derive explicit formula for the coefficients $\bar{a}^{N,k}_{l,n}(x)$ in the case of ULA with constant step size. Suppose we constructed a polynomial approximation for $\widehat{Q}^N_{l,n}(x)$ of the form:

$$\widehat{Q}^N_{l,n}(x) = \sum_{\|\mathbf{s}\| \le m} \hat{\beta}^N_{l,n,\mathbf{s}} x^{\mathbf{s}}, \quad s = (s_1, \dots, s_d)$$

where the coefficients $\hat{\beta}^N_{l,n,\mathbf{s}}$ have been obtained by solving the least-square regression problem on a polynomial basis. Then using the identity

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

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

Eric: simplify this expression

$$\hat{a}^{N,\mathbf{k}}_{l,n}(x) = \mathsf{E}\left[\mathbf{H}_{\mathbf{k}}(x) \widehat{Q}^N_{l,n}(x - \gamma\mu(x) + \sqrt{\gamma}\xi) \,\middle|\, \widetilde{\mathcal{T}}^N_{T,n} \right] = \sum_{\|\mathbf{s}\| \le m} \hat{\beta}^N_{l,n,\mathbf{s}} \prod_{i=1}^{d} E_{i,k_i,s_i}(x)$$

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

Eric: simplify the expression

$$\begin{aligned} E_{i,k,s}(x) &= \mathsf{E}\left[H_k(\xi_i)(x_i - \gamma\mu_i(x) + \sqrt{\gamma}\xi_i)^s \right] \\ &= \sum_{j=0}^{s} \sum_{r=0}^{j/2} j! \frac{1}{2^r} \frac{1}{r!\sqrt{(t-2r)!}} \binom{s}{j} \left[x_i - \gamma\mu_i(x)\right]^{s-j} \gamma^{j/2} \int_{\mathbb{R}} H_k(y) H_{j-2r}(y) \varphi(y)\, dy \end{aligned}$$

and

$$\int_{\mathbb{R}} H_{k_i}(y) H_{j-2r}(y) \varphi(y)\, dy = \delta_{k_i, j-2r}.$$

6.2. Gaussian mixtures

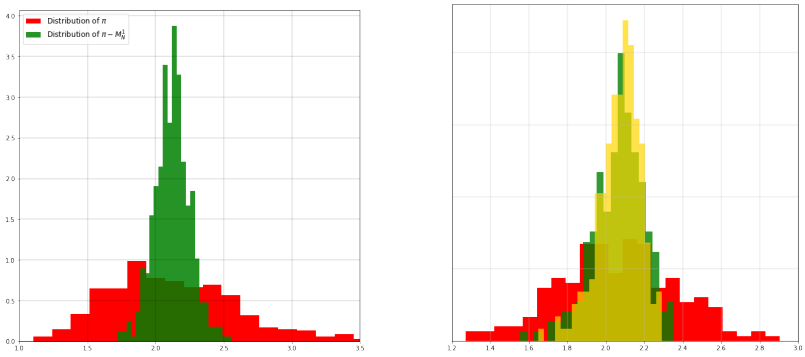
We consider a sample generated by ULA with π given by the mixture of two Gaussian distributions with equal weights:

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

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

Eric: inconsistent notations withe the norm

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



(a) d = 1 (b) d = 2

Figure 1: Histograms for Gaussian mixture. (a) 1-dimensional GM model: histograms of estimators for target function $f(x) = \mathrm{e}^x$ on test sample (200 independent trajectories obtained by ULA algorithm). *Red* bins correspondent to ordinary weighted estimators $\pi_n^N(f)$, *green* - variance-reduced estimators $\pi_{1,n}^N(f)$.
(b) 2-dimensional GM model: histograms of estimators for target function $f(x) = x_1^2 + x_2^2 - \cos(x_1)$, *red* bins: $\pi_n^N(f)$, *green*: $\pi_{1,n}^N(f)$, *yellow*: $\pi_{2,n}^N(f)$.

and

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

respectively. In our experiments we considered dimensions $d = 1$ and $d = 2$ and defined vector a as $((2d)^{-1/2}, \ldots, (2d)^{-1/2})$. For ULA we used constant step sizes $\gamma_i = 0.2$ and $n = 1000$. In order to approximate coefficients $a_{p,l,\mathbf{k}}(x)$, we generated $T = 500$ independent "training" trajectories and solved the least squares problems (14) with polynomial basis functions with maximum degree 5 and 3 for dimensions 1 and 2, respectively. More precisely we used the polynomials

Eric

It is not clear which algorithm you are trying to illustrate; bowplots are better I guess

$$\begin{aligned} \Psi^1 &= \left\{1, x, x^2, x^3, x^4, x^5\right\}, \text{ for } d = 1, \\ \Psi^2 &= \left\{1, x_1, x_2, x_1x_2, x_1^2, x_2^2, x_1^2x_2, x_1x_2^2, x_1^3, x_2^3\right\}, \text{ for } d = 2. \end{aligned}$$

We fixed $K = 1$ for $d = 1$ and $K = 2$ for $d = 2$. To test our variance reduction algorithm, we generated $N_{\text{test}} = 200$ independent paths and computed empirical variance of the new variance reduced estimator $\pi_{K,n}^N(f)$ of target functions $f(x) = \mathrm{e}^x$ for $d = 1$ and $f(x) = x_1^2 + x_2^2 - \cos(x_1)$ for $d = 2$. Figure 1 shows the histograms of weighted average estimator $\pi_n^N(f)$ and variance reduced estimator $\pi_{K,n}^N(f)$ computed on test sample. We have repeated the whole experiment 5 times and presented the results in Table 1. Eventually, we can see that new estimator has considerably reduced variance in comparison with ordinary estimator.

Eric

far from being enough ! make it 100 times at least

In order to illustrate the dependence of variances of the proposed variance reduced estimator on the number of elements in trajectory, we report in Figure

	$d=1$				
$\mathsf{Var}(\pi_n^N)$	0.28641	0.24557	0.26398	0.27346	0.27845
$\mathsf{Var}(\pi_{1,n}^N)$	0.02046	0.02068	0.02789	0.02218	0.01957
	$d=2$				
$\mathsf{Var}(\pi_n^N)$	0.09238	0.10268	0.09458	0.09024	0.09312
$\mathsf{Var}(\pi_{1,n}^N)$	0.01800	0.02156	0.01961	0.01288	0.01754
$\mathsf{Var}(\pi_{2,n}^N)$	0.01155	0.01341	0.01097	0.00842	0.01018

Table 1: Gaussian Mixtures: Empirical variances of ordinary weighted and variance-reduced estimators on test sample.

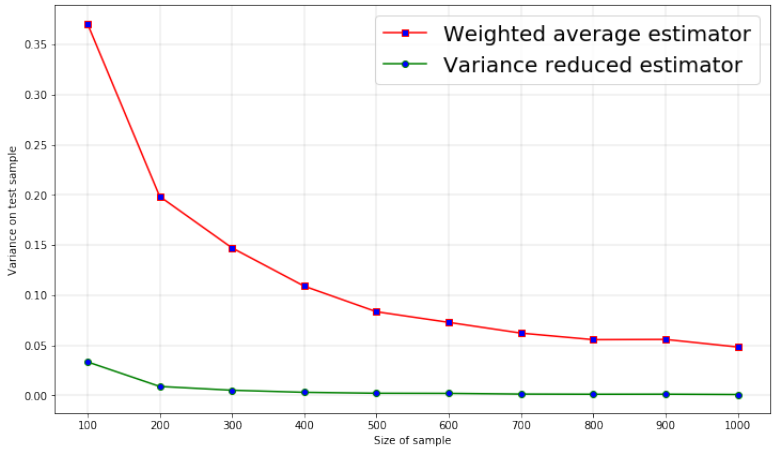


Figure 2: 1-dimensional GM model. Vertical axis is empirical variance on test sample, horizontal axis is the length of test trajectories obtained by ULA algorithm.Red traceplot correspondents to ordinary weighted estimator, green traseplot illustrates empirical variances of variance-reduced estimator for K=1.

2 the traceplots of the empirical variances versus n for the case of the one-dimensional Gaussian mixture and $K = 1$. One may observe that the sample size needed to achieve the "almost zero" variance is much smaller for the variance reduced estimator $\pi_{K,n}^N(f)$ than for the ordinary weighted average estimator $\pi_n^N(f)$.

Eric

what do you mean by that ?

6.3. Binary Logistic regression

Second experiment considers the problem of logistic regression, similar to that considered by Dalalyan [12]. Suppose we have i.i.d. sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^m$ with features $\mathbf{X}_i \in \mathbb{R}^p$ and binary labels $Y_i \in \{0, 1\}$. The binary logistic regression model defines the conditional distribution of Y given $\mathbf{X}$ as Bernoulli random variable with success probability specified by the logistic function

$$r(\theta, \mathbf{X}) = \frac{\mathrm{e}^{\theta^T \mathbf{X}}}{1 + \mathrm{e}^{\theta^T \mathbf{X}}} \tag{28}$$

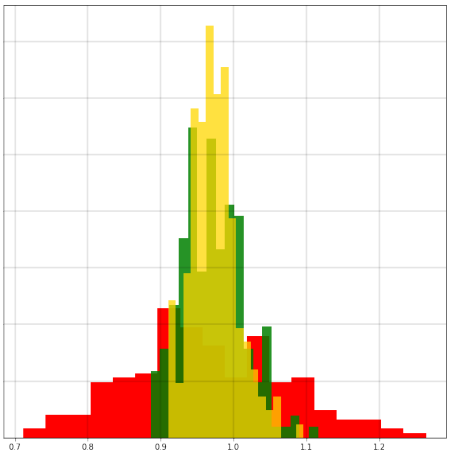


Figure 3: Binary Logistic Regression: Histograms of estimators for target function $f(\theta) = 2\theta_1^2 + 7\theta_2^2$ on test sample. *Red* bins correspondent to ordinary weighted estimators $\pi_n^N(f)$, *green* - variance-reduced estimators $\pi_{1,n}^N(f)$and *yellow* - $\pi_{2,n}^N(f)$.

where θ is parameter of model. We use a Bayesian setting and denote by $\pi_0(\theta)$ the prior distribution. We are willing to sample the posterior density $\pi_m(\theta)$ of the parameter θ given the observations $\{(\mathbf{X}_i, Y_i)\}_{i=1}^m$. Taking a Gaussian prior π_0 with zero mean and covariance matrix proportional to the inverse of the Gram matrix $\Sigma_{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^m \mathbf{X}_i\mathbf{X}_i^T$ the posterior density takes the form

Eric

I think this is called a Zellner prior

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

where $\mathbf{Y} = (Y_1, \ldots, Y_m)^T \in \{0,1\}$ and $\lambda > 0$ additional parameter specified by practitioner. Denote

$$U(\theta) = \mathbf{Y}^T\mathbf{X}\theta + \sum_{i=1}^m \log(1+\mathrm{e}^{-\theta^T\mathbf{X}_i}) + \frac{\lambda}{2}\left\|\Sigma_{\mathbf{X}}^{1/2}\theta\right\|_2^2$$

$$\nabla U(\theta) = \mathbf{X}^T\mathbf{Y} - \sum_{i=1}^m \frac{\mathbf{X}_i}{1+\mathrm{e}^{\theta^T\mathbf{X}_i}} + \lambda\Sigma_{\mathbf{X}}\theta\,.$$

In our second experiment, we randomly generated m independent samples as in paper [12], more precisely features $\mathbf{X}_i$ were generated from a Rademacher distribution and then normalized to have an norm equal to one. Each target variable Y_i has been obtained from a Bernoulli distribution with parameter $r(\theta_{\text{true}}, \mathbf{x})$, where θ_{true} is defined as $(1, \ldots, 1)^T$. We fix $d = 2$ and generated $m = 50$ samples according Rademacher distribution. To construct trajectories of length $n = 500$ we determined constant step size $\gamma_i = 0.02$ for ULA scheme. As

	$d=2$				
$\mathrm{Var}(\pi_n^N)$	0.03727	0.02124	0.04769	0.01147	0.02974
$\mathrm{Var}(\pi_{1,n}^N)$	0.00269	0.00224	0.00306	0.00179	0.00196
$\mathrm{Var}(\pi_{2,n}^N)$	0.00182	0.00117	0.00213	0.00100	0.00107

Table 2: BLR: Empirical variance of ordinary weighted and variance-reduced estimators on test sample.

in previous experiment we use polynomials approximations to explicitly compute $a_{p,l,k}$ and fixed $T = 300$, $N_{test} = 200$ and $K = 2$. The target function defined as

Eric

λ is not specified

$$f(\theta) = 2\theta_1^2 + 7\theta_2^2$$

Table 2 summarizes results of conventional weighted estimator and variance reduced estimator.

7. Proofs

7.1. Proof of Theorem 3

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

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

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

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

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

$$G_{p,l}(x, y_l, \dots, y_p) := \Phi_p(\cdot, y_p) \circ \Phi_{p-1}(\cdot, y_{p-1}) \circ \dots \circ \Phi_l(x, y_l) \tag{29}$$

with

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

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

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

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

$$\partial_{y_l}^{\mathbf{k}} f(X_p) := \partial_{y_l}^{\mathbf{k}} [f \circ G_{p,l}](X_{l-1}, \sqrt{\gamma_l} Z_l, \dots, \sqrt{\gamma_p} Z_p) \tag{30}$$

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

Lemma 7. *Fix $l \le p$ and some $\mathbf{k}, \mathbf{k}' \in \mathbb{N}_0^d$ with $\mathbf{k}' \le \mathbf{k}$ componentwise. Then the following representation holds*

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

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

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

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

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

The last expression yields the result.

For multi-indices $\mathbf{k}, \mathbf{k}' \in \mathbb{N}_0^d$ with $\mathbf{k}' \le \mathbf{k}$ componentwise and $\mathbf{k}' \ne \mathbf{k}$, we get applying first Lemma 7

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

Assume that μ and f are $K \times d$ times continuously differentiable. Then, given $\mathbf{k} \in \mathbb{N}_0^d$, by taking $\mathbf{k}' = \mathbf{k}'(\mathbf{k}) = (K1_{\{k_1>K\}}, \dots, K1_{\{k_d>K\}})$, for each $l \in \{N+1, \dots, N+n\}$, we get

$$\sum_{\mathbf{k}\colon \|\mathbf{k}\|\geq K+1} \left[\bar{a}_{l,\mathbf{k}}^2(X_{l-1})|\mathcal{G}_N\right] = \sum_{\mathbf{k}\colon \|\mathbf{k}\|\geq K+1} \left(\gamma_l^{|\mathbf{k}'|}\frac{(\mathbf{k}-\mathbf{k}')!}{\mathbf{k}!}\right) Q(\mathbf{k}', \mathbf{k}-\mathbf{k}') \quad (31)$$

$$= \sum_{I\subseteq\{1,\dots,d\},\, I\neq\emptyset} \gamma_l^{|I|K} \sum_{\mathbf{m}_I\in\mathbb{N}_I^d} \frac{\mathbf{m}_I!}{(\mathbf{m}_I+\mathbf{K}_I)!} \times \sum_{\mathbf{m}_{I^c}\in\mathbb{N}_{0,I^c}^d,\, \|\mathbf{m}_{I^c}\|\leq K} Q(\mathbf{K}_I, \mathbf{m}_I+\mathbf{m}_{I^c}),$$

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

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

In (31) the first sum runs over all nonempty subsets I of the set $\{1,\dots,d\}$. For any subset I, $\mathbb{N}_I^d$ stands for a set of multi-indices $\mathbf{m}_I$ with elements $m_i = 0$, $i \notin I$, and $m_i \in \mathbb{N}$, $i \in I$. Moreover, $I^c = \{1,\dots,d\}\setminus I$ and $\mathbb{N}_{0,I^c}^d$ stands for a set of multi-indices $\mathbf{m}_{I^c}$ with elements $m_i = 0$, $i \in I$, and $m_i \in \mathbb{N}_0$, $i \notin I$. Finally, the multi-index $\mathbf{K}_I$ is defined as $\mathbf{K}_I = (K1_{\{1\in I\}}, \dots, K1_{\{d\in I\}})$. Applying the estimate

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

we get

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

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

$\{\mathbf{H_m}(Z_l)\}_{m\in\mathbb{N}_0^d}$ is orthonormal conditionally on X_{l-1}), we derive

$$\begin{aligned}\sum_{\mathbf{k}:\ \|\mathbf{k}\|\geq K+1}\left[\bar{a}^2_{l,\mathbf{k}}(X_{l-1})|\mathcal{G}_N\right] &\leq \sum_{I\subseteq\{1,\ldots,d\},\, I\neq\emptyset}\left(\frac{\gamma_l}{2}\right)^{|I|K}\\ &\quad\times\mathsf{E}\left[\mathsf{Var}\left(\sum_{p=l}^{N+n}\omega^N_{p,n}\partial^{\mathbf{K}_I}_{y_l}f\left(X_p\right)\Big|X_{l-1}\right)\Big|\mathcal{G}_N\right]\\ &= \frac{1}{\Gamma^2_{N+2,N+n+1}}\sum_{I\subseteq\{1,\ldots,d\},\, I\neq\emptyset}\left(\frac{\gamma_l}{2}\right)^{|I|K}R^{I,K}_{l,n,N}\end{aligned}$$

with

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

As a result

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

Next result we show that under the conditions of Theorem 3, the quantity $R^{I,K}_{l,n,N}$ is uniformly bounded in l, n, N, I. First we need to prove several auxiliary results.

Lemma 8. *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*

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

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

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

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

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

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

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

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

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

In what follows, we use the notation

$$\alpha_k = 1 - \gamma_k b_\mu, \quad k \in \mathbb{N}. \tag{35}$$

Remark 2. Notice that, under (22), not only (34) but also

$$\sum_{r=j}^{\infty} \gamma_r \prod_{k=j}^{r} \alpha_k \le \overline{C}_2 \tag{36}$$

is satisfied with the same constant $\overline{C}_2$ (which is C of (22)) simultaneously for all $j \in \mathbb{N}$. Below this will allow us to apply Lemma 8 to bound double indexed sequences $(x_{j,p})_{j\ge 1, p\ge j}$ satisfying

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

with suitable $(\epsilon_{j,p})_{j\ge 1, p\ge j+1}$ and the constant $\overline{C}_2$ in (36) being independent of j.

Lemma 9. *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*

$$\left|\partial_{y_l}^r X_p\right| \le C_r \prod_{k=l+1}^{p} \alpha_k \,, \tag{37}$$

where $\partial_{y_l}^r X_p$ is defined in (30)*. Moreover, we can choose $C_1 = 1$.*

Proof. The proof is along the same lines as Lemma 10.

Lemma 10. *Under assumptions of Theorem 3, for all natural $r \le K$, $j \ge l$ and $p > j$, we have*

$$\left|\partial_{y_j} \partial_{y_l}^r X_p\right| \le c_r \prod_{k=l+1}^{p} \alpha_k \tag{38}$$

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

Proof. The last statement is straightforward. We fix natural numbers $j \ge l$ and prove (38) for all $p > j$ by induction in r. First, for $p > j$, we write

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

and differentiate this identity with respect to y_j

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

By Lemma 9, we have

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

with a suitable constant (we can take, e.g., const $= \frac{B_\mu}{(1-\gamma_1 b_\mu)^2}$). By Lemma 8 together with Remark 2 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 8 is zero, while $\overline{C}_1$ in Lemma 8 has the form const $\prod_{k=l+1}^{j}\alpha_k$), we obtain (38) for $r=1$.

The induction hypothesis is now that the inequality

$$\left|\partial_{y_j}\partial_{y_l}^k X_p\right| \le c_k \prod_{s=l+1}^{p}\alpha_s \tag{39}$$

holds for all natural $k < r$ ($\le K$) and $p > j$. We need to show (39) for $k=r$. Faà di Bruno's formula implies for $2\le r\le K$ and $p>l$

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

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

$$1\cdot m_1 + 2\cdot m_2 + \dots + (r-1)\cdot m_{r-1} = r. \tag{41}$$

Notice that we work with $(r-1)$-tuples rather than with r-tuples because the term containing $\partial_{y_l}^r X_{p-1}$ on the right-hand side of (40) is listed separately. For $p>j$, we then have

$$\begin{aligned}\partial_{y_j}\partial_{y_l}^r X_p &= [1-\gamma_p\mu'(X_{p-1})]\,\partial_{y_j}\partial_{y_l}^r X_{p-1} - \gamma_p\mu''(X_{p-1})\partial_{y_l}^r X_{p-1}\partial_{y_j}X_{p-1} \\ &\quad - \gamma_p\sum\frac{r!}{m_1!\dots m_{r-1}!}\mu^{(m_1+\dots+m_{r-1}+1)}(X_{p-1})\partial_{y_j}X_{p-1}\prod_{k=1}^{r-1}\left(\frac{\partial_{y_l}^k X_{p-1}}{k!}\right)^{m_k} \\ &\quad - \gamma_p\sum\frac{r!}{m_1!\dots m_{r-1}!}\mu^{(m_1+\dots+m_{r-1})}(X_{p-1})\partial_{y_j}\left[\prod_{k=1}^{r-1}\left(\frac{\partial_{y_l}^k X_{p-1}}{k!}\right)^{m_k}\right] \\ &= [1-\gamma_p\mu'(X_{p-1})]\,\partial_{y_j}\partial_{y_l}^r X_{p-1} + \gamma_p\epsilon_{l,j,p},\end{aligned} \tag{42}$$

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

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

Using Lemma 9, induction hypothesis (39) and the fact that $m_1+\ldots+m_{r-1}\ge 2$ for $(r-1)$-tuples of nonnegative integers satisfying (41), we can bound $|\epsilon_{l,j,p}|$ as follows

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

with some constant "const", which is, in fact, $\frac{1}{(1-\gamma_1 b_\mu)^2}$ times the expression involving $B_\mu, r, C_1,\ldots,C_r,c_1,\ldots,c_{r-1}$. Thus, (42) now implies

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

We can again apply Lemma 8 and Remark 2 to bound $|\partial_{y_j}\partial^r_{y_l}X_p|$ for $p\ge j+1$ (notice that $\partial_{y_j}\partial^r_{y_l}X_j=0$, that is, $\overline{C}_0$ in Lemma 8 is zero, while $\overline{C}_1$ in Lemma 8 has the form const $\prod_{k=l+1}^{j}\alpha_k$), and we obtain (39) for $k=r$. This concludes the proof.

Lemma 11. *Under assumptions of Theorem 3, for all natural $l\le q$, it holds*

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

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

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

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

By the (conditional) Gaussian Poincaré inequality, we have

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

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

$$\partial_{Z_j} F = \sqrt{\gamma_j}\, \partial_{y_j} F,$$

hence,

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

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

$$\mathsf{Var}\left[\sum_{p=l}^{q} \gamma_{p+1} \partial_{y_l}^K f\left(X_p\right) \Big| X_{l-1}\right] \le \sum_{j=l}^{q} \gamma_j \mathsf{E}\left[\left(\sum_{p=j}^{q} \gamma_{p+1} \partial_{y_j} \partial_{y_l}^K f\left(X_p\right)\right)^2 \Big| X_{l-1}\right]. \tag{43}$$

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

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

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

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

Then

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

Using the bounds of Lemmas 9 and 10, we obtain

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

with a suitable constant A_K. Substituting this in (43), we proceed as follows

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

where C is the bound from (22). The proof is completed.

7.2. Proof of Lemma 6

We have for any $\mathbf{k}\neq\mathbf{0}$ and $x\in\mathbb{R}^d$, we get

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

showing that $\|a_{p,l,\mathbf{k}}\|_\infty \leq \sqrt{\gamma_l}\|\nabla Q_{p,l}\|_\infty$. Since $f,\mu\in C^1(\mathbb{R}^d)$, we have

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

Hence we have $\|\nabla Q_{p,l}\|_\infty \leq (1-b_\mu\gamma_{l+1})\|\nabla Q_{p,l+1}\|_\infty$, which implies that

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

which concludes the proof.

[1] C. Robert, G. Casella, Monte Carlo statistical methods (1999).

[2] R. Y. Rubinstein, D. P. Kroese, Simulation and the Monte Carlo method, Vol. 10, John Wiley & Sons, 2016.

[3] P. Glasserman, Monte Carlo methods in financial engineering, Vol. 53, Springer Science & Business Media, 2013.

[4] P. Dellaportas, I. 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) (2012) 133–161.

[5] A. Mira, R. Solgi, D. Imparato, Zero variance markov chain Monte Carlo for bayesian estimators, Statistics and Computing 23 (5) (2013) 653–662.

[6] N. Brosse, A. Durmus, S. Meyn, E. Moulines, Diffusion approximations and control variates for mcmc, arXiv preprint arXiv:1808.01665.

[7] I. T. Dimov, Monte Carlo methods for applied scientists, World Scientific, 2008.

[8] C. J. Oates, M. Girolami, N. Chopin, Control functionals for monte carlo integration, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79 (3) (2017) 695–718.

[9] D. Belomestny, S. Häfner, M. Urusov, Variance reduction for discretised diffusions via regression, Journal of Mathematical Analysis and Applications 458 (2018) 393–418.

[10] S. G. Henderson, Variance reduction via an approximating markov process, Ph.D. thesis, Stanford University (1997).

[11] A. B. Duncan, T. Lelievre, G. Pavliotis, Variance reduction using nonreversible Langevin samplers, Journal of statistical physics 163 (3) (2016) 457–491.

[12] A. 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) (2017) 651–676.

[13] A. Durmus, E. Moulines, Nonasymptotic convergence analysis for the unadjusted Langevin algorithm, Ann. Appl. Probab. 27 (3) (2017) 1551–1587. doi:10.1214/16-AAP1238.
URL https://doi.org/10.1214/16-AAP1238

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

[15] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of state calculations by fast computing machines, The journal of chemical physics 21 (6) (1953) 1087–1092.

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

[17] J.-Y. Audibert, O. Catoni, et al., Robust linear least squares regression, The Annals of Statistics 39 (5) (2011) 2766–2794.

[18] S. Aida, D. Stroock, Moment estimates derived from poincaré and logarithmic sobolev inequalities, Mathematical Research Letters 1 (1) (1994) 75–86.