National Research University Higher School of Economics

Faculty of computer science

Statistical Learning Theory Programme

Term paper

Variance reduction for Markov Chain Monte Carlo

Student: Shagadatov Nurlan
Supervisor: Denis Belomestny

Moscow
2018

VARIANCE REDUCTION FOR MARKOV CHAIN MONTE CARLO

ABSTRACT. Sampling from various kinds of distributions is an issue of paramount importance in statistics, since it is often the key ingredient for constructing estimators. Usually practitioners prefer to use gradient Markov chain Monte Carlo algorithms, called Langevin MCMC schemes. In this term paper we consider new estimator for such types of algorithms, which is considered to be variance-reduced in relation to classical weighted estimator. We considered several examples to clarify the effectiveness of new estimator.

1. INTRODUCTION

There has recently been a real explosion in the use of the Bayesian posterior inference applied to high-dimensional models often involved in machine learning applications. Bayesian estimators are rarely available in explicit form. Therefore, special algorithms are used to sample random variables from a density which is known only up to a factor.

The problem can be formulated as follows. We aim at sampling a posterior distribution $\pi(x) \propto e^{-U(x)}$, where U is continuously differentiable function.The Langevin diffusion is defined by stochastic differential equation

$$dY_t = -\nabla U(Y_t)dt + \sqrt{2}dW_t, \quad t \geq 0, \tag{1.1}$$

where $\{W_t : t \geq 0\}$ is a Brownian motion. In the case of a strongly convex function U having a Lipschitz continuous gradient, equation (1) has a unique strong solution which is a Markov process, which admits π as its unique invariant distribution [3]. We will consider a Euler-Maruyama discretization of the Langevin diffusion, which is a non-homogeneous discrete-time Markov chain defined by

$$X_{p+1} = X_p - \gamma_{p+1}\mu(X_p) + \sqrt{\gamma_{p+1}}Z_{p+1} \tag{1.2}$$

where $(Z_p)_{p\geq 1}$ is an i.i.d. sequence of standard Gaussian random variables and $(\gamma_p)_{p\geq 1}$ is a sequence of step sizes, which can either be held constant or be chosen to monotonically decrease to 0. This first-order Gaussian approximation of diffusion process is called the Unadjusted Langevin Algorithm (ULA). This stationary Markov chain returns $X_1, X_2, \ldots, X_t, \ldots$ such that X_t is converging to π when $\mu = \frac{\nabla U}{2}$. This means that chain can be considered as a sample, albeit a dependent sample, and approximately distributed from π. Due to the Markovian nature of the sampling, the first values are highly dependent on the starting point X_1 and usually trajectory is considered after burn-in period.

As it is frequently done in optimization theory, one may introduce a preconditioner in the ULA algorithm so that convergence is accelerated. More precisely, it aims to choosing a definite positive matrix $\mathbf{A}$, called preconditioner, and applying the Langevin algorithm to the function $F(y) = U(Ay)$. As a results, if $\{Y_p, p > N\}$ is sequence abtained by the ULA applied to the function F, that is the density of Y_p is close to $\pi_F(y) \propto e^{-F(y)}$, then the sequence $X_p = AY_p$ is approximately sampled from the target density $\pi(x) \propto e^{-U(x)}$. The theoretical guarantees of preconditioned ULA can be found in [2].

In this paper we will consider d-dimensional space and unbiased estimators for the quantity

$$\pi(f) = \int_{\mathbb{R}^d} f(y)\pi(dy), \tag{1.3}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a suitable Borel function. There is the classical estimator which is widely used in the literature

$$\pi_n^N(f) = \sum_{p=N+1}^{N+n} \omega_{p,n}^N f(X_p), \quad \omega_{p,n}^N = \frac{\gamma_{p+1}}{\sum_{i=N+2}^{N+n+1} \gamma_i}, \tag{1.4}$$

where $N \in \mathbb{N}$ denotes the length of burn-in period and n is the number of effective samples.

2. Variance reduced estimator

In this section we will introduce a stochastic representation suggested in [1] which will be used to construct an effective variance reduced estimator. More precisely, we consider the representation for the Markov chain obtained by Unadjasted Langevin Algorithm.

For any $k \in \mathbb{N} \cup \{0\}$ we denote the normalized Hermite polynomials on $\mathbb{R}$ as

$$H_k(x) = \frac{(-1)^k}{\sqrt{k!}} e^{\frac{x^2}{2}} \frac{\partial^k}{\partial x^k} e^{\frac{-x^2}{2}}, \quad x \in \mathbb{R}. \tag{2.1}$$

For a multiindex $\mathbf{k} = (k_i) \in (\mathbb{N} \cup \{0\})^d$, $\mathbf{H_k}$ denotes the normalized polynomial on $\mathbb{R}^d$

$$\mathbf{H_k}(x) = \prod_{i=1}^{d} H_{k_i}(x_i), \quad x = (x_i) \in \mathbb{R}^d \tag{2.2}$$

Proposition 2.1. ***(D.Belomestny)*** *. Let f be a Borel function $\mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}\left[|f(X_p)|^2\right] < \infty$. Then, for $p > j$, the following representation holds*

$$f(X_p) = \mathbb{E}\left[f(X_p)|\zeta_j\right] + \sum_{k \in \mathbb{N}_0^d \setminus \{0\}} \sum_{l=j+1}^{p} a_{p,l,k}(X_{l-1}) \mathbf{H_k}(Z_l), \tag{2.3}$$

where

$$a_{p,l,k}(x) = \mathbb{E}\left[f(X_p)\mathbf{H_k}(Z_l)|X_{l-1} = x\right] \tag{2.4}$$

and $\zeta_j = \sigma(Z_1, \ldots, Z_j), j \in \mathbb{N}$, and $\zeta_0 = triv$

Proposition 2.2. *The coefficients $a_{p,l,k}$ in (2.4) can be alternatively represented as*

$$a_{p,l,k}(x) = \mathbb{E}\left[\mathbf{H_k}(\xi) Q_{p,l}(x - \gamma_l \mu(x) + \sqrt{\gamma_l}\xi)\right] \tag{2.5}$$

with $Q_{p,l}(x) = \mathbb{E}\left[f(X_p)\middle| X_l = x\right]$. The functions $Q_{p,l}(x)$ satisfy the recursion

$$Q_{p,l}(x) = \mathbb{E}\left[Q_{p,l+1}(X_{l+1})\middle| X_l = x\right], \quad Q_{p,p}(x) = f(x), \tag{2.6}$$

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

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

$$M_{K,n}^N = \sum_{0<\|k\|\le K} \sum_{p=N+1}^{N+n} \omega_{p,n}^N \sum_{l=N+1}^{p} a_{p,l,k}(X_{l-1}) \mathbf{H_K}(Z_l) \tag{2.7}$$

with $\|k\| = \max_i k_i$.

Thus, the author suggested the variance-reduced estimator for $\pi(f)$ of (1.3)

$$\pi_{K,n}^N(f) = \pi_n^N(f) - M_{K,n}^N(f) \tag{2.8}$$

In order to approximate the coefficients $a_{p,l,k}$ author suggested to approximate the functions $Q_{p,l}$ from (2.6) using regression on a set of "training" trajectories. More precisely, generate

N_{train} independent paths $X^{(s)}_{N+1}, \dots, X^{(s)}_{N+n}, s = 1, \dots, N_{train}$ of the chain X and solve least squares optimization problems:

$$\hat{Q}_{p,l} = \underset{\psi \in span(\psi_1,\dots,\psi_Q)}{\operatorname{argmin}} \sum_{i=1}^{N_{train}} \mid \hat{Q}_{p,l+1}(X^{(s)}_{l+1}) - \psi(X^{(s)}_l) \mid^2 \tag{2.9}$$

for $l = N+1, \dots, p$, where $\psi_1, \dots, \psi_Q$ is a set of basis functions in $\mathbb{R}^d$.

3. Polynomial approximation

In this section we derive explicit formula for coefficients $a_{p,l,k}(x)$ in case of polynomial approximation for functions $Q_{p,l}(x)$ appearing in (2.6). In particular, we consider Markov chain generated by ULA with constant step size γ. Suppose we constructed a polynomial approximation for $Q_{p,l}(x)$ of the form:

$$\widehat{Q}_{p,l}(x) = \sum_{\|\mathbf{s}\| \le m} \beta_{\mathbf{s}} x^{\mathbf{s}}, \quad s = (s_1, \dots, s_d)$$

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

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

we derive

$$\begin{aligned} \widehat{a}_{p,l,\mathbf{k}}(x) &= \mathsf{E}\left[\mathbf{H}_{\mathbf{k}}(\xi) Q_{p,l}(x - \gamma\mu(x) + \sqrt{\gamma}\xi)\right] = \\ &= \sum_{\|\mathbf{s}\| \le m} \beta_{\mathbf{s}} \, \mathsf{E}\left[\prod_{i=1}^{d} H_{k_i}(\xi_i)(x_i - \gamma\mu_i(x) + \sqrt{\gamma}\xi_i)^{s_i}\right] \\ &= \sum_{\|\mathbf{s}\| \le m} \beta_{\mathbf{s}} \prod_{i=1}^{d} E_i \end{aligned}$$

with

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

and

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

4. Numerical results

To illustrate the results introduced in the previous sections, we conducted some experiments on synthetic data. We considered two examples similar to that considered by Dalalyan [2]. The code used to run the experiments is available at `https://github.com/ShagadatovNurlan/Variance-reduction`.

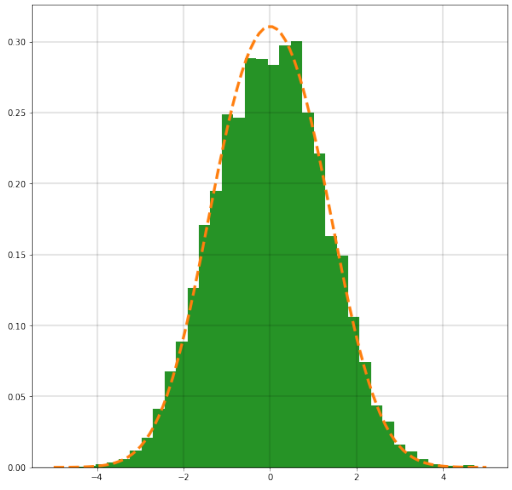


FIGURE 4.1. Histograms of the samples for Gaussian mixture using ULA scheme. The dimesion is $d = 1$ and $n = 5000$ samples.

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

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

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

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

and

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

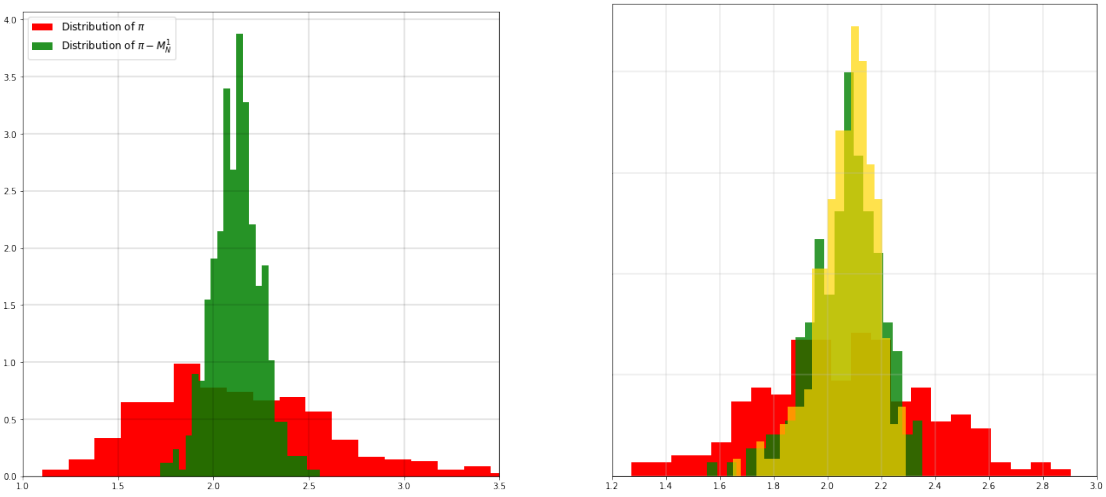
respectively.

In our experiments we considered dimensions $d = 1$ and $d = 2$ and defined vector a as $(\frac{1}{\sqrt{2d}}, \ldots, \frac{1}{\sqrt{2d}})$. 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 $N_{tr} = 500$ independent "training" trajectories and solved the least squares problems (**??**) with polynomial basis functions with maximum degree 5 and 3 for dimensions 1 and 2, respectively. More precisely the polynomials defined as follows:

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

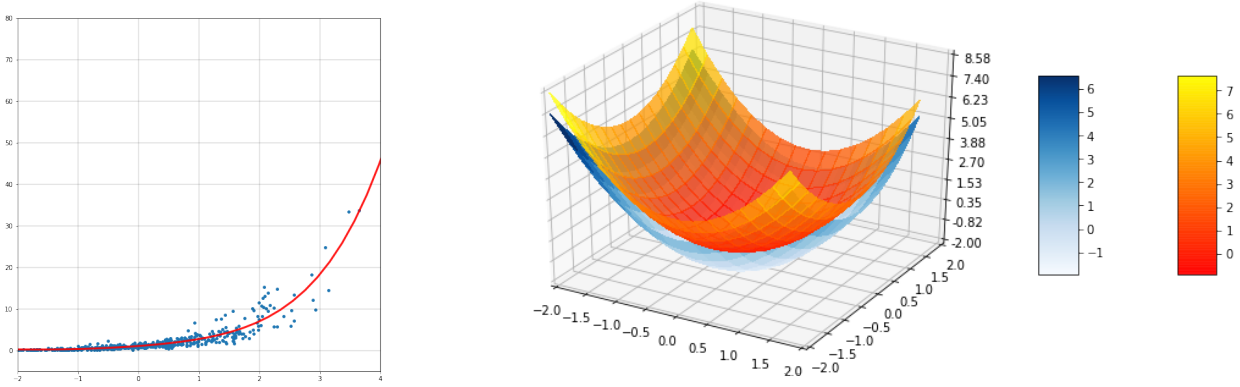
We fixed K = 1 for d=1 and K=2 for d = 2. To test our variance reduction algorithm, we generated $N_{\text{test}} = 200$ independent paths and computed empirical variance of the new variance reduced estimator $\pi^N_{K,n}(f)$ of target functions $f(x) = e^x$ for $d = 1$ and $f(x) = x_1^2 + x_2^2 - \cos(x_1)$ for $d = 2$. Figure 4.3 shows the histograms of weighted average estimator $\pi^N_n(f)$ and variance reduced estimator $\pi^N_{K,n}(f)$ 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.

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 4.4 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 $\pi^N_{K,n}(f)$ than for the ordinary weighted average estimator $\pi^N_n(f)$.



(A) d = 1 (B) d = 2

FIGURE 4.2. Histograms for Gaussian mixture. (a) 1-dimensional GM model: histograms of estimators for target function $f(x) = e^x$ 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*: $\pi_{2,n}^N(f)$.



(A) d = 1 (B) d = 2

FIGURE 4.3. Illustration of approximations. (a) 1-dimensional GM model: approximation of $Q_{105,100}$ for target function $f(x) = e^x$, (b) 2-dimensional GM model: approximation of $Q_{55,50}$ for target function $f(x) = x_1^2 + x_2^2 - \cos(x_1)$.

4.2. **Binary Logistic regression.** Second experiment considers the problem of logistic regression. Suppose we have i.i.d. sample $\{(\mathbf{X}_i, Y_i)\}$ for $i = 1, \ldots, m$ 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 X by a logistic function

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

where θ is parameter of model. In order to estimate θ according to given data, the Bayessian approach introduces prior distribution $\pi_0(\theta)$ and inferences the posterior density $\pi(\theta)$ using Bayes' rule. In case of Gaussian prior π_0 with zero mean and covariance matrix proportional

	$d=1$				
$\mathsf{Var}(\pi_n^N)$	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.

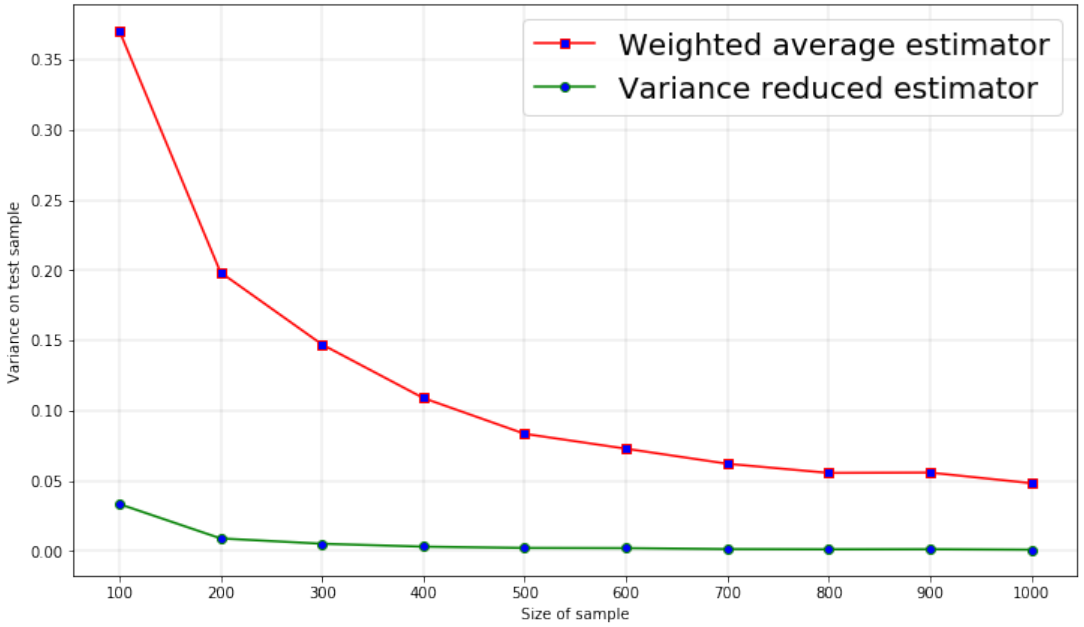


FIGURE 4.4. 1-dimensional GM model. Vertical axis is empirical variance on test sample, horizontal axis is the length of test trajectories obtained by ULA algorithm.Red traceplot correspondents to ordinary weighted estimator, green traseplot illustrates empirical variances of variance-reduced estimator for K=1.

to the inverse of the Gram matrix $\mathbf{\Sigma_X} = \frac{1}{n}\sum_{i=1}^m \mathbf{X}_i\mathbf{X}_i^T$, the posterior density takes the form

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

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

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

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

In our second experiment, we randomly generated m independent samples as in paper [**?**], 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 $(1,\dots,1)^T$. We fix $d=2$ and

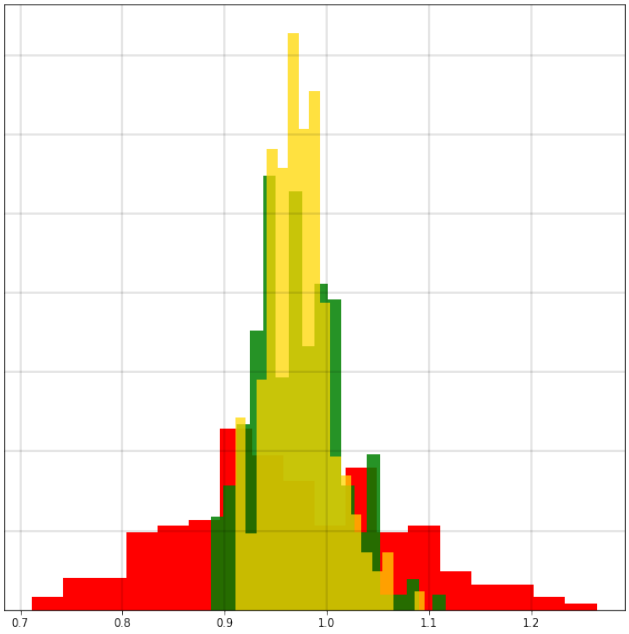


FIGURE 4.5. Binary Logistic Regression: Histograms of estimators for target function $f(\theta) = 2\theta_1^2 + 7\theta_2^2$ on test sample. *Red* bins correspondent to ordinary weighted estimators $\pi_n^N(f)$, *green* - variance-reduced estimators $\pi_{1,n}^N(f)$and *yellow* - $\pi_{2,n}^N(f)$.

	$d = 2$				
$Var(\pi_n^N)$	0.03727	0.02124	0.04769	0.01147	0.02974
$Var(\pi_{1,n}^N)$	0.00269	0.00224	0.00306	0.00179	0.00196
$Var(\pi_{2,n}^N)$	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.

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

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

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

5. Summary and conclusion

In this term paper we considered an efficient variance reduced estimator for Unadjusted Langevin Algorithm. We introduced the stochastic representation for Markov Chain, which allows to contruct the novel variance reduction approach. We provided a scheme to efficiently compute estimator using polynomial regression. Subsequently, we conducted two experiments, which indicate the effectiveness of proposed method.

In conclusion, several open questions arise from our work. It would be interesting to apply the stochastic representation for Metropolis Adjusted Langevin Algorithm and compare the efficiency of suggested approach with existing variance reduction methods. Another open question is to rigorously examine the complexity of proposed variance reduced estimator.

References

[1] Denis Belomestny, Eric Moulines, Nurlan Shagadatov, Mikhail Urusov. *"Variance and complexity reduction for MCMC via regression based control variates"*.
[2] Arnak Dalalyan. *"Theoretical guarantees for approximate sampling from smooth and log-concave densities"*.
[3] Gareth Roberts, Rishard Tweedie. *"Exponential convergence of Langevin distributions and their discrete approximations"*.