National Research University Higher School of Economics

Faculty of Mathematics

as a manuscript

Leonid Iosipoi

Variance reduction for Monte Carlo estimates

Summary of the PhD thesis
for the purpose of obtaining academic degree
Doctor of Philosophy in Mathematics

Academic supervisor:
candidate of sciences,
professor Denis Belomesnty

Moscow – 2021

Introduction

Suppose that we wish to compute an expected value $\pi(f) := \mathsf{E}\,[f(X)]$, where X is a random vector on $\mathsf{X} \subset \mathbb{R}^d$ with a probability density function $\pi(x)$ and $f:\ \mathsf{X} \to \mathbb{R}$ is a function from $L^2(\pi)$. We assume that both the function $f(x)$ and the density $\pi(x)$ are known.

This problem can be viewed as integration of a function $f(x) \cdot \pi(x)$ over the set X. We assume that the function $f(x)$ and the density $\pi(x)$ are complicated enough, so the integral cannot be computed analytically. Hence we have to estimate its value using numerical integration algorithms. It is known that any deterministic (even adaptive) numerical integration algorithm faces the so-called curse of dimensionality which means that complexity of an algorithm grows exponentially with the dimension d, see, for instance, the classic work of N. S. Bakhvalov [B71] and the modern review of E. Novak [N16]. In turn, randomized algorithms such as Monte Carlo schemes are more effective than other approaches to this problem if the dimension d is large and/or the function f is hard to compute, see, for instance, the review of G. Roberts and J. Rosenthal [RR04] or the book by C. Robert and G. Casella [RC99].

Assume first that it is possible to get an independent sample $X_1 \ldots, X_n$ from a distribution with the density $\pi(x)$. A natural estimate for the expected value $\pi(f)$ is a Monte Carlo estimate

$$\pi_n(f) := \frac{1}{n} \sum_{k=1}^{n} f(X_k).$$

This estimate is unbiased, that is, $\mathrm{E}[\pi_n(f)] = \pi(f)$, and has variance $\mathrm{Var}(\pi_n(f)) = V(f)/n$, where, here and subsequently, $V(f)$ denotes the variance of function f with respect to π, i.e.,

$$V(f) := \mathrm{Var}(f(X)), \quad X \sim \pi.$$

Using the central limit theorem, one can construct an asymptotically valid confidence interval for $\pi(f)$ of the form

$$\pi_n(f) \pm c \sqrt{\frac{V(f)}{n}}, \tag{1}$$

where c is a quantile of a normal distribution. From the point of view of applications, an interesting problem is to improve accuracy of the Monte Carlo estimate $\pi_n(f)$, which in turn means reducing the size of the aforementioned asymptotic confidence interval. This can be achieved by increasing the sample size n but in some cases this solution may not be practical (for example, due to the difficulty of generating random variables from the density $\pi(x)$ or the difficulty of computing values of the function $f(x)$). Another approach to reduce the size of the confidence interval is to decrease the value $V(f)$ by considering a new Monte Carlo estimate with the same mean but lower variance. Such techniques are called variance reduction methods.

Extensive literature is available on variance reduction methods, see, for instance, classic books of I. Dimov [D08], P. Glasserman [G13], R. Rubinstein and D. Kroese [RK16]. We should note that the majority of the variance reduction techniques explicitly use the structure of function $f(x)$ or density $\pi(x)$, therefore they are not applicable in relatively general setups. The method of control variates is one of the few variance reduction methods applicable in a general setting. The idea behind this method is to consider a class $\mathcal{G} \subset L^2(\pi)$ of functions $g \in \mathcal{G}$ satisfying $\pi(g) = 0$. Such functions are called control variates. We may then introduce a new Monte Carlo estimate

$$\pi_n\big(f - \bar{g}\big) := \frac{1}{n}\sum_{k=1}^{n}\big(f(X_k) - \bar{g}(X_k)\big),$$

where $\bar{g}$ is a function from $\mathcal{G}$ which minimizes the variance $V\big(f-g\big) = \mathrm{Var}\big(f(X) - g(X)\big)$, $X \sim \pi$, that is,

$$\bar{g} \in \operatorname*{argmin}_{g\in\mathcal{G}} V\big(f-g\big). \tag{2}$$

If the class $\mathcal{G}$ of control variates is chosen well, this approach may significantly reduce the variance. However, there are two fundamental issues related to this approach. The first issue is how to find and constructively describe the class $\mathcal{G}$ of control variates. The second issue is how to solve the optimization problem (2) numerically as the variance $V(f-g)$ is not assumed to be known and its computation is a more difficult problem than the initial one.

The question of constructing control variates $g \in \mathcal{G}$ was previously studied in the works of A. Mira, R. Solgi, D. Imparato [MSI13] and C. Oates, M. Girolami, N. Chopin [OGC17] and [OGC19], therefore we will not dwell on it in detail. We only mention that the popular method for constructing control variates $g \in \mathcal{G}$ is to substitute various functions $\Phi : \mathsf{X} \to \mathbb{R}^d$ into

$$g_\Phi = \langle \Phi, \nabla \log \pi \rangle + \mathrm{div}(\Phi), \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard scalar product in $\mathbb{R}^d$, and $\mathrm{div}(\Phi)$ stands for the divergence of Φ. Under rather mild conditions on π and Φ, it follows from the integration by parts that $\pi(g_\Phi) = 0$ (see Propositions 1 and 2 in [MSI13]). Such control variates are usually called Stein control variates.

Regarding the question of solving the optimization problem (2), it seems natural to replace the true variance $V(f-g)$ by its empirical counterpart $V_n(f-g)$ computed on the sample $X_1, \dots, X_n$, that is,

$$V_n(f-g) := \frac{1}{n-1}\sum_{k=1}^{n}\big(f(X_k) - g(X_k) - \pi_n(f-g)\big)^2. \tag{4}$$

Moreover, we also replace the class $\mathcal{G}$ by its discrete finite approximation $\mathcal{G}'$, since numerical optimization over an infinite set of functions is a complicated problem that cannot be solved in general. Thus a natural question arises: how much will we "lose" by replacing the true variance by its

empirical counterpart and the set $\mathcal{G}$ by its approximation $\mathcal{G}'$? The answer to this question is the main goal of this work.

The present summary is organized as follows. In Section 1, we set up notation and terminology used in the next sections. Section 2 provides a high-probability bound on the excess risk

$$V\big(f-\widehat{g}\,\big)-\inf_{g\in\mathcal{G}}V\big(f-g\big), \tag{5}$$

where

$$\widehat{g}\in\operatorname*{argmin}_{g\in\mathcal{G}'}V_n(f-g),\quad \mathcal{G}' \text{ is an } \varepsilon\text{-net of } \mathcal{G},$$

and the variance $V(f-\widehat{g}\,)$ is taken conditionally on the sample $X_1,\dots,X_n$ used to compute $\widehat{g}$. The estimates of this form are often analyzed in the statistical literature and are referred to as skeleton or sieve estimates (see, for instance, the work of W. H. Wong, X. Shen [WS95], or books by L. Devroye, L. Györfi, G. Lugosi [DGL96] and S. van de Geer [G00]). In Section 3, we consider the case where an independent sample $X_1\dots,X_n$ from the distribution with density $\pi(x)$ cannot be obtained, but $\pi(x)$ itself is known (possibly up to a normalizing constant). In such a scenario, one can often use Markov chain Monte Carlo algorithms (MCMC algorithms) to obtain a Markov chain $\{X_k\}_{k=1}^{\infty}$ with distribution converging to the distribution with density $\pi(x)$. Under certain conditions, the central limit theorem also holds for the Markov chain $\{X_k\}_{k=1}^{\infty}$ and, therefore, it is possible to construct an asymptotic confidence interval

$$\pi_n(f)\pm c\sqrt{\frac{V^{\infty}(f)}{n}}, \tag{6}$$

where $V^{\infty}(f)$ is the asymptotic variance defined as

$$V^{\infty}(f):=\lim_{n\to\infty}n\cdot\mathrm{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n}f(X_k)-\pi(f)\right)^2\right]. \tag{7}$$

Due to the inherent serial correlation in the chain $\{X_k\}_{k=1}^{\infty}$, the asymptotic variance $V^{\infty}(f)$, in general, is not equal to the variance $V(f)$. Therefore, it is more preferable to choose a control variate $g\in\mathcal{G}$ by optimizing the asymptotic variance $V^{\infty}(f-g)$. Since the asymptotic variance is not assumed to be known, we will replace it by its empirical counterpart. In Section 3, we provide a high-probability bound on the excess risk

$$V^{\infty}\big(f-\widehat{g}\,\big)-\inf_{g\in\mathcal{G}}V^{\infty}\big(f-g\big), \tag{8}$$

where for an estimate V_n^{∞} of the asymptotic variance V^{∞}

$$\widehat{g}\in\operatorname*{argmin}_{g\in\mathcal{G}'}V_n^{\infty}\left(f-g\right),\quad \mathcal{G}' \text{ is an } \varepsilon\text{-net of } \mathcal{G},$$

and the variance $V^{\infty}(f-\widehat{g}\,)$ is taken conditionally on the sample $X_1,\dots,X_n$ used to compute $\widehat{g}$. This generalization turns out to be challenging for at least two reasons. First, due to the inherent serial correlation, estimating the asymptotic variance requires specific techniques such as spectral and batch means methods, see J. Flegal and G. Jones [FJ10] for a survey on asymptotic variance estimators and their statistical properties. Second, a non-asymptotic analysis of the estimate $f-\widehat{g}$ requires a generalization of the concentration inequalities used in the proof to Markov chains, which is also not trivial. We perform this analysis for a rather general class of Markov chains including many MCMC algorithms. Finally, in Section 4, we consider an MCMC algorithm named Unadjusted Langevin Algorithm and conduct a simulation study of the results obtained.

Aim of this work

The aim of the present work is to study theoretical properties of the control variates method for Monte Carlo estimates and, in particular, to obtain non-asymptotic bounds on the excess risk (5) and (8) in both independent and dependent cases respectively.

1. Notation and Terminology

Before we proceed to the main results, let us introduce some notation. We will denote the class of functions $h(x)=f(x)-g(x)$ for $g\in\mathcal{G}$ by $\mathcal{H}$, i.e.,

$$\mathcal{H}:=\{f-g:\ g\in\mathcal{G}\},$$

where $\mathcal{G}$ is the class of control variates (that is, the class of functions with $\pi(g)=0$ for all functions $g\in\mathcal{G}$). Note that any function $h\in\mathcal{H}$ has the expected value equal to $\pi(f)$. As was mentioned before, instead of looking for the best control variate in the whole class $\mathcal{H}$ we will do this in its finite approximation. Fix $\varepsilon>0$ and let $r=1$ or $r=2$. Assuming that the class $\mathcal{H}$ is totally bounded, let us denote by $\mathcal{H}_{\varepsilon,r}$ a minimal ε-net in the $L^r(\pi)$-norm, i.e., the smallest possible collection of functions $\mathcal{H}_{\varepsilon,r}=\{h_1,\dots,h_m\}\subset\mathcal{H}$ with the property that for any $h\in\mathcal{H}$ there exists $h_*\in\mathcal{H}_{\varepsilon,r}$ such that the distance between h an h_* in $L^r(\pi)$-norm is less than or equal to ε. The metric entropy of $\mathcal{H}$ is denoted by $H_{L^r(\pi)}(\mathcal{H},\varepsilon):=\log|\mathcal{H}_{\varepsilon,r}|$, where $|\mathcal{H}_{\varepsilon,r}|$ is the cardinality of $\mathcal{H}_{\varepsilon,r}$. Define now by $\gamma_{L^r(\pi)}(\mathcal{H},n)$ the quantity

$$\gamma_{L^r(\pi)}(\mathcal{H},n):=\inf\big\{\eta>0:H_{L^r(\pi)}(\mathcal{H},\eta)\le n\eta^r\big\},$$

which will be used later in the main results. Note that a number $\eta>0$ satisfying $H_{L^r(\pi)}(\mathcal{H},\eta)\le n\eta^r$ exists and is finite because the metric entropy $H_{L^r(\pi)}(\mathcal{H},\eta)$ is decreasing function in η and the

mapping $\eta \mapsto n\eta^r$ is increasing in η. The quantity $\gamma_{L^r(\pi)}(\mathcal{H}, n)$ will be used to control the cardinality of $\mathcal{H}_{\varepsilon,r}$. Indeed, by choosing $\varepsilon \geq \gamma_{L^r(\pi)}(\mathcal{H}, n)$ we get $|\mathcal{H}_{\varepsilon,r}| \leq \mathrm{e}^{n\varepsilon^r}$. It is easily seen from the above definition that $\gamma_{L^r(\pi)}(\mathcal{H}, n)$ is a decreasing function in n.

Let us discuss a typical behaviour of $\gamma_{L^r(\pi)}(\mathcal{H}, n)$ as $n \to \infty$ when $\mathcal{H}$ is a subset of the weighted Sobolev space $W^{s,p}(\mathsf{X}, \langle x\rangle^\beta)$. Let us remind that the Sobolev space $W^{s,p}(\mathsf{X})$ is defined as $W^{s,p}(\mathsf{X}) = \big\{h \in L^p(\lambda) : \ D^\alpha h \in L^p(\lambda), \ \forall |\alpha| \leqslant s\big\}$, where λ is the Lebesgue measure, $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multi-index with $|\alpha| = \alpha_1 + \ldots + \alpha_d$, and D^α stands for the differential operator $D^\alpha = \partial^{|\alpha|}/\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}$. Here all derivatives are understood in the weak sense. The weighted Sobolev space $W^{s,p}(\mathsf{X}, \langle x\rangle^\beta)$ for a polynomial weighting function $\langle x\rangle^\beta = \big(1 + \|x\|^2\big)^{\beta/2}$, $\beta \in \mathbb{R}$, is defined by

$$W^{s,p}(\mathsf{X}, \langle x\rangle^\beta) = \big\{h \in L^p(\lambda) : \ h \cdot \langle x\rangle^\beta \in W^{s,p}(\mathsf{X})\big\}.$$

The norm in this weighted Sobolev space is defined as

$$\|h\|_{W^{s,p}_\beta} = \sum_{|\alpha| \leq s} \big\|D^\alpha\big(h \cdot \langle x\rangle^\beta\big)\big\|_{L^p(\lambda)}.$$

A subset $\mathcal{H} \subset W^{s,p}(\mathsf{X}, \langle x\rangle^\beta)$ is called norm-bounded if $\mathcal{H}$ is totally bounded and there exists $c > 0$, such that $\|h\|_{W^{s,p}_\beta} \leq c$ for any $h \in \mathcal{H}$. The following proposition holds.

Proposition 1. *Let $\mathcal{H}$ be a (non-empty) norm-bounded subset of $W^{s,p}(\mathbb{R}^d, \langle x\rangle^\beta)$, where $1 < p < \infty$, $\beta \in \mathbb{R}$, and $s - d/p > 0$. Suppose also that $\|\langle x\rangle^{\kappa-\beta}\|_{L^r(\pi)} < \infty$ for some $\kappa > 0$. Then for $r = 1, 2$ it holds*

$$\gamma_{L^r(\pi)}(\mathcal{H}, n) \lesssim \begin{cases} n^{-\frac{1}{r+d/s}} & \text{for } \kappa > s - d/p, \\ n^{-\frac{1}{r+(\kappa/d+1/p)^{-1}}} & \text{for } \kappa < s - d/p, \end{cases}$$

where the symbol $\lesssim$ stands for inequality up to a constant not depending on n.

Proposition 1 follows from the general result for the entropy of bounded subsets of $W^{s,p}(\mathsf{X}, \langle x\rangle^\beta)$, see Corollary 4 in R. Nickl, B. Pötscher [NP07]. It is not considered as a result of the present thesis and is given here as an example of an estimate for $\gamma_{L^r(\pi)}(\mathcal{H}, n)$ which will be used later in the main results.

In what follows, we will write π for the distribution with the density $\pi(x)$ and the corresponding probability measure when no confusion can arise. Unless otherwise specified, the symbol $\lesssim$ stands for inequality up to an absolute constant and the symbol $\propto$ stands for equality up to an absolute constant.

2. Variance reduction in independent case

Before we proceed to the main results in the independent case, let us recall that for the class $\mathcal{H} = \{f - g : \ g \in \mathcal{G}\}$ the estimate we study can be written as

$$\widehat{h}_{\varepsilon,r} \in \operatorname*{argmin}_{h \in \mathcal{H}_{\varepsilon,r}} V_n(h),$$

where $V_n(h)$ is the sample variance defined in (4) and $\mathcal{H}_{\varepsilon,r}$ is an ε-net of the set $\mathcal{H}$ in $L^r(\pi)$-norm, $r = 1, 2$. The detailed definitions are given in Section 1. Here and subsequently, we will assume that the minimizer $\widehat{h}_{\varepsilon,r}$ exists as all the following arguments can easily be adapted by considering an approximate minimizer.

Now we can state the main results of the section. We start with the case when functions $h \in \mathcal{H}$ are bounded and the class $\mathcal{H}$ is closed and convex.

Theorem 2. *Assume that $\mathcal{H}$ is closed and convex. Assume also that $\sup_{h\in\mathcal{H}} \sup_{x\in\mathsf{X}} |h(x)| \le b$ for some $b > 0$ and $\pi(h) = c$ for all $h \in \mathcal{H}$ and some $c \in \mathbb{R}$. Then for some $\varepsilon > 0$ (specified in the proof) and any $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$V\Big(\widehat{h}_{\varepsilon,1}\Big) - \inf_{h\in\mathcal{H}} V\big(h\big) \lesssim b^2 \gamma_{L^1(\pi)}(\mathcal{H}, n) + \frac{b^2 \log(1/\delta)}{n},$$

where the first variance is taken conditionally on the sample $X_1, \dots, X_n$ used to compute $\widehat{h}_{\varepsilon,1}$ and the symbol $\lesssim$ stands for inequality up to an absolute constant.

Let us now state a similar result where instead of convexity of the class $\mathcal{H}$ we will assume that $\mathcal{H}$ contains a constant function. Since, by construction, $\pi(h) = \pi(f)$ for all $h \in \mathcal{H}$, this constant must be equal to $\pi(f)$, and hence $\inf_{h\in\mathcal{H}} V(h) = 0$.

Theorem 3. *Assume that $\mathcal{H}$ contains a constant function, that is, $h^*(x) \equiv c$ for some $h^* \in \mathcal{H}$ and some $c \in \mathbb{R}$. Assume also that $\sup_{h\in\mathcal{H}} \sup_{x\in\mathsf{X}} |h(x)| \le b$ for some $b > 0$ and $\pi(h) = c$ for all $h \in \mathcal{H}$ and some $c \in \mathbb{R}$. Then for some $\varepsilon > 0$ (specified in the proof) and any $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$V\Big(\widehat{h}_{\varepsilon,1}\Big) \lesssim b^2 \gamma_{L^1(\pi)}(\mathcal{H}, n) + \frac{b^2 \log(1/\delta)}{n}$$

and

$$V\Big(\widehat{h}_{\varepsilon,2}\Big) \lesssim b^2 \big(\gamma_{L^2(\pi)}(\mathcal{H}, n)\big)^2 + \frac{b^2 \log(1/\delta)}{n},$$

where the variances are taken conditionally on the sample $X_1, \dots, X_n$ used to compute $\widehat{h}_{\varepsilon,1}$ and $\widehat{h}_{\varepsilon,2}$, and the symbol $\lesssim$ stands for inequality up to an absolute constant.

If the assumptions of Theorem 2 or Theorems 3 hold and additionally $\mathcal{H}$ is a bounded subset of weighted Sobolev space $W^{s,p}(\mathsf{X}, \langle x \rangle^{\beta})$, using the result of Proposition 1, we get the following final high-probability bound

$$V\Big(\widehat{h}_{\varepsilon,1}\Big) - \inf_{h \in \mathcal{H}} V\big(h\big) \lesssim n^{-\alpha}, \quad \alpha \in (0,1), \tag{9}$$

where the value of parameter α depends on the dimension d and the parameters of the Sobolev space s, p, and β. In statistical literature, such rates are referred to as fast rates of convergence since they yield a faster convergence rate than the standard rate $1/\sqrt{n}$ in parametric estimation. As a corollary, it is worth noting that the considered variance reduction procedure reduces the length of the asymptotic confidence interval (1) from order $n^{-1/2}$ to $n^{-1/2-\alpha/2}$ if the class $\mathcal{H}$ is chosen so that $\inf_{h \in \mathcal{H}} V\big(h\big) \lesssim n^{-\alpha}$.

To the best of our knowledge, Theorem 2 and Theorem 3 are the first results on minimization of the sample variance $V_n(h)$ available in literature. The main technical difficulty of the considered problem is that the sample variance is not a sum of independent random variables $h(X_1), \ldots, h(X_n)$ contrary to classical problems in empirical processes. Instead, $V_n(h)$ is a particular case of U-statistics. Fast rates in empirical minimization of U-statistics have been previously obtained, as far as we know, only in S. Clémençon, G. Lugosi, N. Vayatis [CLV08] (and some other works of the authors) in the context of ranking problems. In the proofs of Theorem 2 and Theorem 3, by a clever choice of a finite (but representative) subset of $\mathcal{H}$, we avoid many of the technical difficulties that appear in [CLV08].

Now let us formulate two consequences of Theorem 3 with an explicit value of α from (9). For simplicity, we consider the one-dimensional case $\mathsf{X} = \mathbb{R}$ and the Stein control variates, see (3), which can be rewritten as

$$g_{\phi}(x) := \phi(x) \cdot \big(\log \pi(x)\big)' + \phi'(x) = \frac{1}{\pi(x)} \big(\phi(x)\pi(x)\big)',$$

where functions $\phi : \mathbb{R} \to \mathbb{R}$ belong to some class Φ to be specified later. Let us denote the set of all s-times continuously differentiable functions on $\mathbb{R}$ by $C^s(\mathbb{R})$ and the set of functions from $C^s(\mathbb{R})$ with derivatives growing not faster than a polynomial by

$$C^s_{poly}(\mathbb{R}) := \big\{\phi \in C^s(\mathbb{R}) : \ \exists\, m \in \mathbb{N}, \text{ such that } |\phi^{(k)}(x)| \lesssim |x|^m \text{ for } |x| \to \infty, \ \forall k = 0, \ldots, s\big\}.$$

The following corollary deals with the case when the tails of π decay exponentially.

Corollary 4. *Let $\pi(x) \propto e^{-c|x|^q}$ for some $c \in \mathbb{R}$ and some $q \in \mathbb{N}$. Let also $f \in C^s_{poly}(\mathbb{R})$ for some $s \in \mathbb{N}$. Choose any $\Phi \subset C^{s+1}_{poly}(\mathbb{R})$ such that $\mathcal{H} = \{f - g_\phi : \phi \in \Phi\}$ is norm-bounded, contains a constant function, and $\sup_{h\in\mathcal{H}} \sup_{x\in\mathsf{X}} |h(x)| \le b$ for some $b > 0$. Then for some $\varepsilon > 0$ and any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$V\left(\widehat{h}_{\varepsilon,1}\right) \lesssim \left(\frac{1}{n}\right)^{\frac{1}{1+1/s}} + \frac{\log(1/\delta)}{n},$$

where the symbol $\lesssim$ stands for inequality up to a constant not depending on n but possibly depending on other parameters.

Next we consider the case when tails of π are of polynomial decay. To do this, let us denote the set of functions with derivatives growing not faster than a polynomial of a fixed order by

$$C^s_{poly<m}(\mathbb{R}) := \left\{\phi \in C^s(\mathbb{R}) : \ |\phi^{(k)}(x)| \lesssim |x|^m \text{ for } |x| \to \infty, \forall k = 0, \dots, s\right\}.$$

Corollary 5. *Let $\pi(x) \propto (1+x^2)^{-q}$ for some $q \in \mathbb{N}$. Let also $f \in C^s_{poly<m}(\mathbb{R})$ for some $s, m \in \mathbb{N}$. Assume that $2q - m - 3 > 0$. Choose any $\Phi \subset C^{s+1}_{poly}(\mathbb{R})$ such that $\mathcal{H} = \{f - g_\phi : \phi \in \Phi\}$ is norm-bounded, contains a constant function, and $\sup_{h\in\mathcal{H}} \sup_{x\in\mathsf{X}} |h(x)| \le b$ for some $b > 0$. Then for some $\varepsilon > 0$ and any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$V\left(\widehat{h}_{\varepsilon,1}\right) \lesssim \begin{cases} \left(\frac{1}{n}\right)^{\frac{1}{1+1/s}} + \frac{\log(1/\delta)}{n} & \text{for } s \le 2q-m-3, \\ \left(\frac{1}{n}\right)^{\frac{1}{1+(2q-m-3)^{-1}}} + \frac{\log(1/\delta)}{n} & \text{for } s > 2q-m-3, \end{cases}$$

where the symbol $\lesssim$ stands for inequality up to a constant not depending on n but possibly depending on other parameters.

Corollary 4 and Corollary 5 reveal an interesting phenomenon. If π decays exponentially fast and f grows polynomially, the smoother function f is, the faster convergence rates we get. On the contrary, when the decay of π and the growth of f are both polynomial, there is a critical power after which the smoothness of f will not affect the convergence rates any longer.

3. Variance reduction in dependent case

In this section, we formulate analogues of Theorem 2 and Theorem 3 for Markov chains. Let $\{X_k\}_{k=0}^\infty$ be a Markov chain (here we number the sequence for convenience from 0) with Markov kernel P and initial value $X_0 = x_0$ with $x_0 \in \mathsf{X}$. We will assume that the Markov kernel P has a unique invariant distribution π.

As was mentioned in the Introduction, there are several estimates for the asymptotic variance $V^\infty(h)$ available in the literature, see J. Flegal and G. Jones [FJ10]. For the sake of clarity, we consider only the spectral variance estimator which provides the most generic way to estimate $V^\infty(h)$. Nevertheless, similar results to those below can be obtained for other estimates. We start with a definition. Let $\mathcal{H} = \{f - g : \ g \in \mathcal{G}\}$, where $\mathcal{G}$ is the class of control variates (see Section 1). We denote the autocovariance function of the process $\{h(X_k)\}_{k=0}^\infty$ by

$$\rho(s) := \mathrm{E}\big[(h(X_0) - \pi(h))(h(X_k) - \pi(h))\big],$$

where it is assumed that $X_0 \sim \pi$. The corresponding sample autocovariance function is denoted by

$$\hat{\rho}_n(s) := \frac{1}{n} \sum_{k=0}^{n-s-1} \big(h(X_k) - \pi_n(h)\big)\big(h(X_{k+s}) - \pi_n(h)\big).$$

It can be easily seen that the asymptotic variance $V^\infty(h)$ defined in (7) can be rewritten as

$$V^\infty(h) = \sum_{s=-\infty}^{+\infty} \rho(|s|).$$

The spectral variance estimator is based on truncation and weighting of the sample autocovariance function and is given by

$$V_n^\infty(h) := \sum_{s=-(b_n-1)}^{b_n-1} w_n(s)\hat{\rho}_n(|s|), \tag{10}$$

where w_n is the lag window and b_n is the truncation point. The truncation point b_n is a sequence of integers and the lag window w_n is a kernel of the form $w_n(s) = w(s/b_n)$, where w is a symmetric non-negative function supported on $[-1, 1]$ given by

$$w(s) = \begin{cases} 2s + 2, & -1 \le s < -1/2, \\ 1, & -1/2 \le s \le 1/2, \\ -2s + 2, & 1/2 < s \le 1. \end{cases}$$

Other possible choices of $w(s)$ can be considered, see details in [FJ10]. We will examine the following estimate

$$\widehat{h}_{\varepsilon,2} := \operatorname*{argmin}_{h \in \mathcal{H}_{\varepsilon,2}} V_n^\infty(h),$$

where $\mathcal{H}_{\varepsilon,2}$ is an ε-net of the set $\mathcal{H}$ in $L^2(\pi)$-norm, see definitions in Section 1.

Recall that given two Markov kernels P and Q on $\mathsf{X} \times \mathcal{X}$, where $\mathcal{X}$ is the Borel σ-field on X, we define $PQ(x, A) = \int P(x, \mathrm{d}y)Q(y, A)$. We also define P^n inductively by $P^n = PP^{n-1}$.

Let $W : \mathsf{X} \to [1, \infty)$ be a measurable function. The W-norm of a function $h : \mathsf{X} \to \mathbb{R}$ is given by $\|h\|_W = \sup_{x \in \mathsf{X}} \{|h(x)|/W(x)\}$. For any two probability measures μ and ν on $(\mathsf{X}, \mathcal{X})$ satisfying $\mu(W) < \infty$ and $\nu(W) < \infty$, the W-norm of $\mu - \nu$ is defined by $\|\mu - \nu\|_W = \sup_{\|f\|_W \leq 1} |\mu(f) - \nu(f)|$.

Now let us turn to the assumptions needed to state the main results of the section. Our first assumption is the geometric ergodicity of the Markov chain $\{X_k\}_{k=0}^{\infty}$.

(GE) The Markov kernel P admits a unique invariant probability measure π. Moreover, for some measurable function $W : \mathsf{X} \to [1, \infty)$ with $\pi(W) < \infty$ and some $\varsigma > 0$, $\rho \in (0, 1)$, it holds for all $x \in \mathsf{X}$ and any $n \in \mathbb{N}$ that

$$\|P^n(x, \cdot) - \pi\|_W \leq \varsigma W(x) \rho^n.$$

It is worth noting that, given $h \in \mathcal{H}$, (GE) implies the central limit theorem for $\{h(X_k)\}_{k=0}^{\infty}$ if additionally $\pi(h^{2+\kappa}) < \infty$ for some $\kappa > 0$, see I. A. Ibragimov and Y. V. Linnik [IL71] or G. Jones [J04]. In what follows, we will assume that $\mathcal{H} \subset L^{2+\kappa}$ for some $\kappa > 0$, so that the asymptotic confidence interval (6) is well-defined.

Recall that L^2-Wasserstein distance between probability measures μ and ν is given by

$$W_2(\mu, \nu) = \inf_{\zeta} \left(\int_{\mathsf{X} \times \mathsf{X}} \|x - y\|^2 \, \mathrm{d}\zeta(x, y) \right)^{1/2},$$

where $\|\cdot\|$ is the euclidian norm in $\mathbb{R}^d$ and the infimum is taken over all probability measures ζ on the product space $\mathsf{X} \times \mathsf{X}$ with marginal distributions μ and ν. The Kullback-Leibler divergence for μ and ν is defined as

$$\mathrm{KL}(\mu \| \nu) = \begin{cases} \int \log(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}) \mathrm{d}\mu, & \mu \ll \nu, \\ \infty, & \text{otherwise.} \end{cases}$$

We say that probability measure μ satisfies the transportation cost-information inequality $\mathrm{T}_2(C)$ if there exists a constant $C > 0$ such that for any probability measure ν

$$W_2(\mu, \nu) \leq \sqrt{2C \, \mathrm{KL}(\mu \| \nu)}.$$

In the proof of the main results, we need a concentration inequality for $V_n^{\infty}(h)$. It is important to note that $V_n^{\infty}(h)$ is a quadratic form in $\{h(X_j)\}_{j=0}^{n-1}$. We have proved the Gaussian concentration for $V_n^{\infty}(h)$ in the case when functions $h \in \mathcal{H}$ are Lipschitz and the Markov kernel P is a contraction in L^2-Wasserstein distance.

(L) Functions $h \in \mathcal{H}$ are L-Lipschitz, that is, $|h(x) - h(y)| \leq L\|x - y\|$ for some $L > 0$ and all $x, y \in \mathsf{X}$.

(CW) The Markov kernel $P(x,\cdot)$ satisfies $\mathrm{T}_2(C)$ for any $x \in \mathsf{X}$ and some $C > 0$. Moreover, there exists $r \in (0,1)$ such that $W_2(P(x,\cdot), P(y,\cdot)) \le r\|x-y\|$ for any $x, y \in \mathsf{X}$.

Now we can state the following theorem.

Theorem 6. *Assume **(GE)**, **(L)**, and **(CW)**. Set $b_n = 2\log(n)/\log(1/\rho)$. Then for some $\varepsilon > 0$ (specified in the proof), any initial value $x_0 \in X$, and any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$V^\infty\Big(\widehat{h}_{\varepsilon,2}\Big) - \inf_{h\in\mathcal{H}} V^\infty\big(h\big) \lesssim A_1 \log(n)\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H},n) + A_2\frac{\log(n)\log(1/\delta)}{\sqrt{n}},$$

where the first asymptotic variance is taken conditionally on the sample $X_1,\dots,X_n$ used to compute $\widehat{h}_{\varepsilon,2}$, the symbol $\lesssim$ stands for inequality up to an absolute constant, and

$$A_1 = \frac{\sqrt{C}L^2}{(1-r)\log(1/\rho)}, \quad A_2 = \frac{\sqrt{\varsigma}(\pi(W)+W(x_0))}{\sqrt{1-\rho}\log(1/\rho)}\left(\frac{\sqrt{C}L^2}{1-r} + \sup_{h\in\mathcal{H}}\|h\|^2_{W^{1/2}}\right).$$

If the assumptions of Theorem 6 hold and additionally $\mathcal{H}$ is a bounded subset of weighted Sobolev space $W^{s,p}(\mathsf{X}, \langle x\rangle^\beta)$, using the result of Proposition 1, we get the following final high-probability bound

$$V^\infty\Big(\widehat{h}_{\varepsilon,1}\Big) - \inf_{h\in\mathcal{H}} V^\infty\big(h\big) \lesssim n^{-\alpha}, \quad \alpha \in (0,1/2),$$

where the value of parameter α depends on the dimension d and the parameters of the Sobolev space s, p, and β. In statistical literature, such rates are referred to as slow rates of convergence.

This rate can be improved up to $n^{-\alpha}$ for $\alpha \in (0,1)$ by imposing an additional condition on $\mathcal{H}$. To this end, let us consider the case when $\mathcal{H}$ contains a constant function. Since $\pi(h) = \pi(f)$ for all $h \in \mathcal{H}$, this constant must be equal to $\pi(f)$, and hence $\inf_{h\in\mathcal{H}} V^\infty(h) = 0$.

Theorem 7. *Assume **(GE)**, **(L)**, and **(CW)**. Assume also that $\mathcal{H}$ contains a constant function, that is, $h^*(x) \equiv c$ for some $h^* \in \mathcal{H}$ and some $c \in \mathbb{R}$. Set $b_n = 2\log(n)/\log(1/\rho)$. Then for some $\varepsilon > 0$ (specified in the proof), any initial value $x_0 \in X$, and any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$V^\infty\Big(\widehat{h}_{\varepsilon,2}\Big) \lesssim A_1 \log(n)\big(\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H},n)\big)^2 + A_2\frac{\log(n)\log(1/\delta)}{n},$$

where the asymptotic variance is taken conditionally on the sample $X_1,\dots,X_n$ used to compute $\widehat{h}_{\varepsilon,2}$, the symbol $\lesssim$ stands for inequality up to an absolute constant, and

$$A_1 = \frac{CL^2}{(1-r)^2\log(1/\rho)}, \quad A_2 = \frac{CL^2}{(1-r)^2\log(1/\rho)} + \frac{\varsigma(\pi(W)+W(x_0))}{(1-\rho)^{1/2}\log(1/\rho)}\sup_{h\in\mathcal{H}}\|h\|^2_{W^{1/2}}.$$

Assuming additionally that $\mathcal{H}$ is a bounded subset of weighted Sobolev space $W^{s,p}(\mathsf{X}, \langle x\rangle^\beta)$, we obtain with high probability

$$V^\infty\Big(\widehat{h}_{\varepsilon,2}\Big) \lesssim n^{-\alpha}, \quad \alpha \in (0,1).$$

In turn, the considered variance reduction procedure reduces the length of the asymptotic confidence interval (6) from order $n^{-1/2}$ to $n^{-1/2-\alpha/2}$.

The concentration inequality we obtain in the proof of Theorem 6 and Theorem 7 seems to be of independent interest.

Proposition 8. *Let $\{X_k\}_{k=0}^{\infty}$ be a Markov chain with Markov kernel P and initial value $X_0 = x_0$ with $x_0 \in \mathsf{X}$. Assume that there exists $C > 0$ such that $P(x, \cdot)$ satisfies $\mathrm{T}_2(C)$ for any $x \in \mathsf{X}$. Assume also that there exists $r \in (0, 1)$ such that for any $x, y \in \mathsf{X}$*

$$W_2\big(P(x,\cdot), P(y,\cdot)\big) \le r\|x - y\|.$$

For an L-Lipschitz function $h : \mathsf{X} \to \mathbb{R}$ let $Z_n(h) = (h(X_0), \dots, h(X_{n-1}))^\top$. Then for any $n \times n$ matrix A and any $t > 0$

$$\mathrm{P}\Big(\big|Z_n(h)^\top A Z_n(h) - \mathrm{E}\big[Z_n(h)^\top A Z_n(h)\big]\big| > t\Big) \le 2\exp\bigg(-\frac{t^2}{cK^2\left(\mathrm{E}\|AZ_n(h)\|^2 + t\|A\|\right)}\bigg),$$

where $K^2 = CL^2/(1-r)^2$ and $c > 0$ is some universal constant.

The proof of this concentration inequality falls naturally into two steps. First we show, using a result from H. Djellout, A. Guillin and L. Wu [DGW04], that the joint distribution of $\{X_k\}_{k=0}^{n-1}$ satisfies $\mathrm{T}_2(C/(1-r)^2)$ model which implies the Gaussian concentration for all Lipschitz functions. Further, using the ideas from R. Adamczak [A15], we prove the concentration inequality for quadratic forms.

4. Unadjusted Langevin Algorithm

One simple instance of a Markov Chain Monte Carlo algorithm satisfying (GE) and (CW) is the Unadjusted Langevin Algorithm (ULA). Let the probability density function $\pi(x)$ be known up to a normalizing constant $\pi(x) \propto \mathrm{e}^{-U(x)}$ for some non-negative function $U(x)$ called the potential. In the Unadjusted Langevin Algorithm, the discrete-time Markov chain $\{X_k\}_{k\ge 0}$ is defined recurrently as

$$X_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma}\, Z_{k+1}, \tag{11}$$

where $\{Z_k\}_{k\ge 1}$ is an i.i.d. sequence of d-dimensional standard Gaussian random variables and $\gamma > 0$ is a step size. This algorithm is a first-order Euler-Maruyama discretization of the Langevin stochastic differential equation

$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

where $\{B_t\}_{t\geq 0}$ is the standard d-dimensional Brownian motion. Under mild technical conditions, the Langevin diffusion $\{Y_t\}_{t\geq 0}$ admits π as its unique invariant distribution and the Markov chain $\{X_k\}_{k\geq 0}$ converges to a stationary distribution π_γ which is close to π (in a sense that one can bound the distance between π_γ and π in, for example, L^2-Wasserstein distance or total variation). More information on the unadjusted Langevin algorithm and the aforementioned convergence theorems can be found in G. Roberts, R. Tweedie [RT96], A. Dalalyan [D17], A. Durmus, É. Moulines [DM17]. When the density $\pi(x)$ is not known, to apply Unadjusted Langevin Algorithm one can use methods from the papers of D. Belomestny and the author [BI19] and [BI20].

Consider the following conditions on the potential $U(x)$ needed for the assumptions (GE) and (CW) to hold.

(U1) The potential U is 2-times continuously differentiable, $U \in C^2(\mathbb{R}^d)$, and has Lipschitz gradient, that is, there exists $L > 0$ such that for all $x, y \in \mathbb{R}^d$

$$\|\nabla U(x) - \nabla U(y)\| \leq L\|x - y\|.$$

(U2) The potential U is strongly convex, that is, there exists $m > 0$ such that for all $x, y \in \mathbb{R}^d$

$$U(y) \geq U(x) + \langle \nabla U(x), y - x\rangle + \frac{m}{2}\|x - y\|^2.$$

Proposition 9. *Assume (U1) and (U2). Then for any $\gamma \in (0, 2/(m + L))$, the Markov kernel P_γ associated to the chain $\{X_k\}_{k\geq 0}$ from (11) fulfills (GE) and (CW) with some $\rho \in (0, 1)$,*

$$W(x) = \|x\|^2, \quad C = 2\gamma, \quad \textit{and} \quad r = \sqrt{1 - 2\gamma m L/(m + L)}.$$

This fact is a direct consequence of the results of the two papers of A. Durmus, É. Moulines [DM17] and [DM19].

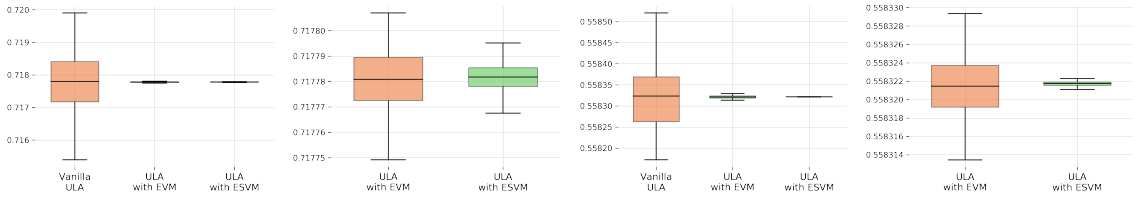
Bayesian Logistic Regression. Let us consider a numerical example for Unadjusted Langevin Algorithm. Our aim is to compare the following two methods to choose a control variate: minimization of Empirical Variance (EVM method), see (4), and minimization of Spectral Variance (ESVM method), see (10). We analyze performance of both algorithms on two classification datasets from the UCI repository. The first dataset, Pima, contains $N = 768$ observations in dimension $d = 9$. The second one, EEG, has $N = 14\,980$ observations in dimension $d = 15$.

In logistic regression, the probability of the i-th output $y_i \in \{-1, 1\}$ for $i = 1, \ldots, N$ is given by $\mathrm{p}(y_i|\mathbf{x}_i, \theta) = (1 + \mathrm{e}^{-y_i\langle\theta,\mathbf{x}_i\rangle})^{-1}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is a vector of predictors and $\theta \in \mathbb{R}^d$ is the vector of unknown regression coefficients. First we split each dataset into a training set $\mathcal{T}^{\text{train}} = \{(y_i, \mathbf{x}_i)\}_{i=1}^{N-K}$ and a test set $\mathcal{T}^{\text{test}} = \{(y_i', \mathbf{x}_i')\}_{i=1}^{K}$ by randomly picking K test points from the data

(and excluding them from the training set). Then we complete the Bayesian model by considering the standard normal prior $\mathrm{p}_0(\theta)$ for θ and use Unadjusted Langevin Algorithm to sample from the posterior distribution $\mathrm{p}(\theta|\mathcal{T}^{\text{train}}) \propto \mathrm{p}_0(\theta)\prod_{(y_i,\mathbf{x}_i)\in\mathcal{T}^{\text{train}}}\mathrm{p}(y_i|\mathbf{x}_i,\theta)$. Given a sample $\{\theta_k\}_{k=0}^{n-1}$, we estimate the predictive distribution $\mathrm{p}(y'|\mathbf{x}') = \int_{\mathbb{R}^d}\mathrm{p}(y'|\mathbf{x}',\theta)\,\mathrm{p}(\theta|\mathcal{T}^{\text{train}})\,\mathrm{d}\theta$ for a fixed test point $(y',\mathbf{x}')$ by computing the estimate $n^{-1}\sum_{k=0}^{n-1} f(\theta_k)$ for $f(\theta) = \mathrm{p}(y'|\mathbf{x}',\theta)$. To get rid of randomness, we also estimate the average predictive distribution for the whole test set $\mathcal{T}^{\text{test}}$ by computing the same estimate for the function $f(\theta) = K^{-1}\sum_{(y_i,\mathbf{x}_i)\in\mathcal{T}^{\text{test}}}\mathrm{p}(y_i'|\mathbf{x}_i',\theta)$. Moreover, we compute similar estimates for the EVM and ESVM methods, where we use the Stein control variates, see (3), with $\Phi(x)\equiv b$ for some vector $b\in\mathbb{R}^d$.

The boxplots of the estimated average predictive distribution are shown in Figure 1. Note that both the EVM and ESVM methods lead to a significant variance reduction with a slightly better performance of ESVM (since the latter method takes into account the Markovian structure of a chain).

Figure 1: Estimation of the average predictive distribution in Bayesian logistic regression. The first two graphs are given for the Pima dataset, the second two graphs are given for the EEG dataset. In the first graph for each dataset, each boxplot corresponds to 100 estimates computed by using the vanilla ULA, ULA with EVM, and ULA with ESVM. We zoom in the last two boxplots on the second graph.



Main results of the present thesis

In the present thesis, the following results are obtained by the author.

1) Nonasymptotic bounds on the excess risk (5) in the independent case for a general class of control variates $\mathcal{G}$. These results are given in Theorem 2 and Theorem 3. Moreover, it was demonstrated what bounds can be obtained when Stein control variates are used, see Corollary 4 and Corollary 5.

2) Nonasymptotic bounds on the excess risk (8) in the case of Markov chains for a general class of control variates $\mathcal{G}$. These results are given in Theorem 6 and Theorem 7. Moreover, it was suggested to minimize an estimate of the asymptotic variance (10) instead of the sample variance (4). This criteria works better in practice when dependent random variables are considered.

3) A new concentration inequality for a quadratic form of a Markov chain when its Markov kernel is a contraction in L^2-Wasserstein distance. This result is given in Proposition 8.

These results are both theoretical and practical. Methods developed in the present work can be used for further exploration of variance reduction algorithms for Monte Carlo estimates. In addition, the obtained results can be useful in several problems in Bayesian inference.

Personal contribution of the author

The main results listed above reflect the author's personal contribution to published works. The contribution of the author was decisive in all the results listed above except Proposition 8. Propositions that are not listed above but are given in the present text are either simple consequences of the results that do not belong to the author and his co-authors or belong to the co-authors and are given here only to make our exposition self-contained.

Approbation of results

The results of the thesis were presented at the following conferences and seminars.

- Conference "Information technologies and systems" at IITP RAS, Saint Petersburg, September 2016. Talk "Concentration for the Euclidean norm of an isotropic log-concave random vector".

- PhD Seminar of Probability Theory Department at MSU, Moscow, December 2017. Talk "Variance reduction for Monte Carlo estimates".
- Conference "New frontiers in high-dimensional probability and statistics" at NRU HSE, Moscow, February 2018. Talk "Variance reduction in Monte Carlo via empirical variance minimization".
- Conference "Lomonosov" at MSU, Moscow, April 2018. Talk "Variance reduction in Monte Carlo via empirical variance minimization".
- Seminar on Probability Theory and Mathematical Statistics at PDMI RAS, Saint Petersburg, September 2018. Talk "Variance reduction in Monte Carlo via empirical variance minimization".
- Winter school "New frontiers in high-dimensional probability and statistics 2" at NRU HSE, Moscow, February 2019. Talk "MCMC estimation for distributions with known characteristic function".
- Seminar of Applied Mathematics Department at École Polytechnique, Paris, June 2019. Talk "Variance reduction via empirical variance minimization".
- Conference "European Meeting of Statisticians" (EMS), organized by Bernoulli Society for Mathematical Statistics and Probability, Palermo, July 2019. Talk "Variance reduction for Markov chains via empirical variance minimization".
- Conference "Structural Inference in High-Dimensional Models 2", Saint Petersburg, September 2019. Talk "MCMC algorithms for heavy-tailed distributions".
- Third winter conference on probability theory and mathematical physics at PDMI RAS, Saint Petersburg, December 2019. Talk "Variance reduction for dependent sequences".

Publications

The present thesis is based on three published papers.

[1] L. Iosipoi, D. Belomestny, N. Zhivotovskiy. *Variance Reduction in Monte Carlo Estimators via Empirical Variance Minimization.* Doklady Mathematics, 98:2, 494–497, 2018.

[2] D. Belomestny, L. Iosipoi. *On density estimation via Fourier series.* Large-Scale Systems Control, 82, 28–43, 2019.

[3] D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, S. Samsonov. *Variance reduction for Markov chains via empirical variance minimization with application to MCMC.* Statistics and Computing, 30:4, 973–997, 2020.

Bibliography

[A15] R. Adamczak. *A note on the Hanson-Wright inequality for random vectors with dependencies.* Electron. Commun. Probab., 20:71, 1–13, 2015.

[B71] N. S. Bakhvalov. *On the approximate calculation of multiple integrals.* Vestnik MGU, Ser. Math. Mech. Astron. Phys. Chem., 4, 3–18, 1959.

[BI19] D. Belomestny, L. Iosipoi. *On density estimation via Fourier series.* Large-Scale Systems Control, 82, 28–43, 2019.

[BI20] D. Belomestny, L. Iosipoi. *Fourier transform MCMC, heavy tailed distributions, and geometric ergodicity.* ArXiv preprint, arXiv:1909.00698, 2020.

[CLV08] S. Clémençon, G. Lugosi, N. Vayatis. *Ranking and Empirical Minimization of U-statistics.* Ann. Statist., 36:2, 844–874, 2008.

[D17] A. Dalalyan. *Theoretical guarantees for approximate sampling from smooth and log-concave densities.* Journal of the Royal Statistical Society Series B (Statistical Methodology), 79:3, 651–676, 2017.

[DGL96] L. Devroye, L. Györfi, G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer, New York, 1996.

[DGW04] H. Djellout, A. Guillin, L. Wu. *Transportation cost-information inequalities and applications to random dynamical systems and diffusions. Ann. Probab.*, 32:3B, 2702–2732, 2004.

[D08] I. Dimov. *Monte Carlo methods for applied scientists.* World Scientific, 2008.

[DM17] A. Durmus, É. Moulines. *Non-asymptotic convergence analysis for the unadjusted Langevin algorithm.* Ann. Appl. Probab., 27:3, 1551–1587, 2017.

[DM19] A. Durmus, É. Moulines. *High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm.* Bernoulli, 4A, 2854–2882, 2019.

[FJ10] J. Flegal, G. Jones. *Batch means and spectral variance estimators in Markov chain Monte Carlo.* Ann. Statist., 38:2, 1034–1070, 2010.

[G00] S. van de Geer. *Empirical Processes in M-Estimation.* Cambridge, 2000.

[G13] P. Glasserman. *Monte Carlo methods in financial engineering.* Springer Science & Business Media, 2013.

[IL71] I. A. Ibragimov, Y. V. Linnik. *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff, Groningen, 1971.

[J04] G. Jones. *On the Markov chain central limit theorem.* Probab. Surveys, 1, 299–320, 2004.

[MSI13] A. Mira, R. Solgi, D. Imparato. *Zero variance Markov chain Monte Carlo for Bayesian Estimators.* Statistics and Computing, 23:5, 653–662, 2013.

[NP07] R. Nickl, B. Pötscher. *Bracketing Metric Entropy Rates and Empirical Central Limit Theorems for Function Classes of Besov- and Sobolev-Type.* Journal of Theoretical Probability, 20:2, 177–199, 2007.

[N16] E. Novak. *Some Results on the Complexity of Numerical Integration.* Monte Carlo and Quasi-Monte Carlo Methods. Springer Proceedings in Mathematics & Statistics, 163, 161–183, 2016.

[OGC17] C. Oates, M. Girolami, N. Chopin. *Control functionals for Monte Carlo integration.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79:3, 695–718, 2017.

[OGC19] C. Oates, M. Girolami, N. Chopin. *Convergence rates for a class of estimators based on Stein's identity.* Bernoulli, 25:2, 1141–1159, 2019.

[RC99] C. Robert, G. Casella. *Monte Carlo Statistical Methods.* Springer, New York, 1999.

[RT96] G. Roberts, R. Tweedie. *Exponential convergence of Langevin distributions and theirdiscrete approximations.* Bernoulli, 2:4, 341–363, 1996.

[RR04] G. Roberts, J. Rosenthal. *General state space Markov chains and MCMC algorithms.* Probab. Surveys, 1, 20–71, 2004.

[RK16] R. Rubinstein, D. Kroese. *Simulation and the Monte Carlo method.* John Wiley & Sons, 2016.

[WS95] W. H. Wong, X. Shen. *Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLES.* Ann. Statist., 23:2, 339–362, 1995.