Variance reduction for MCMC methods via martingale representations

Belomestny D. [1,3], Moulines E.[2,1], **Samsonov S.**[1*], Shagadatov N. [1]

[1]HDI Lab, Higher School of Economics, Moscow
[2]Ecole Polytechnique
[3]University of Duisburg-Essen
[*]Contact email: svsamsonov@hse.ru

HDI Lab

Introduction

MCMC methods are often the only reasonable way to sample from the distribution of interest, especially in high dimensions. But MCMC algorithms are known to suffer from high variance, hence some variance reduction techniques are called for.
In this work we introduce the new variance reduction method for Markov Chains based on discrete time martingale representation. Proposed approach is fully non-asymptotic and does not require any type of ergodicity or special product structure of the underlying density.

MCMC and control variates

- Suppose that we are willing to estimate expectation of some function f w.r.t. measure π:

$$\pi(f) := \int_{\mathbb{R}^d} f(x)\,\pi(dx)$$

- MCMC approach: based on samples $X_1, \dots, X_{N+n}$ from appropriate Markov Kernel P, estimate $\pi(f)$ by the ergodic averages of the form

$$\pi_n^n(f) := \frac{1}{n}\sum_{i=N+1}^{N+n} f(X_i)$$

- Consider the class $\mathcal{G} : \pi(g) = 0, \forall g \in \mathcal{G}$, then it is valid to estimate $\pi(f)$ by

$$\pi_n^N(f-g) := \frac{1}{n}\sum_{i=N+1}^{N+n} f(X_i) - g(X_i)$$

- How to construct g?

Martingale decomposition

- Let $(\xi_p)_{p\geq 1} \in \mathbb{R}^m$ be random vectors with distribution P_ξ, denote $\mathcal{G}_j = \sigma(\xi_1, \dots, \xi_j)$ and $\mathcal{G}_0 = \text{triv}$;
- Let $(\phi_k)_{k\geq 0}$ be a complete orthonormal system in $L^2(\mathbb{R}^m, P_\xi)$ with $\phi_0 = 1$;
- Consider a Markov chain

$$X_p = \Phi_p(X_{p-1}, \xi_p), \quad p = 1, 2, \dots, \quad X_0 = x, \tag{0.1}$$

- For all Borel functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $\mathsf{E}\left[|f(X_p)|^2\right] < \infty$, it holds

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

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

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

Particular case: ULA

- Fix step size $\gamma > 0$. For $U : \mathbb{R}^d \to \mathbb{R}$, consider a Markov chain $(X_p)_{p\geq 0}, X_0 = x$

$$X_{p+1} = X_p - \gamma\nabla U(X_p) + \sqrt{2\gamma}\xi_{p+1}, \tag{0.2}$$

where $(\xi_p)_{p\geq 1}$ is an i.i.d. sequence of d-dimensional standard Gaussian vectors.
- We use chain (0.2) to approximately sample from the density

$$\pi(x) = \text{const}\, e^{-U(x)}, \tag{0.3}$$

- We can use Hermite polynomials

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

as a complete orthonormal system ;
- We define normalized Hermite polynomials in $\mathbb{R}^d$: for $x \in \mathbb{R}^d$ and $\mathbf{k} = (k_1, \dots, k_d)$

$$H_{\mathbf{k}}(x) = \prod_{i=1}^{d} H_{k_i}(x_i)$$

Algorithm

- Generate T training trajectories $\left(X_1^{(s)}, \dots, X_{N+n}^{(s)}\right), s = 1, \dots, T$ using ULA algorithm;
- Estimate functions $Q_r(x)$ using a modified least-squares criteria:

$$\widehat{Q}_r = \arg\min_{\psi\in\Psi}\sum_{s=1}^{T}\sum_{l=N+1}^{N+n-r}\left|f(X_{l+r}^{(s)}) - \psi(X_l^{(s)})\right|^2 \tag{0.4}$$

for $1 \leq r \leq n_{\text{trunc}}$, Ψ - class of polynomials $\Psi = \{\psi(x)|\psi(x) = \sum_{\|\mathbf{s}\|\leq m}\alpha_{\mathbf{s}}x^{\mathbf{s}}\}$
- compute estimates of $a_{r,\mathbf{k}}(x)$ (possible in closed form):

$$\widehat{a}_{r,\mathbf{k}}(x) = \mathsf{E}H_{\mathbf{k}}(\xi)\widehat{Q}_r(x - \gamma\mu(x) + \sqrt{\gamma}\xi), \xi \sim \mathcal{N}(0, I_d)$$

- Estimate $\pi(f)$ by $\pi_n^N(f) - \widehat{M}_{K,n,n_{\text{trunc}}}^N(f)$ where

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

Main Result

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

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

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

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

Theorem. 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 $0 < \|\mathbf{k}\| \leq K$,

$$|\partial^{\mathbf{k}}f(x)| \leq B_f, \quad |\partial^{\mathbf{k}}\mu(x)| \leq B_\mu. \tag{0.5}$$

Then it holds

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

Numerical Experiments

Gaussian Mixture Model. We consider ULA-generated sample with π given by the mixture of two Gaussian distributions with equal weights:

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

with $d = 2$ and $d = 8$ and take $a = ((2d)^{-1/2}, \dots, (2d)^{-1/2})$.
Logistic regression. Suppose we have i.i.d. sample $\{(\mathbf{X}_i, Y_i)\}$ for $i = 1, \dots, m$ with features $\mathbf{X}_i \in \mathbb{R}^d$ and binary labels $Y_i \in \{0, 1\}$. The binary logistic regression model defines the conditional distribution of Y given X by a logistic function

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

where θ is a model parameter. We put prior $\pi_0 \sim \mathcal{N}(0, \sigma^2 I_d)$, the posterior density takes the form:

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

The target function is taken to be $f(\theta) = \sum_{i=1}^{i=d}\theta_i$.
Probit regression The log-likelihood of the model looks as follows

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

where $\theta \in \mathbb{R}^d$ is a model parameter and Φ is a cumulative distribution function of the $\mathcal{N}(0, 1)$. We put prior $\pi_0 \sim \mathcal{N}(0, \sigma^2 I_d)$, the posterior density takes the form

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

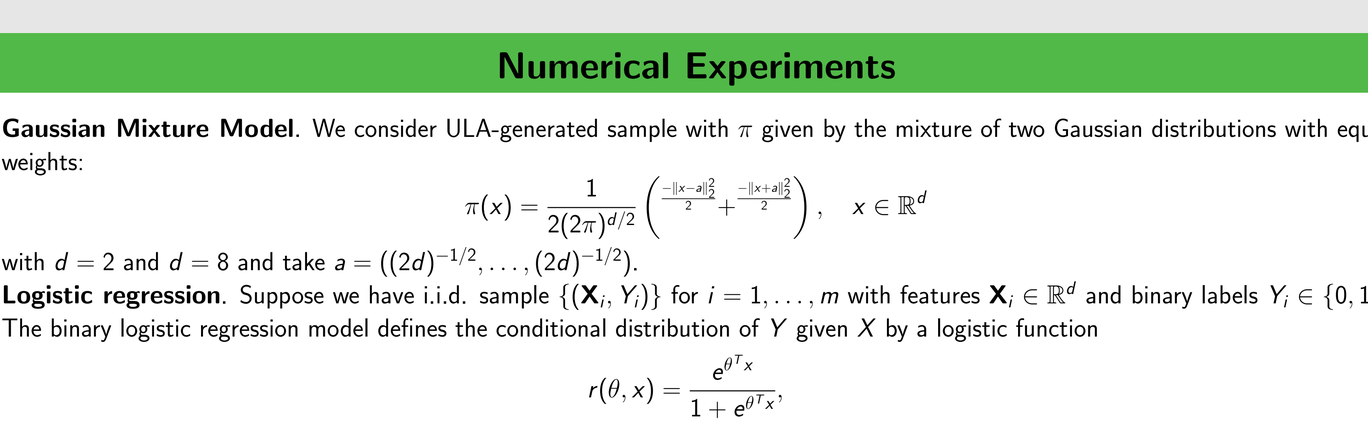

Figure 1: Boxplots of ergodic averages from the variance reduced ULA. Left: Gaussian Mixture in $d = 8$, central: Logistic regression, right: Probit regression. (O) - ordinary empirical average , (CV-B) - our estimator, (ZV) - zero variance estimator, (CV) - diffusion approximation control variates