Variance reduction for Markov chains with application to MCMC

D Belomestny, L Iosipoi, E Moulines, A Naumov, S Samsonov

Variance reduction for Markov chains with application to MCMC

D. Belomestny [*], L. Iosipoi [†], E. Moulines [‡], A. Naumov [§], and S. Samsonov [¶]

Abstract

In this paper we propose a novel variance reduction approach for additive functionals of Markov chains based on minimization of an estimate for the asymptotic variance of these functionals over suitable classes of control variates. A distinctive feature of the proposed approach is its ability to significantly reduce the overall finite sample variance. This feature is theoretically demonstrated by means of a deep non asymptotic analysis of a variance reduced functional as well as by a thorough simulation study. In particular we apply our method to various MCMC Bayesian estimation problems where it favourably compares to the existing variance reduction approaches.

1 Introduction

Variance reduction methods play nowadays a prominent role as a complexity reduction tool in simulation based numerical algorithms like Monte Carlo (MC) or Markov Chain Monte Carlo (MCMC). Introduction to many of variance reduction techniques can be found in Robert and Casella [32], Rubinstein and Kroese [36], Gobet [18], and Glasserman [17]. While variance reduction techniques for MC algorithms are well studied, MCMC algorithms are still waiting for efficient variance reduction methods. Recently one witnessed a revival of interest in this area with numerous applications to Bayesian statistics, see for example Dellaportas and Kontoyiannis [9], Mira et al. [26], Brosse et al. [7], and references therein. The main difficulty in constructing efficient variance reduction methods for MCMC lies in the dependence between the successive values of the underlying Markov chain which can significantly increase the overall variance and needs to be accounted for.

Suppose that we wish to compute $\pi(f) \stackrel{\text{def}}{=} \mathsf{E}_\pi\big[f(X)\big]$, where X is a random vector with a distribution π on $\mathsf{X} \subseteq \mathbb{R}^d$ and $f : \mathsf{X} \to \mathbb{R}$ with $f \in \mathrm{L}^2(\pi)$. Let $(X_k)_{k\geq 0}$ be a time homogeneous Markov chain with values in X. Denote by P its Markov kernel and define for any bounded measurable function f

$$Pf(x) = \int_{\mathsf{X}} P(x, \mathrm{d}y) f(y)\,, \quad x \in \mathsf{X}.$$

Assume that P has the unique invariant distribution π, that is, $\int_{\mathsf{X}} \pi(\mathrm{d}x) P(x, \mathrm{d}y) = \pi(\mathrm{d}y)$. Under appropriate conditions, the Markov kernel P may be shown to converge to the stationary distribution π, that is, for any $x \in \mathsf{X}$,

$$\lim_{n\to\infty} \|P^n(x, \cdot) - \pi\|_{\mathrm{TV}} = 0,$$

[*]Duisburg-Essen University, Germany, and HSE University, Russia, denis.belomestny@uni-due.de.
[†]HSE University, Russia, iosipoileonid@gmail.com.
[‡]Ecole Polytechnique, France, and HSE University, Russia, eric.moulines@polytechnique.edu.
[§]HSE University, Russia, anaumov@hse.ru.
[¶]HSE University, Russia, svsamsonov@hse.ru.

where $\|\mu-\nu\|_{\mathrm{TV}}=\sup_{A\in\mathcal{X}}|\mu(A)-\nu(A)|$ and $\mathcal{X}$ is the Borel σ-field associated to X. More importantly, under rather weak assumptions, the ergodic averages

$$\pi_n(f)\stackrel{\mathrm{def}}{=}n^{-1}\sum_{k=0}^{n-1}f(X_k)$$

satisfy, for any initial distribution, a central limit theorem (CLT) of the form

$$\sqrt{n}\big[\pi_n(f)-\pi(f)\big]=n^{-1/2}\sum_{k=0}^{n-1}\big[f(X_k)-\pi(f)\big]\xrightarrow{\mathcal{D}}\mathcal{N}\big(0,V_\infty(f)\big)$$

with the asymptotic variance $V_\infty(f)$ given by

$$V_\infty(f)\stackrel{\mathrm{def}}{=}\lim_{n\to\infty}n\,\mathsf{E}_\pi\big[\{\pi_n(f)-\pi(f)\}^2\big]=\mathsf{E}_\pi\big[\tilde{f}^2\big]+2\sum_{k=1}^{\infty}\mathsf{E}_\pi\big[\tilde{f}P^k\tilde{f}\big],\tag{1}$$

where $\tilde{f}=f-\pi(f)$. This motivates to use ergodic averages $\pi_n(f)$ as a natural estimate for $\pi(f)$. For a broader discussion of the Markov chain CLT and conditions under which CLT holds, see Jones [23], Roberts and Rosenthal [33], and Douc et al. [12].

One important and widely used class of variance reduction methods for Markov chains is the method of control variates which is based on subtraction of a zero-mean random variable (control variate) from $\pi_n(f)$. There are several methods to construct such control variates. If $\nabla\log\pi$ is known, one can use popular zero-variance control variates based on the Stein's identity, see Assaraf and Caffarel [2] and Mira et al. [26]. A non-parametric extension of such control variates is suggested in Oates et al. [29] and Oates et al. [28]. Control variates can be also obtained using the Poisson equation. Namely, it was observed by Henderson [21] that the function $U_g\stackrel{\mathrm{def}}{=}g-Pg$ has zero mean with respect to π, provided that $\pi(|g|)<\infty$. Then the choice $g=\hat{f}$ with $\hat{f}$ satisfying the so-called Poisson equation $\hat{f}(x)-P\hat{f}(x)=\tilde{f}(x)$ leads to $f-U_{\hat{f}}=f-\hat{f}+P\hat{f}=\pi(f)$ hence yielding a zero-variance control variate for the empirical mean under π. Although the Poisson equation involves the quantity of interest $\pi(f)$ and can not be solved explicitly in most cases, the above idea still can be used to construct some approximations for the zero-variance control variate $\hat{f}(x)-P\hat{f}(x)$. For example, Henderson [21] proposed to compute approximations to the solution of the Poisson equation for specific Markov chains with particular emphasis on models arising in stochastic network theory. In Dellaportas and Kontoyiannis [9] and Brosse et al. [7] regression-type control variates are developed and studied for reversible Markov chains. It is assumed in Dellaportas and Kontoyiannis [9] that the one-step conditional expectations can be computed analytically for a set of basis functions. The authors in Brosse et al. [7] proposed another approach tailored to diffusion setting which does require the computation of integrals of basis functions and only involves the application of the underlying differential generator.

There is a fundamental issue related to the control variates method. Since one usually needs to consider a large class of control variates, one has to choose a criterion to select the "best" control variate from this class. In the literature, such a choice is often based on the least squares criterion or on the sample variance, see, for example, Mira et al. [26], Oates et al. [29], South et al. [37]. Note that such criteria can not properly take into account the correlation structure of the underlying Markov chain and hence can only reduce the first term in (1).

In this paper, we propose a novel variance reduction method for Markov chains based on the *empirical spectral variance minimization*. The proposed method can be viewed as a generalization of the approach in Belomestny et al. [5, 4] to Markov chains. In a nutshell, given a class of control variates $\mathcal{G}$, that is, functions $g\in\mathcal{G}$ with $\pi(g)=0$ we consider the

estimator

$$\pi_n(f - \widehat{g}_n) \stackrel{\text{def}}{=} n^{-1} \sum_{k=0}^{n-1} \{f(X_k) - \widehat{g}_n(X_k)\}$$

with $\widehat{g}_n \stackrel{\text{def}}{=} \operatorname{argmin}_{g \in \mathcal{G}} V_n(f-g)$, where $V_n(f)$ stands for an estimator of the asymptotic variance $V_\infty(f)$ defined in (1). This generalization turns out to be challenging for at least two reasons. First, there is no simple way to estimate the asymptotic variance $V_\infty(f)$ for Markov chains. Due to inherent serial correlation, estimating $V_\infty(f)$ requires specific techniques such as spectral and batch means methods; see Flegal and Jones [15] for a survey on variance estimators and their statistical properties. Second, a nonasymptotic analysis of the estimate $\widehat{g}_n$ is highly nontrivial and requires careful treatment. We perform this analysis for a rather general class of geometrically ergodic Markov chains including the well known Unadjusted Langevin Algorithm (ULA), Metropolis-Adjusted Langevin Algorithm (MALA) and Random Walk Metropolis (RWM). In particular, we show that under some restrictions on $\mathcal{G}$, the rate of the excess for the asymptotic variance can be controlled with high probability as follows:

$$V_\infty(f - \widehat{g}_n) - \inf_{g \in \mathcal{G}} V_\infty(f-g) = O\left(n^{-\alpha}\right)$$

for some $\alpha \in [1/2, 1)$. Let us stress that our results are rather generic and can cover various types of control variates. Apart from a comprehensive theoretical analysis we conduct an extensive simulation study including Bayesian inference via MCMC for logistic regression, Gaussian mixtures and Bayesian inference of ODE models. We show that for various MCMC algorithms our approach leads to a further significant variance reduction as compared to the least-squares-type criteria.

The paper is organised as follows. In Section 2 we introduce a general empirical variance minimisation procedure for Markov chains and analyse its properties. In Section 3 we apply our theoretical results to a widely used ULA and MALA. In Section 4 we conduct a thorough numerical study of the proposed approach. Finally all proofs are collected in Section 5 and Appendix A.

Notations Let $\|\cdot\|$ denote the standard Euclidean norm. We say that $f : \mathbb{R}^d \to \mathbb{R}$ is $L-$Lipschitz function if $|f(x) - f(x')| \le L\|x - x'\|$ for any $x, x' \in \mathbb{R}^d$.

For any probability measure ξ on $(\mathsf{X}, \mathcal{X})$, we denote by P_ξ the unique probability under which $(X_n)_{n \geqslant 0}$ is a Markov chain with Markov kernel P and initial distribution ξ. We denote by E_ξ the expectation under the distribution P_ξ. For ξ a probability measure on $(\mathsf{X}, \mathcal{X})$ and $A \in \mathcal{X}$, we denote by $\xi P(A) = \int \xi(\mathrm{d}x) P(x, A)$; for $h : \mathsf{X} \to \mathbb{R}_+$ a measurable function, we denote by $Ph(x) = \int P(x, \mathrm{d}y) h(y)$. 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 $PQ(x, A) = \iint P(x, \mathrm{d}y) Q(y, A)$. 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 defined as $\|h\|_W = \sup_{x \in \mathsf{X}} \{|h(x)|/W(x)\}$. 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 as $\|\mu - \nu\|_W = \sup_{\|f\|_W \le 1} |\mu(f) - \nu(f)|$.

We also use the 2-Wasserstein distance and the Kullback-Leibler divergence in our analysis. The 2-Wasserstein distance between probability measures μ and ν is denoted by $W_2(\mu, \nu) \stackrel{\text{def}}{=} \inf_\zeta \left(\int_{\mathsf{X} \times \mathsf{X}} \|x - y\|^2 \, \mathrm{d}\zeta(x, y) \right)^{1/2}$, where 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 $\mathrm{KL}(\mu \| \nu) = \mathsf{E}_\mu \left[\log(\mathrm{d}\mu / \mathrm{d}\nu) \right]$ if $\nu \ll \mu$ and $\mathrm{KL}(\mu \| \nu) = \infty$ otherwise. We say that the probability measure μ satisfies the transportation cost-information inequality $\mathrm{T}_2(\alpha)$ if there is a constant $\alpha > 0$ such that for any probability measure ν

$$W_2(\mu, \nu) \le \sqrt{2\alpha \, \mathrm{KL}(\mu \| \nu)}. \tag{2}$$

For a real-valued function h on $\mathsf{X} \subset \mathbb{R}^d$ and a σ-finite measure λ on $(\mathsf{X}, \mathcal{X})$ we write $\|h\|_{\mathrm{L}^p(\lambda)} = (\int_{\mathsf{X}} |h(x)|^p \lambda(\mathrm{d}x))^{1/p}$ with $1 \leq p < \infty$. The set of all functions h with $\|h\|_{\mathrm{L}^p(\lambda)} < \infty$ is denoted by $\mathrm{L}^p(\lambda) = \mathrm{L}^p(\mathsf{X}, \lambda)$.

Finally, the Sobolev space is defined as $W^{s,p}(\mathsf{X}) = \{u \in \mathrm{L}^p(\lambda) : \mathrm{D}^\alpha u \in \mathrm{L}^p(\lambda),\ \forall |\alpha| \leqslant s\}$, 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 differential operator of the form $\mathrm{D}^\alpha = \partial^{|\alpha|}/\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}$. 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 $\langle x \rangle^\beta = (1 + \|x\|^2)^{\beta/2}$, $\beta \in \mathbb{R}$, is defined by

$$W^{s,p}(\mathsf{X}, \langle x \rangle^\beta) = \left\{u :\ u \cdot \langle x \rangle^\beta \in W^{s,p}(\mathsf{X})\right\}. \tag{3}$$

The Sobolev norm is defined as $\|u\|_{W^{s,p}(\mathsf{X}, \langle x \rangle^\beta)} = \sum_{|\alpha| \leq s} \left\|\mathrm{D}^\alpha \left(u \langle x \rangle^\beta\right)\right\|_{\mathrm{L}^p(\lambda)}$. We say that $U \subset W^{s,p}(\mathsf{X}, \langle x \rangle^\beta)$ is norm-bounded if there exists $c > 0$, such that $\|u\|_{W^{s,p}(\mathsf{X}, \langle x \rangle^\beta)} \leq c$ for any $u \in U$.

In what follows, we use the symbol $\lesssim$ for inequality up to an absolute constant.

2 Main results

2.1 Empirical spectral variance minimisation (ESVM)

In this paper, we propose a novel approach to choose a control variate from the set $\mathcal{G}$ referred to as the *Empirical Spectral Variance Minimisation* (ESVM). To shorten notation, let us denote by $\mathcal{H} = \mathcal{H}(\mathcal{G})$ a class of functions $h(x) = f(x) - g(x)$, with $g \in \mathcal{G}$. The main idea of the ESVM approach is to select a control variate which minimizes a finite sample estimate for the asymptotic variance $V_\infty(h)$. There are several estimates for $V_\infty(h)$ available in the literature, see Flegal and Jones [15]. For the sake of clarity we consider only the spectral variance estimator which provides the most generic way to estimate $V_\infty(h)$. It is defined as follows. Let P be a Markov kernel admitting a unique invariant probability π and set $\tilde{h} \stackrel{\text{def}}{=} h - \pi(h)$ (assuming $\pi(|h|) < \infty$). For $s \in \mathbb{Z}_+$, define the stationary lag s autocovariance $\rho_\pi^{(h)}(s) \stackrel{\text{def}}{=} \mathsf{E}_\pi\left[\tilde{h}(X_s)\tilde{h}(X_0)\right]$ and the lag s sample autocovariance via

$$\hat{\rho}_n^{(h)}(s) \stackrel{\text{def}}{=} n^{-1} \sum_{k=0}^{n-s-1} \{h(X_k) - \pi_n(h)\}\{h(X_{k+s}) - \pi_n(h)\}, \tag{4}$$

where $\pi_n(h) \stackrel{\text{def}}{=} n^{-1} \sum_{j=0}^{n-1} h(X_j)$. The spectral variance estimator is based on truncation and weighting of the sample autocovariance function,

$$V_n(h) \stackrel{\text{def}}{=} \sum_{s=-(b_n-1)}^{b_n-1} w_n(s) \hat{\rho}_n^{(h)}(|s|)\,, \tag{5}$$

where w_n is the *lag window* and b_n is the *truncation point*. The truncation point is a sequence of integers and the lag window is a kernel of the form $w_n(s) = w(s/b_n)$, where w is a symmetric non-negative function supported on $[-1, 1]$ which fulfils $|w(s)| \leq 1$ for $s \in [-1, 1]$ and $w(s) = 1$ for $s \in [-1/2, 1/2]$. Other possible choices of the lag window w_n can be considered, see Flegal and Jones [15]. In the ESVM approach we choose a control variate by minimizing the spectral variance

$$\widehat{h} \stackrel{\text{def}}{=} \operatorname*{argmin}_{h \in \mathcal{H}} V_n(h). \tag{6}$$

As the class $\mathcal{H}$ can be too large making the resulting optimization problem (6) computationally intractable, we consider a smaller class. Given $\varepsilon > 0$, let $\mathcal{H}_\varepsilon \subset \mathcal{H}$ consist of centres of the minimal ε-covering net of $\mathcal{H}$ with respect to the $\mathrm{L}^2(\pi)$ distance. Further set

$$\widehat{h}_\varepsilon \stackrel{\text{def}}{=} \operatorname*{argmin}_{h \in \mathcal{H}_\varepsilon} V_n(h). \tag{7}$$

In what follows, we assume that $\mathcal{H}$ is a norm-bounded set in $\mathrm{L}^2(\pi)$. Hence the set $\mathcal{H}_\varepsilon$ is finite. The estimates of the form (7) are referred to as skeleton or sieve estimates in the statistical literature (see, for example, Wong and Shen [39], Devroye et al. [10], and van de Geer [38]).

2.2 Theoretical analysis

In this section, we analyze the proposed ESVM procedure in terms of the excess of the asymptotic variance. Namely, we provide non-asymptotic bounds of the form:

$$V_\infty(\widehat{h}_\varepsilon) - \inf_{h \in \mathcal{H}} V_\infty(h) = O\left(n^{-\alpha}\right), \quad 1/2 < \alpha < 1, \tag{8}$$

holding with high probability.

Before we proceed to theoretical results, let us define a quantity which is used to choose a radius ε of the covering net $\mathcal{H}_\varepsilon$ over which $\widehat{h}_\varepsilon$ is computed. Given any $\varepsilon > 0$, let $H_{\mathrm{L}^2(\pi)}(\mathcal{H}, \varepsilon)$ be a metric entropy of $\mathcal{H}$ in $\mathrm{L}^2(\pi)$, that is, $H_{\mathrm{L}^2(\pi)}(\mathcal{H}, \varepsilon) \stackrel{\text{def}}{=} \log |\mathcal{H}_\varepsilon|$, where $|\mathcal{H}_\varepsilon|$ is cardinality of $\mathcal{H}_\varepsilon$ (which is assumed to be finite). Define by $\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n)$ a so-called *fixed point*

$$\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n) \stackrel{\text{def}}{=} \inf\{\eta > 0 : \; H_{\mathrm{L}^2(\pi)}(\mathcal{H}, \eta) \le n\eta^2\}. \tag{9}$$

Note that a number $\eta > 0$ satisfying $H_{\mathrm{L}^2(\pi)}(\mathcal{H}, \eta) \le n\eta^2$ is finite because of monotonicity of the metric entropy and the mapping $\eta \to n\eta^2$ in η. The quantity $\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n)$ is used to control the cardinality of $\mathcal{H}_\varepsilon$. Indeed by choosing $\varepsilon \ge \gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n)$ we get $|\mathcal{H}_\varepsilon| \le \mathrm{e}^{n\varepsilon^2}$. It is easily seen from the above definition that the fixed point is a decreasing function in n. Let us discuss a typical behaviour of $\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n)$ as $n \to \infty$ when $\mathcal{H}$ is a subset of the weighted Sobolev space $W^{s,p}(\mathsf{X}, \langle x \rangle^\beta)$, see (3) for definition. The following result can be derived from Nickl and Pötscher [27].

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 $s - d/p > 0$. Let also for some $\alpha > 0$, $\|\langle x \rangle^{\alpha - \beta}\|_{\mathrm{L}^2(\pi)} < \infty$. Then it holds*

$$\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n) \lesssim \begin{cases} n^{-\frac{1}{2+d/s}} & \text{for } \alpha > s - d/p, \\ n^{-\frac{1}{2+(\alpha/d+1/p)^{-1}}} & \text{for } \alpha < s - d/p. \end{cases}$$

Now let us turn to assumptions needed for (8) to hold. Our first assumption is the *geometric ergodicity* of the Markov chain $(X_k)_{k \ge 0}$. Let $W : \mathsf{X} \to [1, \infty)$ be a measurable function.

(GE) The Markov kernel P admits a unique invariant probability measure π such that $\pi(W) < \infty$ and there exist $\varsigma > 0, 0 < \rho < 1$ such that for all $x \in \mathsf{X}$ and $n \in \mathbb{N}$,

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

(BR) There exist a non-empty set $\mathsf{S} \subset \mathsf{X}$ and real numbers $u > 1, J > 0$ and $l > 0$ such that

$$\sup_{x \in \mathsf{S}} \mathsf{E}_x[u^{-\sigma}] \le J \quad \text{and} \quad \sup_{x \in \mathsf{S}} W(x) \le l, \tag{10}$$

where σ is the return time to the set S.

Remark 2 *Let us introduce drift and small set conditions.*

(DS) *The Markov kernel P is irreducible, aperiodic and*

- *there exist measurable function $W : \mathsf{X} \to [1, \infty)$, $\lambda \in [0, 1)$, $b < \infty$, and $l < \infty$ such that $\lambda + 2b/(1 + l) < 1$ and*

$$PW \leq \lambda W + b\mathbb{1}_{\{W \leq l\}}. \tag{11}$$

- *there exist $m, \varepsilon > 0$ such that for all $x, x' \in \{W \leq l\}$, $\|\delta_x P^m - \delta_{x'} P\|_{\mathrm{TV}} \leq 2(1-\varepsilon)$.*

It follows from Douc et al. [12, Theorem 19.5.1]) that (DS) implies (GE) and by Douc et al. [12, Proposition 14.1.2]) (DS) implies (BR). Explicit expressions for the constants ς and ρ may be found in Douc et al. [12, Theorem 19.4.1]). Note also that (GE) implies that P is positive, aperiodic and condition (DS) is satisfied for some small set S and some function W_0 verifying $W \leq W_0 \leq \varsigma_0 W$ and constants $\varsigma_0 < \infty$, $b_0 < \infty$, $\lambda_0 \in [0, 1)$. Hence (GE) implies (BR) for some constants $u > 1$ and $J > 0$ (see Douc et al. [12, Theorem 15.2.4].

We also need a *Gaussian concentration* for $V_n(h)$, which requires an additional assumption on the class $\mathcal{H}$. It is important to note that $V_n(h)$ is a quadratic form of $(h(X_j))_{j=0}^{n-1}$. As a result, without much surprise, concentration results for the quadratic forms of Markov Chains shall play a key role in our analysis. We shall consider below two situations. While the first situation corresponds to bounded functions h, the second one deals with Lipschitz continuous functions h. In the second case we additionally assume a contraction in L^2-Wasserstein distance. Thus we assume either

(B) *Bounded case*: There exist $B > 0$ such that $\sup_{h \in \mathcal{H}} |h|_\infty \leq B$ with $|h|_\infty = \sup_{x \in \mathsf{X}} |h(x)|$

or

(L) *Lipschitz case*: Functions $h \in \mathcal{H}$ are L-Lipschitz.

together with

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

The rate of convergence for the variance excess is given in the following theorem.

Theorem 3 *Assume (GE) and either (L)+(CW) or (B)+(BR). Set $b_n = 2(\log(1/\rho))^{-1} \log(n)$ and take $\varepsilon = \gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n)$. Then for any $\delta \in (0, 1)$, there is $n_0 = n_0(\delta) > 0$ such that for any $n \geq n_0$ and $x_0 \in \mathsf{X}_0$ with P_{x_0} −probability at least $1 - \delta$, it holds*

$$V_\infty(\widehat{h}_\varepsilon) - \inf_{h \in \mathcal{H}} V_\infty(h) \lesssim C_1 \log(n) \gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n) + C_2 \frac{\log(n) \log(1/\delta)}{\sqrt{n}},$$

where $\lesssim$ stands for inequality up to an absolute constant,

$$C_1 = \frac{K^2}{\log(1/\rho)} \quad C_2 = \frac{\varsigma^{1/2}(\pi(W) + W(x_0))}{(1-\rho)^{1/2} \log(1/\rho)} \left(K^2 + \sup_{h \in \mathcal{H}} \|h\|^2_{W^{1/2}} \right),$$

$\mathsf{X}_0 = \mathsf{X}$, $K^2 = \sqrt{\alpha} L^2/(1 - r)$ under (L)+(CW) and $\mathsf{X}_0 = \mathsf{S}$, $K^2 = \beta B^2$ under (B)+(BR), with

$$\beta = \frac{\varsigma l}{1 - \rho} \left(\frac{1}{\log u} + \frac{J \varsigma l}{1 - \rho} \right) .$$

In view of Proposition 1, Theorem 3 may be summarized by saying that the excess variance $V_\infty(\widehat{h}_\varepsilon) - \inf_{h \in \mathcal{H}} V_\infty(h)$ is bounded with high probability by a multiple of $n^{-1/2+\eta}$ for some $\eta > 0$ depending on the capacity of the class $\mathcal{H}$. In statistical literature, such rates are referred to as slow rates of convergence. These rates can be improved by imposing

additional conditions on $\mathcal{H}$. To this end let 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 $\inf_{h\in\mathcal{H}} V_n(h) = 0$. In this case, we obtain tighter bounds.

Theorem 4 *Assume (GE), (L), and (CW). Assume also that $\mathcal{H}$ contains a constant function $h^*(x) \equiv const$. Fix the size of the lag window $b_n = 2(\log(1/\rho))^{-1}\log(n)$ and take $\varepsilon = \gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n)$. Then for any $\delta \in (0,1)$, there is $n_0 = n_0(\delta) > 0$ such that for all $n \geq n_0$ and $x_0 \in \mathsf{X}$ it holds with P_{x_0} −probability at least $1-\delta$,*

$$V_\infty(\widehat{h}_\varepsilon) \lesssim C_1 \log(n)\gamma^2_{\mathrm{L}^2(\pi)}(\mathcal{H}, n) + C_2 \frac{\log(n)\log(1/\delta)}{n}, \tag{12}$$

where

$$C_1 = \frac{\alpha L^2}{(1-r)^2\log(1/\rho)} \quad \textit{and} \quad C_2 = \frac{\alpha L^2}{(1-r)^2\log(1/\rho)} + \frac{\varsigma(\pi(W)+W(x_0))}{(1-\rho)^{1/2}\log(1/\rho)} \sup_{h\in\mathcal{H}} \|h\|^2_{W^{1/2}}.$$

In view of Proposition 1, Theorem 4 asserts that under an additional assumption that $\mathcal{H}$ contains a constant function, the excess variance $V_\infty(\widehat{h}_\varepsilon) - \inf_{h\in\mathcal{H}} V_\infty(h)$ can be bounded by a multiple of $n^{-1+\eta}$ for some $\eta > 0$ depending on $\mathcal{H}$.

3 Application to Markov Chain Monte Carlo

In this section we consider the application of the ESVM approach to MCMC-type algorithms. The main goal of MCMC algorithms is to estimate expectations with respect to a probability measure π on $\mathbb{R}^d$, $d \geq 1$, with a density π of the form $\pi(x) = \mathrm{e}^{-U(x)}/\int_{\mathbb{R}^d} \mathrm{e}^{-U(y)}\mathrm{d}y$ with respect to the Lebesgue measure, where U is a nonnegative potential. Let x^* be such that $\nabla U(x^*) = 0$ and without loss of generality we assume $x^* = 0$. Consider the following conditions on the potential U.

(LD1) The function U is continuously differentiable on $\mathbb{R}^d$ with Lipschitz continuous gradient: there exists $\mathrm{L}_U > 0$ such that for all $x, y \in \mathbb{R}^d$,

$$\|\nabla U(x) - \nabla U(y)\| \leq \mathrm{L}_U \|x - y\| .$$

(LD2) U is strongly convex: there exists a constant $m_U > 0$, such that for all $x, y \in \mathbb{R}^d$ it holds that

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

(LD3) There exist $K_1 \geq 0$ and $\tilde{\mathrm{m}}_U > 0$ such that for any $x \in \mathbb{R}^d$ with $\|x\| > K_1$ and any $y \in \mathbb{R}^d$, $\langle \mathrm{D}^2 U(x)y, y\rangle \geq \tilde{\mathrm{m}}_U \|y\|^2$. Moreover, there exists $\mathrm{M}_U \geq 0$ such that for any $x \in \mathbb{R}^d$, $\left\|\mathrm{D}^3 U(x)\right\| \leq \mathrm{M}_U$.

Unadjusted Langevin Algorithm The Langevin stochastic differential equation associated with π is defined by

$$\mathrm{d}Y_t = -\nabla U(Y_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t , \tag{13}$$

where $(B_t)_{t\geq 0}$ is the standard d-dimensional Brownian motion. Under mild technical conditions, the Langevin diffusion admits π as its unique invariant distribution. We consider the sampling method based on the Euler-Maruyama discretization of (13). This scheme referred to as unadjusted Langevin algorithm (ULA), defines the discrete-time Markov chain $(X_k)_{k\geq 0}$ given by

$$X_{k+1} = X_k - \gamma\nabla U(X_k) + \sqrt{2\gamma}Z_{k+1} , \tag{14}$$

where $(Z_k)_{k\geq 1}$ is an i.i.d. sequence of d-dimensional standard Gaussian random variables and $\gamma > 0$ is a step size; see Roberts and Tweedie [34]. We denote by P_γ^{ULA} the Markov kernel associated to the chain (14). It is known that under (LD1) and (LD2) or (LD3), P_γ^{ULA} has a stationary distribution π_γ which is close to π (in a sense that one can bound the distance between π_γ and π, e.g., in total variation and Wasserstein distances, see Dalalyan [8], Durmus and Moulines [14]).

Proposition 5 *1. Assume (LD1), (LD2). Then for any $0 < \gamma < 2/(\mathrm{m}_U + \mathrm{L}_U)$, P_γ^{ULA} satisfies (GE) with the invariant distribution π_γ and $W(x) = \|x\|^2$. Moreover, P_γ^{ULA} fulfils (CW) with*

$$\alpha = 2\gamma \quad \text{and} \quad r = \sqrt{1 - \gamma\, \mathrm{k}_U},$$

where $\mathrm{k}_U \stackrel{def}{=} 2\, \mathrm{m}_U\, \mathrm{L}_U\, /(\mathrm{m}_U + \mathrm{L}_U)$.

2. Assume (LD1), (LD3). Then for any $0 < \gamma < \tilde{\mathrm{m}}_U/(4\, \mathrm{L}_U^2)$, P_γ^{ULA} satisfies (GE), (BR) with the invariant distribution π_γ, $W(x) = \|x\|^2$, and $\mathsf{S} = \left\{x \in \mathbb{R}^d : \|x\| \leq R\right\}$ with sufficiently large radius $R > 0$.

Proof: 1. For the proof of (GE) see Durmus and Moulines [13, Proposition 2] and Durmus and Moulines [14, Theorem 12] and remark 2. To prove (CW) we observe that $P_\gamma^{\mathrm{ULA}}(x, \cdot) = \mathcal{N}(x - \gamma \nabla U(x), 2\gamma\, \mathrm{I}_d)$. Hence, for all $\gamma > 0$, we get using Bakry et al. [3, Theorem 9.2.1], $P_\gamma^{\mathrm{ULA}}(x, \cdot) \in \mathrm{T}_2(2\gamma)$, that is $P_\gamma^{\mathrm{ULA}}(x, \cdot)$ fulfils (2). Assuming that (LD1) and (LD2) hold, we may show using Durmus and Moulines [13, Proposition 3] that for any $0 < \gamma \leq 2/(\mathrm{m}_U + \mathrm{L}_U))$ and any $x, y \in \mathsf{X}$, $W_2(P_\gamma^{\mathrm{ULA}}(x, \cdot), P_\gamma^{\mathrm{ULA}}(y, \cdot)) \leq \sqrt{1 - \gamma\, \mathrm{k}_U}\, d(x, y)$.

2. See Brosse et al. [7, Lemma 19 and Proposition 16].

□

Metropolis Adjusted Langevin Algorithm (MALA) Here we consider a popular modification of ULA called Metropolis Adjusted Langevin Algorithm (MALA). At each iteration, a new candidate Y_{k+1} is proposed according to

$$Y_{k+1} = X_k - \gamma \nabla U(X_k) + \sqrt{2\gamma} Z_{k+1}\,, \tag{15}$$

where $(Z_k)_{k\geq 1}$ is an i.i.d. sequence of d-dimensional standard Gaussian random vectors and $\gamma > 0$ is a step size. This proposal is accepted with probability $\alpha(X_k, Y_{k+1})$, where

$$\alpha(x, y) \stackrel{\mathrm{def}}{=} \min\left(1, \frac{\pi(y) q_\gamma(y, x)}{\pi(x) q_\gamma(x, y)}\right),$$

where $q_\gamma(x, y) = (4\pi\gamma)^{-d/2} \exp(-\|y - x + \gamma \nabla U(x)\|^2/(4\gamma))$. We denote by P_γ^{MALA} the Markov kernel associated to the MALA chain.

Proposition 6 *Assume (LD1), (LD3). Then there exists $\overline{\gamma} > 0$ such that for any $\gamma \in [0, \overline{\gamma}]$, P_γ^{MALA} satisfies (GE), (BR) with the invariant distribution π, $W(x) = \|x\|^2$, and $\mathsf{S} = \left\{x \in \mathbb{R}^d : \|x\| \leq R\right\}$ with sufficiently large radius $R > 0$.*

Proof: See Brosse et al. [7, Proposition 21 and 23]. □

Random Walk Metropolis (RWM) At each iteration, a new candidate Y_{k+1} is proposed according to

$$Y_{k+1} = X_k + \sqrt{\gamma} Z_{k+1}, \tag{16}$$

where $(Z_k)_{k\geq 1}$ is an i.i.d. sequence of d-dimensional standard Gaussian random vectors and $\gamma > 0$. This proposal is accepted with probability $\alpha(X_k, Y_{k+1})$, where

$$\alpha(x, y) = \min\big(\pi(y)/\pi(x), 1\big)$$

We denote by P_γ^{RWM} the Markov kernel associated to the RWM chain. Assumption (GE) is discussed in Roberts and Tweedie [35] and Jarner and Hansen [22] under various conditions. In particular the following result for super-exponential densities holds.

Proposition 7 *Assume (LD1), (LD3). Then P_γ^{RWM} satisfies (GE), (BR) with the invariant distribution π, $W(x) = c\pi^{-1/2}(x)$ for some $c > 0$, and $\mathsf{S} = \big\{x \in \mathbb{R}^d : \|x\| \leq R\big\}$ with sufficiently large radius $R > 0$.*

Proof: See Jarner and Hansen [22, Theorem 4.2]. □

4 Numerical study

In this section we study numerical performance of the ESVM method for simulated and real-world data. Python implementation is available at https://github.com/svsamsonov/esvm.

Following Assaraf and Caffarel [2], Mira et al. [26], Oates et al. [30], we choose $\mathcal{G}$ to be a class of Stein control variates of the form

$$g_\Phi = -\langle \Phi, \nabla U\rangle + \mathrm{div}(\Phi), \tag{17}$$

where $\Phi : \Theta \to \mathbb{R}^d$ with $\Theta \subset \mathbb{R}^d$, $\mathrm{div}(\Phi)$ is the divergence of Φ, and U is the potential associated with π, that is, $\pi(x) \propto \mathrm{e}^{-U(x)}$, see Section 3. Under (LD1) and (LD2), for continuously differentiable functions Φ, $\pi(g_\Phi) = 0$, see Oates et al. [30, Lemma 1]. This suggests to consider a class $\mathcal{H} = \{h = f - g_\Phi : g_\Phi \in \mathcal{G}\}$. Our standard choice will be $\Phi(x) = b$ or $\Phi(x) = Ax + b$, where $A \in \mathbb{R}^{d\times d}$ is a matrix and $b \in \mathbb{R}^d$ is a vector. They will be referred to as the first- and second-order control variates respectively. It is worth noting that polynomial-based control variates are not exhaustive and one can use other control variates. For instance, in the Gaussian mixture model considered below, polynomial-based control variates do not fit structure of the problem, so a class of radial basis functions will be used.

In the ESVM method, we choose the trapezoidal non-negative kernel w supported on $[-1, 1]$:

$$w(s) = \begin{cases} 2s + 2, & -1 \leq s < -1/2, \\ 1, & -1/2 \leq s \leq 1/2, \\ -2s + 2, & 1/2 < s \leq 1. \end{cases} \tag{18}$$

Our experiments with other kernels, for instance, $w(s) = \frac{1}{2} + \frac{1}{2}\cos \pi s$ did not reveal any sensitivity of ESVM to a particular kernel choice. In fact, even the simplest kernel $w(s) = \mathbb{1}_{\{|s|\leq \frac{1}{2}\}}$ showed results comparable with ones for $w(s)$ given in (18). Another parameter of ESVM to be chosen is the lag-window size b_n. In practice, it is not convenient to choose b_n according to Theorem 3 and Theorem 4, since it involves parameters of the Markov chain which are not usually available. Therefore, we choose b_n by analyzing the sample autocorrelation function (ACF) of the Markov chain, see discussion below. Moreover, our experiments show that ESVM is not much sensitive to particular choice of b_n. For a wide range of possible values our procedure shows reasonably good performance.

Numerical study is organized as follows. First we use ULA, MALA, or RWM algorithm to sample a training trajectory of the size $n = n_{\mathrm{burn}} + n_{\mathrm{train}}$. We consider the first n_{burn} observations as a burn-in period, and exclude them from subsequent computations. Then we compute optimal parameters $\hat{A}_{\mathrm{ESVM}}$, $\hat{b}_{\mathrm{ESVM}}$ which minimise the spectral variance $V_n(h)$ with $n = n_{\mathrm{train}}$ and obtain the resulting control variate $\hat{h}_{\mathrm{ESVM}}$. For comparison purposes, we also

compute parameters $\hat{A}_{\text{EVM}}$, $\hat{b}_{\text{EVM}}$ based on minimisation of the empirical variance $V_n'(h) = (n-1)^{-1}\sum_{k=0}^{n-1}\{h(X_k) - \pi_n(h)\}^2$ with $n = n_{\text{train}}$ and obtain the corresponding control variate $\hat{h}_{\text{EVM}}$. Variance reduction using $\hat{h}_{\text{EVM}}$ will be referred to as the EVM algorithm, see Belomestny et al. [4], Mira et al. [26], and Papamarkou et al. [31]. We use the BFGS optimisation method to find the optimal parameters for both ESVM and EVM algorithms.

To evaluate performance of ESVM and EVM, we use the same MCMC algorithm to sample $N_{\text{test}} = 100$ independent training trajectories of size $n = n_{\text{burn}} + n_{\text{test}}$. Then for each trajectory we exclude first n_{burn} observations and compute three different estimates for $\pi(f)$: (i) vanilla estimate (ergodic average of f without variance reduction); (ii) EVM estimate (ergodic average of $\hat{h}_{\text{EVM}}$); (iii) ESVM estimate (ergodic average of $\hat{h}_{\text{ESVM}}$). For each test trajectory, we define the Variance Reduction Factors (VRF) as the ratios $V_n(f)/V_n(\hat{h}_{\text{ESVM}})$ or $V_n(f)/V_n(\hat{h}_{\text{EVM}})$ with $n = n_{\text{test}}$. We report the average VRF over N_{test} trajectories together with the corresponding boxplots of ergodic averages. On these boxplots we display the lower and upper quartiles for each estimation procedure. We will refer to the methods based on the first-order control variates as ESVM-1 and EVM-1, and for the second-order ones as ESVM-2 and EVM-2, respectively. The values b_n, n_{burn}, n_{train}, n_{test} together with parameters of MCMC algorithms for each example considered below are presented in Section 6, Table 6.

Gaussian Mixture Model (GMM) Let π be a mixture of two Gaussian distributions, that is, $\pi = \rho\mathcal{N}(\mu, \Sigma) + (1-\rho)\mathcal{N}(-\mu, \Sigma)$ for $\rho \in [0,1]$. It is straightforward to check that (LD1) holds. If μ and Σ are such that $\|\Sigma^{-1}\mu\|^2 \leq \lambda_{\min}(\Sigma^{-1})$, the density π satisfies (LD2). Otherwise, we have (LD3).

We set $\rho = 1/2$, $d = 2$, $\mu = (0.5, 0.5)^\top$, and consider two instances of the covariance matrix: $\Sigma = \mathrm{I}$ and $\Sigma = \Sigma_0$, where Σ_0 is a randomly initialised symmetric matrix with $\lambda_{\min}(\Sigma_0) \geq 0.1$. The quantities of interest are $\mathsf{E}_\pi[X_1]$ and $\mathsf{E}_\pi[X_1^2]$.

First let us briefly discuss how one can choose the lag-window size b_n. Let us look at the sample ACF plot of the first coordinate given in Figure 1. One may observe that ACF decreases fast enough for any MCMC algorithm, and it seems reasonable to set $b_n = 50$ or close to it. Moreover, we analyse performance of ESVM for different choices of b_n by running the ULA algorithm to estimate $\mathsf{E}_\pi[X_1]$ and letting b_n to run over the values from 1 to 5000. The corresponding VRFs are given also in Figure 1. Here, to compute the spectral variance over test trajectories, we use fixed $b_n^{\text{test}} = n_{\text{test}}^{1/3}$, no matter which value of b_n was used during the training. Note that even for $b_n = 1$ on train (that is, taking into account only the first-order autocovariance) ESVM outperforms EVM, and for values $b_n \in [10, 1000]$ we observe the optimal performance of ESVM.

Numerical results for estimating $\mathsf{E}_\pi[X_1]$ are presented in Table 1. The corresponding boxplots for $\mathsf{E}_\pi[X_1]$ are given in Figure 2, and for $\mathsf{E}_\pi[X_1^2]$ are given in Section 6, Figure 6 and Figure 7. For the sake of convenience, all the estimates are centred by their analytically computed expectations. Note that ESVM outperforms EVM in both cases $\Sigma = \mathrm{I}$ and $\Sigma = \Sigma_0$ and for all samplers used.

Gaussian Mixture with isolated modes Let us now consider the Gaussian mixture model with different means and covariates, $\pi = \rho\mathcal{N}(\mu_1, \sigma_1) + (1-\rho)\mathcal{N}(-\mu_2, \sigma_2)$ with $\rho \in [0,1]$. For simplicity, we let $d = 1$. We are interested in the case when $|\mu_1 - \mu_2| \gg \max\{\sigma_1, \sigma_2\}$. When sampling from π using ULA, MALA, or RWM, the corresponding Markov chain tends to "stuck" at the modes of density π, which leads to slow convergence. We are going to compare the results obtained using ESVM and EVM with the ones from Mijatović and Vogrinc [25] based on a discretized Poisson equation. For comparison purposes, we will reproduce experiments from the aforementioned paper, see Section 5.2.1, and refer to the reported variance reduction factors.

Figure 1: GMM with $\Sigma = \Sigma_0$. Left: Sample autocorrelation function for X_1. Right: average variance reduction factors for different choices of b_n.

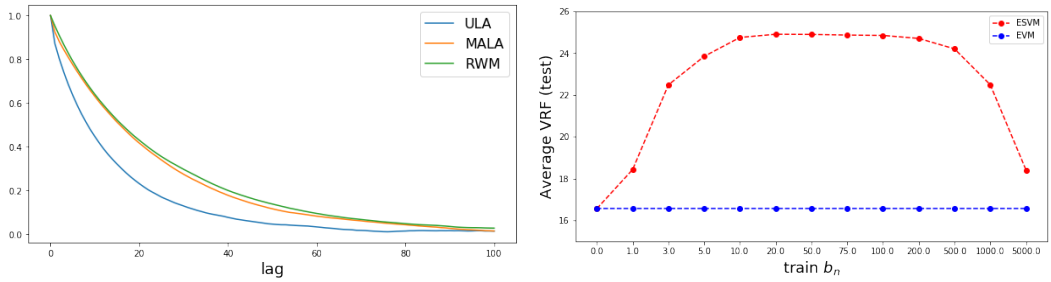
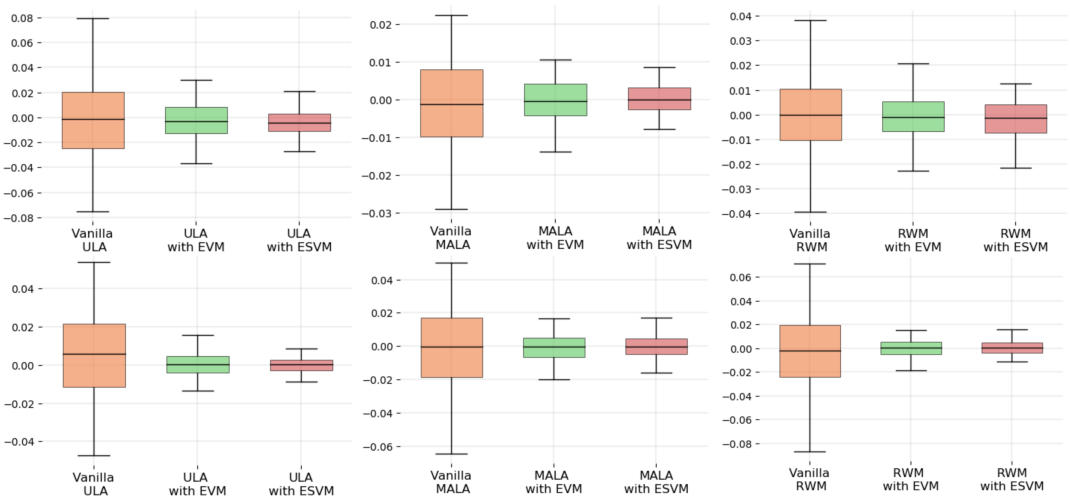


Table 1: Variance Reduction Factors in GMM with $\Sigma = \mathrm{I}$ (top) and $\Sigma = \Sigma_0$ (bottom).

	$\mathsf{E}_\pi[X_1]$			$\mathsf{E}_\pi[X_1^2]$		
Method	ULA	MALA	RWM	ULA	MALA	RWM
ESVM	**9.1**	**6.1**	**8.2**	**609.2**	**319.6**	**531.2**
EVM	4.5	3.6	5.3	607.8	316.3	528.7

	$\mathsf{E}_\pi[X_1]$			$\mathsf{E}_\pi[X_1^2]$		
Method	ULA	MALA	RWM	ULA	MALA	RWM
ESVM	**24.6**	**7.9**	**22.2**	**15.2**	**9.4**	**15.3**
EVM	16.5	7.5	14.3	9.2	5.0	9.3

Figure 2: Estimation of $\mathsf{E}_\pi[X_1]$ in GMM with $\Sigma = \mathrm{I}$ (top row) and $\Sigma = \Sigma_0$ (bottom row). In each row boxplots are given for ULA, MALA, and RWM, respectively.



Our aim is to estimate $\pi(f)$ with $f(x) = x^3$. We fix $\rho = 0.4$, $\mu_1 = -3$, $\mu_2 = 4$, $\sigma_1 = 1$, $\sigma_2 = 0.5$, and use RWM with step size $\gamma = 1.0$ as a generating procedure. Results for the second-order control variates (our standard choice) are reported in Table 2, showing that this class of functions Φ does not allow us to achieve comparable to Mijatović and Vogrinc [25]

variance reduction factors. Let us consider instead the following set of radial basis functions

$$\Phi(x) = \sum_{k=1}^{r} a_k (x - b_k) \exp\left(-\frac{(x-b_k)^2}{2}\right), \tag{19}$$

where a_k, $b_k \in \mathbb{R}$, $k = 1, \ldots, r$. The ESVM algorithm with control variates determined by $\Phi(x)$ from (19) will be referred to as the ESVM-r algorithm. Results for ESVM-r are also given in Table 2 showing comparable results with the Poisson-based approach from Mijatović and Vogrinc [25] (it is referred to as the Poisson-CV) and even outperforming it for large enough train sample size n_{train} and number of basis functions r.

Table 2: Variance Reduction Factors in GMM with isolated modes.

n_{train}	EVM-2	ESVM-2	Poisson-CV	ESVM-r, $r = 4$	ESVM-r, $r = 10$	ESVM-r, $r = 20$
10^4	1.03	1.04	up to 8900	95.8	6457.2	265382.8
10^5	1.92	1.20	up to 13200	98.8	7176.5	378249.0

Banana-shape density The "Banana-shape" distribution, proposed by Haario et al. [19], can be obtained from a d-dimensional Gaussian vector with zero mean and covariance $\text{diag}(p, 1, \ldots, 1)$ by applying transformation

$$\varphi_b(x) : \mathbb{R}^d \to \mathbb{R}^d, \quad \varphi(x) = (x_1,\, x_2 + bx_1^2 - pb,\, x_3, \ldots, x_d),$$

where $p > 0$ and $b > 0$ are parameters; here b controls the curvature of density's level sets. The potential U is given by

$$U(x_1, \ldots, x_d) = x_1^2/2p + (x_2 + bx_1^2 - pb)^2 + \sum_{k=3}^{d} x_k^2/2.$$

As can be easily seen, the assumption (H3) holds. As to the assumption (H1), it is fulfilled only locally. The quantity of interest is $\mathsf{E}_\pi[X_2]$. In our simulations, we set $p = 100$, $b = 0.1$ and consider $d = 2$ and $d = 8$. VRFs are reported in Table 3. Boxplots for $d = 8$ are shown in Figure 3. In this problem, ESVM significantly outperforms EVM both for $d = 2$ and $d = 8$. Because of the curvature of the level sets, the step sizes in all considered methods should be chosen small enough, leading to highly correlated samples. This explains a poor performance of the EVM method in this context.

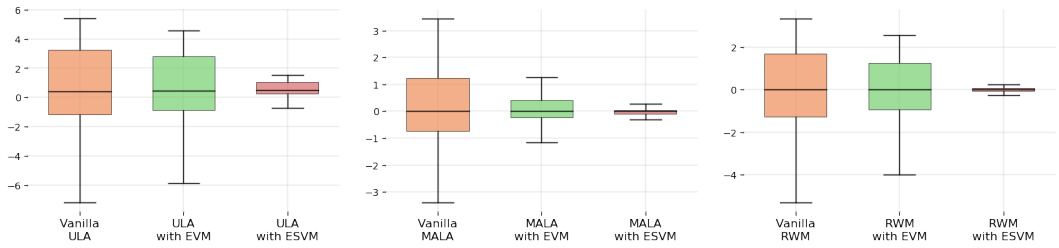
Table 3: Estimation of $\mathsf{E}_\pi[X_2]$ for the banana-shaped density in $d = 2$ and $d = 8$.

	$d = 2$			$d = 8$		
Method	ULA	MALA	RWM	ULA	MALA	RWM
ESVM	**4.7**	**2.7**	**42.4**	**5.3**	**6.5**	**18.5**
EVM	1.4	1.3	1.5	1.4	4.6	1.7

Logistic and probit regression Let $\mathsf{Y} = (\mathsf{Y}_1, \ldots, \mathsf{Y}_n) \in \{0, 1\}^n$ be a vector of binary response variables, $x \in \mathbb{R}^d$ be a vector of regression coefficients, and $\mathsf{Z} \in \mathbb{R}^{N \times d}$ be a design matrix. The log-likelihood and likelihood of i-th point for the logistic and probit regression are given by

$$\ell_{\log}(\mathsf{Y}_i | x, \mathsf{Z}_i) = \mathsf{Y}_i \mathsf{Z}_i^\top x - \ln(1 + \mathrm{e}^{\mathsf{Z}_i^\top x}), \quad p_{\log}(\mathsf{Y}_i | x, \mathsf{Z}_i) = \exp(\ell_{\log}(\mathsf{Y}_i | x, \mathsf{Z}_i)),$$
$$\ell_{\text{pro}}(\mathsf{Y}_i | x, \mathsf{Z}_i) = \mathsf{Y}_i \ln(\Phi(\mathsf{Z}_i^\top x)) + (1 - \mathsf{Y}_i) \ln(\Phi(-\mathsf{Z}_i^\top x)), \quad p_{\text{pro}}(\mathsf{Y}_i | x, \mathsf{Z}_i) = \exp(\ell_{\text{pro}}(\mathsf{Y}_i | x, \mathsf{Z}_i)),$$

Figure 3: Estimation of $\mathsf{E}_\pi[X_2]$ for the banana-shape density in $d = 8$. Boxplots are given for ULA, MALA, and RWM respectively.



where $\mathsf{Z}_i^\top$ is the i-th row of Z for $i \in \{1, \dots, N\}$. We complete the Bayesian model by considering the Zellner g-prior for the regression parameter x, that is, $\mathcal{N}_d(0, g(\mathsf{Z}^\top\mathsf{Z})^{-1})$. Defining $\tilde{x} = (\mathsf{Z}^\top\mathsf{Z})^{1/2}x$ and $\tilde{\mathsf{Z}}_i = (\mathsf{Z}^\top\mathsf{Z})^{-1/2}\mathsf{Z}_i$, the scalar product is preserved, that is $\langle x, \mathsf{Z}_i\rangle = \langle \tilde{x}, \tilde{\mathsf{Z}}_i\rangle$ and, under the Zellner g-prior, $\tilde{x} \sim \mathcal{N}_d(0, gI_d)$. In the sequel, we apply the algorithms in the transformed parameter space with normalized covariates and put $g = 100$.

The unnormalized posterior probability distributions $\pi_{\log}$ and π_{pro} for the logistic and probit regression models are defined for all $\tilde{x} \in \mathbb{R}^d$ by

$$\pi_{\log}(\tilde{x}|\mathsf{Y}, \mathsf{Z}) \propto \exp(-U_{\log}(\tilde{x})) \quad \text{with} \quad U_{\log}(\tilde{x}) = -\sum_{i=1}^{N} \ell_{\log}(\mathsf{Y}_i|\tilde{x}, \mathsf{Z}_i) + (2\sigma^2)^{-1} \|\tilde{x}\|^2,$$
$$\pi_{\text{pro}}(\tilde{x}|\mathsf{Y}, \mathsf{Z}) \propto \exp(-U_{\text{pro}}(\tilde{x})) \quad \text{with} \quad U_{\text{pro}}(\tilde{x}) = -\sum_{i=1}^{N} \ell_{\text{pro}}(\mathsf{Y}_i|\tilde{x}, \mathsf{Z}_i) + (2\sigma^2)^{-1} \|\tilde{x}\|^2.$$

It is straightforward to check that $U_{\log}, U_{\text{pro}}$ satisfy (LD1) and (LD2).

We analyze the performance of ESVM algorithm on two datasets from the UCI repository. The first dataset, Pima[1], contains $N = 768$ observations in dimension $d = 9$. The second one, EEG[2], has dimension $d = 15$, and for our experiments we take randomly selected subset of size 5000 (to speed up sampling procedure). We split each dataset into a training part $\mathcal{T}_N^{\text{train}} = [(y_i, \mathsf{Z}_i)]_{i=1}^N$ and a test part $\mathcal{T}_K^{\text{test}} = [(y_i', \mathsf{Z}_i')]_{i=1}^K$ by randomly picking K test points from the data. Then we use ULA, MALA, and RWM algorithms to sample from $\pi_{\log}(\tilde{x}|\mathsf{Y}, \mathsf{Z})$ and $\pi_{\text{pro}}(\tilde{x}|\mathsf{Y}, \mathsf{Z})$ respectively.

Given the sample $(\tilde{x}_k)_{k=0}^{n-1}$, we aim at estimating the average likelihood over the test set $\mathcal{T}_K^{\text{test}}$, that is,

$$\int_{\mathbb{R}^d} f(\tilde{x})\pi_{\log}(\tilde{x}|\mathsf{Y}, \mathsf{Z})\,\mathrm{d}\tilde{x} \quad \left(\text{or } \int_{\mathbb{R}^d} f(\tilde{x})\pi_{\text{pro}}(\tilde{x}|\mathsf{Y}, \mathsf{Z})\,\mathrm{d}\tilde{x} \text{ for probit regression}\right),$$

where the function f is given by

$$f(\tilde{x}) = K^{-1}\sum_{i=1}^{K} p_{\log}(y_i'|\mathsf{Z}_i', \tilde{x}) \quad \left(\text{or } K^{-1}\sum_{i=1}^{K} p_{\text{pro}}(y_i'|\mathsf{Z}_i', \tilde{x}) \text{ for probit regression}\right).$$

VRFs are reported for first- and second-order control variates. Results for logistic regression are given in Table 4. Boxplots for the average test likelihood estimation using second-order control variates are shown in Figure 4. The same quantities for probit regression are reported in Section 6, see Table 7, Figure 8, and Figure 9.

Note that ESVM also outperforms EVM in this example. It is worth noting that for ULA and RWM, we show up to 100 times better performance in terms of VRF. For MALA, the results for EVM and ESVM are similar since the samples are much less positively correlated.

[1] https://www.kaggle.com/uciml/pima-indians-diabetes-database
[2] https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State

Figure 4: Estimation of the average test likelihood in logistic regression for the Pima dataset (top row) and the EEG dataset (bottom row). In each row boxplots are given for ULA, MALA, and RWM respectively.

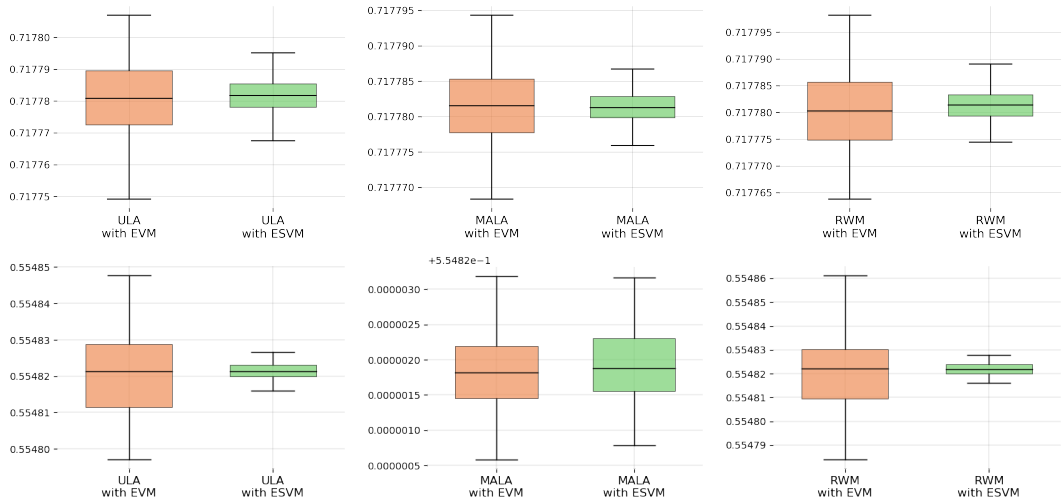


Table 4: Average test likelihood estimation in logistic regression.

	PIMA dataset			EEG dataset		
Method	ULA	MALA	RWM	ULA	MALA	RWM
ESVM-1	347.6	535.6	411.7	542.3	996.6	483.5
EVM-1	347.9	542.1	415.5	548.1	1020.2	508.9
ESVM-2	**11387.3**	**28792.8**	**19503.3**	**11406.6**	**44612.5**	**11324.9**
EVM-2	2704.8	4087.3	5044.1	350.3	39985.4	453.3

Van der Pol oscillator equation The setup of this experiment is much similar to the one reported in South et al. [37]. Here a position $p_x(t) \in \mathbb{R}$ evolves in time t according to the second order differential equation

$$\frac{\mathrm{d}^2 p_x}{\mathrm{d}t^2} - x(1 - p_x^2)\frac{\mathrm{d}p_x}{\mathrm{d}t} + p = 0, \tag{20}$$

where $x \in \mathbb{R}$ is an unknown parameter indicating the non-linearity and the strength of the damping. Letting $q_x = \mathrm{d}p_x/\mathrm{d}t$ we can formulate the oscillator as the first-order system

$$\begin{cases} \frac{\mathrm{d}p_x}{\mathrm{d}t} = q_x, \\ \frac{\mathrm{d}q_x}{\mathrm{d}t} = x(1 - p_x^2)q_x + p_x, \end{cases}$$

where only the first component p_x is observed. This system was solved numerically using $x_\star = 1$ and starting point $p_{x_\star}(0) = 0$, $q_{x_\star}(0) = 2$. Observations $\mathsf{Y}_i = p_{x_\star}(t_i) + \varepsilon_i$ were made at successive time instants $t_i = i$, $i = 1, \ldots, T$, and Gaussian measurement noise ε_i of standard deviation $\sigma = 0.5$ was added. We use a normal prior $\pi_0(x)$ with mean $\mu = 1$ and standard deviation $\sigma_0 = 0.5$. The unnormalized posterior probability distribution is defined for all $x > 0$ by

$$\pi(x|\mathsf{Y}) \propto \exp(-U(x)) \quad \text{with} \quad U(x) = -\log \pi_0(x) + \sum_{i=1}^{T} \frac{(\mathsf{Y}_i - p_x(t_i))^2}{2\sigma^2}.$$

Clearly, U satisfies (LD1) and (LD3). To sample from $\pi(x|\mathsf{Y})$ we use the MALA algorithm. The quantity of interest is the posterior mean $\int_{\mathbb{R}} x\pi(x|\mathsf{Y})\,\mathrm{d}x$. In this example, we use control variates up to degree 3. Results are presented in Section 6 — VRFs are summarized in Table 8 and boxplots for the second-order control variates are given in Figure 10. In this problem, ESVM slightly outperforms EVM in terms of variance reduction factor.

Lotka-Volterra system The Lotka-Volterra model is a well-known system of ODEs describing the joint evolution of two interacting biological populations, predators and preys. Denote the population of preys and predators at moment t by $u(t)$ and $v(t)$ respectively, then the corresponding model can be written as the following first-order system

$$\begin{cases} \frac{\mathrm{d}u}{\mathrm{d}t} = (\alpha - \beta v)u, \\ \frac{\mathrm{d}v}{\mathrm{d}t} = (-\gamma + \delta u)v, \\ u(0) = u_0,\ v(0) = v_0. \end{cases} \tag{21}$$

The parameter vector is given by $x = (\alpha, \beta, \gamma, \delta)$, with all components being non-negative due to the physical meaning of the problem. The system was solved numerically with the true parameters $x_\star = (0.6, 0.025, 0.8, 0.025)$ and starting populations $u_0 = 30.0$, $v_0 = 4.0$. The system is observed at successive time moments $t_i = i$, $i = 1, \ldots, T$, with the lognormal measurements $\mathsf{Y}_i \sim \mathrm{Lognormal}(\log u(t_i), \sigma^2)$, $\mathsf{Z}_i \sim \mathrm{Lognormal}(\log v(t_i), \sigma^2)$ with $\sigma = 0.25$. A weakly informative normal prior $\pi_0(x)$ was used for the model parameters: $\mathcal{N}(1, 0.5)$ for α and γ, $\mathcal{N}(0.05, 0.05)$ for β and δ. The posterior distribution is given by $\pi(x|\mathsf{Y}, \mathsf{Z}) \propto \exp(-U(x))$, where

$$U(x) = -\log \pi_0(x) + \sum_{i=1}^{T} \left(\frac{(\log \mathsf{Y}_i - \log u(t_i))^2 + (\log \mathsf{Z}_i - \log v(t_i))^2}{2\sigma^2} + \log \mathsf{Y}_i + \log \mathsf{Z}_i \right).$$

We use the MALA algorithm to sample from $\pi(x|\mathsf{Y}, \mathsf{Z})$. The quantity of interest is the posterior mean $\int_{\mathbb{R}^4} x\pi(x|\mathsf{Y}, \mathsf{Z})\,\mathrm{d}x$. VRFs are summarized in Table 5 and boxplots for the second-order control variates are given in Figure 5 and Section 6, Figure 11. For some model parameters ESVM significantly outperforms EVM in terms of VRF, for others the results are comparable with slight superiority of ESVM.

Table 5: Estimation of the posterior mean in the Lotka-Volterra model.

Estimated parameter	α	β	δ	γ
ESVM-1	**10.5**	**6.5**	**6.2**	**8.3**
EVM-1	6.6	4.2	4.9	6.0
ESVM-2	**757.6**	**427.8**	**277.2**	**446.6**
EVM-2	642.1	286.0	275.0	429.7

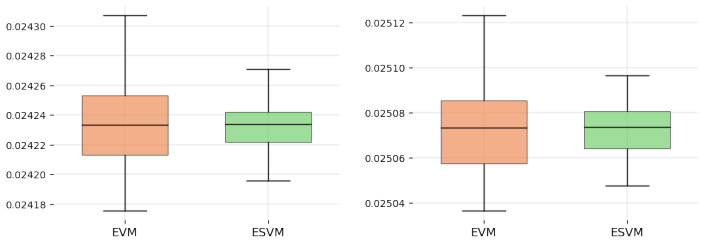
5 Proofs

5.1 Proof of Proposition 1

Before we proceed to the proof of Proposition 1, let us refer to a general result from Nickl and Pötscher [27] which is used below to bound the fixed point of a subset of a weighted Sobolev space. First we need to introduce some notations.

Let μ be a (nonnegative) Borel measure. Given the two functions $l, u : \mathsf{X} \to \mathbb{R}$ in $\mathrm{L}^p(\mu)$, the bracket $[l, u]$ is the set of all functions in $\mathrm{L}^p(\mu)$ with $l \le f \le u$. The $\mathrm{L}^p(\mu)$-size of the

Figure 5: Estimation of the posterior mean of β (left figure) and δ (right figure) in the Lotka-Volterra model.



bracket $[l, u]$ is defined as $\|l - u\|_{\mathrm{L}^p(\mu)}$. The $\mathrm{L}^p(\mu)$-bracketing number $\mathcal{N}^{[]}_{\mathrm{L}^p(\mu)}(F, \varepsilon)$ of a (non-empty) set F is the minimal number of brackets of $\mathrm{L}^p(\mu)$-size less than or equal to $\varepsilon > 0$ necessary to cover F. The logarithm of the bracketing number is called the $\mathrm{L}^p(\mu)$-bracketing metric entropy $H^{[]}_{\mathrm{L}^p(\mu)}(F, \varepsilon)$.

Theorem 8 ([27, Corollary 4]) *Let $1 < p < \infty$, $\beta \in \mathbb{R}$, and $s - d/p > 0$. Let F be a (non-empty) norm-bounded subset of $W^{s,p}(\mathbb{R}^d, \langle x\rangle^\beta)$. Suppose M is a (non-empty) family of Borel measures on $\mathbb{R}^d$ such that the condition $\sup_{\mu\in M} \|\langle x\rangle^{\alpha-\beta}\|_{L^r(\mu)} < \infty$ holds for some $1 \le r \le \infty$ and for some $\alpha > 0$. Then*

$$\sup_{\mu\in M} H^{[]}_{L^r(\mu)}(F, \varepsilon) \lesssim \begin{cases} \varepsilon^{-d/s} & \text{for } \alpha > s - d/p, \\ \varepsilon^{-(\alpha/d+1/p)^{-1}} & \text{for } \alpha < s - d/p. \end{cases}$$

Proof of Proposition 1. We first bound the metric entropy of $\mathcal{H}$ by the bracketing metric entropy. If $h \in \mathcal{H}$ is in the 2ε-bracket $[l, u]$, $l, u \in \mathcal{H}$, then it is in the ball of radius ε around $(l + u)/2$. So,

$$H_{\mathrm{L}^2(\pi)}(\mathcal{H}, \varepsilon) \le H^{[]}_{\mathrm{L}^2(\pi)}(\mathcal{H}, 2\varepsilon).$$

Now our aim is apply Theorem 8 to $\mathcal{H}$ which is a norm-bounded subset of $W^{s,p}(\mathbb{R}^d, \langle x\rangle^\beta)$ by assumption. For $M = \{\pi\}$ and $r = 2$, the condition $\sup_{\mu\in M} \|\langle x\rangle^{\alpha-\beta}\|_{L^r(\mu)} < \infty$ also holds by assumption. Hence,

$$H_{\mathrm{L}^2(\pi)}(\mathcal{H}, \varepsilon) \lesssim \begin{cases} \varepsilon^{-d/s} & \text{for } \alpha > s - d/p, \\ \varepsilon^{-(\alpha/d+1/p)^{-1}} & \text{for } \alpha < s - d/p. \end{cases}$$

Now we turn to the bound for the fixed point $\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n)$ (see (9)). Consider first the case $\alpha > s - d/p$. The solution to the inequality $\varepsilon^{-d/s} \lesssim n\varepsilon^2$ is $\varepsilon \gtrsim n^{-\frac{1}{2+d/s}}$. Taking $\varepsilon_0 \sim n^{-\frac{1}{2+d/s}}$, where $\sim$ stands for equality up to a constant, yields

$$H_{\mathrm{L}^2(\pi)}(\mathcal{H}, \varepsilon_0) \lesssim n\varepsilon_0^2, \quad \text{for } \alpha > s - d/p.$$

Since $\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n)$ is the infimum over all such $\varepsilon > 0$, it holds $\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n) \lesssim n^{-\frac{1}{2+d/s}}$. Repeated computations for $\alpha < s - d/p$ give us $\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n) \lesssim n^{-\frac{1}{2+(\alpha/d+1/p)^{-1}}}$. Combining these two bounds, we have

$$\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n) \lesssim \begin{cases} n^{-\frac{1}{2+d/s}} & \text{for } \alpha > s - d/p, \\ n^{-\frac{1}{2+(\alpha/d+1/p)^{-1}}} & \text{for } \alpha < s - d/p, \end{cases}$$

which is the desired conclusion. □

5.2 Spectral variance estimator

We investigate properties of the spectral variance $V_n(h)$ defined in (5). Note that $V_n(h)$ can be represented as a quadratic form $Z_n(h)^\top A_n Z_n(h)$, where $Z_n(h) = (h(X_0), \dots, h(X_{n-1}))^\top$ and A_n is an $n \times n$ symmetric matrix. Namely, let I_n be the identity $n \times n$ matrix and $1_n = (1, \dots, 1)^\top \in \mathbb{R}^n$. Given the lag window w_n, we denote the weight matrix by $W_n = (w_n(j-i))_{i,j=1}^n$. By rearranging the summations in (5), we have

$$V_n(h) = n^{-1} \sum_{k=0}^{n-1} \sum_{j=0}^{n-1} w_n(k-j) \Big(h(X_k) - \pi_n(h) \Big) \Big(h(X_j) - \pi_n(h) \Big),$$

Hence the spectral variance can be represented as

$$V_n(h) = Z_n(h)^\top A_n Z_n(h) \quad \text{for } A_n = \frac{1}{n} \left(\mathrm{I}_n - \frac{1}{n} 1_n 1_n^\top \right)^\top W_n \left(\mathrm{I}_n - \frac{1}{n} 1_n 1_n^\top \right). \tag{22}$$

In the following lemma we provide an upper bound on the operator norm of A_n.

Lemma 9 *If the truncation point b_n of the lag window w_n satisfies $b_n \le n$, then $\|A_n\| \le 2b_n/n$.*

Proof: Denote $P = \mathrm{I}_n - n^{-1} 1_n 1_n^\top$. Since P is an orthonormal projector, we get

$$\|A_n\| = \frac{1}{n} \|P W_n P\| \le \frac{1}{n} \|W_n\|.$$

To bound the operator norm of W_n (which is a Toeplitz matrix), we use the standard technique based on the discrete-time Fourier transform of the sequence $w : [-b_n, b_n] \to [0, 1]$, defined, for $\lambda \in [-\pi, \pi)$ by

$$\hat{w}_n(\lambda) = \sum_{k=-b_n}^{b_n} w_n(k) \mathrm{e}^{-\mathrm{i}k\lambda}.$$

Obviously, $|\hat{w}_n(\lambda)| \le 2b_n$. We have $\|W_n\| = \sup_{\|x\|=1} x^\top W_n x$. Moreover, for any unit vector $u = (u_1, \dots, u_n)^\top$ it holds

$$u^\top W_n u = \sum_{k,j=1}^n \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}(k-j)\lambda} \hat{w}_n(\lambda) \mathrm{d}\lambda \right) u_k u_j = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \sum_{k=1}^n \mathrm{e}^{\mathrm{i}k\lambda} u_k \right|^2 \hat{w}_n(\lambda) \mathrm{d}\lambda \le 2b_n.$$

Hence $\|W_n\| \le 2b_n$ and $\|A_n\| \le 2b_n/n$. The lemma is proved. □

In the next lemma we prove several technical results on expectation of the operator norm of $Z_n(h)$ and $V_n(h)$ which hold under (GE) assumption.

Lemma 10 *Under (GE), it holds for any $h, h' \in \mathcal{H}$*

$$\mathsf{E}_{x_0} \Big[\|Z_n(h)\|^2 \Big] \le n \|h\|^2_{\mathrm{L}^2(\pi)} + \frac{\varsigma W(x_0)}{1-\rho} \|h\|^2_{W^{1/2}},$$

and

$$\mathsf{E}_{x_0} \Big[\|Z_n(h) - Z_n(h')\|^2 \Big] \le n \|h - h'\|^2_{\mathrm{L}^2(\pi)} + \frac{\varsigma W(x_0)}{1-\rho} \|h - h'\|^2_{W^{1/2}}.$$

Moreover, for any $h \in \mathcal{H}$, this bound implies

$$\mathsf{E}_{x_0} \big[V_n(h) \big] \le 2b_n \|h\|^2_{\mathrm{L}^2(\pi)} + \frac{2\|h\|^2_{W^{1/2}} \varsigma W(x_0)}{1-\rho} \frac{b_n}{n}.$$

Proof: We first observe that

$$\mathsf{E}_{x_0}\Big[\|Z_n(h)\|^2\Big] = \mathsf{E}_{x_0}\left[\sum\nolimits_{k=0}^{n-1} h^2(X_k)\right] = \sum\nolimits_{k=0}^{n-1} \|h\|^2_{\mathrm{L}^2(P^k(x_0,\cdot))}.$$

Now each summand can be bounded in the following way,

$$\begin{aligned}\|h\|^2_{\mathrm{L}^2(P^k(x_0,\cdot))} &= \|h\|^2_{\mathrm{L}^2(\pi)} + \Big(\|h\|^2_{\mathrm{L}^2(P^k(x_0,\cdot))} - \|h\|^2_{\mathrm{L}^2(\pi)}\Big) \\ &\le \|h\|^2_{\mathrm{L}^2(\pi)} + \int |h(x)|^2 |P^k(x_0,\cdot) - \pi|(\mathrm{d}x) \\ &\le \|h\|^2_{\mathrm{L}^2(\pi)} + \|h\|^2_{W^{1/2}} \|P^k(x_0,\cdot) - \pi\|_W.\end{aligned}$$

This inequality and (GE) together imply

$$\mathsf{E}_{x_0}\Big[\|Z_n(h)\|^2\Big] \le n\|h\|^2_{\mathrm{L}^2(\pi)} + \frac{\|h\|^2_{W^{1/2}} \varsigma W(x_0)}{1-\rho},$$

which proves the first inequality. Repeated computations for $Z_n(h) - Z_n(h')$ yield

$$\mathsf{E}_{x_0}\Big[\|Z_n(h) - Z_n(h')\|^2\Big] \le n\|h - h'\|^2_{\mathrm{L}^2(\pi)} + \frac{\varsigma W(x_0)}{1-\rho}\|h - h'\|^2_{W^{1/2}}.$$

The first statement is proved. To prove the second statement we note that

$$\mathsf{E}_{x_0}\big[V_n(h)\big] = \mathsf{E}_{x_0}\Big[Z_n(h)^\top A_n Z_n(h)\Big] \le \|A_n\| \mathsf{E}_{x_0}\Big[\|Z_n(h)\|^2\Big].$$

By Lemma 9 we have $\|A_n\| \le 2b_n/n$. Substituting this we deduce our claim. □

It is known that the spectral variance $V_n(h)$ is a biased estimate of the asymptotic variance $V_\infty(h)$. In the following proposition we show how close is the expected value of $V_n(h)$ to $V_\infty(h)$.

Proposition 11 *Assume (GE). Then for any $h \in \mathcal{H}$ and any $x_0 \in \mathsf{X}$,*

$$\Big|\mathsf{E}_{x_0}\big[V_n(h)\big] - V_\infty(h)\Big| \le \frac{\varsigma^{1/2}\pi(W)\|\tilde{h}\|^2_{W^{1/2}}}{1-\rho^{1/2}}\left(\frac{9\varsigma W(x_0)}{(1-\rho)\pi(W)}\frac{b_n}{n^2} + \frac{9b_n}{n} + 2\rho^{b_n/2}\right),$$

where $\tilde{h} = h - \pi(h)$. Moreover, if $n \ge \varsigma W(x_0)/((1-\rho)\pi(W))$ then

$$\Big|\mathsf{E}_{x_0}\big[V_n(h)\big] - V_\infty(h)\Big| \le \frac{20\varsigma^{1/2}\pi(W)\|\tilde{h}\|^2_{W^{1/2}}}{1-\rho^{1/2}}\left(\frac{b_n}{n} \vee \rho^{b_n/2}\right),$$

where $a \vee b \stackrel{def}{=} \max\{a,b\}$.

Proof: Recall that the asymptotic variance $V_\infty(h)$ may be written as $V_\infty(h) = \sum_{|s|\ge 0} \rho^{(h)}_\pi(|s|)$ with $\rho^{(h)}_\pi(s) = \mathsf{E}_\pi\big[\tilde{h}(X_0)\tilde{h}(X_s)\big]$ and, by definition, $V_n(h) = \sum_{|s|<b_n} w_n(s)\hat{\rho}^{(h)}_n(|s|)$, where the lag s empirical autocovariance coefficient $\hat{\rho}^{(h)}_n(s)$ is given in (4). We have

$$\begin{aligned}\Big|\mathsf{E}_{x_0}\big[V_n(h)\big] - V_\infty(h)\Big| \le 2\sum_{s=0}^{b_n-1} w_n(s)\Big|\mathsf{E}_{x_0}\big[\hat{\rho}^{(h)}_n(s)\big] - \rho^{(h)}_\pi(s)\Big| \\ + 2\sum_{s=0}^{b_n-1}|1-w_n(s)||\rho^{(h)}_\pi(s)| + 2\sum_{s=b_n}^{\infty}|\rho^{(h)}_\pi(s)|. \qquad (23)\end{aligned}$$

To bound each summand in this decomposition, we need the following lemma.

Lemma 12 *Assume (GE). Then for any $h \in \mathcal{H}$, $x \in \mathcal{X}$, and $s \in \mathbb{Z}_+$,*

$$\left|\mathsf{E}_x\left[\tilde{h}(X_0)\tilde{h}(X_s)\right]\right| \le \varsigma^{1/2}\rho^{s/2}W(x)\|\tilde{h}\|^2_{W^{1/2}}\,, \tag{24}$$

and

$$\left|\rho_\pi^{(\tilde{h})}(s)\right| \le \varsigma^{1/2}\rho^{s/2}\pi(W)\|\tilde{h}\|^2_{W^{1/2}}\,. \tag{25}$$

Proof: The proof is straightforward. Since $\pi(\tilde{h}) = 0$, we have

$$\begin{aligned}\left|\mathsf{E}_x\left[\tilde{h}(X_0)\tilde{h}(X_s)\right]\right| &\le \left|\tilde{h}(x)\right|\left|\int_{\mathsf{X}} \tilde{h}(y)\big(P^s(x,\cdot)-\pi\big)(\mathrm{d}y)\right| \\ &\le \|\tilde{h}\|^2_{W^{1/2}}W^{1/2}(x)\int_{\mathsf{X}} W^{1/2}(y)\left|P^s(x,\cdot)-\pi\right|(\mathrm{d}y)\,.\end{aligned}$$

By Hölder's inequality,

$$\begin{aligned}\int_{\mathsf{X}} W^{1/2}(y)\left|P^s(x,\cdot)-\pi\right|(\mathrm{d}y) &\le \left|P^s(x,\mathsf{X})-\pi(\mathsf{X})\right|^{1/2}\left(\int_{\mathsf{X}} W(y)\left|P^s(x,\cdot)-\pi\right|(\mathrm{d}y)\right)^{1/2} \\ &\le \|P^s(x,\cdot)-\pi\|_W^{1/2}\,.\end{aligned}$$

Combining these bounds and using (GE), we conclude

$$\left|\mathsf{E}_x\left[\tilde{h}(X_0)\tilde{h}(X_s)\right]\right| \le \varsigma^{1/2}\rho^{s/2}W(x)\|\tilde{h}\|^2_{W^{1/2}}\,,$$

and (24) is proved. Integrating this relation with respect to the stationary distribution π, we obtain the second inequality. The lemma is proved. □

Let us first bound the last two summands in the decomposition (23). By definition, $w_n(s) = 1$ for all $s \in [-b_n/2, b_n/2]$. From (25) we have the second summand

$$\sum_{s=0}^{b_n-1}|1-w_n(s)||\rho_\pi^{(h)}(s)| \le \sum_{s=\lceil b_n/2\rceil}^{b_n-1}|\rho_\pi^{(h)}(s)| \le \varsigma^{1/2}\pi(W)\|\tilde{h}\|^2_{W^{1/2}}\frac{\rho^{b_n/2}}{1-\rho^{1/2}}\,. \tag{26}$$

where $\lceil b_n/2\rceil$ is the nearest integer greater than or equal to $b_n/2$. Similar arguments apply to the last summand in (23),

$$\sum_{s=b_n}^{\infty}|\rho_\pi^{(h)}(s)| \le \varsigma^{1/2}\pi(W)\|\tilde{h}\|^2_{W^{1/2}}\frac{\rho^{b_n}}{1-\rho^{1/2}} \le \varsigma^{1/2}\pi(W)\|\tilde{h}\|^2_{W^{1/2}}\frac{\rho^{b_n/2}}{1-\rho^{1/2}}\,. \tag{27}$$

It remains to bound the first summand in (23). We note that lag s empirical autocovariance coefficient satisfies $\hat{\rho}_n^{(h)}(s) = \hat{\rho}_n^{(\tilde{h})}(s)$. Moreover, for any $s < n$, it may be decomposed as $\hat{\rho}_n^{(\tilde{h})}(s) = \sum_{i=1}^3 A_{n,i}(s)$, where

$$A_{n,1}(s) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{k=0}^{n-s-1}\tilde{h}(X_k)\tilde{h}(X_{k+s}), \quad A_{n,2}(s) \stackrel{\text{def}}{=} \frac{\pi_n(\tilde{h})}{n}\left\{\sum_{k=0}^{n-s-1}\tilde{h}(X_k)+\sum_{k=s}^{n-1}\tilde{h}(X_k)\right\},$$

and $A_{n,3}(s) \stackrel{\text{def}}{=} (1-s/n)\pi_n^2(\tilde{h})$. Since $|w_n(s)| \le 1$ by definition, it holds by the triangle inequality

$$\begin{aligned}\sum_{s=0}^{b_n-1} w_n(s)\left|\mathsf{E}_{x_0}\left[\hat{\rho}_n^{(\tilde{h})}(s)\right]-\rho_\pi^{(h)}(s)\right| &\le \sum_{s=0}^{b_n-1}\left|\mathsf{E}_{x_0}\left[A_{n,1}(s)\right]-\rho_\pi^{(h)}(s)\right| \\ &\quad + \sum_{s=0}^{b_n-1}\left|\mathsf{E}_{x_0}\left[A_{n,2}(s)\right]\right| + \sum_{s=0}^{b_n-1}\left|\mathsf{E}_{x_0}\left[A_{n,3}(s)\right]\right|\,.\end{aligned} \tag{28}$$

For any $s \in \{0, \dots, n-1\}$, by the Markov property, (GE), and (24) we obtain

$$\begin{aligned}\left|\mathsf{E}_{x_0}\left[\tilde{h}(X_k)\tilde{h}(X_{k+s})\right] - \rho_\pi^{(h)}(s)\right| &= \left|\int \mathsf{E}_x\left[\tilde{h}(X_0)\tilde{h}(X_s)\right](P^k(x_0,\cdot) - \pi)(\mathrm{d}x)\right| \\ &\le \varsigma^{1/2}\rho^{s/2}\|\tilde{h}\|_{W^{1/2}}^2\|P^k(x_0,\cdot) - \pi\|_W \le \varsigma^{3/2}\rho^{s/2+k}W(x_0)\|\tilde{h}\|_{W^{1/2}}^2 . \qquad (29)\end{aligned}$$

Therefore by (25) and (29),

$$\begin{aligned}\sum_{s=0}^{b_n-1}\left|\mathsf{E}_{x_0}\left[A_{n,1}(s)\right] - \rho_\pi^{(h)}(s)\right| &\le n^{-1}\sum_{s=0}^{b_n-1}\sum_{k=0}^{n-s-1}\left|\mathsf{E}_{x_0}\left[\tilde{h}(X_k)\tilde{h}(X_{k+s})\right] - \rho_\pi^{(h)}(s)\right| + n^{-1}\sum_{s=0}^{b_n-1} s|\rho_\pi^{(h)}(s)| \\ &\le \frac{\varsigma^{3/2}W(x_0)\|\tilde{h}\|_{W^{1/2}}^2}{n}\sum_{s=0}^{b_n-1}\sum_{k=0}^{n-s-1}\rho^{s/2+k} + \frac{\varsigma^{1/2}\pi(W)\|\tilde{h}\|_{W^{1/2}}^2}{n}\sum_{s=0}^{b_n-1} s\rho^{s/2} \\ &\le \frac{\varsigma^{3/2}W(x_0)\|\tilde{h}\|_{W^{1/2}}^2}{n(1-\rho)(1-\rho^{1/2})} + \frac{b_n\varsigma^{1/2}\pi(W)\|\tilde{h}\|_{W^{1/2}}^2}{n(1-\rho^{1/2})} .\end{aligned}$$

Note that (29) also yields

$$\begin{aligned}\mathsf{E}_{x_0}\left[\pi_n^2(\tilde{h})\right] &\le 2n^{-2}\sum_{k=0}^{n-1}\sum_{s=0}^{n-k-1}\left|\mathsf{E}_{x_0}\left[\tilde{h}(X_k)\tilde{h}(X_{k+s})\right]\right| \\ &\le 2n^{-2}\sum_{k=0}^{n-1}\sum_{s=0}^{n-k-1}\varsigma^{3/2}\rho^{s/2+k}W(x_0)\|\tilde{h}\|_{W^{1/2}}^2 + 2n^{-2}\sum_{k=0}^{n-1}\sum_{s=0}^{n-k-1}\left|\rho_\pi^{(h)}(s)\right| \\ &\le \frac{2\varsigma^{3/2}W(x_0)\|\tilde{h}\|_{W^{1/2}}^2}{n^2(1-\rho)(1-\rho^{1/2})} + \frac{2\varsigma^{1/2}\pi(W)\|\tilde{h}\|_{W^{1/2}}^2}{n(1-\rho^{1/2})} . \qquad (30)\end{aligned}$$

We now turn to $A_{n,2}(s)$. By the Cauchy-Schwarz inequality and similar argument to (30),

$$\begin{aligned}\left|\mathsf{E}_{x_0}\left[A_{n,2}(s)\right]\right| &\le \frac{\sqrt{2}}{n}\left\{\mathsf{E}_{x_0}\left[\pi_n^2(\tilde{h})\right]\right\}^{1/2}\left\{\mathsf{E}_{x_0}\left[\left(\sum\nolimits_{k=0}^{n-s-1}\tilde{h}(X_k)\right)^2\right] + \mathsf{E}_{x_0}\left[\left(\sum\nolimits_{k=s}^{n-1}\tilde{h}(X_k)\right)^2\right]\right\}^{1/2} \\ &\le \frac{4\sqrt{2}\varsigma^{3/2}W(x_0)\|\tilde{h}\|_{W^{1/2}}^2}{n^2(1-\rho)(1-\rho^{1/2})} + \frac{4\sqrt{2}\varsigma^{1/2}\pi(W)\|\tilde{h}\|_{W^{1/2}}^2}{n(1-\rho^{1/2})} .\end{aligned}$$

This gives

$$\sum_{s=0}^{b_n-1}\left|\mathsf{E}_{x_0}\left[A_{n,2}(s)\right]\right| \le \frac{4\sqrt{2}b_n\varsigma^{3/2}W(x_0)\|\tilde{h}\|_{W^{1/2}}^2}{n^2(1-\rho)(1-\rho^{1/2})} + \frac{4\sqrt{2}b_n\varsigma^{1/2}\pi(W)\|\tilde{h}\|_{W^{1/2}}^2}{n(1-\rho^{1/2})} .$$

Finally, for $A_{n,3}(s)$ it follows from (30) that

$$\sum_{s=0}^{b_n-1}\left|\mathsf{E}_{x_0}\left[A_{n,3}(s)\right]\right| \le \frac{2b_n\varsigma^{3/2}W(x_0)\|\tilde{h}\|_{W^{1/2}}^2}{n^2(1-\rho)(1-\rho^{1/2})} + \frac{2b_n\varsigma^{1/2}\pi(W)\|\tilde{h}\|_{W^{1/2}}^2}{n(1-\rho^{1/2})} .$$

Substituting these bounds into (28) we obtain

$$\sum_{s=0}^{b_n-1} w_n(s)\left|\mathsf{E}_{x_0}\left[\hat{\rho}_n^{(\tilde{h})}(s)\right] - \rho_\pi^{(h)}(s)\right| \le \frac{9b_n\varsigma^{3/2}W(x_0)\|\tilde{h}\|_{W^{1/2}}^2}{n^2(1-\rho)(1-\rho^{1/2})} + \frac{9b_n\varsigma^{1/2}\pi(W)\|\tilde{h}\|_{W^{1/2}}^2}{n(1-\rho^{1/2})} . \qquad (31)$$

Collecting the estimates (26), (27), (31) and substituting them into (23) we conclude

$$\left|\mathsf{E}_{x_0}\left[V_n(h)\right] - V_\infty(h)\right| \le \frac{\varsigma^{1/2}\pi(W)\|\tilde{h}\|_{W^{1/2}}^2}{1-\rho^{1/2}}\left(\frac{9\varsigma W(x_0)}{(1-\rho)\pi(W)}\frac{b_n}{n^2} + \frac{9b_n}{n} + 2\rho^{b_n/2}\right),$$

which is our claim. If additionally $n \geq \varsigma W(x_0)/((1-\rho)\pi(W))$ then

$$\left|\mathsf{E}_{x_0}\left[V_n(h)\right] - V_\infty(h)\right| \leq \frac{20\varsigma^{1/2}\pi(W)\|\tilde{h}\|^2_{W^{1/2}}}{(1-\rho^{1/2})}\left(\frac{b_n}{n} \vee \rho^{b_n/2}\right),$$

and the proof is complete. □

5.3 Proof of Theorem 3

For simplicity of notation, without loss of generality, we assume that functions $h \in \mathcal{H}$ are zero-mean, since, by definition, $V_n(h) = V_n(h - \pi(h))$ and hence h may be replaced by $\tilde{h} = h - \pi(h)$ which also satisfies assumptions imposed on h. Further, we write $\overline{V}_n(h) = \mathsf{E}_{x_0}\left[V_n(h)\right]$ and set

$$H \stackrel{\text{def}}{=} \sup_{h\in\mathcal{H}} \|h\|_{\mathrm{L}^2(\pi)} \quad \text{and} \quad M \stackrel{\text{def}}{=} \sup_{h\in\mathcal{H}} \|h\|_{W^{1/2}}. \tag{32}$$

Without loss of generality we may assume that $M < \infty$ since otherwise the statement of the theorem is obviously true.

It follows from Proposition 11 that if $n \geq \varsigma W(x_0)/((1-\rho)\pi(W))$ then

$$\sup_{h\in\mathcal{H}}\left|V_\infty(h) - \overline{V}_n(h)\right| \lesssim G\left(\frac{b_n}{n} \vee \rho^{b_n/2}\right), \quad \text{where} \quad G \stackrel{\text{def}}{=} \frac{\varsigma^{1/2}M^2\pi(W)}{1-\rho^{1/2}}.$$

Hence

$$V_\infty(\widehat{h}_\varepsilon) - \inf_{h\in\mathcal{H}} V_\infty(h) \leq \overline{V}_n(\widehat{h}_\varepsilon) - \inf_{h\in\mathcal{H}} \overline{V}_n(h) + 2G\left(\frac{b_n}{n} \vee \rho^{b_n/2}\right). \tag{33}$$

We are reduced to bounding the difference $\overline{V}_n(\widehat{h}_\varepsilon) - \inf_{h\in\mathcal{H}} \overline{V}_n(h)$. Let us denote by h^* a function in $\mathcal{H}$ minimizing $\overline{V}_n(h)$, that is,

$$h^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{h\in\mathcal{H}} \overline{V}_n(h). \tag{34}$$

We assume that such a minimizer exists (a simple modification of the proof is possible if h^* is an approximate solution of (34)). Let also $h^*_\varepsilon \in \mathcal{H}_\varepsilon$ be the closest point to $h^* \in \mathcal{H}$ in $\mathrm{L}^2(\pi)$. By the definition of $\widehat{h}_\varepsilon$, $V_n(\widehat{h}_\varepsilon) - V_n(h^*_\varepsilon) < 0$. We have

$$\begin{aligned}
\overline{V}_n(\widehat{h}_\varepsilon) - \overline{V}_n(h^*) &\leq \overline{V}_n(\widehat{h}_\varepsilon) - \overline{V}_n(h^*) - \left(V_n(\widehat{h}_\varepsilon) - V_n(h^*_\varepsilon)\right) \\
&= \overline{V}_n(\widehat{h}_\varepsilon) - \overline{V}_n(h^*) - \left(V_n(\widehat{h}_\varepsilon) - V_n(h^*)\right) + \left(V_n(h^*_\varepsilon) - V_n(h^*)\right) \\
&\leq \sup_{h\in\mathcal{H}_\varepsilon}\left\{\overline{V}_n(h) - V_n(h)\right\} + \left(V_n(h^*) - \overline{V}_n(h^*)\right) + \left(V_n(h^*_\varepsilon) - V_n(h^*)\right).
\end{aligned} \tag{35}$$

It remains to bound each summand in the right hand side of the decomposition (35). To do this, we need an exponential concentration for $V_n(h)$. Let us remind that we consider two cases, Lipschitz and bounded functions $h \in \mathcal{H}$. Depending on the case we consider, it follows from Theorem 19 (equation (52)) or Theorem 20 that, for a fixed $\tau > 0$, for all $t < \tau$, and all $h \in \mathcal{H}$,

$$\mathrm{P}_{x_0}\left(\left|V_n(h) - \overline{V}_n(h)\right| > t\right) \leq 2\exp\left(-\frac{t^2 n}{cK_\tau^2 b_n^2}\right), \tag{36}$$

where $c > 0$ is an absolute constant and

$$K_\tau^2 \stackrel{\text{def}}{=} \frac{\alpha L^2}{(1-r)^2}\left(H^2 + \frac{\varsigma M^2 W(x_0)}{1-\rho} + \frac{\tau}{b_n}\right) \quad \text{or} \quad K_\tau^2 \stackrel{\text{def}}{=} \beta^2 B^4$$

in the Lipschitz and bounded cases correspondingly. Note that K_τ does not depend on τ in the bounded case. The value of $\tau > 0$ is specified later. For the first summand in the decomposition (35), using the union bound and the concentration inequality (36), we obtain

$$\begin{aligned}\mathrm{P}_{x_0}\left(\sup_{h\in\mathcal{H}_\varepsilon}\left\{\overline{V}_n(h) - V_n(h)\right\} > t\right) &\le |\mathcal{H}_\varepsilon| \sup_{h\in\mathcal{H}_\varepsilon} \mathrm{P}_{x_0}\Big(\overline{V}_n(h) - V_n(h) > t\Big) \\ &\le 2|\mathcal{H}_\varepsilon| \sup_{h\in\mathcal{H}_\varepsilon} \exp\left(-\frac{nt^2}{cK_\tau^2 b_n^2}\right).\end{aligned}$$

For any $\varepsilon \ge \gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n)$ it holds $|\mathcal{H}_\varepsilon| \le \mathrm{e}^{n\varepsilon^2}$. We can select $t = \sqrt{c}K_\tau b_n\big(\varepsilon + n^{-1/2}\log^{1/2}(8/\delta)\big)$ to obtain

$$\mathrm{P}_{x_0}\left(\sup_{h\in\mathcal{H}_\varepsilon}\left\{\overline{V}_n(h) - V_n(h)\right\} > t\right) \le \delta/4\,. \tag{37}$$

In the same manner we can bound the second term in the right hand side of the decomposition (35). For $t = \sqrt{c}K_\tau b_n n^{-1/2}\log^{1/2}(8/\delta)$, it holds

$$\mathrm{P}_{x_0}\left(V_n(h^*) - \overline{V}_n(h^*) > t\right) \le \delta/4\,. \tag{38}$$

It remains to estimate the last summand in (35). This term is small since h^*_ε is ε-close to h^* in $\mathrm{L}^2(\pi)$. We represent this summand in the following way

$$V_n(h^*) - V_n(h^*_\varepsilon) = V_n(h^*) - V_n(h^*_\varepsilon) - \Big[\overline{V}_n(h^*) - \overline{V}_n(h^*_\varepsilon)\Big] + \Big[\overline{V}_n(h^*) - \overline{V}_n(h^*_\varepsilon)\Big].$$

Now we have by the union bound and the concentration result (36),

$$\mathrm{P}_{x_0}\Big(V_n(h^*) - V_n(h^*_\varepsilon) - \overline{V}_n(h^*) - \overline{V}_n(h^*_\varepsilon) > t\Big) \le \frac{\delta}{2} \tag{39}$$

for $t = \sqrt{c}K_\tau b_n n^{-1/2}\log^{1/2}(8/\delta)$. Furthermore, let us represent $V_n(h)$ as a quadratic form $Z_n(h)^\top A_n Z_n(h)$ with $\|A_n\| \le 2b_n/n$, see Section 5.2 for details. It holds by the Cauchy-Schwarz inequality

$$\begin{aligned}\overline{V}_n(h^*) - \overline{V}_n(h^*_\varepsilon) &= \mathsf{E}_{x_0}\big[Z_n(h^*)^\top A_n Z_n(h^*) - Z_n(h^*_\varepsilon)^\top A_n Z_n(h^*_\varepsilon)\big] \\ &= \mathsf{E}_{x_0}\big[Z_n(h^*)^\top A_n\big(Z_n(h^*) - Z_n(h^*_\varepsilon)\big) + \big(Z_n(h^*) - Z_n(h^*_\varepsilon)\big)^\top A_n Z_n(h^*_\varepsilon)\big] \\ &\le \|A_n\|\,\Big(\mathsf{E}_{x_0}\big[\|Z_n(h^*) - Z_n(h^*_\varepsilon)\|^2\big]\Big)^{1/2} \cdot 2\sup_{h\in\mathcal{H}}\big(\mathsf{E}_{x_0}\|Z_n(h)\|^2\big)^{1/2}.\end{aligned}$$

Let $R^2 \stackrel{\mathrm{def}}{=} \varsigma M^2 W(x_0)(1-\rho)^{-1}$. Then Lemma 10 yields

$$\overline{V}_n(h^*) - \overline{V}_n(h^*_\varepsilon) \le 4b_n\left(\varepsilon + \frac{\sqrt{2}R}{\sqrt{n}}\right)\left(H + \frac{R}{\sqrt{n}}\right). \tag{40}$$

Combining the bounds (37), (38), (39), and (40) for all summands and substituting them into (35), we can assert that for $\varepsilon \ge \gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n)$, with probability at least $1-\delta$,

$$\overline{V}_n(\widehat{h}_\varepsilon) - \overline{V}_n(h^*) \lesssim \left(K_\tau + H + \frac{R}{\sqrt{n}}\right)b_n\varepsilon + \frac{b_n R}{\sqrt{n}}\left(H + \frac{R}{\sqrt{n}}\right) + K_\tau\frac{b_n\log^{1/2}(\frac{8}{\delta})}{\sqrt{n}},$$

where $\lesssim$ stands for inequality up to an absolute constant. Now we can set τ to be an upper bound for the chosen t, namely, $\tau = \sqrt{c}K_\tau b_n\big(\varepsilon + n^{-1/2}\log^{1/2}(8/\delta)\big)$. In the bounded case, K_τ does not depend on τ, but in the Lipschitz case this choice leads to a quadratic equation

$$K_\tau^2 = \frac{\alpha L^2}{(1-r)^2}\left(H^2 + \frac{R^2}{n} + K_\tau\sqrt{c}\left(\varepsilon + \frac{\log^{1/2}(8/\delta)}{\sqrt{n}}\right)\right),$$

For a large $c > 0$, this quadratic equation always has a solution which may be written as $K_\tau \lesssim \frac{\sqrt{\alpha}L}{1-r}\big(H + Rn^{-1/2} + \big(\varepsilon + n^{-1/2}\log^{1/2}(8/\delta)\big)\big)$. Let $n \geq n_0$, where n_0 satisfies

$$n_0 \geq \frac{\varsigma W(x_0)}{(1-\rho)\pi(W)}, \quad n_0 \geq \frac{\max\{R^2, \log(8/\delta)\}}{H^2}, \quad \text{and} \quad \gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n_0) \leq H.$$

Then $K_\tau \lesssim \sqrt{\alpha}HL/(1-r)$ (in the Lipschitz case) and $H + Rn^{-1/2} \lesssim H$. We set $\varepsilon = \gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n)$ and obtain

$$\overline{V}_n(\widehat{h}_\varepsilon) - \overline{V}_n(h^*) \lesssim (K_\tau + H) b_n \gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n) + (K_\tau + HR)\frac{b_n \log(1/\delta)}{\sqrt{n}}.$$

Substituting this into (33) and taking $b_n = 2(\log(1/\rho))^{-1}\log(n)$, we conslude

$$V_\infty(\widehat{h}_\varepsilon) - \inf_{h\in\mathcal{H}} V_\infty(h) \lesssim (\log(1/\rho))^{-1}(K_\tau + H)\log(n)\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H}, n)$$
$$+ (\log(1/\rho))^{-1}\left(K_\tau + \frac{\varsigma^{1/2}MHW(x_0)}{(1-\rho)^{1/2}} + \frac{\varsigma^{1/2}M^2\pi(W)}{\sqrt{n}(1-\rho^{1/2})}\right)\frac{\log(n)\log(1/\delta)}{\sqrt{n}},$$

Note that $H \lesssim L$ or $H \lesssim B$ in the Lipschitz and bounded cases correspondingly, and $H \lesssim H^2 \lesssim K_\tau$ in both cases. Taking $K^2 = K_\tau$ and simplifying last expression, we get the desired conclusion.

5.4 Proof of Theorem 4

As above, we assume that functions $h \in \mathcal{H}$ are zero-mean and set $\overline{V}_n(h) = \mathsf{E}_{x_0}\big[V_n(h)\big]$. It follows from Proposition 11 that if $n \geq \varsigma W(x_0)/((1-\rho)\pi(W))$ then

$$\sup_{h\in\mathcal{H}}\big|V_\infty(h) - \overline{V}_n(h)\big| \lesssim G\left(\frac{b_n}{n} \vee \rho^{b_n/2}\right), \quad \text{where } G \stackrel{\text{def}}{=} \frac{\varsigma^{1/2}M^2\pi(W)}{1-\rho^{1/2}},$$

where M is defined in (32). Hence

$$V_\infty(\widehat{h}_\varepsilon) \leq \overline{V}_n(\widehat{h}_\varepsilon) + G\left(\frac{b_n}{n} \vee \rho^{b_n/2}\right). \tag{41}$$

We are reduced to bounding $\overline{V}_n(\widehat{h}_\varepsilon)$. Let us denote by h^* a constant function in $\mathcal{H}$ exising by assumption. Let also $h^*_\varepsilon \in \mathcal{H}_\varepsilon$ be the closest point to h^* in $\mathcal{H}_\varepsilon$ in $\mathrm{L}^2(\pi)$. By the definition of $\widehat{h}_\varepsilon$, $V_n(\widehat{h}_\varepsilon) - V_n(h^*_\varepsilon) < 0$. We have for any $c > 0$,

$$\overline{V}_n(\widehat{h}_\varepsilon) \leq \overline{V}_n(\widehat{h}_\varepsilon) - (1+c)\big(V_n(\widehat{h}_\varepsilon) - V_n(h^*_\varepsilon)\big) = \overline{V}_n(\widehat{h}_\varepsilon) - (1+c)V_n(\widehat{h}_\varepsilon) + (1+c)V_n(h^*_\varepsilon)$$
$$\leq \sup_{h\in\mathcal{H}_\varepsilon}\Big\{\overline{V}_n(h) - (1+c)V_n(h)\Big\} + (1+c)V_n(h^*_\varepsilon). \tag{42}$$

We take $c = 1$ and bound the two summands in the right hand side of (42) separately. To do this, we need an exponential concentration for $V_n(h)$. It follows from Theorem 19 (equation (51)) that, for all $t > 0$ and for all $h \in \mathcal{H}$,

$$\mathrm{P}_{x_0}\Big(\big|V_n(h) - \overline{V}_n(h)\big| > t\Big) \leq 2\exp\left(-\frac{nt^2}{cK^2\, b_n\big(\overline{V}_n(h) + t\big)}\right), \tag{43}$$

where $c > 0$ is some universal constant, $K^2 = \alpha L^2/(1-r)^2$, and b_n is the size of the lag window. For the first summand in the right hand side of the decomposition (42), using the

union bound and the concentration inequality (43), we obtain

$$\mathrm{P}_{x_0}\Big(\sup_{h\in\mathcal{H}_\varepsilon}\Big\{\overline{V}_n(h)-2V_n(h)\Big\}>t\Big)\le|\mathcal{H}_\varepsilon|\sup_{h\in\mathcal{H}_\varepsilon}\mathrm{P}_{x_0}\Big(\overline{V}_n(h)-2V_n(h)>t\Big)$$
$$\le 2|\mathcal{H}_\varepsilon|\sup_{h\in\mathcal{H}_\varepsilon}\exp\Big(-\frac{n\,(t+V_n(h))}{cK^2b_n}\Big)\le 2|\mathcal{H}_\varepsilon|\exp\Big(-\frac{nt}{cK^2b_n}\Big),$$

where the last inequality holds since $V_n(h)\ge 0$. For any $\varepsilon\ge\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H},n)$ it holds $|\mathcal{H}_\varepsilon|\le e^{n\varepsilon^2}$. Hence we can select $t=cK^2b_n\left(\varepsilon^2+n^{-1}\log(4/\delta)\right)$ to obtain

$$\mathrm{P}_{x_0}\Big(\sup_{h\in\mathcal{H}_\varepsilon}\Big\{\overline{V}_n(h)-2V_n(h)\Big\}>t\Big)\le\delta/2. \tag{44}$$

The second term in (42) is small since h^*_ε is ε-close to h^* in $\mathrm{L}^2(\pi)$. First we note that

$$V_n(h^*_\varepsilon)=V_n(h^*_\varepsilon)-2\overline{V}_n(h^*_\varepsilon)+\overline{V}_n(h^*_\varepsilon).$$

By the union bound and the concentration inequality (43), we have

$$\mathrm{P}_{x_0}\Big(V_n(h^*_\varepsilon)-2\overline{V}_n(h^*_\varepsilon)>t\Big)\le 2\exp\Big(-\frac{n\,(t+V_n(h^*_\varepsilon))}{cK^2b_n}\Big)\le 2\exp\Big(-\frac{nt}{cK^2b_n}\Big). \tag{45}$$

Hence for $t=cK^2b_nn^{-1}\log(4/\delta)$ this probability is bounded by $\delta/2$. Furthermore, let us represent $V_n(h)$ as a quadratic form $Z_n(h)^\top A_nZ_n(h)$ (see Section 5.2 for details). By assumption, h^* is a constant function, and hence $A_nZ_n(h^*)$ is the zero vector. Since $\|A_n\|\le 2b_n/n$ (see Lemma 9), it holds

$$\overline{V}_n(h^*_\varepsilon)=\mathsf{E}_{x_0}\left[Z_n(h^*_\varepsilon)^\top A_nZ_n(h^*_\varepsilon)\right]=\mathsf{E}_{x_0}\left[(Z_n(h^*_\varepsilon)-Z_n(h^*))^\top A_n(Z_n(h^*_\varepsilon)-Z_n(h^*))\right]$$
$$\le\frac{2b_n}{n}\mathsf{E}_{x_0}\left[\|Z_n(h^*_\varepsilon)-Z_n(h^*)\|^2\right]. \tag{46}$$

Let $R^2\stackrel{\mathrm{def}}{=}\varsigma M^2W(x_0)(1-\rho)^{-1}$. Then Lemma 10 yields

$$\overline{V}_n(h^*_\varepsilon)\le 2b_n\varepsilon^2+8R^2\frac{b_n}{n}. \tag{47}$$

Combining the bounds (44), (45) and (47) for all summands and substituting them into (42), we can assert that for $\varepsilon\ge\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H},n)$, with probability at least $1-\delta$, we have

$$\overline{V}_n(\widehat{h}_\varepsilon)\lesssim K^2b_n\varepsilon^2+(K^2+R^2)\frac{b_n\log(\frac{4}{\delta})}{n}.$$

Substituting this bound into (41) with $\varepsilon=\gamma_{\mathrm{L}^2(\pi)}(\mathcal{H},n)$ and $b_n=2(\log(1/\rho))^{-1}\log(n)$ yields

$$V_\infty(\widehat{h}_\varepsilon)\lesssim\frac{K^2}{\log(1/\rho)}\log(n)\gamma^2_{\mathrm{L}^2(\pi)}(\mathcal{H},n)+\frac{K^2+R^2+G}{\log(1/\rho)}\cdot\frac{\log(n)\log(\frac{1}{\delta})}{n}$$
$$\lesssim\frac{\alpha L^2}{(1-r)^2\log(1/\rho)}\log(n)\gamma^2_{\mathrm{L}^2(\pi)}(\mathcal{H},n)$$
$$+\left(\frac{\alpha L^2}{(1-r)^2\log(1/\rho)}+\frac{\varsigma M^2(\pi(W)+W(x_0))}{(1-\rho)^{1/2}\log(1/\rho)}\right)\frac{\log(n)\log(\frac{1}{\delta})}{n},$$

which is the desired conclusion.

6 Tables and Figures

Table 6: Experimental setup details.

Experiment	n_{burn}	n_{train}	n_{test}	γ_{ULA}	γ_{MALA}	γ_{RWM}	b_n
GMM, $\mathsf{E}_\pi[X_2]$, $\Sigma = \mathrm{I}$	10^4	10^5	10^5	0.1	1.0	0.5	50
GMM, $\mathsf{E}_\pi[X_2]$, $\Sigma = \Sigma_0$	10^4	10^5	10^5	0.1	0.2	0.1	50
GMM, $\mathsf{E}_\pi[X_2^2]$, $\Sigma = \mathrm{I}$	10^4	10^5	10^5	0.1	1.0	0.5	50
GMM, $\mathsf{E}_\pi[X_2^2]$, $\Sigma = \Sigma_0$	10^4	10^5	10^5	0.1	0.1	0.1	50
Banana-shape, $d = 2$	10^5	10^6	10^6	0.01	0.5	0.5	300
Banana-shape, $d = 8$	10^5	10^6	10^6	0.01	0.2	0.1	300
Logistic and probit regression, Pima	10^3	10^4	10^4	0.1	0.5	0.5	10
Logistic regression, EEG	10^3	10^4	10^4	0.1	1.0	0.1	10
Probit regression, EEG	10^3	10^4	10^4	0.1	0.5	0.1	10
Van der Pol oscillator	10^2	10^3	10^3	–	10^{-3}	–	10
Lotka-Volterra model	10^3	10^4	10^4	–	5×10^{-6}	–	10

Table 7: Variance Reduction Factors in probit regression, average test likelihood.

	PIMA dataset			EEG dataset		
Method	ULA	MALA	RWM	ULA	MALA	RWM
ESVM-1	263.2	419.7	251.4	1317.0	1515.0	938.5
EVM-1	270.1	430.1	261.6	1331.6	1572.7	948.1
ESVM-2	**26835.7**	**55373.7**	**28905.0**	**45059.2**	45964.5	**34957.1**
EVM-2	6660.7	29710.4	14187.1	29620.4	**71095.6**	6340.1

Table 8: Variance Reduction Factors for Van der Pol oscillator, posterior mean estimation.

Method	1st order CV	2nd order CV	3rd order CV
ESVM	30.7	**49.1**	**243.2**
EVM	**33.9**	44.1	183.7

Figure 6: Estimation of $\mathsf{E}_\pi[X_2^2]$ in GMM with $\Sigma = \mathrm{I}$. Left figure: boxplot for ULA estimates compared to the corresponding boxplots for EVM and ESVM estimates. Next three figures: boxplots for EVM and ESVM estimates for ULA, MALA, and RWM with second-order control variates being used.

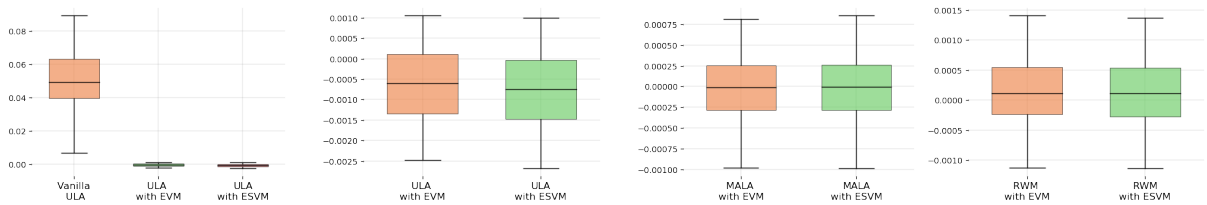


Figure 7: Estimation of $\mathsf{E}_\pi[X_2^2]$ in GMM with $\Sigma = \Sigma_0$. Left figure: boxplot for ULA estimates compared to the corresponding boxplots for EVM and ESVM estimates. Next three figures: boxplots for EVM and ESVM estimates for ULA, MALA, and RWM with second-order control variates being used.

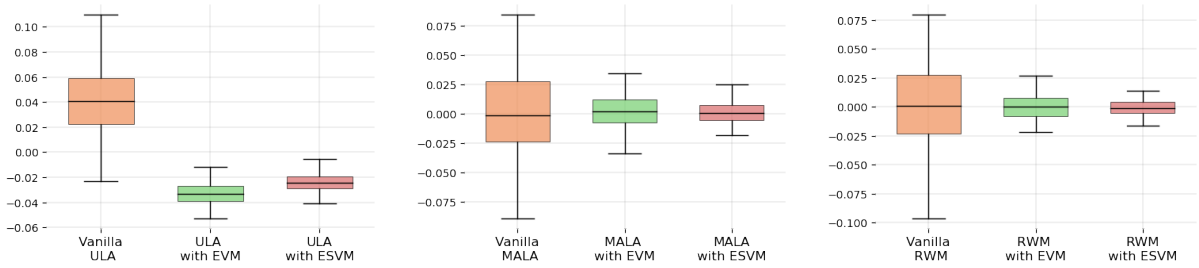


Figure 8: Estimation of the average test likelihood in probit regression for the Pima dataset. Left figure: boxplot for ULA estimates compared to the corresponding boxplots for EVM and ESVM estimates. Next three figures: boxplots for EVM and ESVM estimates for ULA, MALA, and RWM with second-order control variates being used.

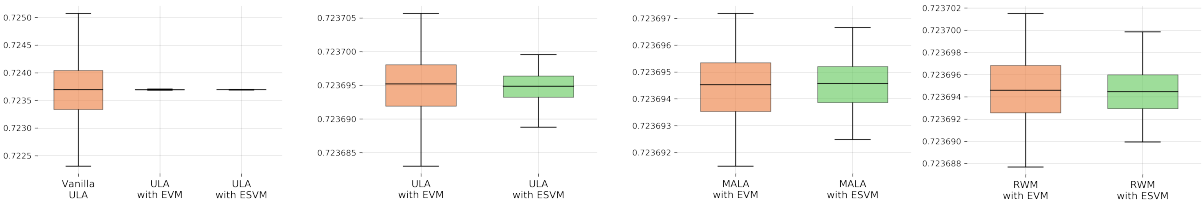


Figure 9: Estimation of the average test likelihood in probit regression for the EEG dataset. Left figure: boxplot for ULA estimates compared to the corresponding boxplots for EVM and ESVM estimates. Next three figures: boxplots for EVM and ESVM estimates for ULA, MALA, and RWM with second-order control variates being used.

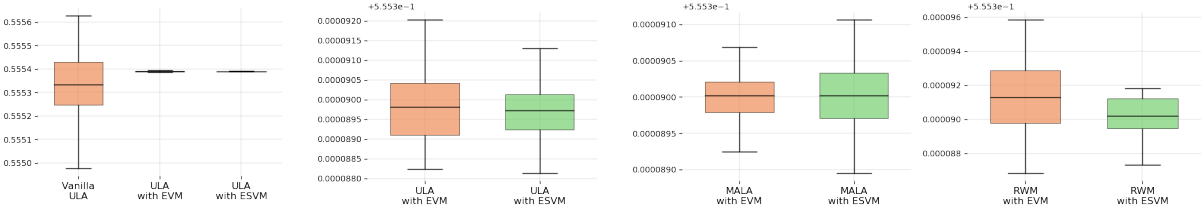


Figure 10: Estimating the mean of the posterior distribution in the Van der Pol model. From left to right: boxplots for vanilla estimates and the corresponding EVM and ESVM estimates with third-order polynomials being used as control variates, EVM and ESVM comparison for second-order polynomials, and EVM and ESVM comparison for third-order polynomials.

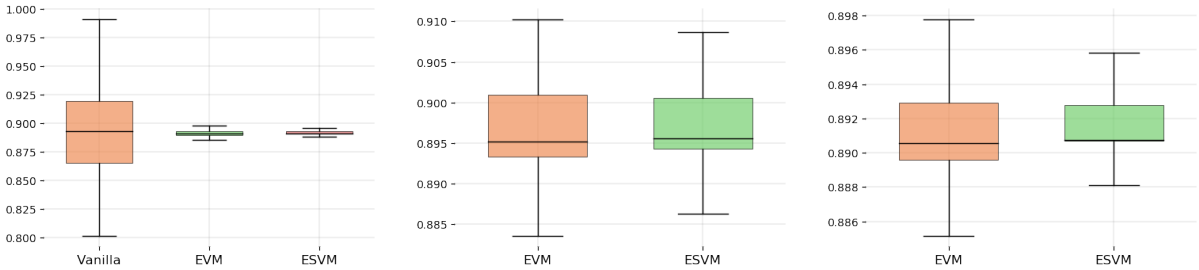
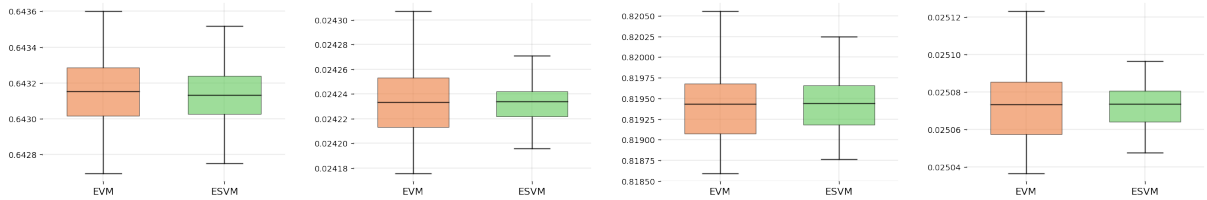


Figure 11: Estimating the mean of the posterior distribution in the Lotka-Volterra model. From left to right: posterior mean for parameters α, β, γ, and δ.



A Appendix

A.1 Concentration of the spectral variance estimator for Lipschitz functions

The proof of a concentration inequality for Lipschitz functions falls naturally into three steps. First we show, using a result from Djellout et al. [11], that the joint distribution of $(X_k)_{k=0}^{n-1}$ satisfies $\mathrm{T}_2(\alpha)$ model. Then we note that $\mathrm{T}_2(\alpha)$ implies Gaussian concentration for all Lipschitz functions. And, finally, this Gaussian concentration property implies a concentration inequality for quadratic forms from Adamczak [1], which we apply to the spectral variance estimator. For the sake of completeness we provide all necessary details below.

Tensorization of $\mathrm{T}_2(\alpha)$ for Markov chains. Let $\mathrm{P}_{x_0}^n$ be the joint distribution of the Markov chain $(X_k)_{k=0}^{n-1}$ with the Markov kernel P under P_{x_0}. Since here we consider distributions on the product space X^{n-1}, additional definitions are needed. We define the distance between points $x^{n-1} = (x_1, \dots, x_{n-1}) \in \mathsf{X}^{n-1}$ and $y^{n-1} = (y_1, \dots, y_{n-1}) \in \mathsf{X}^{n-1}$ by

$$\mathrm{d}_2(x^{n-1}, y^{n-1}) \stackrel{\mathrm{def}}{=} \left(\sum_{j=1}^{n-1} \|x_j - y_j\|^2 \right)^{1/2}, \tag{48}$$

The L^p-Wasserstein distance between probability measures μ and ν on X^{n-1} with respect to the metric d_2 is given by

$$W_p^{\mathrm{d}_2}(\mu, \nu) \stackrel{\mathrm{def}}{=} \inf_{\zeta} \left(\int_{\mathsf{X}^{n-1} \times \mathsf{X}^{n-1}} \mathrm{d}_2^p(x, y) \, \mathrm{d}\zeta(x, y) \right)^{1/p},$$

where the infimum is taken over all probability measures ζ on the product space $\mathsf{X}^{n-1} \times \mathsf{X}^{n-1}$ with marginal distributions μ and ν. And finally, we say that the probability measure μ on X^{n-1} satisfies $\mathrm{T}_p(\alpha)$ if there is a constant $\alpha > 0$ such that for any probability measure ν on X^{n-1}

$$W_p^{\mathrm{d}_2}(\mu, \nu) \le \sqrt{2\alpha \, \mathrm{KL}(\mu \| \nu)}.$$

The following theorem provides sufficient conditions for the measure $\mathrm{P}_{x_0}^n$ to satisfy $\mathrm{T}_2(\alpha)$.

Theorem 13 (Djellout et al. [11, Theorem 2.5]) *Assume that there exists $\alpha > 0$, such that $P(x, \cdot) \in \mathrm{T}_2(\alpha)$ for any $x \in \mathsf{X}$, and there exists $0 < r < 1$, such that for any $x, y \in \mathsf{X}$,*

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

Then for any probability measure Q on X^{n-1}, the product measure $\mathrm{P}_{x_0}^n$ satisfies $\mathrm{T}_2(\alpha/(1-r)^2)$, i.e.

$$W_2^{\mathrm{d}_2}(\mathrm{Q}, \mathrm{P}_{x_0}^n) \le \frac{1}{1-r} \sqrt{2\alpha \, \mathrm{KL}(\mathrm{Q} \, \| \, \mathrm{P}_{x_0}^n)}.$$

Gaussian concentration for Lipschitz functions. A probability measure which satisfies $\mathrm{T}_2(\alpha)$ inequality is known to satisfy Gaussian concentration inequality for all Lipschitz functions. Together with Theorem 13 this implies the following result.

Theorem 14 *Assume that P satisfies (CW). Then for any L-Lipschitz function $\phi : \mathsf{X}^{n-1} \to \mathbb{R}$ with respect to the metric d_2 from (48), it holds*

$$\mathrm{P}_{x_0}\left(\left| \phi(X_0, \dots, X_{n-1}) - \mathsf{E}_{x_0}\left[\phi(X_0, \dots, X_{n-1}) \right] \right| \ge t \right) \le 2 \exp\left(- \frac{t^2}{2\alpha L^2/(1-r)^2} \right). \tag{49}$$

Proof: It follows from Bakry et al. [3, Section 9.2] that $\mathrm{T}_2(\alpha)$ implies $\mathrm{T}_1(\alpha)$ with the same constant $\alpha > 0$ and with respect to the same metric d_2. In its turn $\mathrm{T}_1(C)$ imply the Gaussian concentration (49) due to the result of Bobkov and Götze [6]. It remains to note that $\mathrm{P}_{x_0}^n$ satisfies $\mathrm{T}_2(\alpha/(1-r)^2)$ by Theorem 13. $\square$

Gaussian concentration for quadratic forms. Once we have proved the Gaussian concentration for Lipschitz functions, we can obtain the Bernstein-type inequality for quadratic forms. This idea is due to Adamczak [1], but since we use a modified version of the inequality, we provide the details for readers convenience.

Definition 15 (Concentration property) *Let Z be a random vector in $\mathbb{R}^n$. We say that Z has the concentration property with constant K if for every 1-Lipschitz function $\phi : \mathbb{R}^n \to \mathbb{R}$, we have $\mathsf{E}|\phi(X)| < \infty$ and for every $t > 0$,*

$$\mathrm{P}\Big(\big|\phi(Z) - \mathsf{E}\big[\phi(Z)\big]\big| \geq t\Big) \leq 2\exp\left(-t^2/K^2\right).$$

The following theorem shows that the concentration property implies a concentration inequality for quadratic forms.

Theorem 16 *Let Z be a random vector in $\mathbb{R}^n$. If Z has the concentration property with constant K, then for any $n \times n$ matrix A and every $t > 0$,*

$$\mathrm{P}\Big(\big|Z^\top AZ - \mathsf{E}\big[Z^\top AZ\big]\big| > t\Big) \leq 2\exp\left(-\frac{t^2}{cK^2\left(\mathsf{E}\big[\|AZ\|^2\big] + t\|A\|\right)}\right),$$

where $c > 0$ is a universal constant.

Proof: Without loss of generality one may assume that A is symmetric and positively semidefinite. Let $\varphi(z) \stackrel{\mathrm{def}}{=} z^\top Az$, $z \in \mathbb{R}^n$. Define $\psi(z) \stackrel{\mathrm{def}}{=} \|\nabla\varphi(z)\|$. Since $\|\nabla\varphi(z)\| \leq 2\|A\|\|z\|$, the function ψ is $(2\|A\|)$-Lipschitz. By the concentration property

$$\mathrm{P}\big(\big|\psi(Z) - \mathsf{E}\big[\psi(Z)\big]\big| \geq t\big) \leq 2\exp\left(-\frac{t^2}{4K^2\|A\|^2}\right).$$

Note that $\mathsf{E}\big[\psi(Z)\big] = 2\mathsf{E}\big[\|AZ\|\big]$ and set for $t > 0$,

$$B_t \stackrel{\mathrm{def}}{=} \big\{z \in \mathbb{R}^n : \psi(z) \leq 2\mathsf{E}\big[\|AZ\|\big] + \sqrt{t\|A\|}\big\}.$$

It holds

$$\mathrm{P}(Z \notin B_t) \leq 2\exp\left(-\frac{t}{4K^2\|A\|}\right).$$

Define $\widetilde{\varphi}(z) \stackrel{\mathrm{def}}{=} \sup_{y \in B_t}(\langle\nabla\varphi(y), z-y\rangle + \varphi(y))$. This function is Lipschitz, since for any $z, x \in B_t$, $|\widetilde{\varphi}(z_1) - \widetilde{\varphi}(z_2)| \leq \sup_{y\in B_t}\|\nabla\varphi(y)\|\|z_1 - z_2\| \leq M\|z_1 - z_2\|$ with $M \stackrel{\mathrm{def}}{=} 2\mathsf{E}\big[\|AZ\|\big] + \sqrt{t\|A\|}$. Hence, again by the concentration property, for any $s > 0$,

$$\begin{aligned}\mathrm{P}\big(\big|\widetilde{\varphi}(Z) - \mathsf{E}\big[\widetilde{\varphi}(Z)\big]\big| \geq s\big) &\leq 2\exp\left(-\frac{s^2}{K^2(2\mathsf{E}\big[\|AZ\|\big] + \sqrt{t\|A\|})^2}\right)\\ &\leq 2\exp\left(-\frac{s^2}{4K^2(\mathsf{E}\big[\|AZ\|\big] + \sqrt{t\|A\|})^2}\right).\end{aligned}$$

Moreover, by convexity of φ, we have $\widetilde{\varphi}(z) \leq \varphi(z)$ and for $z \in B_t$, $\widetilde{\varphi}(z) = \varphi(z)$. Consider two random variables $Y = \varphi(Z)$ and $\widetilde{Y} = \widetilde{\varphi}(Z)$. We have proved that Y and $\widetilde{Y}$ coincide on the set B_t of large probability and $\widetilde{Y}$ has the concentration property. It follows from

Lemma 17 (given below) that in this case we have the Gaussian concentration for Y around median $\operatorname{Med} Y$ of the form

$$\begin{aligned}\mathrm{P}\Big(\big|Z^\top AZ - \operatorname{Med}[Z^\top AZ]\big| \ge t\Big) &\le 2\exp\left(-\frac{t^2}{cK^2(\mathsf{E}\big[\|AZ\|\big] + \sqrt{t\|A\|})^2}\right)\\ &\le 2\exp\left(-\frac{t^2}{2cK^2(\mathsf{E}\big[\|AZ\|^2\big] + t\|A\|)}\right).\end{aligned}$$

By a standard argument (see, for example, Adamczak [1, Lemma 3.2]), we replace the median by the mean at the cost of a universal factor. This completes the proof for a new absolute constant $c > 0$. $\square$

Lemma 17 *Assume that there exist positive constants $a, b, t > 0$ such that for any $s > 0$ random variables Y, $\widetilde{Y}$ satisfy*

$$\mathrm{P}\left(\big|\widetilde{Y} - \mathsf{E}\widetilde{Y}\big| \ge s\right) \le 2\exp\left(-s^2/(a + b\sqrt{t})^2\right)$$

and $\mathrm{P}\left(\widetilde{Y} \ne Y\right) \le 2\exp\big(-t/b\big)$. Then for some positive constant $c > 0$ and all $t > 0$,

$$\mathrm{P}(|Y - \operatorname{Med} Y| \ge t) \le 2\exp\left(-t^2/\{c(a + b\sqrt{t})^2\}\right).$$

Proof: This lemma is proved in Adamczak [1, Lemma 3.2]. We just note that the quantity $-\min\big(t^2/a^2, t/b\big)$, which appears in the result of Adamczak [1], is bounded by the quantity $-t^2/(a + b\sqrt{t})^2$. $\square$

We have arrived at the following concentration result for quadratic forms of Lipschitz function of a Markov chain. This result is of independent interest.

Corollary 18 *Assume that there exists $\alpha > 0$, such that $P(x, \cdot) \in \mathrm{T}_2(\alpha)$ for any $x \in \mathsf{X}$, and there exists $0 < r < 1$, such that for any $x, y \in \mathsf{X}$,*

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

Let also $h : \mathsf{X} \to \mathbb{R}$ be a L-Lipschitz function. Denote $Z_n(h) \stackrel{def}{=} (h(X_0), \dots, h(X_{n-1}))^\top$. Then for any $n \times n$ matrix A and any $t > 0$,

$$\begin{aligned}&\mathrm{P}_{x_0}\Big(\big|Z_n(h)^\top AZ_n(h) - \mathsf{E}_{x_0}\big[Z_n(h)^\top AZ_n(h)\big]\big| > t\Big)\\ &\qquad\le 2\exp\left(-\frac{t^2}{cK^2\left(\mathsf{E}_{x_0}\big[\|AZ_n(h)\|^2\big] + t\|A\|\right)}\right), \quad (50)\end{aligned}$$

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

Proof: The statement follows from the fact $Z_n(h)$ has the concentration property with $K = 2\alpha L^2/(1 - r)^2$. Indeed, for any 1-Lipschitz function $\phi : \mathbb{R}^n \to \mathbb{R}$ and any $x^{n-1} \stackrel{\text{def}}{=} (x_1, \dots, x_{n-1}) \in \mathsf{X}^{n-1}$, $y^{n-1} \stackrel{\text{def}}{=} (y_1, \dots, y_{n-1}) \in \mathsf{X}^{n-1}$, it holds

$$\begin{aligned}\big|\phi(h(x_0), \dots, h(x_{n-1})) - \phi(h(y_0), \dots, h(y_{n-1}))\big| &\le \left(\sum_{j=1}^{n-1}(h(x_j) - h(y_j))^2\right)^{1/2}\\ &\le Ld(x^{n-1}, y^{n-1}).\end{aligned}$$

Hence the concentration property follows from Theorem 14. Application of Theorem 16 to $Z_n(h)$ finishes the proof. $\square$

Gaussian concentration of the spectral variance estimator The main result of this section is the following.

Theorem 19 *Assume that functions $h \in \mathcal{H}$ and the Markov kernel P satisfy (L) and (CW) with parameters $L > 0$, $\alpha > 0$, and $0 < r < 1$. Then for all $t > 0$,*

$$\mathrm{P}_{x_0}\Big(\big|V_n(h) - \mathsf{E}_{x_0}\big[V_n(h)\big]\big| > t\Big) \le 2\exp\left(-\frac{nt^2}{cK^2\, b_n\big(\mathsf{E}_{x_0}\big[V_n(h)\big] + t\big)}\right), \tag{51}$$

where $c > 0$ is some universal constant, $K^2 = \alpha L^2/(1-r)^2$, and b_n is the size of the lag window. Moreover, if additionally $(X_k)_{k=0}^{n-1}$ satisfies (GE) with parameters ς, ρ, and function W, then for all $t < \tau$,

$$\mathrm{P}_{x_0}\Big(\big|V_n(h) - \mathsf{E}_{x_0}\big[V_n(h)\big]\big| > t\Big) \le 2\exp\left(-\frac{nt^2}{cK_\tau^2 b_n^2}\right), \tag{52}$$

where

$$K_\tau^2 \stackrel{def}{=} \frac{\alpha L^2}{(1-r)^2}\left(\|h\|^2_{\mathrm{L}^2(\pi)} + \frac{\varsigma W(x_0)\|h\|^2_{W^{1/2}}}{1-\rho} + \frac{\tau}{b_n}\right).$$

Proof: The proof is straightforward. We have showed that the spectral variance estimator can be represented as a quadratic form $V_n(h) = Z_n(h)^\top A_n Z_n(h)$ with $\|A_n\| \le 2b_n/n$, see Section 5.2 and Lemma 9 therein. Now Corollary 18 yields for $K^2 = \alpha L^2/(1-r)^2$ and all $t > 0$, that

$$\begin{aligned}\mathrm{P}_{x_0}\Big(\big|V_n(h) - \mathsf{E}_{x_0}\big[V_n(h)\big]\big| > t\Big) &\le 2\exp\left(-\frac{t^2}{cK^2\big(\mathsf{E}_{x_0}\big[\|A_n Z_n(h)\|^2\big] + t\|A_n\|\big)}\right)\\ &\le 2\exp\left(-\frac{nt^2}{2cK^2 b_n\,\big(\mathsf{E}_{x_0}\big[V_n(h)\big] + t\big)}\right),\end{aligned}$$

which establishes (51) for a new absolute constant $c > 0$. To prove the second inequality we note that by Lemma 9 and Lemma 10,

$$\mathsf{E}_{x_0}\big[V_n(h)\big] \le \|A_n\|\mathsf{E}_{x_0}\big[\|Z_n(h)\|^2\big] \le 2b_n\|h\|^2_{\mathrm{L}^2(\pi)} + \frac{2\varsigma W(x_0)\|h\|^2_{W^{1/2}}}{1-\rho}\frac{b_n}{n}.$$

Hence for any $0 < t < \tau$, we have

$$\mathsf{E}_{x_0}\big[V_n(h)\big] + t \le b_n\left(2\|h\|^2_{\mathrm{L}^2(\pi)} + \frac{2\varsigma W(x_0)\|h\|^2_{W^{1/2}}}{1-\rho}\frac{1}{n} + \frac{\tau}{b_n}\right).$$

Substituting this into (51) we deduce

$$\mathrm{P}_{x_0}\Big(\big|V_n(h) - \mathsf{E}_{x_0}\big[V_n(h)\big]\big| > t\Big) \le 2\exp\left(-\frac{nt^2}{cK_\tau^2 b_n^2}\right)$$

for a new absolute constant $c > 0$ and

$$K_\tau^2 \stackrel{\mathrm{def}}{=} \frac{\alpha L^2}{(1-r)^2}\left(\|h\|^2_{\mathrm{L}^2(\pi)} + \frac{\varsigma W(x_0)\|h\|^2_{W^{1/2}}}{1-\rho} + \frac{\tau}{b_n}\right),$$

which completes the proof. □

A.2 Concentration of the spectral variance estimator for bounded functions

Theorem 20 *Assume that P satisfies (GE) and (BR) with parameters $\varsigma, \rho, l > 0$, function W, and set S. Assume also that functions $h \in \mathcal{H}$ satisfy (B) with parameter $B > 0$. Then for $x_0 \in \mathsf{S}$, for all functions $h \in \mathcal{H}$, and all $t > 0$,*

$$\mathrm{P}_{x_0}\Big(\big|V_n(h) - \mathsf{E}_{x_0}\big[V_n(h)\big]\big| > t\Big) \le 2\exp\left(-\frac{t^2 n}{cK^2 b_n^2}\right), \tag{53}$$

where b_n is the size of the lag window, $K = \beta B^2$, and β is given by

$$\beta = \frac{\varsigma l}{1-\rho}\left(\frac{1}{\log u} + \frac{J\varsigma l}{1-\rho}\right). \tag{54}$$

Proof: The main idea of the proof is to show that the spectral variance satisfies the bounded difference property. First we rewrite the lag s sample autocovariance function as

$$\begin{aligned}\hat{\rho}_n^{(h)}(s) &= \frac{1}{n}\sum_{k=0}^{n-s-1}\Big(h(X_k) - \pi_n(h)\Big)\Big(h(X_{k+s}) - \pi_n(h)\Big) \\ &= \frac{1}{n}\sum_{k=0}^{n-s-1} h(X_k)h(X_{k+s}) - \frac{\pi_n(h)}{n}\sum_{k=s}^{n-s-1} h(X_k).\end{aligned}$$

Let $\hat{\rho}_n^{(h,i)}(s)$ and $V_n^{(i)}(h)$ be the sample autocovariance function and the spectral variance determined on another sample $X_0, \ldots, X_{i-1}, X_i', X_{i-1}, \ldots, X_{n-1}$, where we have replaced X_i by X_i'. It holds

$$\big|\hat{\rho}_n^{(h)}(s) - \hat{\rho}_n^{(h,i)}(s)\big| \le 2B^2 + \frac{2(n-2s+n)}{n^2}B^2 \le \frac{6B^2}{n},$$

and since $|w_n(s)| \le 1$ by definition,

$$\big|V_n(h) - V_n^{(i)}(h)\big| \le 2b_n \sup_s |w_n(s)| \cdot |\hat{\rho}_n^{(h)}(s) - \hat{\rho}_n^{(h,i)}(s)| \le \frac{12 b_n B^2}{n}.$$

The bounded differences inequality for Markov chains from Douc et al. [12, Theorem 23.3.1]) with explicit constants from Havet et al. [20] yields

$$\mathrm{P}_{x_0}\Big(\big|V_n(h) - \mathsf{E}_{x_0}\big[V_n(h)\big]\big| > t\Big) \le 2\exp\left(-\frac{t^2 n}{144\beta B^4 b_n^2}\right), \text{ with } \beta = \frac{\varsigma l}{1-\rho}\left(\frac{1}{\log u} + \frac{J\varsigma l}{1-\rho}\right).$$

which completes the proof. □

References

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

[2] Roland Assaraf and Michel Caffarel. Zero-variance principle for Monte Carlo algorithms. *PHYSICAL REVIEW LETTERS*, 83(23):4682–4685, DEC 6 1999.

[3] Dominique Bakry, Ivan Gentil and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 348. Springer Science & Business Media, 2013.

[4] Denis Belomestny, Leonid Iosipoi and Nikita Zhivotovskiy. Variance reduction via empirical variance minimization: convergence and complexity. *arXiv:1712.04667*, 2017.

[5] Denis Belomestny, Leonid Iosipoi and Nikita Zhivotovskiy. Variance reduction in monte carlo estimators via empirical variance minimization. *Doklady Mathematics*, 98(2):494–497, 2018.

[6] Sergey Bobkov and Friedrich Götze. Exponential integrability and transportation cost related to logarithmic sobolev inequalities. *Journal of Functional Analysis*, 163(1):1–28, 4 1999. ISSN 0022-1236. doi: 10.1006/jfan.1998.3326.

[7] Nicolas Brosse, Alain Durmus, Sean Meyn, Eric Moulines and Anand Radhakrishnan. Diffusion approximations and control variates for MCMC. *arXiv:1808.01665*, 2019.

[8] Arnak 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.

[9] Petros Dellaportas and Ioannis 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):133–161, 2012.

[10] Luc Devroye, László Györfi and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[11] Hacéne Djellout, Arnaud Guillin and Liming Wu. Transportation cost-information inequalities and applications to random dynamical systems and diffusions. *Ann. Probab.*, 32(3B):2702–2732, 2004. ISSN 0091-1798. doi: 10.1214/009117904000000531. URL https://doi.org/10.1214/009117904000000531.

[12] Randal Douc, Eric Moulines, Pierre Priouret and Philippe Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2018. ISBN 978-3-319-97703-4; 978-3-319-97704-1. doi: 10.1007/978-3-319-97704-1. URL https://doi.org/10.1007/978-3-319-97704-1.

[13] Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *arXiv:1605.01559*, 2016.

[14] Alain Durmus and Éric Moulines. Non-asymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.

[15] James Flegal and Galin Jones. Batch means and spectral variance estimators in Markov chain monte carlo. *Ann. Statist.*, 38(2):1034–1070, 04 2010. doi: 10.1214/09-AOS735. URL https://doi.org/10.1214/09-AOS735.

[16] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari and Donald Rubin. *Bayesian data analysis*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, third edition, 2014.

[17] Paul Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53. Springer Science & Business Media, 2013.

[18] Emmanuel Gobet. *Monte-Carlo methods and stochastic processes*. CRC Press, Boca Raton, FL, 2016.

[19] Heikki Haario, Eero Saksman and Johanna Tamminen. Adaptive proposal distribution for random walk metropolis algorithm. *Computational Statistics*, 14(3):375–395, Sep 1999. ISSN 0943-4062. doi: 10.1007/s001800050022. URL https://doi.org/10.1007/s001800050022.

[20] Antoine Havet, Matthieu Lerasle, Eric Moulines and Elodie Vernet. A quantitative Mc Diarmid's inequality for geometrically ergodic Markov chains. *arXiv: 1907.02809*, 2019.

[21] Shane Henderson. *Variance reduction via an approximating Markov process*. PhD thesis, Stanford University, 1997.

[22] Søren Fiig Jarner and Ernst Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.*, 85(2):341–361, 2000. ISSN 0304-4149. doi: 10.1016/S0304-4149(99)00082-4. URL https://doi.org/10.1016/S0304-4149(99)00082-4.

[23] Galin Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004.

[24] Jean-Michel Marin and Christian Robert. *Bayesian core: a practical approach to computational Bayesian statistics*. Springer Texts in Statistics. Springer, New York, 2007. ISBN 978-0-387-38979-0; 0-387-38979-2.

[25] Mijatovi, Aleksandar and Vogrinc, Jure. On the Poisson equation for Metropolis-Hastings chains. *Bernoulli*, 24(3):2401–2428, 2018. URL https://doi.org/10.3150/17-BEJ932.

[26] Antonietta Mira, Reza Solgi and Daniele Imparato. Zero variance Markov chain Monte Carlo for Bayesian estimators. *Statistics and Computing*, 23(5):653–662, 2013.

[27] Richard Nickl and Benedikt 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.

[28] Chris Oates, Jon Cockayne, François-Xavier Briol and Mark Girolami. Convergence Rates for a Class of Estimators Based on Stein's Identity. *arXiv:1603.03220*, 2016.

[29] Chris Oates, Mark Girolami and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017.

[30] Chris Oates, Jon Cockayne, François-Xavier Briol and Mark Girolami. Convergence rates for a class of estimators based on Stein's method. *Bernoulli*, 25(2):1141–1159, 2019. ISSN 1350-7265. doi: 10.3150/17-bej1016. URL https://doi.org/10.3150/17-bej1016.

[31] Theodore Papamarkou, Antonietta Mira and Mark Girolami. Zero variance differential geometric Markov chain monte carlo algorithms. *Bayesian Anal.*, 9(1):97–128, 03 2014. doi: 10.1214/13-BA848. URL https://doi.org/10.1214/13-BA848.

[32] Christian Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, New York, 1999.

[33] Gareth Roberts and Jeffrey Rosenthal. General state space Markov chains and MCMC algorithms. *Probab. Surveys*, 1:20–71, 2004. doi: 10.1214/154957804100000024. URL https://doi.org/10.1214/154957804100000024.

[34] Gareth Roberts and Richard Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. ISSN 1350-7265. doi: 10.2307/3318418. URL http://dx.doi.org/10.2307/3318418.

[35] Gareth Roberts and Richard Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1): 95–110, 1996. ISSN 0006-3444. doi: 10.1093/biomet/83.1.95. URL https://doi.org/10.1093/biomet/83.1.95.

[36] Reuven Rubinstein and Dirk Kroese. *Simulation and the Monte Carlo Method*, volume 10. John Wiley & Sons, 2016.

[37] Leah South, Chris Oates, Antonietta Mira and Christopher Drovand i. Regularised Zero-Variance Control Variates for High-Dimensional Variance Reduction. *arXiv:1811.05073*, 2018.

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

[39] Wing Wong and Xiaotong Shen. Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLES. *The Annals of Statistics*, 23(2):339–362, 1995.