SIAM/ASA J. UNCERTAINTY QUANTIFICATION
Vol. 9, No. 2, pp. 507–535

Variance Reduction for Dependent Sequences with Applications to Stochastic Gradient MCMC*

Denis Belomestny†, Leonid Iosipoi‡, Eric Moulines§, Alexey Naumov‡, and Sergey Samsonov‡

Abstract. In this paper we propose a novel and practical variance reduction approach for additive functionals of dependent sequences. Our approach combines the use of control variates with the minimization of an empirical variance estimate. We analyze finite sample properties of the proposed method and derive finite-time bounds of the excess asymptotic variance to zero. We apply our methodology to stochastic gradient Markov chain Monte Carlo (SGMCMC) methods for Bayesian inference on large data sets and combine it with existing variance reduction methods for SGMCMC. We present empirical results carried out on a number of benchmark examples showing that our variance reduction method achieves significant improvement as compared to state-of-the-art methods at the expense of a moderate increase of computational overhead.

Key words. MCMC algorithms, variance reduction, stochastic gradient, generative models

AMS subject classifications. 60J20, 65C40, 65C60

DOI. 10.1137/19M1301199

1. Introduction.

Variance reduction aims at reducing the stochastic error of a Monte Carlo estimate; see [39], [42], [25], and [24] for an introduction to this field. Recently one witnessed a revival of interest in variance reduction techniques for dependent sequences with applications to Bayesian inference and reinforcement learning among others; see, for instance, [33], [28], [15], [11], [2], and references therein.

Suppose that we wish to compute the integral of an arbitrary function $f : \mathsf{X} \mapsto \mathbb{R}$ with respect to a probability measure π on a general state-space $(\mathsf{X}, \mathcal{X})$, that is, $\pi(f) = \int_{\mathsf{X}} f(x)\pi(\mathrm{d}x)$. If drawing an independent and identically distributed (i.i.d.) sample from π is an option, a natural estimator for $\pi(f)$ is the sample mean

$$\pi_N(f) := N^{-1} \sum_{k=0}^{N-1} f(X_k), \quad N \in \mathbb{N},$$

where $(X_k)_{k=0}^{N-1}$ is an i.i.d. sample from π. Using the central limit theorem, one can construct an asymptotically valid confidence interval for the value $\pi(f)$ of the form $\pi_N(f) \pm$

*Received by the editors November 20, 2019; accepted for publication (in revised form) January 22, 2021; published electronically May 3, 2021.
https://doi.org/10.1137/19M1301199
Funding: This work was prepared within the framework of the HSE University Basic Research Program. Results of section 3 were obtained under support of the RSF grant 19-71-30020 (HSE University).

†Duisburg-Essen University, Essen, Germany, and HSE University, Moscow, Russia (denis.belomestny@uni-due.de).
‡HSE University, Moscow, Russia (liosipoi@hse.ru, anaumov@hse.ru, svsamsonov@hse.ru).
§Ecole Polytechnique, Paris, France, and HSE University, Moscow, Russia (eric.moulines@polytechnique.edu).

$\mathfrak{q}\, N^{-1/2}(\mathrm{Var}_\pi(f))^{1/2}$, where $\mathfrak{q}$ is a quantile of a normal distribution, and $\mathrm{Var}_\pi(f) = \int_{\mathsf{X}}\{f(x) - \pi(f)\}^2\pi(\mathrm{d}x)$. A general way to reduce the variance $\mathrm{Var}_\pi(f)$ is to select another function g in a set $\mathcal{G}$ such that $\pi(g) = 0$ and $\mathrm{Var}_\pi(f-g) \ll \mathrm{Var}_\pi(f)$. Such a function g is called a *control variate*. A natural approach to learn $g \in \mathcal{G}$ is to minimize the empirical variance

$$D_n(f-g) = (n-1)^{-1}\sum_{k=0}^{n-1}\big(f(X_k) - g(X_k) - \pi_n(f-g)\big)^2 \tag{1.1}$$

constructed using a new independent learning sample $(X_k)_{k=0}^{n-1}$. This leads to the empirical variance minimization (EVM) method recently studied in [6] and [7]. In many problems of interest, drawing an i.i.d. sample from π is not an option, yet it is possible to obtain a nonstationary dependent sequence $(X_k)_{k=0}^{\infty}$ whose marginal distribution converges to π. This situation is typical in Bayesian statistics, where π represents a posterior distribution and $(X_k)_{k=0}^{\infty}$ is sampled using Markov chain Monte Carlo (MCMC) methods. Under appropriate conditions, the central limit theorem also holds, and therefore, it is possible to construct the asymptotic confidence interval for $\pi(f)$ of the form

$$\left[\pi_N(f) - \mathfrak{q}\sqrt{\frac{V_\infty(f)}{N}}, \pi_N(f) + \mathfrak{q}\sqrt{\frac{V_\infty(f)}{N}}\right], \tag{1.2}$$

where $V_\infty(f)$ is the asymptotic variance defined as

$$V_\infty(f) := \lim_{N\to\infty} N\cdot\mathsf{E}\Big[\big(\pi_N(f) - \pi(f)\big)^2\Big]. \tag{1.3}$$

A sensible approach is to select a control variate $g \in \mathcal{G}$ by minimizing an estimate for the asymptotic variance $V_\infty(f-g)$. When the spectral estimate of $V_\infty(f-g)$ is used, this leads to the empirical spectral variance minimization (ESVM); see [5].

In this paper, special attention is paid to the case when $\mathsf{X} = \mathbb{R}^d$ and π admits a smooth and everywhere positive density (also denoted by π) w.r.t. the Lebesgue measure such that the gradient $\nabla U := -\nabla\log\pi$ can be evaluated. We study below sampling methods derived from the discretization of the overdamped Langevin dynamics. It is defined by the following stochastic differential equation:

$$\mathrm{d}Y_t = -\nabla U(Y_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t\,, \tag{1.4}$$

where $(W_t)_{t\ge0}$ is the standard Brownian motion. Note that ∇U does not depend on the normalizing constant of π which is typically unknown in Bayesian inference. Under some technical conditions, the distribution of Y_t converges to π as $t\to\infty$; see [40]. The gradient-based MCMC algorithms are based on a time-discretized version of (1.4). In the Bayesian setting, a computational bottleneck of these algorithms is that the complexity of the gradient ∇U evaluation scales proportionally to the number of observations (sample size) K which can be very time consuming in the "big data" limit. To alleviate this problem, [46] proposed to replace the "full" gradient ∇U by a stochastic gradient estimate based on sums over random *minibatches*. This algorithm, stochastic gradient Langevin dynamics (SGLD), has emerged as a key MCMC algorithm in Bayesian inference for large scale datasets. The analysis of SGLD and its finite sample performance has attracted a wealth of contributions; see, for example, [30], [45], [34], [14], and the references therein. These works show that the use of stochastic

gradient comes at a price: while the resulting estimate of the gradient is still unbiased, its variance might annihilate the computational advantages of SGLD [14]. Several proposals have been made to reduce the variance of the stochastic gradient estimate of the "full" gradient, inspired by several methods, proposed for incremental stochastic optimization; see [41], [28], and [15]. [20] has investigated the properties of the Stochastic Average Gradient (SAGA) and Stochastic Variance Reduced Gradient (SVRG) estimators for Langevin dynamics. These results have been later completed and sharpened by [14], [11], [10]. Other variance reduction approaches include various subsampling schemes and constructing alternative estimates for the gradient (see, for instance, [2] and [47]).

The paper is organized as follows. In section 2, we analyze the ESVM approach for general dependent sequences. In particular, the ESVM method is described in subsection 2.1. In subsection 2.2, we study the theoretical properties of the ESVM method for asymptotically stationary dependent sequences. Here we provide a bound for the excess risk $V_\infty(f-\widehat{g}_n) - \inf_{g\in\mathcal{G}} V_\infty(f-g)$, where a control variate $\widehat{g}_n \in \mathcal{G}$ is chosen by minimization of the spectral variance V_n based on $(X_k)_{k=0}^{n-1}$, that is, $\widehat{g}_n \in \arg\min V_n(f-g)$. The precise definition of V_n will be given in subsection 2.1. In section 3, we apply these results to Markov chains which are uniformly geometrically ergodic in Wasserstein distance. While subsection 3.1 is devoted to the (unadjusted) Langevin dynamics, in subsection 3.2 we use the ESVM approach for variance reduction in SGLD-type algorithms. We show that in both cases, the excess variance can be bounded, with high probability and up to logarithmic factors, as

$$V_\infty(f-\widehat{g}_n) - \inf_{g\in\mathcal{G}} V_\infty(f-g) = O\big(n^{-1/2}\big).$$

This implies asymptotically valid confidence intervals (conditional on the sample used to learn $\widehat{g}_n$) of the form

$$\pi_N(f-\widehat{g}_n) \pm \mathsf{q}\sqrt{\frac{\inf_{g\in\mathcal{G}} V_\infty(f-g) + Cn^{-1/2}}{N}}$$

for some constant $C>0$. Note that these intervals can be much tighter than ones in (1.2), provided that n is large and $\inf_{g\in\mathcal{G}} V_\infty(f-g)$ is small. The latter condition is satisfied if the class $\mathcal{G}$ is rich enough. In section 4, we illustrate performance of the proposed variance reduction method on various benchmark problems.

Notations. Let (X,d) be a complete separable metric space. Define the Lipschitz norm of a real-valued function h by $\|h\|_{\mathsf{Lip}} := \sup_{x\neq y\in\mathsf{X}}\{|h(y)-h(x)|/\mathsf{d}(x,y)\}$. We denote by $\mathrm{Lip}_{\mathsf{d}}(L)$ and $\mathrm{Lip}_{b,\mathsf{d}}(L,B)$ the class of Lipschitz (resp., bounded Lipschitz) functions on X with $\|h\|_{\mathsf{Lip}} \le L$ (resp., $\|h\|_{\mathsf{Lip}} \le L$ and $|h|_\infty \le B$). Further, let $\mathbb{M}_1(\mathsf{X})$ be a set of probability measures on X. We denote for $p\ge 1$, $\mathbb{S}_p(\mathsf{X},\mathsf{d}) := \{\lambda\in\mathbb{M}_1(\mathsf{X}) : \int_{\mathsf{X}} \mathsf{d}^p(x,y)\lambda(\mathrm{d}y) < \infty \text{ for all } x\in\mathsf{X}\}$. For $\lambda,\nu\in\mathbb{M}_1(\mathsf{X})$, we denote their coupling set by $\Pi(\lambda,\nu)$, i.e., $\xi\in\Pi(\lambda,\nu)$ is the measure on $\mathsf{X}\times\mathsf{X}$ satisfying for all $A\in\mathcal{B}(\mathsf{X})$, $\xi(A,\mathsf{X}) = \lambda(A)$ and $\xi(\mathsf{X},A) = \nu(A)$. For $p\ge 1$ and $\lambda,\nu\in\mathbb{S}_p(\mathsf{X},\mathsf{d})$, let $W_p^{\mathsf{d}}(\lambda,\nu) := \inf_{\Pi(\lambda,\nu)}\{\int_{\mathsf{X}\times\mathsf{X}} \mathsf{d}^p(x,y)\,\xi(\mathrm{d}x,\mathrm{d}y)\}^{1/p}$ be the Wasserstein distance of order p between λ and ν. For $\lambda,\nu\in\mathbb{M}_1(\mathsf{X})$, let $\mathrm{KL}(\lambda|\nu)$ be the Kullback–Leibler divergence of λ with respect to ν, i.e., $\mathrm{KL}(\lambda|\nu) = \int \log(\mathrm{d}\lambda/\mathrm{d}\nu)\mathrm{d}\lambda$ if $\lambda\ll\nu$ and $\mathrm{KL}(\lambda|\nu)=\infty$ otherwise. Finally, unless otherwise specified, the symbol $\lesssim$ stands for an inequality up to an absolute constant not depending on parameters of the problem.

2. Empirical spectral variance minimization (ESVM).

2.1. Method. Let $(\Omega, \mathfrak{F}, (\mathfrak{F}_k)_{k\geq 0}, \mathsf{P})$ be a filtered probability space and $(X_k)_{k=0}^{\infty}$ be a random process adapted to the filtration $(\mathfrak{F}_k)_{k\geq 0}$ and taking values in X. Let $f : \mathsf{X} \to \mathbb{R}$ be a function such that $\pi(f^2) < \infty$ and $\mathsf{E}[f^2(X_k)] < \infty$ for all $k \in \mathbb{N}$. Let also $\mathcal{G}$ be a set of control variates, that is, functions $g \in \mathcal{G}$ satisfying $\pi(g^2) < \infty$, $\pi(g) = 0$, and $\mathsf{E}[g^2(X_k)] < \infty$ for all $k \in \mathbb{N}$. Particular examples of classes $\mathcal{G}$ are given below in section 3. Denote the class of functions $h = f - g$ for $g \in \mathcal{G}$ by $\mathcal{H}$,

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

To shorten notation, we write $\tilde{h} = h - \pi(h)$ for $h \in \mathcal{H}$.

We impose the following covariance stationarity condition on $(X_k)_{k=0}^{\infty}$ to ensure that the asymptotic variance $V_\infty(h)$ from (1.3) is well-defined for any $h \in \mathcal{H}$.

(CS) For any $h \in \mathcal{H}$, there exists a symmetric, summable, and positive semidefinite sequence $(\rho^{(h)}(\ell))_{\ell\in\mathbb{Z}}$ satisfying the following conditions:

(1) $\rho^{(h)}(0) = \mathrm{Var}_\pi(h)$;

(2) for any $\ell \in \mathbb{N}_0$ and constant $R > 0$ independent of h and ℓ,
$\sum_{k\in\mathbb{N}_0} |\mathsf{E}[\tilde{h}(X_k)\tilde{h}(X_{k+\ell})] - \rho^{(h)}(\ell)| \leq R$;

(3) $\lim_{\ell\to\infty} \sum_{k\in\mathbb{N}_0} |\mathsf{E}[\tilde{h}(X_k)\tilde{h}(X_{k+\ell})] - \rho^{(h)}(\ell)| = 0$.

Proposition 2.1. *Assume that the condition (CS) holds. Then, for all $h \in \mathcal{H}$, the asymptotic variance $V_\infty(h)$ defined in (1.3) exists and can be represented as*

$$V_\infty(h) = \sum_{\ell\in\mathbb{Z}} \rho^{(h)}(\ell). \tag{2.1}$$

Proof. See Appendix A.1. ∎

The spectral variance estimator $V_n(h)$ is based on truncation and weighting of the sample autocovariance functions:

$$V_n(h) := \sum_{|\ell|<b_n} w_n(\ell)\rho_n^{(h)}(\ell), \tag{2.2}$$

where w_n is the lag window, b_n is the truncation point, and $\rho_n^{(h)}(\ell)$ is the sample autocovariance function given, for $\ell \in \mathbb{N}_0$, by

$$\rho_n^{(h)}(\ell) = \rho_n^{(h)}(-\ell) := n^{-1} \sum_{k=0}^{n-\ell-1} \big(h(X_k) - \pi_n(h)\big)\big(h(X_{k+\ell}) - \pi_n(h)\big). \tag{2.3}$$

Here the truncation point b_n is an integer depending on n, and the lag window w_n is a kernel of the form $w_n(\ell) = w(\ell/b_n)$, where w is a symmetric nonnegative function supported on $[-1, 1]$ such that $\sup_{y\in[0,1]} |w(y)| \leq 1$ and $w(y) = 1$ for $y \in [-1/2, 1/2]$. There are several other estimates for the asymptotic variance $V_\infty(h)$; see [22] and the references therein. The ESVM estimator is obtained by

$$\widehat{h}_n \in \arg\min_{h\in\mathcal{H}} V_n\big(h\big). \tag{2.4}$$

The ESVM method is summarized in Algorithm 2.1.

Algorithm 2.1. ESVM method

Input: Two independent sequences: $\mathbf{X}_n = (X_k)_{k=0}^{n-1}$ and $\mathbf{X}'_N = (X'_k)_{k=0}^{N-1}$.

1. Choose a class $\mathcal{G}$ of functions with $\pi(g) = 0$ for all functions $g \in \mathcal{G}$.

2. Find $\widehat{g}_n \in \arg\min_{g\in\mathcal{G}} V_n(f-g)$, where V_n is computed based on $\mathbf{X}_n$.

Output: $\pi_N(f - \widehat{g}_n)$ computed based on $\mathbf{X}'_N$.

2.2. Theoretical analysis. For our theoretical analysis, instead of looking for a function with the smallest spectral variance in the whole class $\mathcal{H}$ we will perform optimization over a finite approximation (net) of $\mathcal{H}$. It turns out that both estimators have similar theoretical properties. Fix some $\varepsilon > 0$. Assuming that the class $\mathcal{H}$ is totally bounded, let $\mathcal{H}_\varepsilon$ be a minimal ε-net in the $L^2(\pi)$-norm, that is, the smallest possible (finite) collection of functions $\mathcal{H}_\varepsilon \subset \mathcal{H}$ with the property that for any $h \in \mathcal{H}$ there exists $h_\varepsilon \in \mathcal{H}_\varepsilon$ such that the distance between h an h_ε in $L^2(\pi)$-norm is less than or equal to ε. The cardinality of $\mathcal{H}_\varepsilon$ is called the covering number and is denoted by $|\mathcal{H}_\varepsilon|$. Define

$$\widehat{h}_{n,\varepsilon} \in \arg\min_{h\in\mathcal{H}_\varepsilon} V_n(h).$$

To obtain a quantitative bound for the asymptotic variance of $\widehat{h}_{n,\varepsilon}$, we need to specify the decay rate of the sequence $(\rho^{(h)}(\ell))_{\ell\in\mathbb{Z}}$ from (CS).

(CD) There exist $\varsigma > 0$ and $\lambda \in [0,1)$ such that, for any $h \in \mathcal{H}$ and $\ell \in \mathbb{N}_0$,

$$\left|\rho^{(h)}(\ell)\right| \le \varsigma\lambda^\ell.$$

The following theorem provides a general bound on the excess of asymptotic variance.

Theorem 2.2. *Assume that the conditions (CS) and (CD) hold. Assume additionally that for any $n \in \mathbb{N}$ there exists a decreasing continuous function α_n satisfying*

$$\sup_{h\in\mathcal{H}} \mathsf{P}\Big(\big|V_n(h) - \mathsf{E}[V_n(h)]\big| > t\Big) \le \alpha_n(t), \quad t > 0.$$

Then, for any $\delta \in (0,1)$ and $\varepsilon > 0$, it holds with probability at least $1-\delta$ that

$$V_\infty(\widehat{h}_{n,\varepsilon}) - \inf_{h\in\mathcal{H}} V_\infty(h) \lesssim \alpha_n^{-1}\left(\frac{\delta}{2|\mathcal{H}_\varepsilon|}\right) + \big(\sqrt{R}n^{-1/2} + \sqrt{D}\big)b_n\varepsilon + \sqrt{RD}\, b_n n^{-1/2} + \big(R + \varsigma(1-\lambda)^{-1}\big)b_n n^{-1} + \varsigma(1-\lambda)^{-2}n^{-1} + \varsigma(1-\lambda)^{-1}\lambda^{b_n/2},$$

where α_n^{-1} is an inverse function for α_n and $D = \sup_{h\in\mathcal{H}} \mathrm{Var}_\pi(h)$.

Proof. See Appendix A.2. ∎

Under some additional assumptions on the covering number of $\mathcal{H}$ and the function $\alpha_n(t)$, a suitable choice of the size of ε-net and the truncation point b_n yields the following high-probability bound:

$$V_\infty(\widehat{h}_{n,\varepsilon}) - \inf_{h\in\mathcal{H}} V_\infty(h) \lesssim n^{-1/(2+\rho)} \quad \text{for some } \rho > 0,$$

where $\lesssim$ stands for inequality up to a constant depending on λ, R, D, and ς. In the next section we shall apply Theorem 2.2 to the analysis of the ESVM algorithm for dependent sequences in the unadjusted Langevin algorithm (ULA) and SGLD.

3. Applications. In general, Theorem 2.2 can be applied to different types of dependent sequences satisfying conditions (CS) and (CD). In what follows, we let (X,d) be a complete separable metric space (equipped with its Borel σ-algebra $\mathcal{X}$) and consider P to be a Markov kernel on $(\mathsf{X},\mathcal{X})$. Let $\Omega = \mathsf{X}^{\mathbb{N}}$ be the set of X-valued sequences endowed with the σ-field $\mathfrak{F} = \mathcal{X}^{\mathbb{N}}$, $(X_k)_{k=0}^{\infty}$ be the coordinate process, and $\mathfrak{F}_k = \sigma(X_\ell, \ell \le k)$ be the canonical filtration. For every probability measure ξ on $(\mathsf{X},\mathcal{X})$ there exists a unique probability P_ξ on $(\mathsf{X}^{\mathbb{N}}, \mathcal{X}^{\otimes\mathbb{N}})$ such that the coordinate process $(X_k)_{k=0}^{\infty}$ is a Markov chain with Markov kernel P and initial distribution ξ. We denote by E_ξ the associated expectation. We focus below on the case where P is W_p^{d}-uniformly ergodic for $p=1$ or $p=2$.

(WE)-p There exist $x_0 \in \mathsf{X}$ such that $\int_{\mathsf{X}} \mathsf{d}(x_0,x)P(x_0,\mathrm{d}x) < \infty$ and a constant $\Delta_p(P) \in [0,1)$ such that

$$\sup_{(x,x')\in\mathsf{X}^2,\, x\neq x'} \frac{W_p^{\mathsf{d}}(\delta_x P, \delta_{x'} P)}{\mathsf{d}(x,x')} = \Delta_p(P)\,.$$

[18, Theorem 20.3.4] shows that if (WE)-p holds for some $p \ge 1$, then P admits a unique invariant probability measure which is denoted by π below. Moreover, $\pi \in \mathbb{S}_p(\mathsf{X},\mathsf{d})$ and for any $\xi \in \mathbb{S}_p(\mathsf{X},\mathsf{d})$,

$$W_p^{\mathsf{d}}(\xi P^n, \pi) \le \Delta_p^n(P) W_p^{\mathsf{d}}(\xi,\pi)\,, \quad n \in \mathbb{N}. \tag{3.1}$$

If there is no risk of confusion, we denote for simplicity $\Delta_p = \Delta_p(P)$. Let us start with a general result for Markov kernels satisfying (WE)-2. We show below that this assumption implies (CS) and (CD) when $\mathcal{H}$ is a subset of Lipschitz functions, and establish an exponential concentration inequality for $V_n(h)$, $h \in \mathcal{H}$. As was emphasized in [31] and [17], powerful tools for exploring concentration properties of W_2^{d}-ergodic Markov kernels are the transportation cost-information inequalities.

Definition 3.1. *For $p \ge 1$, we say that $\mu \in \mathbb{M}_1(\mathsf{X})$ satisfies L^p-transportation cost-information inequality with constant $\alpha > 0$ if for any $\nu \in \mathbb{M}_1(\mathsf{X})$, $W_p^{\mathsf{d}}(\mu,\nu) \le \sqrt{2\alpha\,\mathrm{KL}(\nu|\mu)}$. We write briefly $\mu \in T_p(\alpha)$ for this relation.*

L^p-transportation cost-information inequalities are well-studied in the literature; see, for instance, [4] and references therein. The cases $p=1$ and $p=2$ are of particular interest. Relations between $T_1(\alpha)$ and concentration inequalities are covered in [29] and [8]. In particular, $T_1(\alpha)$ is known to be equivalent to Gaussian concentration for all Lipschitz functions; see [8]. In turn $T_2(\alpha)$ is a stronger inequality than $T_1(\alpha)$. It was first established for the standard Gaussian measure on $\mathbb{R}^d$ by Talagrand in [44]. Moreover, the celebrated result by Bakry and Émery [3] implies that the measure $\pi(dx) = \mathrm{e}^{-U(x)}\mathrm{d}x$ satisfies $T_2(\alpha)$ if $\nabla^2 U \ge \alpha^{-1}\mathrm{I}$; see [4, Chapter 9.6]. We are especially interested in $T_2(\alpha)$, since it is known to be stable under both independent and Markovian tensorizations; see [37] and [17].

Our results on W_2^{d}-ergodic Markov kernels are summarized below.

Proposition 3.2. *Let $\mathcal{H} \subseteq \mathrm{Lip}_{\mathsf{d}}(L)$, and assume that (WE)-2 holds. Then, for any initial distribution $\xi \in \mathbb{S}_2(\mathsf{X},\mathsf{d})$, (CS) is satisfied with*

$$\rho^{(h)}(\ell) = \mathsf{E}_\pi\big[\tilde{h}(X_0)\tilde{h}(X_{|\ell|})\big], \quad R = A_1 L^2 (1-\Delta_2)^{-1} W_2(\xi,\pi), \tag{3.2}$$

where A_1 is a constant given in (A.12), and (CD) is satisfied with

$$\varsigma = L\sqrt{D}\left[\int \{W_2^{\mathsf{d}}(\delta_x,\pi)\}^2\pi(\mathrm{d}x)\right]^{1/2}, \quad \lambda = \Delta_2, \quad D = \sup_{h\in\mathcal{H}}\mathrm{Var}_\pi(h). \tag{3.3}$$

Moreover, if $P(x,\cdot)\in T_2(\alpha)$ for any $x\in\mathsf{X}$ and some $\alpha>0$, then, for any initial distribution $\xi\in T_2(\alpha)$, $n\in\mathbb{N}$, and $t>0$,

$$\mathsf{P}_\xi\big(\big|V_n(h)-\mathsf{E}_\xi[V_n(h)]\big|\ge t\big)\le 2\exp\left(-\frac{(1-\Delta_2)^2nt^2}{c\alpha L^2b_n^2\big(D+Rn^{-1}+t\big)}\right), \tag{3.4}$$

where $c>0$ is an absolute constant.

Proof. See Appendix A.3. ∎

It is also possible to remove a quite restrictive assumption $P(x,\cdot)\in T_2(\alpha)$ and to relax (WE)-2 to (WE)-1, but in this case (CS) and (CD) can be verified only for $\mathcal{H}$ being a subset of bounded Lipschitz functions. As a price for such a generalization, the exponential concentration bound is replaced by a polynomial one.

Proposition 3.3. *Let $\mathcal{H}\subset \mathrm{Lip}_{b,\mathsf{d}}(L,B)$, and assume that (WE)-1 holds. Then for any initial distribution $\xi\in\mathbb{S}_1(\mathsf{X},\mathsf{d})$, (CS) is satisfied with*

$$\rho^{(h)}(\ell)=\mathsf{E}_\pi\big[\tilde h(X_0)\tilde h(X_{|\ell|})\big], \quad R=A_2B\left(1-\Delta_1^{1/2}\right)^{-1}, \tag{3.5}$$

where A_2 is a constant given in (A.18), and (CD) is satisfied with

$$\varsigma = 2LB\int W_1^{\mathsf{d}}(\delta_x,\pi)\pi(\mathrm{d}x), \quad \lambda=\Delta_1, \quad D=\sup_{h\in\mathcal{H}}\mathrm{Var}_\pi(h). \tag{3.6}$$

Moreover, for any $p\in\mathbb{N}$,

$$\mathsf{P}_\xi\big(\big|V_n(h)-\mathsf{E}_\xi[V_n(h)]\big|\ge t\big)\le \frac{C_{\mathsf{R},1}^pB^{2p}b_n^{3p/2}p^p}{n^{p/2}t^p}+\frac{C_{\mathsf{R},2}^pB^{2p}b_n^{2p}p^{2p}}{n^{p-1}t^p}, \tag{3.7}$$

where constants $C_{\mathsf{R},1}$ and $C_{\mathsf{R},2}$ are given in (A.28).

Proof. See Appendix A.4. ∎

3.1. Langevin dynamics. In this case, $\mathsf{X}=\mathbb{R}^d$, and we assume that π has an everywhere positive density w.r.t. the Lebesgue measure, i.e., $\pi(\theta)=Z^{-1}\mathrm{e}^{-U(\theta)}$, where $Z=\int \mathrm{e}^{-U(\vartheta)}\mathrm{d}\vartheta$ is the normalization constant. Consider the first-order Euler–Maruyama discretization of the Langevin dynamics from (1.4),

$$\theta_{k+1}=\theta_k-\gamma\nabla U(\theta_k)+\sqrt{2\gamma}\,\xi_{k+1}, \tag{3.8}$$

where $\gamma>0$ is a step size and $(\xi_k)_{k=1}^\infty$ is an i.i.d. sequence of the standard Gaussian d-dimensional random vectors. The idea of using (3.8) to approximately sample from π has been advocated in [40], whose authors coined the term unadjusted Langevin algorithm (ULA). Consider the following assumption on U.

(ULA) The function U is continuously differentiable on $\mathbb{R}^d$ with gradient ∇U satisfying the following two conditions.

(1) Lipschitz gradient: there exists $L_U > 0$ such that for all $\theta, \theta' \in \mathbb{R}^d$ it holds that $\|\nabla U(\theta) - \nabla U(\theta')\| \le L_U \|\theta - \theta'\|$.

(2) Strong convexity: there exists a constant $m_U > 0$ such that for all $\theta, \theta' \in \mathbb{R}^d$ it holds that $U(\theta') \ge U(\theta) + \langle \nabla U(\theta), \theta' - \theta \rangle + (m_U/2)\|\theta' - \theta\|^2$.

The ULA has been widely studied under the above assumptions; see, for example, [21] and [13]. As known from [21], under (ULA) the associated Markov kernel, denoted by $P_\gamma^{(\mathsf{ULA})}$, is W_2^{d}-uniformly ergodic. For completeness, we state below the following proposition [21, Proposition 3].

Proposition 3.4. *Assume (ULA), and set $\kappa = 2m_U L_U/(m_U + L_U)$. Then for any step size $\gamma \in (0, 2/(m_U + L_U))$, $P_\gamma^{(\mathsf{ULA})}$ satisfies (WE)-2 with $\mathsf{d}(\vartheta, \vartheta') = \|\vartheta - \vartheta'\|$ and $\Delta_2 = \sqrt{1 - \kappa\gamma}$. Moreover, $P_\gamma^{(\mathsf{ULA})}$ has a unique invariant measure $\pi_\gamma^{(\mathsf{ULA})}$.*

It is shown in [21, Corollary 7] that, for any step size $\gamma \in (0, 2/(m_U + L_U))$,

$$W_2^{\mathsf{d}}\big(\pi, \pi_\gamma^{(\mathsf{ULA})}\big) \le \sqrt{2}\kappa^{-1/2} L_U \gamma^{1/2} \left\{\kappa^{-1} + \gamma\right\}^{1/2} \left\{2d + dL_U^2\gamma/m_U + dL_U^2\gamma^2/6\right\}^{1/2}.$$

We define the asymptotic variance as

$$V_\infty^{(\mathsf{ULA})}(h) := \sum_{\ell \in \mathbb{Z}} \mathsf{E}_{\pi_\gamma^{(\mathsf{ULA})}}\Big[\big(h(X_0) - \pi_\gamma^{(\mathsf{ULA})}(f)\big)\big(h(X_{|\ell|}) - \pi_\gamma^{(\mathsf{ULA})}(f)\big)\Big].$$

At each iteration of the algorithm, ∇U is computed. Hence it is an appealing option to use this gradient to construct Stein control variates (see, for instance, [1], [33], and [36]), given by

$$g_\phi(\theta) = -\langle \phi(\theta), \nabla U(\theta) \rangle + \operatorname{div}\big(\phi(\theta)\big), \tag{3.9}$$

where $\phi : \mathsf{X} \to \mathbb{R}^d$ is a continuously differentiable Lipschitz function, $\langle \cdot, \cdot \rangle$ is the standard scalar product in $\mathbb{R}^d$, and $\operatorname{div}(\phi)$ is the divergence of ϕ. Under rather mild conditions on π and ϕ, it follows from the integration by parts that $\pi(g_\phi) = 0$ (see [33, Propositions 1 and 2]). Note that if $\phi(\theta) \equiv b$, $b \in \mathbb{R}^d$, we get $g_b(\theta) = -\langle b, \nabla U(\theta) \rangle$. Then for a parametric class $\mathcal{H} = \{f - g_b : \|b\| \le B\}$, assuming that $f \in \mathrm{Lip}_{\mathsf{d}}(L_1)$ and that condition (ULA) holds, we get $\mathcal{H} \subset \mathrm{Lip}_{\mathsf{d}}(\max(L_1, BL_U))$. For other approaches to construct control variates we refer the reader to [27], [16], and [9]. The next result follows now from Theorem 2.2 and Proposition 3.2.

Theorem 3.5. *Let $\mathcal{H} \subset \mathrm{Lip}_{\mathsf{d}}(L)$, and assume that (ULA) holds. Assume additionally that $\xi \in T_2(\beta)$ for some $\beta > 0$. Fix any $\gamma \in (0, 2/(m_U + L_U))$, and set $b_n = 2\lceil \log(n)/\log(1/\Delta_2) \rceil$ with $\Delta_2 = \sqrt{1 - \kappa\gamma}$ and $\kappa = 2m_U L_U/(m_U + L_U)$. Then, for any $\varepsilon > 0$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$,*

$$V_\infty^{(\mathsf{ULA})}(\widehat{h}_{n,\varepsilon}) - \inf_{h\in\mathcal{H}} V_\infty^{(\mathsf{ULA})}(h)$$
$$\lesssim C_1\,\varepsilon\log(n) + C_2\sqrt{\frac{\log^2(n)\log(|\mathcal{H}_\varepsilon|/\delta)}{n}} + C_3\,\frac{\log^2(n)\log\left(|\mathcal{H}_\varepsilon|/\delta\right)}{n},$$

where

$$C_1 = \frac{\sqrt{R}+\sqrt{D}}{\kappa\gamma},\ C_2 = \frac{L\sqrt{(\beta\vee\gamma)(D+R)}}{\kappa^2\gamma^2} + \frac{\sqrt{DR}}{\kappa\gamma},\ C_3 = \frac{L^2(\beta\vee\gamma)}{\kappa^4\gamma^4} + \frac{R}{\kappa\gamma} + \frac{\varsigma}{\kappa^2\gamma^2}$$

with R, ς *from Proposition 3.2 and* $D = \sup_{h\in\mathcal{H}} \mathrm{Var}_{\pi_\gamma^{(\mathsf{ULA})}}(h)$.

Proof. The Markov kernel associated to ULA can be written as $P_\gamma^{(\mathsf{ULA})}(\theta,\cdot) = \mathcal{N}(\theta - \gamma\nabla U(\theta), 2\gamma I_d)$. Hence, by [4, Theorem 9.2.1], $P_\gamma^{(\mathsf{ULA})}(\theta,\cdot) \in \mathrm{T}_2(2\gamma)$ for any $\gamma > 0$. By Proposition 3.4, (WE) holds with $\Delta_2 = \sqrt{1-\gamma\kappa}$. Hence Proposition 3.2 applies with $\alpha = 2(\beta\vee\gamma)$. Direct computation of the inverse function in the right-hand side of (3.4) leads to

$$\alpha_n^{-1}\left(\frac{\delta}{2|\mathcal{H}_\varepsilon|}\right) \le \frac{4b_n^2L^2(\beta\vee\gamma)\log\left(4|\mathcal{H}_\varepsilon|/\delta\right)}{(1-\Delta_2)^2 n} + \frac{4b_nL\sqrt{(\beta\vee\gamma)(D+R)\log\left(4|\mathcal{H}_\varepsilon|/\delta\right)}}{(1-\Delta_2)\sqrt{n}}.$$

■

Corollary 3.6. *Under the assumptions of Theorem 3.5, the following holds.*

(1) *If class* $\mathcal{H}$ *is parametric, that is,* $|\mathcal{H}_\varepsilon| \le C_\rho\varepsilon^{-\rho}$ *for all* $\varepsilon\in(0,1)$ *and some constants* $C_\rho,\rho>0$, *then it holds for any* $\varepsilon\in(0,1/\sqrt{n})$ *with probability at least* $1-1/n$,

$$V_\infty^{(\mathsf{ULA})}(\widehat{h}_{n,\varepsilon}) - \inf_{h\in\mathcal{H}} V_\infty^{(\mathsf{ULA})}(h) \lesssim n^{-1/2}\log^{1/2}(n).$$

(2) *If class* $\mathcal{H}$ *is nonparametric, that is,* $|\mathcal{H}_\varepsilon| \le C_\rho\exp(\varepsilon^{-\rho})$ *for all* $\varepsilon\in(0,1)$ *and some constants* $C_\rho,\rho>0$, *then it holds for any* $\varepsilon\in(0,1/\sqrt{n})$ *with probability at least* $1-1/n$,

$$V_\infty^{(\mathsf{ULA})}(\widehat{h}_{n,\varepsilon}) - \inf_{h\in\mathcal{H}} V_\infty^{(\mathsf{ULA})}(h) \lesssim n^{-1/(2+\rho)}.$$

Here $\lesssim$ *stands for inequality up to a constant depending on* ρ *and other constants from Theorem 3.5. Moreover, if additionally the constant* $\pi_\gamma^{(\mathsf{ULA})}(f)$ *is in the class* $\mathcal{H}$, *then* $\inf_{h\in\mathcal{H}} V_\infty^{(\mathsf{ULA})}(h) = 0$ *and these bounds hold for the asymptotic variance itself.*

Discussion. It is well-known that if $\hat{f}$ satisfies the so-called Poisson equation $P_\gamma^{(\mathsf{ULA})}\hat{f} - \hat{f} = -f + \pi_\gamma^{(\mathsf{ULA})}(f)$, then by taking $g^\star = \hat{f} - P_\gamma^{(\mathsf{ULA})}\hat{f}$ as a control variate, we get $\pi_\gamma^{(\mathsf{ULA})}(f-g^\star) = \pi_\gamma^{(\mathsf{ULA})}(f)$ and $V_\infty^{(\mathsf{ULA})}(f-g^\star) = 0$. The property $h^\star = f - g^\star = \pi_\gamma^{(\mathsf{ULA})}(f) \in \mathcal{H}$ can be achieved by taking, for example, $\mathcal{H}$ to be a ball in a Sobolev space. Namely, let $W_2^s = \left\{h \in L^2(\lambda) : D^\alpha h \in L^2(\lambda),\ \forall|\alpha| \leqslant s\right\}$ be the Sobolev space; here λ is the Lebesgue measure on $\mathbb{R}^d$, $\alpha = (\alpha_1,\dots,\alpha_d)$ is a multi-index with $|\alpha| = \alpha_1 + \cdots + \alpha_d$, and D^α stands for the differential operator $D^\alpha = \partial^{|\alpha|}/\partial x_1^{\alpha_1}\dots\partial x_d^{\alpha_d}$. The weighted Sobolev space $W_2^s(\langle x\rangle^\beta)$, $\beta\in\mathbb{R}$, for a polynomial weighting function $\langle x\rangle^\beta = (1+\|x\|^2)^{\beta/2}$ is defined by $W_2^s(\langle x\rangle^\beta) = \{h : h\cdot\langle x\rangle^\beta \in W_2^s\}$. Let $\mathcal{H}$ be a norm-bounded subset of $W_2^s(\langle x\rangle^\beta)$ with $\beta\in\mathbb{R}$ and $s - d/2 > 0$. Suppose also that $\|\langle x\rangle^{\alpha-\beta}\|_{L^2(\pi_\gamma^{(\mathsf{ULA})})} < \infty$ for some $\alpha > 0$. Then $|\mathcal{H}_\varepsilon| \lesssim \exp(\varepsilon^{-d/s})$, provided

that $\alpha > s - d/2$; see [35, Corollary 4]. Note that $h^\star = \pi_\gamma^{(\mathsf{ULA})}(f) \in W_2^s(\langle x\rangle^\beta)$ for any $s > 0$ and any $\beta < -1$ so that we can take $\mathcal{H}$ as a norm-bounded subset of $W_2^s(\langle x\rangle^\beta)$ for arbitrary large $s > 0$. Since $\pi_\gamma^{(\mathsf{ULA})}$ and all its derivatives have exponentially decaying tails (see [32]), $\|\langle x\rangle^{\alpha-\beta}\|_{L^2(\pi_\gamma^{(\mathsf{ULA})})} < \infty$ for any $\alpha > 0$, and one can achieve that $|\mathcal{H}_\varepsilon| \lesssim \exp(\varepsilon^{-\delta})$ for arbitrary small $\delta > 0$ and at the same time $h^\star \in W_2^s(\langle x\rangle^\beta)$. Practically one can use Stein control variates of the form (3.9) with infinitely smooth and compactly supported functions ϕ. This will guarantee that $f - g_\phi \in W_2^s(\langle x\rangle^\beta)$ for some $s > 0$, provided that U is smooth enough and $f \in W_2^s(\langle x\rangle^\beta)$.

3.2. Extension to the SGLD. In this section, we shall consider the situations where the target π is given by the posterior distribution in the Bayesian inference problem, that is, $\pi(\theta) \propto \exp(-U(\theta))$, where $U(\theta) = U_0(\theta) + \sum_{i=1}^K U_i(\theta)$ with K being a number of observations. Computing $\nabla U(\theta)$ requires a computational budget that scales linearly with K. Hence it is often impossible to apply procedures based on discretization of Langevin dynamics directly. One possible solution advocated by [46] is to replace $\nabla U(\theta)$ by an unbiased estimate. This gives rise to the SGLD algorithm, where the parameters are updated according to

$$\begin{aligned} \theta_{k+1} &= \theta_k - \gamma G(\theta_k, S_{k+1}) + \sqrt{2\gamma}\,\xi_{k+1}, \\ G(\theta, S) &= \nabla U_0(\theta) + KM^{-1} \sum\nolimits_{i\in S} \nabla U_i(\theta), \end{aligned} \tag{3.10}$$

where each S_{k+1} is a random batch taking values in S_M (here S_M is the set of all subsets S of $\{1,\dots,K\}$ with $|S| = M$) which is sampled from a uniform distribution over S_M independently of $\mathcal{F}_k$ (here $(\mathcal{F}_k)_{k\ge 0}$ is the filtration generated by $\{(\theta_\ell, S_\ell)\}_{\ell\ge 0}$). Note that $\mathsf{E}[G(\theta_k, S_{k+1})|\mathcal{F}_k] = \nabla U(\theta_k)$ and therefore $G(\theta_k, S_{k+1})$ is an unbiased estimate of $\nabla U(\theta_k)$. The available variance reduction techniques for SGLD usually replace the stochastic gradient in (3.10) with more sophisticated estimates which preserve unbiasedness but have lower variance.

The simplest variance reduction technique is the fixed-point method (SGLD-FP) proposed in [2]. This method is applicable when the posterior distribution is strongly log-concave. We set $\hat\theta \in \Theta$ to be a fixed value of the parameter, typically chosen to be close to the mode of posterior distribution. We estimate the gradient $\nabla U(\theta)$ by

$$G_{\mathrm{FP}}(\theta, S) = \nabla U_0(\theta) + KM^{-1} \sum\nolimits_{i\in S} \big(\nabla U_i(\theta) - \nabla U_i(\hat\theta)\big) + \sum_{i=1}^K \nabla U_i(\hat\theta). \tag{3.11}$$

The SGLD-FP algorithm is obtained by plugging this approximation into (3.10).

More sophisticated variance reduction methods typically use reference values $(g_k^i)_{i=1}^K$ of the gradient $(\nabla U_i)_{i=1}^K$ from previous iterates (and not only the last iterate); as a result, constructed sequence $(\theta_k)_{k=0}^\infty$ is often not Markovian. One particular example is SAGA-LD method, adapted from [41], [15]. If $i \in S_k$, the reference value is updated, that is, $g_{k+1}^i = \nabla U_i(\theta_k)$. Otherwise, the reference value is simply propagated, that is, $g_{k+1}^i = g_k^i$. One then considers the following gradient estimator:

$$G_{\mathrm{SAGA}}^k(\theta, S) = \nabla U_0(\theta) + KM^{-1} \sum\nolimits_{i\in S} \big(\nabla U_i(\theta) - g_k^i\big) + g_k\,, \quad g_k = \sum_{i=1}^K g_k^i\,. \tag{3.12}$$

The recursion is initialized with $g_0^i = \nabla U_i(\theta_0)$, $i \in \{1, \dots, K\}$, and $g_0 = \sum_{i=1}^K g_0^i$. Finally, the gradient is computed according to (3.12) and plugged into (3.10).

For theoretical analysis of SGLD and SGLD-FP algorithms we need the following assumptions on U. Without loss of generality, we consider only SGLD; the same reasoning applies to SGLD-FP.

(SGLD) The function $U(\theta) = U_0(\theta) + \sum_{i=1}^K U_i(\theta)$ satisfies the following conditions.

(1) Lipschitz gradient: for any $i \in \{0, \dots, K\}$, U_i is continuously differentiable on $\mathbb{R}^d$ with $\widetilde{L}_U$-Lipschitz gradient.
(2) Convexity: for any $i \in \{0, \dots, K\}$, U_i is convex.
(3) Strong convexity: there exists a constant $m_U > 0$ such that for any $\theta, \theta' \in \mathbb{R}^d$ it holds that $U(\theta') \geq U(\theta) + \langle \nabla U(\theta), \theta' - \theta \rangle + (m_U/2)\|\theta' - \theta\|^2$.

Note that using Stein control variates with SGLD-based sampling procedure (3.10) eliminates benefits of using $G(\theta, S)$ instead of exact gradient $\nabla U(\theta)$. Following [23], we replace ∇U by its stochastic counterpart. More precisely, for the kth iteration of SGLD algorithm, we consider the control variates of the form

$$g_\phi(\theta, S) = -\langle \phi(\theta), G(\theta, S) \rangle + \operatorname{div}\big(\phi(\theta)\big). \tag{3.13}$$

The control variate g_ϕ depends now on the pair (θ, S). Let $\mathcal{H} = \{f(\theta) - g_\phi(x) : \phi \in \Phi\}$, where $x = (\theta, S) \in \mathsf{X} = \Theta \times \mathsf{S}_M$. Consider another sequence $\big(\tilde{S}_k\big)_{k=0}^\infty$ of independent batches uniformly distributed over S_M such that for any k, $\tilde{S}_k$ is independent of $\mathcal{F}_k$. Denote by P_{SGLD} the transition kernel of SGLD, and let Υ_M be a uniform distribution over S_M. Set $\overline{P} := P_{\mathsf{SGLD}} \otimes \Upsilon_M$ and $X_k = (\theta_k, \tilde{S}_k)$.

Proposition 3.7. *Assume* (SGLD). *Then for any step size* $\gamma \in \big(0, \widetilde{L}_U^{-1}(K+1)^{-1}\big)$, $\overline{P}$ *satisfies* (WE)-2 *with* $\Delta_2 = \sqrt{1 - \gamma m_U}$ *and* $\mathsf{d}(x, x') = \|\vartheta - \vartheta'\| + \mathbb{1}_{\{S \neq S'\}}$ *for any* $x = (\vartheta, S)$ *and* $x' = (\vartheta', S')$. *Moreover,* $\overline{P}$ *has a unique invariant measure* $\overline{\pi} = \pi_\gamma^{(\mathsf{SGLD})} \otimes \Upsilon_M$.

Proof. See Appendix A.5. ∎

Similarly to Langevin dynamics, we define

$$V_\infty^{(\mathsf{SGLD})}(h) := \sum\nolimits_{\ell \in \mathbb{Z}} \mathsf{E}_{\overline{\pi}}\Big[\big(h(X_0) - \overline{\pi}(f)\big)\big(h(X_{|\ell|}) - \overline{\pi}(f)\big)\Big].$$

Theorem 3.8. *Let* $\mathcal{H} \subseteq \mathrm{Lip}_{b,\mathsf{d}}(L, B)$, *and assume that* (SGLD) *holds. Fix any* $\gamma \in \big(0, \widetilde{L}_U^{-1}(K+1)^{-1}\big)$, *and set* $b_n = 2\lceil \log(n)/\log(1/\Delta_1) \rceil$ *with* $\Delta_1 = \sqrt{1 - \gamma m_U}$. *Then, for any* $\varepsilon > 0$ *and* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$,

$$\begin{aligned} &V_\infty^{(\mathsf{SGLD})}(\widehat{h}_{n,\varepsilon}) - \inf\nolimits_{h \in \mathcal{H}} V_\infty^{(\mathsf{SGLD})}(h) \\ &\quad \lesssim C_4\, \varepsilon \log(n) + C_5 \sqrt{\frac{\log^5(n)}{n}} \left(\frac{|\mathcal{H}_\varepsilon|}{\delta}\right)^{1/\log(n)} + C_6\, \frac{\log n}{n}, \end{aligned}$$

where

$$C_4 = \frac{\sqrt{R} + \sqrt{D}}{m_U \gamma},\ C_5 = \frac{B^2 R_1(L, \xi)}{(m_U\gamma)^2} + \frac{B^2 R_2(L, \xi)}{(m_U\gamma)^{4 + 2/\log n}} + \frac{\sqrt{RD}}{m_U\gamma},\ C_6 = \frac{D(m_U\gamma) + \varsigma}{(m_U\gamma)^2}$$

with R, ς from Proposition 3.3, $D = \sup_{h\in\mathcal{H}} \mathrm{Var}_{\pi_\gamma^{(\mathsf{SGLD})}}(h)$, and constants $R_1(L,\xi)$, $R_2(L,\xi)$ which can be tracked from (A.27).

Proof. By Proposition 3.7, (WE)-2 holds with $\Delta_2 = \sqrt{1-\gamma m_U}$, and, by Lyapunov inequality, (WE)-1 also holds with $\Delta_1 = \Delta_2$. Hence, the second part of Proposition 3.3 can be applied with $p = \log n$. The remaining part follows from Theorem 2.2 with computation of the inverse function in the right-hand side of (3.7). ∎

Corollary 3.9. *Under the assumptions of Theorem 3.8, if class $\mathcal{H}$ is parametric, that is, $|\mathcal{H}_\varepsilon| \le C_\rho \varepsilon^{-\rho}$ for all $\varepsilon \in (0,1)$ and some constants $C_\rho, \rho > 0$. Then for any $\varepsilon \in (0, 1/\sqrt{n})$ it holds with probability at least $1 - 1/n$ that*

$$V_\infty^{(\mathsf{SGLD})}(\widehat{h}_{n,\varepsilon}) - \inf_{h\in\mathcal{H}} V_\infty^{(\mathsf{SGLD})}(h) \lesssim n^{-1/2}\log^{5/2}(n),$$

where $\lesssim$ stands for inequality up to a constant depending on ρ and other constants from Theorem 3.8. Moreover, if additionally $\overline{\pi}(f) \in \mathcal{H}$, then $\inf_{h\in\mathcal{H}} V_\infty^{(\mathsf{SGLD})}(h) = 0$, and these bounds hold for the asymptotic variance itself.

Remark 3.10. If the class $\mathcal{H}$ is constructed using Stein control variates, we can ensure the inclusion $\mathcal{H} \subseteq \mathrm{Lip}_{b,\mathsf{d}}(L,B)$ by taking smooth and compactly supported functions ϕ. This in turn can be achieved by multiplying a given smooth function ϕ with a mollifier function, that is, an infinitely smooth compactly supported function.

4. Experiments.

In this section, we numerically compare the following two methods to choose control variates: the EVM method, where a control variate is determined by minimizing the marginal variance (see (1.1)), and the ESVM method, where a control variate is determined by minimizing the spectral variance (see (2.2)). Implementation is available at https://github.com/svsamsonov/vr_sg_mcmc.

4.1. Toy example.

We first consider a multimodal distribution in $\mathbb{R}^2$ from [38]. Namely, let $\pi(x_1,x_2) = Z^{-1}\mathrm{e}^{-U(x_1,x_2)}$, where Z is the normalization constant and

$$U(x_1,x_2) = \frac{(\|x\|-\mu)^2}{2M^2} - \log\Big(\mathrm{e}^{-(x_1-\mu)^2/2\sigma^2} + \mathrm{e}^{-(x_1+\mu)^2/2\sigma^2}\Big).$$

We choose $M = 1$ and $\mu = \sigma = 3$; the respective density profile is presented in Figure 1. Our aim is to estimate $\pi(f)$ with $f(x_1,x_2) = x_1 + x_2$ using ULA. The parametric class g_φ in (3.9) is generated by $\varphi(x) = \sum_{k=1}^p \beta_k \psi_k(x)$, where $\psi_k = e^{-\|x-\mu_k\|^2/2\sigma_\psi^2}$ with all μ_k regularly spaced in $[-3,3]\times[-3,3]$ and $\sigma_\psi = 2$. Details on the step size γ of the ULA, length of the burn-in period and test trajectories are summarized in Table 1. Boxplots displaying variation of 100 estimates for EVM and ESVM are presented in the same Figure 1. Furthermore, we compute sample autocovariance functions for a trajectory with and without adding ESVM and EVM control variates. The results reflect a spectacular decrease in high-order autocovariance for ESVM; see Figure 1. Note that EVM aims at minimizing only the lag-zero autocovariance; that is why the autocovariance function for ESVM-adjusted trajectory decreases much faster.

4.2. Gaussian mixture model.

We consider posterior mean estimation for unknown parameter μ in a Bayesian setup with normal prior $\mu \sim \mathcal{N}(0,\sigma_\mu^2)$, $\sigma_\mu^2 = 100$, and sample $(X_k)_{k=0}^{K-1}$, $K = 100$, drawn from the Gaussian mixture model

Table 1
Experimental hyperparameters.

Experiment	n_{burn}	n_{test}	γ	Batch size
Toy example, subsection 4.1	10^3	10^4	0.1	-
Gaussian Mixture, subsection 4.2	10^4	10^5	0.01	10

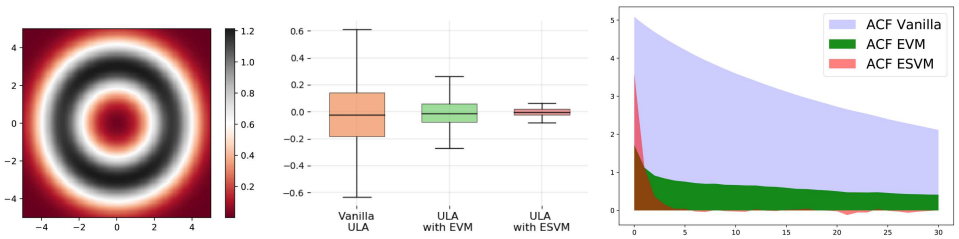

Figure 1. *Toy example from subsection 4.1. From left to right:* (1) *density profile,* (2) *boxplots displaying variation of* 100 *estimates for vanilla ULA, ULA with EVM, and ULA with ESVM,* (3) *sample autocovariance functions (ACFs) for a trajectory with and without ESVM and EVM.*

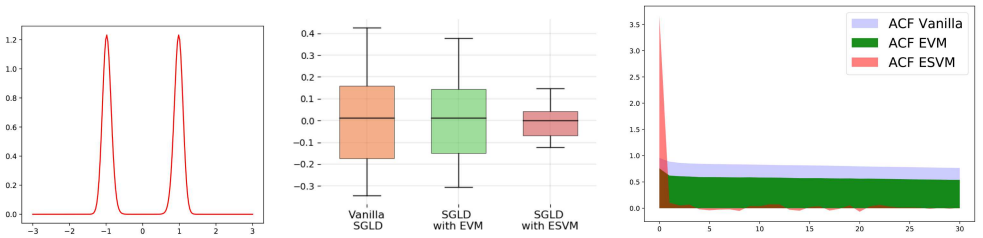

Figure 2. *Gaussian mixture model from subsection 4.2. From left to right:* (1) *density of the posterior distribution,* (2) *boxplots displaying variation of* 100 *estimates for vanilla SGLD, SGLD with EVM, and SGLD with ESVM,* (3) *sample autocovariance functions (ACFs) for a trajectory with and without ESVM and EVM.*

$$0.5\,\mathcal{N}(-\mu,\sigma^2)+0.5\,\mathcal{N}(\mu,\sigma^2) \quad \text{with } \mu=1,\ \sigma^2=1.$$

The density of the posterior distribution over μ is given in Figure 2. It has 2 modes roughly corresponding to $\mu=1$ and $\mu=-1$. To generate data from this posterior distribution and estimate posterior mean, we use SGLD. The parametric class g_φ in (3.13) is generated by $\varphi(x)=\beta_0x^2+\beta_1x+\beta_2$. Boxplots displaying variation of 100 estimates for EVM and ESVM and respective sample autocovariance functions are also presented in Figure 2. Note that the increase in lag-zero autocovariance for ESVM is explained by the additional randomness in (3.13). On contrary, EVM favors far too small coefficients to overcome this additional randomness, which leads to poor variance reduction.

4.3. Bayesian logistic regression. The probability of the ith output $y_i\in\{-1,1\}$, $i=1,\dots,K$, is given by $\mathrm{p}(y_i|\mathbf{x}_i,\theta)=(1+\mathrm{e}^{-y_i\langle\theta,\mathbf{x}_i\rangle})^{-1}$, where $\mathbf{x}_i$ is a $d\times1$ vector of predictors and θ is the vector of unknown regression coefficients. We complete the Bayesian model by considering the Zellner g-prior $\mathcal{N}_d(0,g(\mathbf{X}^\top\mathbf{X})^{-1})$ for θ where $\mathbf{X}=[\mathbf{x}_1,\dots,\mathbf{x}_N]$ is an $K\times d$ design matrix; see [26, section 2]. Normalizing the covariates, for $\tilde{\mathbf{x}}_i=(\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{x}_i$ and $\tilde\theta=(\mathbf{X}^\top\mathbf{X})^{1/2}\theta$, we get $\langle\theta,\mathbf{x}_i\rangle=\langle\tilde\theta,\tilde{\mathbf{x}}_i\rangle$, under the Zellner g-prior, $\tilde\theta\sim\mathcal{N}_d(0,g\mathbf{I}_d)$. In our

Table 2
Experimental hyperparameters.

Experiment	n_{burn}	n_{train}	n_{test}	γ	Batch size
Logistic regression, EEG dataset	10^4	10^4	10^5	0.1	15
Logistic regression, SUSY dataset	10^5	10^5	10^6	0.1	50

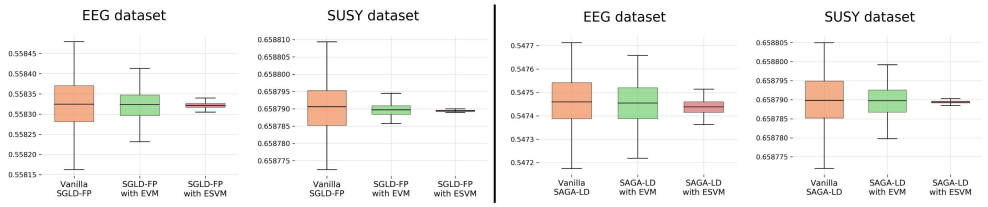

Figure 3. *Bayesian logistic regression for EEG and SUSY datasets from subsection 4.3. Boxplots displaying variation of* 100 *estimates of average predictive distribution for* (1) *left panel: vanilla SGLD-FP, SGLD-FP with EVM, and SGLD-FP with ESVM,* (2) *right panel: vanilla SAGA-LD, SAGA-LD with EVM, and SAGA-LD with ESVM.*

experiments we have not noticed significant impact of particular g value on EVM and ESVM performance and used $g = 100$ as a default choice.

We analyze the performance of EVM and ESVM methods on two datasets from the UCI repository. The first dataset, EEG,[1] contains $K = 14\,980$ observations in dimension $d = 15$; the second dataset, SUSY,[2] has $K = 500\,000$ observations in dimension $d = 19$. The data is first split into a training set $\mathcal{T}_N^{\text{train}} = \{(y_i, \mathbf{x}_i)\}_{i=1}^K$ and a test set $\mathcal{T}_K^{\text{test}} = \{(y_i', \mathbf{x}_i')\}_{i=1}^K$ by randomly picking $K = 100$ test points from the data. We use the SGLD-FP and SAGA-LD algorithms to approximately sample from the posterior distribution $\mathrm{p}(\tilde{\theta}|\mathcal{T}_N^{\text{train}})$. Given a sample $(\tilde{\theta}_k)_{k=0}^{n-1}$, we can estimate the predictive distribution for a fixed test point $(y', \mathbf{x}')$, that is, $\mathrm{p}(y'|\mathbf{x}') = \int_{\mathbb{R}^d} \mathrm{p}(y'|\mathbf{x}', \tilde{\theta})\, \mathrm{p}(\tilde{\theta}|\mathcal{T}_N^{\text{train}})\, \mathrm{d}\tilde{\theta}$, by computing the ergodic mean $n^{-1} \sum_{k=0}^{n-1} f(\tilde{\theta}_k)$ for $f(\tilde{\theta}) = \mathrm{p}(y'|\mathbf{x}', \tilde{\theta})$. To get rid of randomness caused by the random choice of a test point, we estimate the average predictive distribution for the whole test set $\mathcal{T}_K^{\text{test}}$ by computing the ergodic mean for the function $f(\tilde{\theta}) = K^{-1} \sum_{i=1}^K \mathrm{p}(y_i'|\mathbf{x}_i', \tilde{\theta})$. Details on the step size γ, length of the burn-in period and test trajectories, and batch size are summarized in Table 2. Boxplots for the estimation of average predictive distribution are shown in Figure 3. Note that ESVM leads to a significant variance reduction for both SGLD-FP and SAGA-LD.

Further, for the EEG dataset we plot in Figure 4 a part of the trajectory $f(\tilde{\theta}_m) = K^{-1} \sum_{i=1}^K \mathrm{p}(y_i'|\mathbf{x}_i', \tilde{\theta}_m)$ for 500 consecutive sample values $\tilde{\theta}_m$ with and without adding the ESVM control variate. These trajectories are accompanied by the sample autocovariance functions for vanilla and variance-reduced samples for both EVM and ESVM. Again, since EVM aims at minimizing only lag-zero autocovariance, the decrease in autocovariance function for this method is smaller than for ESVM. We also report in Figure 5 how autocovariance functions change with batch sizes. Note that for small batch sizes ESVM still manages to remove correlations, while EVM almost fails.

[1] https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State.
[2] https://archive.ics.uci.edu/ml/datasets/SUSY.

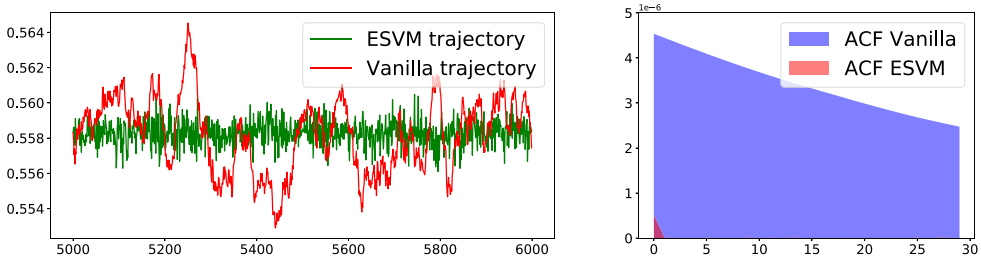

Figure 4. *Bayesian logistic regression for the EEG dataset from subsection 4.3 with the batch size* 150. *From left to right:* (1) *part of a trajectory with and without ESVM,* (2) *sample autocovariance functions (ACFs) for a trajectory with and without ESVM.*

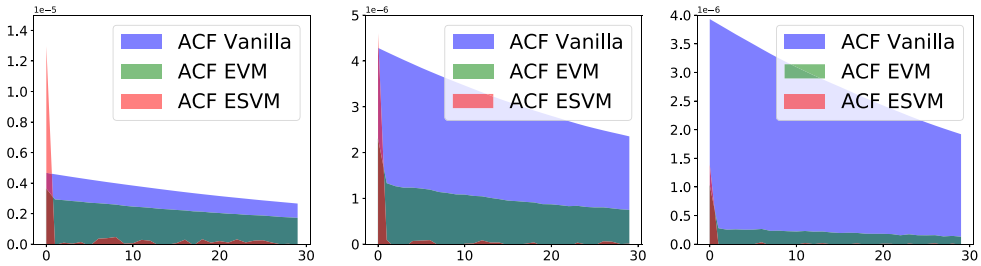

Figure 5. *Bayesian logistic regression for the EEG dataset from subsection 4.3. Comparison of sample autocovariance functions (ACFs) for different batch sizes. From left to right: batch size* 5, 10, 50, *respectively.*

4.4. Bayesian probabilistic matrix factorization. A typical problem in recommendation systems is to predict a user's rating for a particular item given other users' ratings of this item and how a given user evaluated other items. A common approach to this problem is probabilistic matrix factorization via Bayesian inference; see [43]. Namely, we are interested in approximating matrix $R \in \mathbb{R}^{M\times N}$, where M is a number of users, N is a number of rated items, and $R_{i,j}$ stands for the rating assigned by the ith user to the jth item. Due to natural limitations (user is unlikely to rate all possible items), we observe only a some small subset of elements of R and want to predict ratings of the hidden part. In probabilistic matrix factorization, we aim at representing R as a product $R = U^\top V + C$, where $U \in \mathbb{R}^{D\times M}$, $V \in \mathbb{R}^{D\times N}$, and $C \in \mathbb{R}^{M\times N}$ is a matrix of biases with elements $C_{i,j} = a_i + b_j$, $a \in \mathbb{R}^M$, $b \in \mathbb{R}^N$. In the subsequent experiments we assume that rank parameter $D = 10$ is fixed. The naive solution would be to find

$$U, V, a, b = \arg\min_{U,V,a,b} \sum_{(i,j)\in I_{\text{train}}} \big(R_{i,j} - \langle U_i, V_j\rangle - a_i - b_j\big)^2,$$

where I_{train} is a train subset of ratings. Unfortunately, optimizing this criterion leads to a significantly overfitted model. One possible approach to overcome overfitting is to consider the penalized model

$$U, V, a, b = \arg\min_{U,V,a,b} \sum_{(i,j)\in I_{\text{train}}} \big(R_{i,j} - \langle U_i, V_j\rangle - a_i - b_j\big)^2 + \lambda_U\|U\|^2 + \lambda_V\|V\|^2 + \lambda_a\|a\|^2 + \lambda_b\|b\|^2,$$

but it requires careful tuning of penalization coefficients $\lambda_U, \lambda_V, \lambda_a, \lambda_b$. We thus would benefit a lot from a Bayesian approach for tuning weights; this was pointed out in [43]. We follow a slightly simplified formulation proposed by [12]; that is, we consider

$$\lambda_U, \lambda_V, \lambda_a, \lambda_b \sim \Gamma(1,1), \quad U_{k,i} \sim \mathcal{N}\big(0, \lambda_U^{-1}\big), \quad V_{k,j} \sim \mathcal{N}\big(0, \lambda_V^{-1}\big),$$
$$a_i \sim \mathcal{N}\big(0, \lambda_a^{-1}\big), \quad b_i \sim \mathcal{N}\big(0, \lambda_b^{-1}\big), \quad R_{i,j}|U,V \sim \mathcal{N}\big(\langle U_i, V_j\rangle + a_i + b_j, \tau^{-1}\big).$$

In order to sample from the posterior distribution which we denote by $p(\Theta|R)$, where $\Theta = \{U, V, a, b, \lambda_U, \lambda_V, \lambda_a, \lambda_b\}$, we use the following two-step procedure:

1. Sample from $p(U, V, a, b|R, \lambda_U, \lambda_V, \lambda_a, \lambda_b)$ using SGLD or SGLD-FP with a minibatch size of 5000 observations with a step size $\gamma = 10^{-4}$. Sample for 1000 steps before updating the weights $\lambda_U, \lambda_V, \lambda_a, \lambda_b$;
2. Sample new λ from $p(\lambda_U, \lambda_V, \lambda_a, \lambda_b|U, V, a, b)$ using the Gibbs sampler.

The experiments are performed on the MovieLens dataset $ml - 100k$ (link to dataset). We apply our control variates procedure as a postprocessing step following [2]. The functional of interest is the mean squared error over the test subsample, $f(U, V, a, b) = \sum_{(i,j)\in I_{\text{test}}} (R_{i,j} - \langle U_i, V_j\rangle - a_i - b_j)^2$. Since the dimension of parameter space is very high, first-order control variates are the only option among Stein's control variates. Parts of SGLD- and SGLD-FP-based trajectories before and after using control variates, and confidence intervals for estimation of f, are presented in Figure 6.

Appendix A. Supplementary material for variance reduction for dependent sequences with applications to stochastic gradient MCMC.

A.1. Proof of Proposition 2.1. With notation $\tilde{h} = h - \pi(h)$, we can represent the variance of $\pi_n(h)$, $h \in \mathcal{H}$, as

$$\mathsf{E}\left[\left(\frac{1}{n}\sum_{k=0}^{n-1} \tilde{h}(X_k)\right)^2\right] = \frac{1}{n^2}\left\{\sum_{k=0}^{n-1} \mathsf{E}\big[\tilde{h}^2(X_k)\big] + 2\sum_{\ell=1}^{n-1}\sum_{k=0}^{n-\ell-1} \mathsf{E}\big[\tilde{h}(X_k)\tilde{h}(X_{k+\ell})\big]\right\}.$$

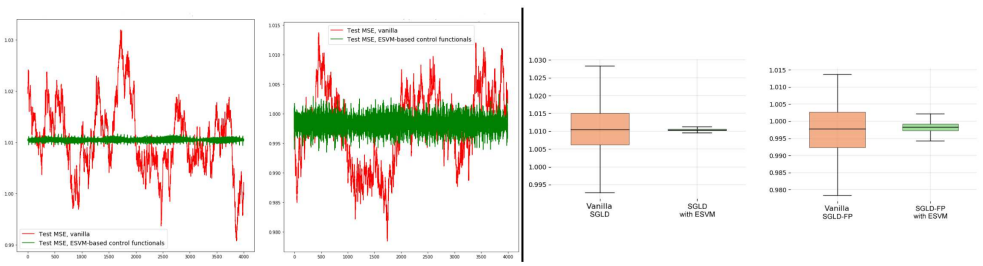

Figure 6. *Bayesian probabilistic matrix factorization from subsection 4.4. Left panel: test mean squared error (MSE) trajectory for SGLD (left) and SGLD-FP (right) with and without ESVM. Right panel: confidence intervals for test mean squared error trajectory for SGLD (left) and SGLD-FP (right).*

Multiplying the both sides by n and subtracting $\rho^{(h)}(0) + 2\sum_{\ell=1}^{n-1}(1-\ell n^{-1})\rho^{(h)}(\ell)$, we get

$$n\,\mathsf{E}\left[\left(\frac{1}{n}\sum_{k=0}^{n-1}\tilde{h}(X_k)\right)^2\right] - \rho^{(h)}(0) - 2\sum_{\ell=1}^{n-1}\left(1-\frac{\ell}{n}\right)\rho^{(h)}(\ell)$$
$$= \frac{1}{n}\sum_{k=0}^{n-1}\Big(\mathsf{E}\big[\tilde{h}^2(X_k)\big] - \rho^{(h)}(0)\Big) + \frac{2}{n}\sum_{\ell=1}^{n-1}\sum_{k=0}^{n-\ell-1}\Big(\mathsf{E}\big[\tilde{h}(X_k)\tilde{h}(X_{k+\ell})\big] - \rho^{(h)}(\ell)\Big).$$

It follows from the Cesàro mean theorem and (CS) that the right-hand side tends to zero as $n\to\infty$. Similarly, $\rho^{(h)}(0) + 2\sum_{\ell=0}^{n-1}(1-\ell n^{-1})\rho^{(h)}(\ell) \to \sum_{\ell\in\mathbb{Z}}\rho^{(h)}(\ell)$ as $n\to\infty$.

A.2. Proof of Theorem 2.2. Let us first start with a technical lemma the proof of which we postpone to the end of the section. In what follows, set $\overline{V}_n(h) := \mathsf{E}[V_n(h)]$.

Lemma A.1. *Let $\mathcal{H}$ be a class of functions with constant mean and assume that (CS) and (CD) hold. Then, for any $h\in\mathcal{H}$ and any $h_1, h_2\in\mathcal{H}$ with $\|h_1-h_2\|_{L^2(\pi)}\le\varepsilon$,*

(1) $\big|V_\infty(h) - \overline{V}_n(h)\big| \lesssim \big(R + \varsigma(1-\lambda)^{-1}\big)b_n n^{-1} + \varsigma(1-\lambda)^{-2}n^{-1} + \varsigma(1-\lambda)^{-1}\lambda^{b_n/2}$,

(2) $\big|\overline{V}_n(h_1) - \overline{V}_n(h_2)\big| \lesssim \sqrt{RD}\,b_n n^{-1/2} + \big(R+\varsigma(1-\lambda)^{-1}\big)b_n n^{-1}$
$+ \big(\sqrt{R}n^{-1/2} + \sqrt{D}\big)b_n\varepsilon$.

Let h^* be a function in $\mathcal{H}$ leading to the smallest $\overline{V}_n(h)$, that is,

$$h^* \in \arg\min_{h\in\mathcal{H}}\overline{V}_n(h).$$

For simplicity, we assume that h^* exists as all the following arguments can easily be adapted by considering an approximate minimizer. We decompose the excess of the asymptotic variance as

$$\begin{aligned} &V_\infty(\widehat{h}_{n,\varepsilon}) - \inf_{h\in\mathcal{H}}V_\infty(h)\\ &\quad= V_\infty(\widehat{h}_{n,\varepsilon}) - \overline{V}_n(\widehat{h}_{n,\varepsilon}) + \overline{V}_n(\widehat{h}_{n,\varepsilon}) - \overline{V}_n(h^*) + \overline{V}_n(h^*) - \inf_{h\in\mathcal{H}}V_\infty(h)\\ &\quad\le 2\sup_{h\in\mathcal{H}}\big|V_\infty(h) - \overline{V}_n(h)\big| + \overline{V}_n(\widehat{h}_{n,\varepsilon}) - \overline{V}_n(h^*). \end{aligned} \tag{A.1}$$

To bound the first term in (A.1), we apply Lemma A.1 and obtain

$$\begin{aligned} &\sup_{h\in\mathcal{H}}\big|V_\infty(h) - \overline{V}_n(h)\big|\\ &\quad\lesssim \big(R+\varsigma(1-\lambda)^{-1}\big)b_n n^{-1} + \varsigma(1-\lambda)^{-2}n^{-1} + \varsigma(1-\lambda)^{-1}\lambda^{b_n/2}. \end{aligned} \tag{A.2}$$

It remains to bound the second term in (A.1). Let $h^*_\varepsilon\in\mathcal{H}_\varepsilon$ be any closest to h^* point in $L^2(\pi)$-distance. By the definition of $\widehat{h}_{n,\varepsilon}$, $V_n(\widehat{h}_{n,\varepsilon}) - V_n(h^*_\varepsilon)\le 0$. Hence,

$$\begin{aligned} &\overline{V}_n(\widehat{h}_{n,\varepsilon}) - \overline{V}_n(h^*)\\ &\quad\le \overline{V}_n(\widehat{h}_{n,\varepsilon}) - \overline{V}_n(h^*) - \big(V_n(\widehat{h}_{n,\varepsilon}) - V_n(h^*_\varepsilon)\big)\\ &\quad= \overline{V}_n(\widehat{h}_{n,\varepsilon}) - \overline{V}_n(h^*) - \big(V_n(\widehat{h}_{n,\varepsilon}) - V_n(h^*)\big) + \big(V_n(h^*_\varepsilon) - V_n(h^*)\big)\\ &\quad\le \sup_{h\in\mathcal{H}_\varepsilon}\big\{\overline{V}_n(h) - V_n(h)\big\} + \big(V_n(h^*) - \overline{V}_n(h^*)\big) + \big(V_n(h^*_\varepsilon) - V_n(h^*)\big). \end{aligned} \tag{A.3}$$

By assumption and the union bound, it holds for the first term in (A.3) that

$$\mathsf{P}\left(\sup_{h\in\mathcal{H}_\varepsilon}\left\{\overline{V}_n(h)-V_n(h)\right\}>t\right)\le|\mathcal{H}_\varepsilon|\sup_{h\in\mathcal{H}_\varepsilon}\mathsf{P}\Big(\overline{V}_n(h)-V_n(h)>t\Big)\le|\mathcal{H}_\varepsilon|\,\alpha_n(t).$$

The second term in (A.3) can be handled in the same way,

$$\mathsf{P}\big(V_n(h^*)-\overline{V}_n(h^*)>t\big)\le\alpha_n(t).$$

The last term in (A.3) we represent as

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

Now the union bound implies

$$\mathsf{P}\Big(V_n(h^*)-V_n(h^*_\varepsilon)-\big(\overline{V}_n(h^*)-\overline{V}_n(h^*_\varepsilon)\big)>t\Big)\le2\alpha_n(t).$$

Furthermore, using Lemma A.1 and the fact that h^*_ε is ε-close to h^* in $L^2(\pi)$-distance,

$$\begin{aligned}&\big|\overline{V}_n(h^*)-\overline{V}_n(h^*_\varepsilon)\big|\\&\quad\lesssim\sqrt{RD}b_nn^{-1/2}+\big(R+\varsigma(1-\lambda)^{-1}\big)b_nn^{-1}+\big(\sqrt{R}n^{-1/2}+\sqrt{D}\big)b_n\varepsilon.\end{aligned}$$

Combining these inequalities and substituting them into (A.3), we obtain, with probability at least $1-(|\mathcal{H}_\varepsilon|+3)\alpha_n(t)$,

$$\begin{aligned}&\overline{V}_n(\widehat{h}_{n,\varepsilon})-\overline{V}_n(h^*)\\&\quad\lesssim t+\sqrt{RD}b_nn^{-1/2}+\big(R+\varsigma(1-\lambda)^{-1}\big)b_nn^{-1}+\big(\sqrt{R}n^{-1/2}+\sqrt{D}\big)b_n\varepsilon.\end{aligned}\tag{A.4}$$

Substituting (A.2) and (A.4) into (A.1) we conclude that, with the same probability,

$$\begin{aligned}V_\infty(\widehat{h}_{n,\varepsilon})-\inf_{h\in\mathcal{H}}V_\infty(h)\lesssim t+\big(\sqrt{R}n^{-1/2}+\sqrt{D}\big)b_n\varepsilon+\sqrt{RD}\,b_nn^{-1/2}\\+\big(R+\varsigma(1-\lambda)^{-1}\big)b_nn^{-1}+\varsigma(1-\lambda)^{-2}n^{-1}+\varsigma(1-\lambda)^{-1}\lambda^{b_n/2},\end{aligned}$$

where we have used the notation of Theorem 2.2. The proof is completed by taking $t=\alpha_n^{-1}\big(\delta/2|\mathcal{H}_\varepsilon|\big)$ and assuming that $|\mathcal{H}_\varepsilon|\ge3$ (this involves no loss of generality). We are left with the task of proving Lemma A.1.

Proof of Lemma A.1. Let us first find a leading term in sample autocovariance function. Recall that for any $h\in\mathcal{H}$, $\tilde{h}=h-\pi(h)$. By expanding the brackets and adding/subtracting $\pi(h)$ in the definition (2.3), we get, for any $|\ell|\le b_n$,

$$\rho_n^{(h)}(\ell)=A_{n,1}^{(h)}(\ell)+A_{n,2}^{(h)}(\ell)+A_{n,3}^{(h)}(\ell),\tag{A.5}$$

where, for $0 \le \ell \le b_n$,

$$A_{n,1}^{(h)}(\ell) = A_{n,1}^{(h)}(-\ell) := n^{-1} \sum_{k=0}^{n-\ell-1} \tilde{h}(X_k)\tilde{h}(X_{k+\ell}),$$

$$A_{n,2}^{(h)}(\ell) = A_{n,2}^{(h)}(-\ell) := -n^{-1}\pi_n(\tilde{h}) \left\{ \sum_{k=0}^{n-\ell-1} \tilde{h}(X_k) + \sum_{k=\ell}^{n-1} \tilde{h}(X_k) \right\},$$

$$A_{n,3}^{(h)}(\ell) = A_{n,3}^{(h)}(-\ell) := (1 - \ell/n)\, \pi_n^2(\tilde{h}).$$

It follows that the leading term in this decomposition is $A_{n,1}^{(h)}(\ell)$. The remainder terms $A_{n,2}^{(h)}(\ell)$ and $A_{n,3}^{(h)}(\ell)$ can be bounded, under assumptions (CS) and (CD), as follows:

$$\begin{aligned} \left| \mathsf{E}\big[A_{n,3}^{(\tilde{h})}(\ell)\big] \right| &\le n^{-2} \left\{ \sum_{k=0}^{n-1} \mathsf{E}\big[\tilde{h}^2(X_k)\big] + 2 \sum_{\ell=0}^{n-1} \sum_{k=1}^{n-\ell-1} \mathsf{E}\big[\tilde{h}(X_k)\tilde{h}(X_{k+\ell})\big] \right\} \\ &\le n^{-1} \left\{ \rho^{(h)}(0) + 2 \sum_{\ell=0}^{n-1} (1 - \ell n^{-1}) \rho^{(h)}(\ell) \right\} + 3Rn^{-1} \\ &\le Cn^{-1}, \end{aligned} \tag{A.6}$$

where $C := 2\varsigma(1-\lambda)^{-1} + 3R$. In the same manner we conclude that

$$\begin{aligned} \left| \mathsf{E}\big[A_{n,2}^{(h)}(\ell)\big] \right| &\le n^{-1} \mathsf{E}^{1/2}\big[\pi_n^2(\tilde{h})\big] \left\{ \mathsf{E}^{1/2} \left[\left(\sum_{k=0}^{n-\ell-1} \tilde{h}(X_k) \right)^2 + \left(\sum_{k=\ell}^{n-1} \tilde{h}(X_k) \right)^2 \right] \right\} \\ &\le 2Cn^{-1}. \end{aligned} \tag{A.7}$$

The last two bounds show that the last two terms in (A.5) are of order n^{-1}. Having disposed of this preliminary step, we can now return to statements of the lemma. ∎

Statement 1. From decomposition (A.5) and bounds (A.6), (A.7), we deduce that

$$\begin{aligned} \left| \overline{V}_n(h_1) - \overline{V}_n(h_2) \right| &= \sum\nolimits_{|\ell| \le b_n} w_n(\ell)\, \mathsf{E}\Big[\rho_n^{(h_1)}(\ell) - \rho_n^{(h_2)}(\ell) \Big] \\ &\le 2b_n \max\nolimits_{|\ell| \le b_n} \left| \mathsf{E}\Big[A_{n,1}^{(h_1)}(\ell) - A_{n,1}^{(h_2)}(\ell) \Big] \right| + 12Cb_n n^{-1}. \end{aligned}$$

With notation $\tilde{h}_{12} = \tilde{h}_1 - \tilde{h}_2$, it follows, for any $0 \le \ell \le b_n$, that

$$A_{n,1}^{(h_1)}(\ell) - A_{n,1}^{(h_2)}(\ell) = n^{-1} \sum_{k=0}^{n-\ell-1} \Big(\tilde{h}_1(X_k)\tilde{h}_{12}(X_{k+\ell}) + \tilde{h}_{12}(X_k)\tilde{h}_2(X_{k+\ell}) \Big).$$

Using the Cauchy–Schwarz inequality (twice) and (CS), we have

$$\begin{aligned} \mathsf{E} \left[\sum_{k=0}^{n-\ell-1} \tilde{h}_1(X_k)\tilde{h}_{12}(X_{k+\ell}) \right] &\le \mathsf{E}^{1/2} \left[\sum_{k=0}^{n-\ell-1} \tilde{h}_1^2(X_k) \right] \mathsf{E}^{1/2} \left[\sum_{k=0}^{n-\ell-1} \tilde{h}_{12}^2(X_{k+\ell}) \right] \\ &\le \sqrt{R + (n-l)\rho^{(\tilde{h}_1)}(0)} \sqrt{R + (n-l)\rho^{(\tilde{h}_{12})}(0)}. \end{aligned}$$

We now apply this argument again and obtain

$$\mathsf{E}\Big[A_{n,1}^{(h_1)}(\ell) - A_{n,1}^{(h_2)}(\ell)\Big] \lesssim Rn^{-1} + \sqrt{RD}n^{-1/2} + \big(\sqrt{R}n^{-1/2} + \sqrt{D}\big)\big\|\tilde{h}_1 - \tilde{h}_2\big\|_{L^2(\pi)}.$$

Finally, since $\|\tilde{h}_1 - \tilde{h}_2\|_{L^2(\pi)} \le 2\|h_1 - h_2\|_{L^2(\pi)}$, we conclude

$$\begin{aligned}&\big|\overline{V}_n(h_1) - \overline{V}_n(h_2)\big| \\ &\quad\lesssim \big(R + C\big)b_n n^{-1} + \sqrt{RD}b_n n^{-1/2} + \big(\sqrt{R}n^{-1/2} + \sqrt{D}\big)b_n\|h_1 - h_2\|_{L^2(\pi)}.\end{aligned}$$

Statement 2. Let us denote

$$V_{n,\rho}(h) = \sum\nolimits_{|\ell|\le b_n} w_n(\ell)\rho^{(h)}(\ell).$$

With this notation, we have the following decomposition:

$$\big|V_\infty(h) - \overline{V}_n(h)\big| \le \big|V_\infty(h) - V_{n,\rho}(h)\big| + \big|V_{n,\rho}(h) - \overline{V}_n(h)\big|. \tag{A.8}$$

To bound the first term in the right-hand side of (A.8), we represent it as

$$|V_{n,\rho}(h) - V_\infty(h)| \le \sum\nolimits_{|\ell|\le b_n} |1 - w_n(\ell)|\,\big|\rho^{(h)}(\ell)\big| + \sum\nolimits_{|\ell|>b_n}\big|\rho^{(h)}(\ell)\big|.$$

Using (CD) and the fact that $w_n(\ell) = 1$ for $\ell \in [-b_n/2, b_n/2]$, we obtain

$$\sum\nolimits_{|\ell|\le b_n} |1 - w_n(\ell)|\big|\rho^{(h)}(\ell)\big| = 2\sum_{\ell=b_n/2}^{b_n} |1 - w_n(\ell)|\,\big|\rho^{(h)}(\ell)\big| \le 2\varsigma(1-\lambda)^{-1}\lambda^{b_n/2}.$$

In the same manner we can see that

$$\sum\nolimits_{|s|>b_n}\big|\rho^{(h)}(s)\big| \le 2\varsigma(1-\lambda)^{-1}\lambda^{b_n}.$$

Combining the last two bounds we conclude that

$$|V_{n,\rho}(h) - V_\infty(h)| \le 4\varsigma(1-\lambda)^{-1}\lambda^{b_n/2}. \tag{A.9}$$

Now let us turn to the second term in the right-hand side of (A.8). The decomposition (A.5) and the bounds (A.6), (A.7) yield

$$\begin{aligned}|\overline{V}_n(h) - V_{n,\rho}(h)| &= \sum\nolimits_{|\ell|\le b_n} w_n(\ell)\Big(\mathsf{E}\big[\rho_n^{(h)}(\ell)\big] - \rho^{(h)}(\ell)\Big) \\ &\le \sum\nolimits_{|\ell|\le b_n}\Big|\mathsf{E}\big[A_{n,1}^{(h)}(\ell)\big] - \rho^{(h)}(\ell)\Big| + 6Cb_n n^{-1}.\end{aligned}$$

Using (CS) and (CD), it follows that

$$\begin{aligned}\Big|\mathsf{E}\big[A_{n,1}^{(h)}(\ell)\big] - \rho^{(h)}(\ell)\Big| &\le n^{-1}\sum_{k=0}^{n-\ell-1}\Big|\mathsf{E}\big[\tilde{h}(X_k)\tilde{h}(X_{k+\ell})\big] - \rho^{(h)}(\ell)\Big| + \ell n^{-1}\rho^{(h)}(\ell) \\ &\le Rn^{-1} + \varsigma\ell\lambda^\ell n^{-1}.\end{aligned}$$

Combining these, we get

$$|\overline{V}_n(h) - V_{n,\rho}(h)| \le \big(12\varsigma(1-\lambda)^{-1} + 20R\big)b_n n^{-1} + 2\varsigma(1-\lambda)^{-2}n^{-1}. \tag{A.10}$$

Finally, we obtain the desired conclusion by substituting (A.9) and (A.10) into (A.8).

A.3. Proof of Proposition 3.2. 1. The sequence $(\rho^{(h)}(\ell))_{\ell=0}^{\infty}$ is symmetric and positive semidefinite by construction. By the Markov property, for any $k, \ell \in \mathbb{N}$,

$$\mathsf{E}_\xi\big[\tilde{h}(X_k)\tilde{h}(X_{k+\ell})\big] - \rho^{(h)}(\ell) = \bar{\mathsf{E}}_\zeta\big[\tilde{h}(X)\phi_\ell(X) - \tilde{h}(X')\phi_\ell(X')\big],$$

where we denote $\phi_\ell(x) := P^\ell \tilde{h}(x)$ and $\zeta \in \Pi(\xi P^k, \pi)$ is the optimal coupling of ξP^k and π in W_2^{d}-distance, $(X, X') \sim \zeta$. Note that

$$\begin{aligned}\big|\bar{\mathsf{E}}_\zeta[\tilde{h}(X)\phi_\ell(X) - \tilde{h}(X')\phi_\ell(X')]\big| &\le \{\bar{\mathsf{E}}_\zeta[\{\tilde{h}(X) - \tilde{h}(X')\}^2]\}^{1/2}\{\bar{\mathsf{E}}_\zeta[\{\phi_\ell(X')\}^2]\}^{1/2} \\ &\quad + \{\bar{\mathsf{E}}_\zeta[\{\phi_\ell(X) - \phi_\ell(X')\}^2]\}^{1/2}\{\bar{\mathsf{E}}_\zeta[\{\tilde{h}(X)\}^2]\}^{1/2}.\end{aligned}$$

It is easy to check that ϕ_ℓ is a Lipschitz function,

$$|\phi_\ell(x) - \phi_\ell(x')| \le L W_2^{\mathsf{d}}(\delta_x P^\ell, \delta_{x'} P^\ell) \le L\Delta_2^\ell \mathsf{d}(x, x')\,.$$

Since the Markov kernel P is W_2-geometrically ergodic, we get

$$\begin{aligned}&\big|\bar{\mathsf{E}}_\zeta[\tilde{h}(X)\phi_\ell(X) - \tilde{h}(X')\phi_\ell(X')]\big| \\ &\quad \le L W_2^{\mathsf{d}}(\xi P^k, \pi)\{\bar{\mathsf{E}}_\zeta[\{\phi_\ell(X')\}^2]\}^{1/2} + L\Delta_2^\ell W_2^{\mathsf{d}}(\xi P^k, \pi)\{\bar{\mathsf{E}}_\zeta[\{\tilde{h}(X)\}^2]\}^{1/2}.\end{aligned}$$

Let us compute $\bar{\mathsf{E}}_\zeta[\{\phi_\ell(X')\}^2] = \pi(\phi_\ell^2)$. Since P is W_2^{d}-geometrically ergodic, we have $W_2^{\mathsf{d}}(\delta_y P^\ell, \pi) \le \Delta_2^\ell W_2^{\mathsf{d}}(\delta_y, \pi)$. Note also that $\pi(\tilde{h}) = 0$ implies $\pi(\phi_\ell) = 0$; hence

$$\pi(\phi_\ell^2) = \int \left[\phi_\ell(y) - \int \phi_\ell(x)\pi(\mathrm{d}x)\right]^2 \pi(\mathrm{d}y) \le L^2\Delta_2^{2\ell}\int \{W_2^{\mathsf{d}}(\delta_y, \pi)\}^2 \pi(\mathrm{d}y). \tag{A.11}$$

Finally, we need to compute $\bar{\mathsf{E}}_\zeta[\{\tilde{h}(X)\}^2] = \xi P^k\big(\tilde{h}^2\big)$. For an arbitrary $\hat{x} \in \mathsf{X}$,

$$\big|\tilde{h}(x)\big|^2 = \left|\int \{h(x) - h(y)\}\,\pi(\mathrm{d}y)\right|^2 \le 2L^2\left(\mathsf{d}^2(x, \hat{x}) + \int \mathsf{d}^2(x, \hat{x})\pi(\mathrm{d}x)\right).$$

In order to bound $\xi P^k(\mathsf{d}^2(x, \hat{x}))$, we write

$$\begin{aligned}\int \mathsf{d}^2(x, \hat{x})\xi P^k(\mathrm{d}x) &= \iint \mathsf{d}^2(x, \hat{x})\zeta(\mathrm{d}x\mathrm{d}x') \le 2\iint \mathsf{d}^2(x, x')\zeta(\mathrm{d}x\mathrm{d}x') \\ &\quad + 2\int \mathsf{d}^2(x, \hat{x})\pi(\mathrm{d}x) \le 2\Delta_2^{2k}\{W_2^{\mathsf{d}}(\xi, \pi)\}^2 + 2\int \mathsf{d}^2(x, \hat{x})\pi(\mathrm{d}x)\,.\end{aligned}$$

Hence

$$\xi P^k\big(\tilde{h}^2\big) \le 4L^2\int \mathsf{d}^2(x, \hat{x})\pi(\mathrm{d}x) + 2L^2\Delta_2^{2k}\{W_2^{\mathsf{d}}(\xi, \pi)\}^2,$$

and $\big|\mathsf{E}_\xi[\tilde{h}(X_k)\tilde{h}(X_{k+\ell})] - \rho^{(h)}(\ell)\big| \le A_1 L^2 \Delta_2^{k+\ell} W_2^{\mathsf{d}}(\xi, \pi)$, where

$$A_1 := 2\inf_{\hat{x}\in\mathsf{X}} W_2^{\mathsf{d}}(\delta_{\hat{x}}, \pi) + 2W_2^{\mathsf{d}}(\xi, \pi) + \left[\int \{W_2^{\mathsf{d}}(\delta_y, \pi)\}^2 \pi(\mathrm{d}y)\right]^{1/2}. \tag{A.12}$$

Summing the last inequality with respect to k, we obtain

$$\sum_{k=0}^{\infty}\left|\mathsf{E}_\xi[\tilde{h}(X_k)\tilde{h}(X_{k+\ell})]-\rho^{(h)}(\ell)\right|\le A_1L^2(1-\Delta_2)^{-1}\Delta_2^\ell W_2(\xi,\pi)\,. \tag{A.13}$$

Hence, the second assumption in (CS) holds with R defined in (3.2). The third assertion clearly follows from (A.13). To check (CD), we write

$$\begin{aligned}|\rho^{(h)}(\ell)| &= \left|\int \tilde{h}(x)\big[\delta_xP^\ell\big(h\big)-\pi(h)\big]\pi(\mathrm{d}x)\right| \le L\int|\tilde{h}(x)|W_2^{\mathsf{d}}(\delta_xP^\ell,\pi)\pi(\mathrm{d}x)\\ &\le L\Delta_2^\ell\int|\tilde{h}(x)|W_2^{\mathsf{d}}(\delta_x,\pi)\pi(\mathrm{d}x)\le L\Delta_2^\ell\sqrt{D}\left[\int\{W_2^{\mathsf{d}}(\delta_x,\pi)\}^2\pi(\mathrm{d}x)\right]^{1/2}. \end{aligned}\tag{A.14}$$

Hence (CD) holds with $\lambda=\Delta_2$ and $\varsigma=L\sqrt{D}[\int\{W_2^{\mathsf{d}}(\delta_x,\pi)\}^2\pi(\mathrm{d}x)]^{1/2}$.

2. The proof essentially relies on [17]. Denote $Z_n(h):=(h(X_0),\dots,h(X_{n-1}))$, and recall the representation (2.2). It follows from [5, section 5.2] that $V_n(h)$ can be represented as a quadratic form

$$V_n(h)=\langle A_nZ_n(h),Z_n(h)\rangle\,,$$

where $A_n=n^{-1}(\mathrm{I}-n^{-1}E)W(\mathrm{I}-n^{-1}E)$, E is an $n\times n$ matrix with elements $E_{j,k}=1$ for any $1\le j,k\le n$, and W is a Toeplitz matrix with elements $W_{j,k}=w_n(j-k)$. Note that $V_n(h)$ is invariant to shifts and, in particular, $V_n(h)=V_n(\tilde{h})$. It is straightforward to show that $\|A_n\|\le 2b_nn^{-1}$; see [5, Lemma 9]. Furthermore, [5, Corollary 18] implies

$$\mathsf{P}_\xi\big(\big|V_n(h)-\mathsf{E}_\xi[V_n(h)]\big|\ge t\big)\le 2\exp\left(-\frac{(1-\Delta_2)^2t^2}{c\alpha L^2\big(\mathsf{E}_\xi[\|A_nZ_n(\tilde{h})\|^2]+t\|A_n\|\big)}\right), \tag{A.15}$$

where $c>0$ is some universal constant. By the Cauchy–Schwarz inequality, $\|A_nZ_n(h)\|^2\le\|A_n\|^2\|Z_n(h)\|^2$. Moreover, using (CS), we get

$$\mathsf{E}_\xi[\|Z_n(\tilde{h})\|^2]\le R+n\,\mathrm{Var}_\pi[h]\le R+n\sup_{h\in\mathcal{H}}\mathrm{Var}_\pi[h]=R+nD. \tag{A.16}$$

The statement follows from substitution (A.16) into (A.15).

A.4. Proof of Proposition 3.3. 1. Proceeding similarly to Appendix A.3, we use the Markov property to write, for $k,\ell\in\mathbb{N}$,

$$\begin{aligned}\left|\mathsf{E}_\xi[\tilde{h}(X_k)\tilde{h}(X_{k+\ell})]-\rho^{(h)}(\ell)\right| &\le \{\bar{\mathsf{E}}_\zeta[\{\tilde{h}(X)-\tilde{h}(X')\}^2]\}^{1/2}\{\bar{\mathsf{E}}_\zeta[\{\phi_\ell(X')\}^2]\}^{1/2}\\ &\quad+\{\bar{\mathsf{E}}_\zeta[\{\phi_\ell(X)-\phi_\ell(X')\}^2]\}^{1/2}\{\bar{\mathsf{E}}_\zeta[\{\tilde{h}(X)\}^2]\}^{1/2}\,,\end{aligned}$$

where $\zeta\in\Pi(\xi P^k,\pi)$ is the optimal coupling of ξP^k and π in W_1^{d} distance, $(X,X')\sim\zeta$. Since function $\tilde{h}$ is bounded and Lipschitz,

$$\bar{\mathsf{E}}_\zeta\big[\{\tilde{h}(X)-\tilde{h}(X')\}^2\big]\le 2B\bar{\mathsf{E}}_\zeta\big[\big|\tilde{h}(X)-\tilde{h}(X')\big|\big]\le 2LB\Delta_1^kW_1^{\mathsf{d}}(\xi,\pi).$$

Similarly, using that ϕ_ℓ is bounded and Lipschitz (see Appendix A.3 for the details),

$$\bar{\mathsf{E}}_\zeta\big[\{\phi_\ell(X)-\phi_\ell(X')\}^2\big]\le 2LB\Delta_1^{k+\ell}W_1^{\mathsf{d}}(\xi,\pi).$$

Proceeding as in (A.11), we obtain

$$\bar{\mathsf{E}}_\zeta[\{\phi_\ell(X')\}^2] \le \iint \big[\phi_\ell(x) - \phi_\ell(y)\big]^2 \pi(\mathrm{d}x)\pi(\mathrm{d}y) \le 2BL\Delta_1^\ell \int W_1^{\mathsf{d}}(\delta_y, \pi)\, \pi(\mathrm{d}y).$$

Using the simple bound $\bar{\mathsf{E}}_\zeta[\{\tilde{h}(X)\}^2] \le 4B^2$, it holds that

$$\big|\mathsf{E}_\xi[\tilde{h}(X_k)\tilde{h}(X_{k+\ell})] - \rho^{(h)}(\ell)\big| \le 2BC_2'\Delta_1^{(k+\ell)/2} \tag{A.17}$$

with

$$C_2' = 2L\{W_1^{\mathsf{d}}(\xi,\pi)\}^{1/2}\left\{\int W_1^{\mathsf{d}}(\delta_x,\pi)\pi(\mathrm{d}x)\right\}^{1/2} + \big\{2LBW_1^{\mathsf{d}}(\xi,\pi)\big\}^{1/2}. \tag{A.18}$$

Hence, the second assumption in (CS) holds with

$$R = 2BC_2'\big(1 - \Delta_1^{1/2}\big)^{-1}, \tag{A.19}$$

and the third one follows from (A.17). Proceeding as in (A.14),

$$\big|\rho^{(h)}(\ell)\big| \le 2LB\Delta_1^\ell \int W_1^{\mathsf{d}}(\delta_x,\pi)\pi(\mathrm{d}x)\,,$$

and (CD) holds with $\lambda = \Delta_1$ and

$$\varsigma = 2LB \int W_1^{\mathsf{d}}(\delta_x,\pi)\pi(\mathrm{d}x). \tag{A.20}$$

2. Without loss of generality, we assume that $\|h\|_\infty \le 1$. By Minkowski's inequality,

$$\big\|V_n(h) - \mathsf{E}_\xi V_n(h)\big\|_{\xi,p} \le \sum\nolimits_{\ell=-b_n+1}^{b_n} w_n(\ell)\big\|\rho_n^{(h)}(\ell) - \mathsf{E}_\xi[\rho_n^{(h)}(\ell)]\big\|_{\xi,p},$$

where $\|\cdot\|_{\xi,p} := \big(\mathsf{E}_\xi[\cdot]^p\big)^{1/p}$. For $\ell \in \mathbb{N}_0$, we get

$$\begin{aligned}\rho_n^{(h)}(\ell) &= \frac{1}{n}\sum_{k=0}^{n-\ell-1}\big(h(X_k) - \pi_n(h)\big)\big(h(X_{k+\ell}) - \pi_n(h)\big) = \frac{1}{n}\sum_{k=0}^{n-\ell-1}\tilde{h}(X_k)\tilde{h}(X_{k+\ell}) \\ &\quad - \frac{1}{n}\big(\pi_n(h) - \pi(h)\big)^2 + \frac{\pi_n(h) - \pi(h)}{n}\sum_{k=0}^{\ell-1}\big[\tilde{h}(X_k) + \tilde{h}(X_{n-\ell+k})\big] \\ &=: T_1 + T_2 + T_3.\end{aligned}$$

Hence,

$$\rho_n^{(h)}(\ell) - \mathsf{E}_\xi[\rho_n^{(h)}(\ell)] = \big(T_1 - \mathsf{E}_\xi[T_1]\big) + \big(T_2 - \mathsf{E}_\xi[T_2]\big) + \big(T_3 - \mathsf{E}_\xi[T_3]\big) =: \bar{T}_1 + \bar{T}_2 + \bar{T}_3.$$

Now we proceed with estimating $\|\bar{T}_1\|_{\xi,p}$. By [19, Theorem 2] and Lemma A.2, setting $\beta_{r,\ell} = 1 \wedge \Delta_1^{r-\ell}$ and $A_1 = 4 \vee 256L\{W_1(\xi,\pi) + 2\inf_{\hat{x}\in\mathsf{X}} W_1(\delta_{\hat{x}},\pi)\}$,

$$\|\bar{T}_1\|_{\xi,p}^p \le \frac{(2p-2)!}{(p-1)!}\mathrm{e}^p \left[\left(nA_1\sum_{r=0}^{n-1}\beta_{r,\ell}\right)^{p/2} \vee nA_1 16^{p-2}\sum_{r=0}^{n-1}(r+1)^{p-2}\beta_{r,\ell}\right]. \tag{A.21}$$

It can be easily seen that

$$\sum_{r=0}^{n-1}(r+1)^{p-1}\beta_{r,\ell} \le \frac{\ell^{p-1}}{p-1} + \frac{p!}{\Delta_1^2}\left[\frac{1}{\log^p(1/\Delta_1)} + \ell^p\right], \quad \sum_{r=0}^{n-1}\beta_{r,\ell} \le \frac{2\ell}{(1-\Delta_1)}. \tag{A.22}$$

Substituting (A.22) into (A.21) and using Stirling's formula,

$$\|\bar{T}_1\|_{\xi,p}^p \le 2^{2p}p^p\left[\left(\frac{2A_1\ell n}{1-\Delta_1}\right)^{p/2} + nA_1 16^p\left\{\ell^p + \frac{2^p p^p p^{1/2}}{\mathrm{e}^p\Delta_1^2}\left(\frac{1}{\log^p(1/\Delta_1)} + \ell^p\right)\right\}\right].$$

Since $\ell \le b_n$, we obtain the following final bound on $\bar{T}_1$,

$$\|\bar{T}_1\|_{\xi,p} \le 4p\left[\left(\frac{2b_nA_1}{n(1-\Delta_1)}\right)^{1/2} + 32b_n n^{1/p-1}A_1^{1/p}\left(1 + \frac{p(1+\log(1/\Delta_1))}{\mathrm{e}\Delta_1^{2/p}\log(1/\Delta_1)}\right)\right]. \tag{A.23}$$

Let us consider now $\bar{T}_2$ and $\bar{T}_3$. Using (WE),

$$\|\pi_n(h) - \pi(h)\|_{\xi,p} \le n^{-1}\left\|\sum_{k=0}^{n-1} h(X_k) - \mathsf{E}_\xi[h(X_k)]\right\|_{\xi,p} + \frac{LW_1(\xi,\pi)}{n(1-\Delta)}.$$

By Lemma A.3 and [19, Theorem 2], setting $A_2 = 4L\{W_1(\xi,\pi) + 2\inf_{\hat{x}\in\mathsf{X}} W_1(\delta_{\hat{x}},\pi)\}$,

$$\left\|\sum_{k=0}^{n-1} h(X_k) - \mathsf{E}_\xi[h(X_k)]\right\|_{\xi,p} \le 4p\left[\frac{n^{1/2}A_2^{1/2}}{\sqrt{1-\Delta_1}} \vee \frac{2pn^{1/p}A_2^{1/p}}{\Delta_1^{1/p}\mathrm{e}\log(1/\Delta_1)}\right].$$

Now it holds for $\bar{T}_2$,

$$\|\bar{T}_2\|_{\xi,p} \le 2\big\|\pi_n(h) - \pi(h)\big\|_{\xi,2p}^2 \le \frac{4}{n^2}\left\|\sum_{k=0}^{n-1} h(X_k) - \mathsf{E}_\xi[h(X_k)]\right\|_{\xi,2p}^2 + \frac{6L^2W_1^2(\xi,\pi)}{n^2(1-\Delta)^2}$$

$$\le 2^6p^2\left[\frac{A_2}{n(1-\Delta_1)} \vee \frac{4p^2n^{2/p}A_2^{2/p}}{n^2\Delta^{2/p}\mathrm{e}^2\log^2(1/\Delta_1)}\right] + \frac{4L^2W_1^2(\xi,\pi)}{n^2(1-\Delta_1)^2}. \tag{A.24}$$

Finally, since $\ell \le b_n$ and h is bounded,

$$\|\bar{T}_3\|_{\xi,p} \le 16b_nn^{-1}. \tag{A.25}$$

Using (A.23), (A.24), and (A.25), we get

$$\big\|V_n(h) - \mathsf{E}_\xi[V_n(h)]\big\|_{\xi,p} \le 2b_n\left[\frac{C_1'pb_n^{1/2}}{n^{1/2}} + \frac{C_2'p^2b_n}{n^{1-1/p}} + \frac{C_3'p^4}{n^{2-2/p}}\right], \tag{A.26}$$

where

$$C_1' = \frac{4(2A_1)^{1/2}}{\sqrt{1-\Delta_1}}, \quad C_2' = \frac{2^7 A_1^{1/p}(1+2\log(1/\Delta_1))}{\Delta_1^{2/p}\log(1/\Delta_1)} + \frac{2^6 A_2}{1-\Delta_1} + 12,$$
$$C_3' = \frac{2^8 A_2^{2/p}}{e^2 \Delta_1^{2/p}\log^2(1/\Delta_1)} + \frac{A_2^2}{4(1-\Delta_1)^2}. \tag{A.27}$$

Under the assumption $p < n^{1/2}$, we obtain

$$\left\|V_n(h) - \mathsf{E}_\xi[V_n(h)]\right\|_{\xi,p} \le b_n \|h\|_\infty^2 \left[\frac{C_{R,1} p b_n^{1/2}}{n^{1/2}} + \frac{C_{R,2} p^2 b_n}{n^{1-1/p}}\right] \tag{A.28}$$

with $C_{R,1} = 2C_1'$, $C_{R,2} = 2C_2' + 2C_3'$. Now the statement follows from Markov's inequality.

Lemma A.2. *Assume* (WE)-1. *Let* $h \in \mathrm{Lip}_{b,\mathsf{d}}(L,B)$. *For* $i,m \in \mathbb{N}_0$, *we define* $\tilde{g}_{i,m}(x,x') = \tilde{h}(x)\tilde{h}(x') - c_{\xi,i,m}$, *where* $c_{\xi,i,m} := \mathsf{E}_\xi[g_{i,m}(X_i, X_{i+m})]$. *For* $p,r,m \in \mathbb{N}_0$, *let*

$$C_{p,r,m}^{(h)} := \sup\left|\mathsf{cov}_\xi\left(\prod_{k=1}^u \tilde{g}_{i_k,m}(X_{i_k}, X_{i_k+m}), \prod_{k=1}^v \tilde{g}_{j_k,m}(X_{j_k}, X_{j_k+m})\right)\right|,$$

where the supremum is taken over all $0 \le i_1 \le \cdots \le i_u < i_u + r \le j_1 \le \cdots \le j_v \le n$ *with* $u+v=p$. *Then, for any* $p,r \in \mathbb{N}$,

$$C_{p,r,m}^{(h)} \le \begin{cases} 2^{2p+2}B^{2p}, & r \le m, \\ L\{W_1(\xi,\pi) + 2\inf_{\hat{x}\in\mathsf{X}} W_1(\delta_{\hat{x}},\pi)\} v 2^{4p} B^{2p-1} \Delta_1^{r-m}, & r > m. \end{cases}$$

Proof. Define the function

$$G_{i_1,\dots,i_u,m}(x_{i_1}, x_{i_1+m}, \dots, x_{i_u}, x_{i_u+m}) := \prod_{k=1}^u \tilde{g}_{i_k,m}(x_{i_k}, x_{i_k+m}).$$

Let $\mathsf{D}_{i_1,\dots,i_u,m} = G_{i_1,\dots,i_u,m}(X_{i_1}, X_{i_1+m}, \dots, X_{i_u}, X_{i_u+m})$. Since $\|G_{i_1,\dots,i_u,m}\|_\infty \le (2B)^{2u}$ and $\|G_{j_1,\dots,j_v,m}\|_\infty \le (2B)^{2v}$, we get $C_{p,r,m}^{(h)} \le 2^{2p+2}B^{2p}$. Now let $m < r$. Using Markov's property,

$$\mathsf{cov}_\xi\left(\mathsf{D}_{i_1,\dots,i_u,m}, \mathsf{D}_{j_1,\dots,j_v,m}\right)$$
$$= \mathsf{E}_\xi\left[\left(\mathsf{D}_{i_1,\dots,i_u,m} - \mathsf{E}_\xi[\mathsf{D}_{i_1,\dots,i_v,m}]\right)\left(P^{j_1-i_u-m}\varphi(X_{i_u+m}) - \pi(\varphi)\right)\right],$$

where

$$\varphi(x) := \mathsf{E}_x\left[G_{j_1,\dots,j_v,m}(x, X_m, X_{j_2-j_1}, X_{j_2-j_1+m}, \dots, X_{j_v-j_1}, X_{j_v-j_1+m})\right].$$

It follows from Lemma A.4 that $\|\varphi\|_{\mathsf{Lip}} \le Lv2^{4v-1}B^{2v-1}$. By [18, Theorem 20.1.2] $\left\|P^{j_1-i_u-m}\varphi\right\|_{\mathsf{Lip}} \le \Delta_1^{j_1-i_u-m}\|\varphi\|_{\mathsf{Lip}}$, and hence

$$|P^{j_1-i_u-m}\varphi(x) - \pi(\varphi)| \le Lv2^{4v-1}B^{2v-1}\Delta_1^{j_1-i_u-m} W_1(\delta_x,\pi).$$

This yields

$$\left|\mathsf{cov}_\xi\left(\mathsf{D}_{i_1,\dots,i_u,m}, \mathsf{D}_{j_1,\dots,j_v,m}\right)\right| \le Lp2^{4p}B^{2p-1}\Delta_1^{j_1-i_u-m}\mathsf{E}_\xi\left[W_1(\delta_{X_{i_u+m}},\pi)\right].$$

For a fixed $\hat{x} \in \mathsf{X}$, by the triangle inequality,

$$W_1(\delta_x, \pi) \le W_1(\delta_x, \delta_{\hat{x}}) + W_1(\delta_{\hat{x}}, \pi) = \mathsf{d}(x, \hat{x}) + W_1(\delta_{\hat{x}}, \pi).$$

Since $\mathsf{E}_\xi[\mathsf{d}(X_{i_u+m}, \hat{x})] \le W_1(\delta_\xi P^{i_u+m}, \delta_{\hat{x}})$, we get

$$\mathsf{E}_\xi\left[\mathsf{d}(X_{i_u+m}, \hat{x})\right] \le W_1(\xi P^{i_u+m}, \pi) + W_1(\delta_{\hat{x}}, \pi) \le \Delta^{i_u+m} W_1(\xi, \pi) + W_1(\delta_{\hat{x}}, \pi),$$

showing that $\mathsf{E}_\xi\left[W_1(\delta_{X_{i_u+m}}, \pi)\right] \le W_1(\xi, \pi) + 2W_1(\delta_{\hat{x}}, \pi)$. The proof is complete. ∎

Lemma A.3. *Assume* (WE)-1. *Let* $h \in \mathrm{Lip}_{b,\mathsf{d}}(L, B)$. *For* $p, r \in \mathbb{N}_0$, *we define*

$$C_{p,r} := \sup \left| \mathsf{cov}_\xi \left(\prod_{k=1}^{u} \left(h(X_{i_k}) - \mathsf{E}_\xi[h(X_{i_k})]\right), \prod_{k=1}^{v} \left(h(X_{j_k}) - \mathsf{E}_\xi[h(X_{j_k})]\right) \right) \right|,$$

where the supremum is taken over all $0 \le i_1 \le \cdots \le i_u < i_u + r \le j_1 \le \cdots \le j_v \le n$ *with* $u + v = p$. *Then for any* $p, r \in \mathbb{N}$,

$$C_{p,r} \le L\left\{W_1(\xi, \pi) + 2 \inf_{\hat{x} \in \mathsf{X}} W_1(\delta_{\hat{X}}, \pi)\right\} p 2^{2p} B^{2p-1} \Delta_1^r.$$

Proof. The proof is along the same lines as Lemma A.2 and is omitted. ∎

Lemma A.4. *Assume* (WE)-1. *Set*

$$\varphi(x) = \mathsf{E}_x\left[G_{j_1,\dots,j_v,m}(x, X_m, X_{j_2-j_1}, X_{j_2-j_1+m}, \dots, X_{j_v-j_1}, X_{j_v-j_1+m})\right],$$

where $G_{j_1,\dots,j_v,m}$ *defined in Lemma A.2. Then*

$$\|\varphi\|_{\mathsf{Lip}} \le L v 2^{4v-1} B^{2v-1}.$$

Proof. We split the proof into two parts. First, we estimate Lipschitz constant of $g(x) = \mathsf{E}_x\left[\prod_{k=1}^{v} \tilde{h}(X_{i_k}) \tilde{h}(X_{i_{k+m}})\right]$ for $0 = i_1 \le i_2 \le \cdots \le i_v$ and $m > 0$. Note that $g(x) = \tilde{h}^{n_1}(x) \mathsf{E}_x[\prod_{k=2}^{b} \tilde{h}^{n_k}(X_{m_k})]$, where $0 = m_1 < m_2 < \cdots < m_b$ are distinct indices among $(i_1, i_1 + m, \dots, i_v, i_v + m)$ and $(n_1, \dots, n_b)$ are their associated multiplicities ($\sum_{k=1}^{b} n_k = 2v$). Hence, applying Lemma A.5 with $f_i = \tilde{h}$ and $K = 2B$, we get $\|g\|_{\mathsf{Lip}} \le 2Lv(2B)^{2v-1}$. Now we estimate the Lipschitz constant of

$$\varphi(x) = \mathsf{E}_x\left[\prod\nolimits_{k=1}^{v} \left(\tilde{h}(X_{j_k-j_1}) \tilde{h}(X_{j_k-j_1+m}) - c_{\xi,j_k,m}\right)\right], \tag{A.29}$$

where $c_{\xi,j_k,m} := \mathsf{E}_\xi[\tilde{h}(X_{j_k}) \tilde{h}(X_{j_k+m})]$. Expanding (A.29), we obtain

$$\varphi(x) = \sum\nolimits_{(\delta_1,\dots,\delta_v)} (-1)^{\sum_k \delta_k} \mathsf{E}_x\left[\prod\nolimits_{k=1}^{v} \tilde{h}^{\delta_k}(X_{j_k}) \tilde{h}^{\delta_k}(X_{j_k+m})\right] c_{\xi,j_k,m}^{1-\delta_k}, \tag{A.30}$$

where the sum is taken w.r.t. all $(\delta_1, \dots, \delta_v)$ with $\delta_i \in \{0, 1\}$. Note that all terms in the decomposition (A.30) are Lipschitz. Since $|c_{\xi,j_k,m}| \le 4B^2$, we get

$$\begin{aligned} \|\varphi\|_{\mathsf{Lip}} &\le \sum_{s=1}^{v} 2Ls \binom{v}{s} (2B)^{2s-1} (2B)^{2v-2s} = 2Lv(2B)^{2v-1} \sum_{s=1}^{v} \frac{(v-1)!}{(s-1)!(v-s)!} \\ &= Lv 2^{4v-1} B^{2v-1}. \end{aligned}$$

∎

Lemma A.5. *Assume* (WE)-1. *For any* $b, v \geq 1$, $0 = i_1 < \cdots < i_b \leq n$ *and* $n_k \in \mathbb{N}, \sum_{k=1}^{b} n_k = v$ *we define* $g(x) = \mathsf{E}_x\big[\prod_{k=1}^{b} f_{i_k}^{n_k}(X_{i_k})\big]$, *where* $f_{i_k} \in \mathrm{Lip}_{b,\mathsf{d}}(L, K)$. *Then* $\|g\|_{\mathsf{Lip}} \leq LvK^{v-1}$.

Proof. Note that $g(x) = f_0^{n_1}(x)\mathsf{E}_x\big[\prod_{k=2}^{b} f_{i_k}^{n_k}(X_{i_k})\big]$. We proceed by induction in the number of distinct indices b. If $b = 1$, then, for any $v \in \mathbb{N}$,

$$|f_0^v(x) - f_0^v(y)| = |f_0(x) - f_0(y)| \cdot \Big|\sum_{k=0}^{v-1} f_0^k(x) f_0^{v-k-1}(y)\Big| \leq vLK^{v-1}\mathsf{d}(x, y)\,.$$

Assume $b > 1$. Since $g(x) = f_0^{n_1}(x)P^{i_2}g_1(x)$ with $g_1(x) = \mathsf{E}_x[\prod_{k=2}^{b} f_{i_k}^{n_k}(X_{i_k - i_2})]$,

$$|g(x) - g(y)| \leq \big|f_0^{n_1}(x) - f_0^{n_1}(y)\big|\big|P^{i_2}g_1(x)\big| + \big|f_0^{n_1}(x)\big|\big|P^{i_2}g_1(x) - P^{i_2}g(y)\big|.$$

The function g_1 depends on $b-1$ indices and $\sum_{k=2}^{b} n_k = v - n_1$. The induction assumption and [18, Theorem 20.1.2] show under (WE)-1 that $\|P_1^{i_2}g_1\|_{\mathsf{Lip}} \leq \|g_1\|_{\mathsf{Lip}} \leq L(v - n_1)K^{v-n_1-1}$. Observe that

$$\|g\|_{\mathsf{Lip}} \leq n_1 LK^{n_1-1}K^{v-n_1} + L(v - n_1)K^{v-n_1-1}K^{n_1}\,,$$

and the proof is complete. ∎

A.5. Proof of Proposition 3.7. We provide the proof only for SGLD, since its adaptation to SGLD-FP is straightforward. Let $x = (\theta_0^{(1)}, \tilde{S}_0^{(1)})$, $y = (\theta_0^{(2)}, \tilde{S}_0^{(2)})$. We use the standard synchronous coupling technique adapted from [10, Lemma 1]. Let $(\xi_k)_{k\geq 0}$ be a sequence of i.i.d. d-dimensional Gaussian random variables, $(S_k)_{k\geq 0}$ and $(\tilde{S}_k)_{k\geq 0}$ be independent mini-batches with $|S_k| = |\tilde{S}_k| = M$. Set $(\theta_0^{(1)}, \theta_0^{(2)}) = (x, y)$, and define recursively for $k \geq 0$,

$$\theta_k^{(i)} = \theta_{k-1}^{(i)} - \gamma G(\theta_{k-1}^{(i)}, S_k) + \sqrt{2\gamma}\xi_k;$$
$$G(\theta, S) = \nabla U_0(\theta) + NM^{-1}\sum\nolimits_{i\in S} \nabla U_i(\theta).$$

Finally, define the sequences $(X_n^{(i)})_{n\geq 0}$, $i = 1, 2$, as $X_n^{(i)} = (\theta_n^{(i)}, \tilde{S}_n)$ for any $n \geq 0$. Since $X_k^{(1)}$ and $X_k^{(2)}$ are distributed according to $\delta_x\overline{P}^k$ and $\delta_y\overline{P}^k$, respectively,

$$W_2^2(\delta_x\overline{P}^k, \delta_y\overline{P}^k) \leq \mathsf{E}\big[\mathsf{d}^2(X_k^{(1)}, X_k^{(2)})\big] = \mathsf{E}\big[\|\theta_k^{(1)} - \theta_k^{(2)}\|^2\big].$$

The rest of the proof follows [10, Lemma 1] and is omitted.

Acknowledgments. The authors are greatly indebted to the Associate Editor and the reviewers for the careful reading of the manuscript and pertinent comments. Their constructive feedback helped to improve the quality of this work and shape its final form.

REFERENCES

[1] R. Assaraf and M. Caffarel, *Zero-variance principle for Monte Carlo algorithms*, Phys. Rev. Lett., 83 (1999), pp. 4682–4685.

[2] J. Baker, P. Fearnhead, E. B. Fox, and C. Nemeth, *Control variates for stochastic gradient MCMC*, Stat. Comput., 29 (2019), pp. 599–615.

[3] D. Bakry and M. Émery, *Diffusions hypercontractives*, Sém. Probab. Strasbourg, 19 (1985), pp. 177–206.

[4] D. Bakry, I. Gentil, and M. Ledoux, *Analysis and Geometry of Markov Diffusion Operators*, Springer Science & Business Media, New Yrok, 2013.
[5] D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, and S. Samsonov, *Variance reduction for Markov chains with application to MCMC*, Stat. Comput., 30 (2020), pp. 973–997.
[6] D. Belomestny, L. Iosipoi, and N. Zhivotovskiy, *Variance Reduction via Empirical Variance Minimization: Convergence and Complexity*, preprint, arXiv:1712.04667 [math. NA], 2017.
[7] D. V. Belomestny, L. S. Iosipoi, and N. K. Zhivotovskiy, *Variance reduction in Monte Carlo estimators via empirical variance minimization*, Dokl. Math., 98 (2018), pp. 494–497.
[8] S. Bobkov and F. Götze, *Exponential integrability and transportation cost related to logarithmic Sobolev inequalities*, J. Funct. Anal., 163 (1999), pp. 1–28.
[9] N. Brosse, A. Durmus, S. Meyn, and E. Moulines, *Diffusion Approximations and Control Variates for MCMC*, preprint, arXiv:1808.01665 [stat. ME], 2018.
[10] N. Brosse, A. Durmus, and E. Moulines, *The promises and pitfalls of stochastic gradient Langevin dynamics*, in Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, 2018, pp. 8278–8288.
[11] N. S. Chatterji, N. Flammarion, Y.-A. Ma, P. L. Bartlett, and M. I. Jordan, *On the theory of variance reduction for stochastic gradient Monte Carlo*, in Proceedings of Machine Learning Research, Vol. 80, 2018, pp. 764–773.
[12] T. Chen, E. B. Fox, and C. Guestrin, *Stochastic gradient Hamiltonian Monte Carlo*, in Proceedings of the 31st International Conference on International Conference on Machine Learning, ICML'14, 2014, pp. 1683–1691.
[13] A. S. Dalalyan, *Theoretical guarantees for approximate sampling from smooth and log-concave densities*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 79 (2017), pp. 651–676.
[14] A. S. Dalalyan and A. G. Karagulyan, *User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient*, Stochastic Process. Appl., 129 (2019), pp. 5278–5311.
[15] A. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, in Advances in Neural Information Processing Systems, 2014, pp. 1646–1654.
[16] P. Dellaportas and I. Kontoyiannis, *Control variates for estimation based on reversible Markov chain Monte Carlo samplers*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 74 (2012), pp. 133–161.
[17] H. Djellout, A. Guillin, and L. Wu, *Transportation cost-information inequalities and applications to random dynamical systems and diffusions*, Ann. Probab., 32 (2004), pp. 2702–2732.
[18] R. Douc, E. Moulines, P. Priouret, and P. Soulier, *Markov Chains*, Springer Ser. Oper. Res. Financ. Eng., Springer, Cham, 2018.
[19] P. Doukhan and S. Louhichi, *A new weak dependence condition and applications to moment inequalities*, Stochastic Process. Appl., 84 (1999), pp. 313–342.
[20] K. A. Dubey, S. J. Reddi, S. A. Williamson, B. Poczos, A. J. Smola, and E. P. Xing, *Variance reduction in stochastic gradient Langevin dynamics*, in Advances in Neural Information Processing Systems, 2016, pp. 1154–1162.
[21] A. Durmus and E. Moulines, *High-dimensional Bayesian inference via the unadjusted Langevin algorithm*, Bernoulli, 25 (2019), pp. 2854–2882.
[22] J. M. Flegal and G. L. Jones, *Batch means and spectral variance estimators in Markov chain Monte Carlo*, Ann. Statist., 38 (2010), pp. 1034–1070.
[23] N. Friel, A. Mira, and C. J. Oates, *Exploiting multi-core architectures for reduced-variance estimation with intractable likelihoods*, Bayesian Anal., 11 (2015), pp. 215–245.
[24] P. Glasserman, *Monte Carlo Methods in Financial Engineering*, Springer Science & Business Media, New York, 2013.
[25] E. Gobet, *Monte-Carlo Methods and Stochastic Processes*, CRC Press, Boca Raton, FL, 2016.
[26] T. E. Hanson, A. J. Branscum, and W. O. Johnson, *Informative g-priors for logistic regression*, Bayesian Anal., 9 (2014), pp. 597–612.
[27] S. G. Henderson, *Variance Reduction via an Approximating Markov Process*, Ph.D. thesis, Stanford University, 1997.
[28] R. Johnson and T. Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, in Advances in Neural Information Processing Systems, 2013, pp. 315–323.

[29] M. Ledoux, *The Concentration of Measure Phenomenon*, Surveys Math. Monogr., American Mathematical Society, Providence, RI, 2001.
[30] Y.-A. Ma, T. Chen, and E. Fox, *A complete recipe for stochastic gradient MCMC*, in Advances in Neural Information Processing Systems, 2015, pp. 2917–2925.
[31] K. Marton, *Bounding $\bar{d}$-distance by informational divergence: A method to prove measure concentration*, Ann. Probab., 24 (1996), pp. 857–866.
[32] S. Menozzi and V. Lemaire, *On some non asymptotic bounds for the Euler scheme*, Electron. J. Probab., 15 (2010), pp. 1645–1681.
[33] A. Mira, R. Solgi, and D. Imparato, *Zero variance Markov chain Monte Carlo for Bayesian estimators*, Stat. Comput., 23 (2013), pp. 653–662.
[34] T. Nagapetyan, A. B. Duncan, L. Hasenclever, S. J. Vollmer, L. Szpruch, and K. Zygalakis, *The True Cost of Stochastic Gradient Langevin Dynamics*, preprint, arXiv:1706.02692 [math. NA], 2017.
[35] R. Nickl and B. M. Pötscher, *Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov- and Sobolev-type*, J. Theoret. Probab., 20 (2007), pp. 177–199.
[36] C. J. Oates, M. Girolami, and N. Chopin, *Control functionals for Monte Carlo integration*, J. R. Stat. Soc. Ser. B Stat. Methodol., 79 (2017), pp. 695–718.
[37] F. Otto and C. Villani, *Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality*, J. Funct. Anal., 173 (2000), pp. 361–400.
[38] D. J. Rezende and S. Mohamed, *Variational Inference with Normalizing Flows*, preprint, arXiv:1505.05770 [stat. ML], 2015.
[39] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, New York, 1999.
[40] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363.
[41] N. L. Roux, M. Schmidt, and F. R. Bach, *A stochastic gradient method with an exponential convergence rate for finite training sets*, in Advances in Neural Information Processing Systems, Vol. 25, 2012, pp. 2663–2671.
[42] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo Method*, John Wiley & Sons, Hoboken, NJ, 2016.
[43] R. Salakhutdinov and A. Mnih, *Bayesian probabilistic matrix factorization using Markov Chain Monte Carlo*, in Proceedings of the 25th International Conference on Machine Learning (ICML-08), 2008, pp. 880–887.
[44] M. Talagrand, *Transportation cost for Gaussian and other product measures*, Geom. Funct. Anal., 6 (1996), pp. 587–600.
[45] Y. W. Teh, A. H. Thiery, and S. J. Vollmer, *Consistency and fluctuations for stochastic gradient Langevin dynamics*, J. Mach. Learn. Res., 17 (2016), pp. 193–225.
[46] M. Welling and Y. W. Teh, *Bayesian learning via stochastic gradient Langevin dynamics*, in Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 681–688.
[47] D. Zou, P. Xu, and Q. Gu, *Subsampled stochastic variance-reduced gradient Langevin dynamics*, in Proceedings of the International Conference on Uncertainty in Artificial Intelligence, 2018.