National Research University Higher School of Economics

Faculty of Mathematics

As a manuscript

Sergey Samsonov

Concentration inequalities for functionals of Markov chains with applications to variance reduction in MCMC

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

Academic Supervisor
Doctor of Science in Computer Science
Alexey Naumov

Moscow – 2024

Contents

Introduction

With the outstanding interest in the development of novel methods in machine learning, there is growing interest in mathematical tools that provide a framework for understanding and evaluating the performance of algorithms when the observable sample size is finite. Various results based on the concentration of measure phenomenon [1, 2] have proved to be the right instrument for obtaining non-asymptotic guarantees for various algorithms in the fields of reinforcement learning [3], optimization [4], learning theory [5], Monte-Carlo and Markov Chain Monte Carlo methods (MCMC, [6, 7]), and many others. Concentration inequalities for functionals of independent random variables or martingales are relatively well understood, as seen in [2, 8, 9]. At the same time, the situation is different when considering concentration inequalities for functions of dependent random variables. While a wealth of results exists for weakly dependent processes with different types of mixing conditions [10, 11], their application even to the natural setting of additive functionals of Markov chains is challenging. In particular, they are either not quantitative or not precise enough in terms of important problem characteristics, such as the variance of the additive functional in the Bernstein type inequalities (see Chapter 1 for the relevant definitions). This drawback is shared by many existing results obtained specifically for functionals of Markov chains [12–15]. However, it is Markovian stochasticity that appears in the vast majority of machine learning algorithms. Markov chains naturally arise in the non-asymptotic analysis of algorithms in the fields of stochastic approximation [16, 17] or reinforcement learning [3, 18].

In Chapter 1 of this thesis, we obtain new counterparts of the classical Rosenthal and Bernstein inequalities for geometrically ergodic Markov chains with explicit dependence on the mixing time of the underlying chains. We consider an additive functional

$$S_n = \sum_{\ell=0}^{n-1} \{g(X_\ell) - \pi(g)\}, \tag{1}$$

where g is an integrable measurable function and $(X_\ell)_{\ell=0}^{\infty}$ is a Markov chain with a Markov kernel P, which admits π as unique invariant distribution. We obtain concentration inequalities for the additive functional S_n, similar to those presented in [12, 14, 19, 20]. We refine the dependency of the new estimates on the variance of S_n and the mixing time of the underlying chain. Our proof is based on the cumulant expansion techniques outlined in [21] and the Leonov-Shiryaev formula [22] relating moments and cumulants.

In the subsequent parts of the thesis, we apply concentration inequalities to the non-asymptotic analysis of variance reduction techniques [23, 24], and propose new variance reduction methods for sequences of dependent random variables. The primary aim of variance reduction is to reduce the stochastic error in Monte Carlo estimates. Classical contributions to this field, including those by [25] and [26], have extensively explored variance reduction techniques, with a primary focus on modeling based on sequences of independent and identically distributed (i.i.d.) random variables (see e.g. [27]). However, in many scenarios, generating i.i.d. observations is not feasible, especially in cases of high problem dimension, and statistical inference must rely on dependent observations. These observations often form a Markov chain, as is the case of MCMC algorithms [6]. Furthermore, the application of variance reduction extends to optimization methods and

reinforcement learning, see e.g. [28–31], and references therein.

In Chapter 2, we propose a practical approach to variance reduction for additive functionals of dependent random variables. This approach extends the one introduced in [32] and is applicable to a broader class of Markov chains satisfying the ergodicity condition in the first-order Kantorovich-Wasserstein metric, and to sequences of dependent random variables satisfying the covariance stationarity assumption. The proposed method is based on using the control variates together with minimizing the empirical estimate of the respective asymptotic variance. We provide estimates for the rate of decrease in excess asymptotic variance with the growth of the training sample size. The proposed approach has been applied to MCMC estimates based on the Stochastic Gradient Langevin Dynamics (SGLD, [33]).

In Chapter 3, we consider the problem of variance reduction for additive functionals of Markov chains in the setting where the analytical expression for the invariant distribution of the underlying chain is unknown. In such a setting, we suggest a variance reduction approach based on discrete-time martingale representation, which generalizes the control variates using orthogonal polynomials expansion [34]. This approach does not require knowledge of the chain's stationary distribution or its specific structure. We analyze the algorithm under a normal noise model (see Section 3.4), which particularly covers the celebrated Unadjusted Langevin Algorithm [35–37].

Goals and objectives of the study

The goal of the study is to obtain a new analytical tools for studying concentration properties of functionals of Markov chains and to apply them for theoretical analysis of post-processing methods for MCMC estimates, which are based on control variates. To solve this problem, we consider the following steps:

1. Derive upper bound on cumulants of additive functionals of geometrically ergodic Markov chains, tracing explicit dependence on the parameters of the underlying Markov kernel;

2. Use the bound above to get new counterparts of Rosenthal inequality and Bernstein inequality, keeping precise dependence on the variance of S_n from (1) and mixing time of the kernel;

3. Generalize the above versions of Rosenthal inequality for quadratic forms of functions of Markov chains, converging geometrically fast to the invariant distribution in terms of first-order Kantorovich-Wasserstein metric;

4. Develop a method for selecting the control variates to adjust MCMC estimates, based on minimizing a certain estimate of the asymptotic variance. Study the statistical properties of the suggested method;

5. Develop a variance reduction method for additive functionals of Markov chain, which does not require to know analytically the invariant distribution of the underlying chain. Provide bounds on the variance of adjusted estimates compared to the variance of non-adjusted functional in the normal noise model described in Section 3.4.

Scientific novelty of the results

All results submitted for defense are new. New concentration inequalities of the Rosenthal and Bernstein types have been obtained for additive functionals of Markov chains. These inequalities generalize known estimates in the literature. Moreover, this work provides an original extension of Bernstein inequality to Markov kernels under the condition of ergodicity in the general weighted Kantorovich-Wasserstein metric. Additionally, a novel non-asymptotic analysis of the performance of several variance reduction methods for MCMC algorithms has been conducted, resulting in estimates for the rate of decrease in excess asymptotic variance with the growth of the training sample size. Suggested method of constructing control variates based on discrete martingale decomposition is new and can be used in several settings, when classical techniques, in particular, the ones based on Stein operator, are not directly applicable.

Theoretical and practical significance of the results

The presented results have both theoretical and methodological significance. The theoretical findings introduce new concentration inequalities for additive functionals of Markov chains, which may be valuable for studying Markov Chain Monte Carlo (MCMC) methods. From a methodological perspective, new variance reduction techniques for MCMC algorithms are proposed, which can be applied, in particular, in Bayesian statistics.

Methodology and research methods

The work extensively employs the analytical tools of probability theory, particularly the coupling methods and the method of cumulants, in particular, relations between cumulant bounds and concentration inequalities discussed in Chapter 1. The proofs of the main results rely on the theory of Markov chains and concentration inequalities.

Publications based on research results

The main contributions of the thesis have been published in three peer-reviewed journal articles [38–40]. All three articles are included in the Scopus and Web of Science databases.

1. A.Durmus, E. Moulines, A. Naumov, S. Samsonov. *Probability and Moment Inequalities for Additive Functionals of Geometrically Ergodic Markov Chains*, Journal of Theoretical Probability, 2024. https://doi.org/10.1007/s10959-024-01315-7;

2. D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, S. Samsonov. *Variance reduction for dependent sequences with applications to stochastic gradient MCMC*, SIAM/ASA Journal on Uncertainty Quantification, 9(2), 507-535, 2021. https://doi.org/10.1137/19M1301199;

3. D. Belomestny, E., Moulines, S. Samsonov. *Variance reduction for additive functionals of Markov chains via martingale representations*, Statistics and Computing, 32(1), 16, 2022. https://doi.org/10.1007/s11222-021-10073-z

Approbation of work

Main results of the thesis were presented at the following conferences, schools, and seminars:

1. Winter school and conference "New frontiers in high-dimensional probability and statistics 2", Moscow, February 22 – 23, 2019. Talk: "Concentration inequalities for functionals of Markov Chains with applications to variance reduction";

2. Conference "Structural Inference in High-Dimensional Models 2". Pushkin, Saint-Petersburg, 26 – 30 August 2019. Poster: "Variance Reduction for Dependent Sequences via Empirical Variance Minimisation";

3. Research seminar "Structural Learning", Faculty of Computer science, HSE, Moscow, October 15, 2019. Talk: "Variance reduction for dependent sequences with applications to Stochastic Gradient MCMC";

4. HSE-Yandex Autumn School on Generative Models, Moscow, November 26 – 29, 2019. Poster: "Variance reduction for MCMC algorithms";

5. Winter school "Math of Machine Learning 2020", Sochi, Sirius, February 19 – 22, 2020. Poster: "Variance Reduction for Dependent Sequences via Empirical Variance Minimisation";

6. City seminar on probability theory and mathematical statistics, Saint-Petersburg, POMI RAS, October 09, 2020. Talk: "Variance reduction methods for MCMC algorithms";

7. Conference "New Trends in Mathematical Stochastics", 30.08.2021-03.09.2021, talk "Probability and moment inequalities for additive functionals of geometrically ergodic Markov chains";

8. Research seminar "Structural Learning", Faculty of Computer science, HSE, Moscow, February 28, 2023. Talk: "Rosenthal type inequalities for Markov chains and their applications to Linear Stochastic Approximation".

Theses submitted for defense

1. In Chapter 1 we obtain new counterparts of Rosenthal and Bernstein inequalities for additive functionals of ergodic Markov chains that converge to the stationary distribution exponentially fast either in V-total variation norm or in the Kantorovich-Wasserstein semi-metric. The proof method we employ is based on the cumulant expansion techniques and the connections between cumulants and centered moments established through the Leonov-Shiryaev formula.

2. In Chapter 2 we propose an extension of the variance reduction method using control variates for the case of dependent random sequences that satisfy the covariance stationarity assumption. We obtain estimates for the rate of decrease in excess asymptotic variance with the growth of the training sample size. We derive concentration inequalities for quadratic forms

of functions of Markov chains satisfying the contraction condition in the Kantorovich-Wasserstein metric and apply these results to MCMC estimates based on the Stochastic Gradient Langevin Dynamics (SGLD).

3. In Chapter 3 we propose a novel variance reduction approach for additive functionals of Markov chains based on a discrete-time martingale representation. We study the variance reduction achieved by our method in a special setting of the normal noise model, covering the Unadjusted Langevin Algorithm (ULA), and show its gain over the non-adjusted estimates without variance reduction.

Reliability of results

All results of the dissertation are justified by mathematical proofs. The findings of the dissertation were presented at conferences and scientific seminars.

Structure and scope of work

The thesis consists of introduction, notation section, three chapters, conclusion, and bibliography. The thesis is 113 pages long, including 105 pages of the text, 2 tables, and 12 figures. The bibliography is 8 pages long and includes 119 items.

Author's personal contribution

The author's contribution is primary in the results of Chapter 1 and Chapter 3. Presented results of these sections were obtained personally by the author, apart from the result of Theorem 5. The latter one is the result of a joint work of the doctoral candidate and other co-authors of [38]. For completeness, Chapter 2 includes results obtained jointly with co-authors, namely the results of Section 2.4: Algorithm 1 and Theorem 6. They are obtained jointly by the doctoral candidate and other co-authors of [39]. The author's primary contribution to Chapter 2 are the results on the concentration of quadratic forms for Markov chains under contractive condition in the first-order Kantorovich-Wasserstein metric with applications to Stochastic Gradient Langevin Dynamics (SGLD). These results are presented in Section 2.5 and Section 2.6. Furthermore, the proof idea for Proposition 2 is attributed to A. Naumov.

Notations and definitions.

We assume by default that all random variables in Chapters 1 and 2 of this thesis take values in a complete separable metric space (X, d), equipped with Borel sigma-algebra $\mathcal{X}$. Situations requiring $\mathsf{X} = \mathbb{R}^d$ are specifically mentioned in the text. For a (signed) measure ξ on $(\mathsf{X}, \mathcal{X})$ and a measurable function $g : \mathsf{X} \to \mathbb{R}$, we use a notation

$$\xi(g) = \int_{\mathsf{X}} g(x)\xi(\mathrm{d}x)\,.$$

For a measurable function $V : \mathsf{X} \to [1, \infty)$, we define L_V as a set of all measurable functions $g : \mathsf{X} \to \mathbb{R}$, such that $\|g\|_V = \sup_{x\in\mathsf{X}}\{\frac{|g(x)|}{V(x)}\} < \infty$. The V-norm (also referred to as V-total variation norm) of a signed measure ξ is defined as

$$\|\xi\|_V = \int_{\mathsf{X}} V(x)\,|\xi|\,(\mathrm{d}x)\,,$$

where $|\xi|$ is the total variation of ξ. In the case $V \equiv 1$, the V-norm is the total variation norm and is denoted by $\|\cdot\|_{\mathrm{TV}}$. Equivalently, we can define $\|\xi\|_V = \sup\{\xi(g) : \|g\|_V \le 1\}$ (see [41, Theorem D.3.2] for details). We write $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

When we consider the Markov chain $\{X_n\}_{n\in\mathbb{N}_0}$ with initial distribution ξ and Markov kernel P on $(\mathsf{X}, \mathcal{X})$, we assume without loss of generality that $\{X_n\}_{n\in\mathbb{N}_0}$ is the associated canonical process defined on the canonical space $(\mathsf{X}^{\mathbb{N}_0}, \mathcal{X}^{\otimes\mathbb{N}_0})$. For any probability measure ξ on $(\mathsf{X}, \mathcal{X})$ we denote by P_ξ and E_ξ respective probability and expected value under initial distribution ξ. We set $\mathsf{E}_x = \mathsf{E}_{\delta_x}$ and $\mathsf{P}_x = \mathsf{P}_{\delta_x}$ for all $x \in \mathsf{X}$. We use the following definitions, which can be found e.g. in [41]:

Definition 1. *A set C is called a small set for the Markov kernel* P*, if there exist $\epsilon \in (0, 1]$, $m \in \mathbb{N}$, and a probability measure ν, such that for all $x \in C$ and $\mathsf{A} \in \mathcal{X}$, it holds that*

$$\mathrm{P}^m(x, \mathsf{A}) \ge \epsilon\nu(\mathsf{A})\,. \tag{2}$$

The set C is then said to be an $(m, \epsilon\nu)$-small set.

When we do not need to mention the associated measure ν above, we simply write that C is an (m, ϵ)-small set. For any $\mathsf{A} \in \mathcal{X}$ we define the return time σ_A of the set A as $\sigma_\mathsf{A} = \inf\{n \ge 1 : X_n \in \mathsf{A}\}$. Set A is called an *accessible set* for the kernel P, if $\mathsf{P}_x(\sigma_\mathsf{A} < \infty) > 0$ for any $x \in \mathsf{X}$. Based on the definitions above, we introduce the definition of *strongly aperiodic* kernel:

Definition 2. *The Markov kernel* P *is called strongly aperiodic, if it admits an accessible $(1, \epsilon\nu)$ - small set C with $\nu(C) > 0$.*

Am important class of Markov kernels P considered in Chapter 1 is the class of V-geometrically ergodic kernels with $V : \mathsf{X} \to [1, \infty)$ being a measurable function.

Definition 3. *The Markov kernel* P *is called V-geometrically ergodic, if* P *admits a unique invariant distribution π, such that $\pi(V) < \infty$, and there exist constants $c > 0$ and $\rho \in (0, 1)$, such that for all $x \in \mathsf{X}$*

$$\|\delta_x\mathrm{P}^n - \pi\|_V \le c\rho^n V(x)\,. \tag{3}$$

If P is V-geometrically ergodic, we also say with slight abuse of terminology that the corresponding Markov chain $\{X_n\}_{n\in\mathbb{N}_0}$ is V-geometrically ergodic.

For two probability measures ξ and ξ' on $(\mathsf{X},\mathcal{X})$, we say that a probability measure ν on $(\mathsf{X}^2,\mathcal{X}^{\otimes 2})$ is a coupling of ξ and ξ', if for each $\mathsf{A}\in\mathcal{X}$, $\nu(\mathsf{A}\times\mathsf{X})=\xi(\mathsf{A})$ and $\nu(\mathsf{X}\times\mathsf{A})=\xi'(\mathsf{A})$. Denote by $\Pi(\xi,\xi')$ the set of couplings of ξ and ξ' on $(\mathsf{X},\mathcal{X})$. Let $\mathsf{c}:\mathsf{X}\times\mathsf{X}\to\mathbb{R}_+$ be a lower semi-continuous symmetric function such that $\mathsf{c}(x,x')=0$ for $x=x'$, and there exists $p_{\mathsf{c}}\in\mathbb{N}$ such that for any $x,x'\in\mathsf{X}$, $(\mathsf{d}(x,x')\wedge 1)^{p_{\mathsf{c}}}\le\mathsf{c}(x,x')$. Then the Kantorovich-Wasserstein semi-metric $\mathbf{W}_{\mathsf{c}}(\xi,\xi')$, associated with the cost function c, is defined as

$$\mathbf{W}_{\mathsf{c}}(\xi,\xi')=\inf_{\nu\in\Pi(\xi,\xi')}\int_{\mathsf{X}\times\mathsf{X}}\mathsf{c}(x,x')\nu(\mathrm{d}x\mathrm{d}x')\,. \tag{4}$$

In English-language literature, this object is commonly referred to as the Wasserstein semi-metric. $\mathbf{W}_{\mathsf{c}}(\xi,\xi')$ is a semi-metric in the sense that it satisfies the axioms of symmetry, non-negativity, and identity, but, generally speaking, it does not satisfy the triangle inequality. Denote a set of probability measures on X as $\mathbb{M}_1(\mathsf{X})$, and for $p\ge 1$, let $\mathbb{S}_p(\mathsf{X},\mathsf{d})$ be probability measures with finite p-th moment:

$$\mathbb{S}_p(\mathsf{X},\mathsf{d}):=\{\xi\in\mathbb{M}_1(\mathsf{X}):\int_{\mathsf{X}}\mathsf{d}^p(x,x')\xi(\mathrm{d}x')<\infty\text{ for all }x\in\mathsf{X}\}\,.$$

For $p\ge 1$ and $\xi,\xi'\in\mathbb{S}_p(\mathsf{X},\mathsf{d})$, define the p-th order Kantorovich-Wasserstein distance between ξ and ξ' as

$$\mathbf{W}_{\mathsf{d},p}(\xi,\xi'):=\inf_{\nu\in\Pi(\xi,\xi')}\Big\{\int_{\mathsf{X}\times\mathsf{X}}\mathsf{d}^p(x,x')\,\nu(\mathrm{d}x\mathrm{d}x')\Big\}^{1/p}.$$

Note that $\mathbf{W}_{\mathsf{d},p}(\xi,\xi')$ is a distance on $\mathbb{S}_p(\mathsf{X},\mathsf{d})$. For a measurable function $\mathcal{W}:\mathsf{X}\to[1,\infty)$, set $\bar{\mathcal{W}}(x,y)=(\mathcal{W}(x)+\mathcal{W}(y))/2$, and for $\beta\ge 0$, define its weighted Lipschitz norm as

$$[f]_{\beta,\mathcal{W}}=\max\left\{\sup_{x,x'\in\mathsf{X},\,x\neq x'}\frac{|f(x)-f(x')|}{\mathsf{c}^{1/2}(x,x')\bar{\mathcal{W}}^{\beta}(x,x')},\sup_{x\in\mathsf{X}}\frac{|f(x)|}{\mathcal{W}^{\beta}(x)}\right\}. \tag{5}$$

The corresponding class of functions is denoted by $\mathcal{L}_{\beta,\mathcal{W}}=\{f:\mathsf{X}\to\mathbb{R}:[f]_{\beta,\mathcal{W}}<\infty\}$. For a function $h:\mathsf{X}\to\mathbb{R}$ we define its Lipschitz norm as $\|h\|_{\mathsf{Lip}}:=\sup_{x\neq y\in\mathsf{X}}\{|h(y)-h(x)|/\mathsf{d}(x,y)\}$. We denote by $\mathrm{Lip}_{\mathsf{d}}(L)$ and $\mathrm{Lip}_{b,\mathsf{d}}(L,\mathsf{B})$ the class of Lipschitz (resp. bounded Lipschitz) functions on X with $\|h\|_{\mathsf{Lip}}\le L$ (resp. $\|h\|_{\mathsf{Lip}}\le L$ and $|h|_\infty\le\mathsf{B}$).

Denote for any $q\in[1,\infty)$ the $2q$-th moment of the standard Gaussian distribution on $\mathbb{R}$ by

$$\mathrm{m}_{\mathrm{G},q}=(2q)!/(q!2^q)=2^q\mathbf{\Gamma}((2q+1)/2)/\pi^{1/2}\,, \tag{6}$$

where $\mathbf{\Gamma}$ is the Gamma function. For a multi-index $\mathbf{k}=(k_1,\ldots,k_d)$ we use the notation $\|\mathbf{k}\|=\max_{i\in\{1,\ldots,d\}}k_i$, $|\mathbf{k}|=\sum_{i=1}^d k_i$ and $\mathbf{k}!:=k_1!\ldots k_d!$. For a function $f:\mathbb{R}^d\to\mathbb{R}$ we write $\nabla f(x)$ and $\nabla^2 f(x)$ for its gradient and Hessian at point x, respectively. In the present text, the following abbreviations are frequently used: "w.r.t." stands for "with respect to", "i.i.d." stands for "independent and identically distributed".

Chapter 1

Rosenthal and Bernstein inequalities for additive functionals of geometrically ergodic Markov chains

1.1. Introduction and problem setup

In the present chapter of the thesis we consider the concentration properties of additive functionals of Markov chains, that is, sums of the form

$$S_n = \sum_{\ell=0}^{n-1} \{g(X_\ell) - \pi(g)\}\,, \tag{1.1}$$

where $\{X_\ell\}_{\ell=0}^{\infty}$ is a Markov chain with the Markov kernel P, which has a unique invariant distribution π, and $g : \mathsf{X} \to \mathbb{R}$ is a measurable function, satisfying $\pi(|g|) < \infty$. We assume that X_0 follows some distribution ξ, which might be different from π. We aim to recover counterparts of concentration inequalities of Bernstein and Rosenthal type for S_n. This chapter is based on the results published in [38].

We begin with a short discussion on the mentioned inequalities for sums of independent random variables. Assume that $(Y_\ell)_{\ell=0}^{n-1}$ are independent random variables with $\mathsf{E}[Y_\ell] = 0$ and set

$$\bar{S}_n = \sum_{\ell=0}^{n-1} Y_\ell\,. \tag{1.2}$$

Assume that for some $c > 0$, for all $\ell \in \{0, \dots, n-1\}$ and integers $k \geq 3$ it holds that $|\mathsf{E}[Y_\ell^k]| \leq (k!/2)\,\mathrm{Var}(Y_\ell)c^{k-2}$. This condition is known as *Bernstein's condition* and yields the corresponding inequality: for any $t > 0$ and $n \in \mathbb{N}$,

$$\mathsf{P}(|\bar{S}_n| \geq t) \leq 2\exp\left\{-\frac{t^2/2}{\mathrm{Var}(\bar{S}_n) + ct}\right\}. \tag{1.3}$$

Bernstein's condition and inequality can be further generalized by cumulant expansion, as suggested by [21]. Recall that the k-th cumulant of a random variable Y is defined as

$$\Gamma_k(Y) = \frac{1}{\mathrm{i}^k}\,\frac{\mathrm{d}^k}{\mathrm{d}t^k}\big(\log \mathsf{E}[\mathrm{e}^{\mathrm{i}tY}]\big)\Big|_{t=0}.$$

Bentkus in [21] provides the generalization of Bernstein's inequality under the following condition: if there exist $\gamma \geq 0$ and $B \geq 0$ such that for any $k \in \mathbb{N}\,, k \geq 2$, it holds

$$|\Gamma_k(\bar{S}_n)| \leq (k!/2)^{1+\gamma}\,\mathrm{Var}(\bar{S}_n)B^{k-2}\,, \tag{1.4}$$

then for all $t \geq 0$, it holds that

$$\mathsf{P}(|\bar{S}_n| \geq t) \leq 2\exp\left\{-\frac{t^2/2}{\mathrm{Var}(\bar{S}_n) + B^{1/(1+\gamma)}t^{2-1/(1+\gamma)}}\right\}. \tag{1.5}$$

It can be shown that the condition (1.4) also follows from some generalizations of the Bernstein

condition, see, for example, [42, theorem 3.1]. Note that Bernstein's inequality implies an exponential rate of decay for the distribution tails of $\bar{S}_n$. Conversely, it is important to explore properties of $\bar{S}_n$ being a sum of random variables, which have only a finite number of moments. In this scenario, one is typically interested to establish the moment bounds for $\bar{S}_n$, and the following version of Rosenthal's inequality is of great importance (see [43] and also the original paper [44]): for $q \geq 2$,

$$\mathsf{E}[|\bar{S}_n|^q] \leq C^q \big\{ q^{q/2} \operatorname{Var}(\bar{S}_n)^{q/2} + q^q \mathsf{E}[|\max_{\ell \in \{0,\dots,n-1\}} Y_\ell|^q] \big\}, \tag{1.6}$$

where C is an absolute constant. If (1.6) is satisfied for all $q \geq 2$, one can show that $\bar{S}_n$ satisfies the Bernstein-type inequality, provided that $|\max_{\ell \in \{1,\dots,n\}} Y_\ell|$ has a finite Orlicz ψ_1-norm, see [12].

1.2. Contributions

In the present chapter of the thesis we obtain counterparts of Rosenthal's and Bernstein's inequalities (1.6) and (1.5) for the additive functionals of the form (1.1). We consider the Markov kernels P, such that their iterates $\xi \mathrm{P}^n$ converge to the stationary distribution π at the exponential rate either in the V-total variation norm or in the Kantorovich-Wasserstein semi-metric (see, respectively, Section 1.4 and Section 1.5). The proof method we employ is based on the cumulant expansion techniques outlined in [21], [42], and further developed in [45] for weakly dependent processes. In the stationary case (when the initial distribution ξ coincides with π), a key step in the proof involves bounding the centered moments associated with $\{g(X_\ell)\}_{\ell=0}^{n-1}$. This connection is established through the Leonov-Shiryaev formula [22]. Our technique allows to obtain results with explicit and computable constants. Finally, we also cover the case of an arbitrary initial distribution ξ. Results for the non-stationary case are derived using coupling methods, see [46] and [41, Chapter 19].

1.3. Literature review

Concentration inequalities for additive functionals of Markov chains have been studied using a wealth of different techniques. In the list of papers below we provide a selection of existing results and related theoretical tools. A number of results [47–49] are devoted to the inequalities of Azuma-Hoeffding type for (1.1). Probability bounds for Markov kernels that are contractive with respect to Kantorovich-Wasserstein distance are presented in [50], yet the results of [50] require to verify additional conditions, involving quantities such as *granularity* and *local dimension*, that are difficult to evaluate in most applications. The authors in [51–53] establish Hoeffding and Bernstein inequalities using spectral methods for Markov chains for bounded functions g under the assumption that P admits a positive spectral gap. However, respective bounds depend on a proxy for the variance, and not the exact variance of S_n. Note also that the V-geometric ergodicity of P does not necessarily imply the existence of a spectral gap (see [54]).

A popular approach, developed in [12, 14, 19, 20, 55, 56], is to use regenerative decomposition to obtain moment bounds and Bernstein inequalities under the geometric ergodicity assumption (see Section 1.4). These techniques are based on the Numellin splitting construction [57] and [58], which allows to split the sum S_n into a random number of one-dependent blocks of random lengths. The closest counterparts of our results are the ones obtained in [14] and [20]. [14,

Theorem 1] provides a Bernstein-type inequality for a V-geometrically ergodic strongly aperiodic Markov kernels (see Definition 2) and unbounded functions. [20, Theorem 1] extends the result to aperiodic Markov chains, but is restricted to bounded functions and does not provide explicit constants. Moreover, mentioned results can't be applied to the setting of Markov chains, which are geometrically ergodic in a sense of Kantorovich-Wasserstein semi-metric. Hence, they do not cover the results presented in Section 1.5.

Moment bounds and Bernstein-type inequalities were also obtained under different weak-dependence or mixing conditions; see [10, 45, 59]. These results are in general not directly comparable with the ones presented in this chapter, because the bounds depend on different types of weak-dependence or mixing coefficients instead of drift conditions and local minorization conditions.

1.4. V-geometrically ergodic Markov chains

First, we consider the case where the Markov kernel P is V-geometrically ergodic (see Definition 3). Namely, we impose the following assumptions on P:

A 1. *There exist a measurable function $V : \mathsf{X} \to [\mathrm{e}, \infty)$, $\lambda \in (0,1)$, and $b \geq 0$ such that for any $x \in \mathsf{X}$, $\mathrm{P}V(x) \leq \lambda V(x) + b$.*

A 2. *There are an integer $m \geq 1$, $\epsilon \in (0,1)$, and $d \geq 0$, such that the level set $\{x \in \mathsf{X} : V(x) \leq d\}$ is (m,ϵ)-small and $\lambda + 2b/(1+d) < 1$. Here λ and b are defined in **A** 1.*

The definition of (m,ϵ)-small set is provided in Definition 1. In contrast to the usual definition of Lyapunov functions in the Markov chain literature, we assume here that V takes values in $[\mathrm{e}, +\infty)$, rather than in $[1, +\infty)$. Such a choice avoids technical issues when considering norms associated with $W(x) = \log V(x)$. Under assumptions **A** 1 and **A** 2, the Markov kernel P is V-geometrically ergodic. Let π denote its unique invariant distribution. Then, according to [41, Theorem 19.4.1], for any probability measure ξ satisfying $\xi(V) < \infty$, and for all $n \in \mathbb{N}$,

$$\|\xi \mathrm{P}^n - \pi\|_{\mathrm{TV}} \leq \|\xi \mathrm{P}^n - \pi\|_V \leq c\{\xi(V) + \pi(V)\}\rho^n \,. \tag{1.7}$$

Explicit expressions for ρ and c depending on parameters from **A** 1 and **A** 2 can be found e.g. in [41, Theorem 19.4.1]. Before proceeding with our main results, we introduce some additional quantities. For each $q \in \mathbb{N}, u \in \{1, \ldots, q-1\}$ and $\gamma \geq 0$, we set

$$\mathrm{B}_\gamma(u,q) = \frac{(2q)!}{u!} \sum_{(k_1,\ldots,k_u) \in \mathscr{E}_{u,q}} \prod_{i=1}^{u} (k_i!)^{\gamma+2} \,, \tag{1.8}$$

where $\mathscr{E}_{u,q} = \{(k_1, \ldots, k_u) \in \mathbb{N}^u : \sum_{i=1}^u k_i = 2q \,,\, k_i \geq 2\}$. We proceed with a Rosenthal-type bound for V-geometrically ergodic Markov chains, with the leading term of the bound being equal to $\mathrm{Var}_\pi(S_n)$ scaled by the corresponding moment of the Gaussian distribution $\mathrm{m}_{\mathrm{G},q}$ (see (6)). Here and further in this chapter we set $\bar{g}(x) = g(x) - \pi(g)$.

Theorem 1. *Assume **A** 1, **A** 2, and let $q \in \mathbb{N}$. Then, for any function $g \in \mathrm{L}_{V^{1/(2q)}}$,*

$$\mathsf{E}_\pi[|S_n|^{2q}] \leq \mathrm{m}_{\mathrm{G},q}\{\mathrm{Var}_\pi(S_n)\}^q + \mathsf{C}_0^{2q} \|\bar{g}\|^{2q}_{V^{1/(2q)}} \sum\nolimits_{u=1}^{q-1} \frac{\mathrm{B}_0(u,q) n^u}{\rho^{u/2} \log^{2q-u}(1/\rho)} \,, \tag{1.9}$$

where $\mathsf{C}_0 = 2c\pi(V)$*, and* c *is defined in* (1.7)*. Moreover, for any probability measure* ξ *on* $(\mathsf{X}, \mathcal{X})$ *satisfying* $\xi(V) < \infty$*, it holds that*

$$\mathsf{E}_\xi\big[\big|S_n\big|^{2q}\big] \le 2^{2q-1}\mathsf{E}_\pi\big[\big|S_n\big|^{2q}\big] + 2^{6q-1}\|\bar{g}\|^{2q}_{V^{1/(2q)}} c\{\xi(V) + \pi(V)\}\frac{q^{2q}}{\rho(\log(1/\rho))^{2q}}\,. \tag{1.10}$$

Our proof strategy is inspired by the cumulant method, explored in [45] for the case of weakly dependent sequences. Same type of techniques was studied for sums of independent random variables in [21] and [42]. The extension to arbitrary initial distribution (1.10) is done using the construction of the exact distributional coupling [41, Chapter 19]. It is worth noting that in our approach it is not necessary to assume that the Markov kernel P is strongly aperiodic, unlike in [14]. Recall that P being strongly aperiodic implies that **A** 2 holds with $m = 1$. The gap with strong aperiodicity was previously closed by [20], but only for the bounded functions g.

In the above results, we have set $q \in \mathbb{N}$ and considered a function $g \in \mathrm{L}_{V^{1/(2q)}}$. With these assumptions we can not control the exponential moments of S_n. Next, we consider the case of the function $g \in \mathrm{L}_{W^\gamma}$, where $W = \log V$ and $\gamma \ge 0$. In this case, in addition to the Rosenthal-type bound (1.9), we can formulate a counterpart of the Bernstein-type bound (1.5).

Theorem 2. *Assume* ***A*** *1,* ***A*** *2 and let* $\gamma \ge 0$*. Then for any* $g \in \mathrm{L}_{W^\gamma}$ *and* $t \ge 0$*, it holds that*

$$\mathsf{P}_\pi(|S_n| \ge t) \le 2\exp\left\{-\frac{t^2/2}{\mathrm{Var}_\pi(S_n) + \mathrm{J}^{1/(\gamma+3)}_{n,W^\gamma}\, t^{2-1/(\gamma+3)}}\right\}, \quad \textit{where } \mathrm{J}_{n,W^\gamma} \textit{ is given by} \tag{1.11}$$

$$\mathrm{J}_{n,W^\gamma} = \left(\frac{n\rho^{-1/2}\{\log(1/\rho)\}^{-1}\mathsf{C}_0^2\|\bar{g}\|^2_{W^\gamma}}{\mathrm{Var}_\pi(S_n)} \vee 1\right)\frac{2^{1+3\gamma}\gamma^{3\gamma}\mathsf{C}_0\|\bar{g}\|_{W^\gamma}}{\log(1/\rho)}\,. \tag{1.12}$$

Moreover, for any initial distribution ξ *with* $\xi(V) < \infty$*, it holds for* g *with* $\|\bar{g}\|_{W^\gamma} = 1$*, that*

$$\mathsf{P}_\xi(|S_n| \ge t) \le \mathsf{P}_\pi(|S_n| \ge \tfrac{t}{4}) + \big(\tfrac{\exp\{-h_1(\gamma,\rho)t^{\varpi\gamma}\}}{\rho^{1/2}} + \tfrac{\exp\{-h_2(\gamma)t^{\varpi\gamma}\}}{1-\rho}\big)c\{\xi(V) + \pi(V)\}\,, \tag{1.13}$$

where $\varpi_\gamma = 1/(1+\gamma)$*,* $h_1(\gamma,\rho) = \log(1/\rho)/(4^{1+\varpi_\gamma}\varpi_\gamma)$*, and* $h_2(\gamma) = (1+\gamma)/(2^{1+2\varpi_\gamma}\gamma)$*.*

Comparing (1.11) with (1.5), one can see that in the subexponential regime $t^{1/(\gamma+1)}$ is replaced by $t^{1/(\gamma+3)}$, as in [45]. The expression for J_{n,W^γ} is not distribution-free, moreover, compared to (1.5), it is possible that J_{n,W^γ} scales with n. This drawback is shared by other results obtained with cumulant expansion [45]. The proof of Bernstein-type bound for arbitrary initial distribution (1.13) uses the distributional coupling argument [41, Chapter 19].

One could compare the results of Theorem 2 with those obtained in [14, Theorem 1.1-1.3], provided that we additionally assume that P is strongly aperiodic. Then (1.11) provides a version of the Bernstein inequality with the exact constant 2 in front of the variance term and offers explicit dependence on parameters from **A** 1 and **A** 2, which improves over [14, Theorem 1.1] Yet the bound of [14] is tighter for large t and decreases with $\exp\{-t^{1/(1+\gamma)}\}$ compared to $\exp\{-t^{1/(3+\gamma)}\}$ in (1.11). It is worth noting that the exponent of the terms reflecting the dependence on the initial condition is $1/(1+\gamma)$ in (1.13), as in [14, Theorem 1.1]. But in contrast to [14, Theorem 1.1], dependence on the initial condition appears as a multiplicative factor, not in the exponential rate.

1.5. Geometrically ergodic Markov chains with respect to Kantorovich-Wasserstein semi-metric

In this chapter we extend the results obtained in Section 1.4 to the case of Markov kernels that are geometrically ergodic for a Kantorovich-Wasserstein semi-metric. This setting covers cases where the Markov kernel P is not irreducible. This means that we no longer impose the assumption **A** 2, and the regeneration methods studied in [14, 20] are no longer applicable. Examples of Markov chains with such kernels P are typical in an infinite-dimensional setting, see e.g. [60, 61] and [41, Chapter 20]. In Section 1.4, it was the combination of **A** 1 and **A** 2 that allowed us to show the existence and uniqueness of an invariant distribution π for P, together with the geometric rate convergence of the iterates $\xi \mathrm{P}^n$ to π. To extend these results without relying on **A** 2, we need to introduce several objects associated with the Kantorovich-Wasserstein semi-metric below. Consider the cost function $\mathsf{c} : \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$, which satisfies the following condition:

C 1. c *is a lower semi-continuous symmetric function such that* $\mathsf{c}(x, x') = 0$ *for* $x = x'$. *Also, there exists* $p_{\mathsf{c}} \in \mathbb{N}$ *such that for any* $x, x' \in \mathsf{X}$, $(\mathsf{d}(x, x') \wedge 1)^{p_{\mathsf{c}}} \leq \mathsf{c}(x, x') \leq 1$.

We say that K is a kernel coupling of P if for all $(x, x') \in \mathsf{X}^2$ and $\mathsf{A} \in \mathcal{X}$, $\mathrm{K}((x, x'), \mathsf{A} \times \mathsf{X}) = \mathrm{P}(x, \mathsf{A})$ and $\mathrm{K}((x, x'), \mathsf{X} \times \mathsf{A}) = \mathrm{P}(x', \mathsf{A})$. We next consider the following assumption, which weakens the small set condition **A** 2:

A 3. *There exist a kernel coupling* K *of* P, $m \in \mathbb{N}, \varepsilon \in (0, 1)$, $\kappa_{\mathrm{K}} \geq 1$ *such that*

$$\mathrm{K}\mathsf{c}(x, x') \leq \kappa_{\mathrm{K}} \mathsf{c}(x, x') \,, \qquad \mathrm{K}^m \mathsf{c}(x, x') \leq (1 - \varepsilon \mathbb{1}_{\bar{\mathsf{C}}}(x, x')) \mathsf{c}(x, x') \,, \tag{1.14}$$

where $\bar{\mathsf{C}} = \{V(x) \leq d\} \times \{V(x) \leq d\}$, *and parameter* d *satisfies* $\lambda + 2b/(1 + d) < 1$. *Here* λ *and* b *are given in* ***A*** *1, and* c *is defined in* ***C*** *1.*

Next we show that the assumptions **A** 1 and **A** 3 imply the existence and uniqueness of an invariant distribution π and the geometric convergence rate of $\xi \mathrm{P}^n$ to π for any initial distribution ξ for the semi-metric $\mathbf{W}_{\mathsf{c}^{1/2} \bar{V}^{1/2}}$.

Proposition 1. *Assume* ***A*** *1,* ***A*** *3, and* ***C*** *1. Then* P *admits a unique invariant probability measure* π *satisfying* $\pi(V) < \infty$. *Moreover, for all initial distributions* ξ *and* $n \in \mathbb{N}$,

$$\mathbf{W}_{\mathsf{c}} \left(\xi \mathrm{P}^n, \pi \right) \leq \mathbf{W}_{\mathsf{c}^{1/2} \bar{V}^{1/2}} \left(\xi \mathrm{P}^n, \pi \right) \leq c_1 \varrho^n \left[\xi(V^{1/2}) + \pi(V^{1/2}) \right] \,, \tag{1.15}$$

where quantitative expressions for $c_1 > 0$ *and* $\varrho \in (0; 1)$ *can be found in [38, Propostion 1].*

The first main result of this section is a Rosenthal-type inequality. Now we have to consider functions g from a weighted Lipschitz class $\mathcal{L}_{\beta, V}$ with suitable $\beta > 0$ (see the definition of $\mathcal{L}_{\beta, V}$ in (5)). Requiring only a finite V^β-norm as in Section 1.4 will not be enough. Similar to Theorem 1, the leading term of the bound is the stationary variance $\mathrm{Var}_\pi(S_n)$ multiplied by the moment of a Gaussian random variable.

Theorem 3. *Assume* ***A*** *1,* ***A*** *3,* ***C*** *1, and let* $q \in \mathbb{N}$. *Then for any* $g \in \mathcal{L}_{1/(4q),V}$,

$$\mathsf{E}_\pi[|S_n|^{2q}] \le \mathrm{m}_{\mathrm{G},q}\{\mathrm{Var}_\pi(S_n)\}^q + \mathsf{C}_1^{2q}[\bar{g}]^{2q}_{1/(4q),V} \sum\nolimits_{u=1}^{q-1} \frac{\mathrm{B}_0(u,q) n^u}{\varrho^{u/2}\{\log(1/\varrho)\}^{2q-u}}\,, \tag{1.16}$$

where $\mathrm{B}_0(u,q)$ *is defined in* (1.8), *and* $\mathsf{C}_1 = 4c_1\{\pi(V)\}^{1/2}$.

The proof is based on a suitable inequality for centered moments that is adapted to the Kantorovich-Wasserstein semi-metric. We can now extend this result to the non-stationary case in a similar way to (1.10), but using a coupling kernel instead of the distributional coupling argument. We present this result in the setting of Bernstein-type inequality.

Theorem 4. *Assume* ***A*** *1,* ***A*** *3,* ***C*** *1. Then, for any* $\gamma \ge 0$, $g \in \mathcal{L}_{1,W^\gamma}$, *and* $t \ge 0$,

$$\mathsf{P}_\pi(|S_n| \ge t) \le 2\exp\left\{-\frac{t^2/2}{\mathrm{Var}_\pi(S_n) + \mathfrak{J}_{n,W^\gamma}^{1/(\gamma+3)} t^{2-1/(\gamma+3)}}\right\}, \text{ where } \mathfrak{J}_{n,W^\gamma} \text{ is given by}$$

$$\mathfrak{J}_{n,W^\gamma} = \left(\frac{n\varrho^{-1/2}\{\log(1/\varrho)\}^{-1}\mathsf{C}_1^2(2\gamma)^{4\gamma}[\bar{g}]^2_{1,W^\gamma}}{\mathrm{Var}_\pi(S_n)} \vee 1\right)\frac{2(2\gamma)^{2\gamma}\mathsf{C}_1[\bar{g}]_{1,W^\gamma}}{\log(1/\varrho)}\,. \tag{1.17}$$

Theorem 5. *Under the assumptions of Theorem 4, for any probability measure* ξ *on* $(\mathsf{X},\mathcal{X})$ *satisfying* $\xi(V^{1/2}) < \infty$, *it holds, setting* $\varpi_\gamma = 1/(1+\gamma)$ *and* $\upsilon_\gamma = 1 \wedge (2\gamma)^{-1}$, *that*

$$\begin{aligned}\mathsf{P}_\xi(|S_n| \ge t) \le \mathsf{P}_\pi(|S_n| \ge t/2) &+ \exp\left(-\frac{\log(1/\varrho)t^{\varpi_\gamma}}{2^{3+\varpi_\gamma}[\bar{g}]^{\varpi_\gamma}_{1,W^\gamma}\varpi_\gamma}\right)c_1^{1/2}\{\pi(V^{1/2}) + \xi(V^{1/2})\}^{1/2}h_1(\varrho) \\ &+ \exp\left(-\frac{(1+\gamma)\upsilon_\gamma t^{\varpi_\gamma}}{2^{5+\varpi_\gamma}[\bar{g}]^{\varpi_\gamma}_{1,W^\gamma}\gamma}\right)c_1^{\upsilon_\gamma}\{\pi(V^{1/2}) + \xi(V^{1/2})\}^{\upsilon_\gamma}h_2(\varrho)\,,\end{aligned}$$

where $h_1(\varrho)$ *and* $h_2(\varrho)$ *are defined in [38, Theorem 12].*

To the author's knowledge, Theorem 4 establishes for the first time a Bernstein-type inequality for functions g from the weighted Lipschitz class $\mathcal{L}_{1,W^\gamma}$ without the condition **A** 2 or its analogs. Previous results of this type for unbounded functions and weakly dependent sequences [10] covers only functions g with linear growth. Note that the result of Theorem 5 allows for the same rate in the term reflecting the dependence on the initial conditions (that is, $\exp\{-t^{1/(1+\gamma)}\}$), as it was obtained before for V-geometrically ergodic setting considered in (1.13). This is the first result of this kind for kernels, which are geometrically ergodic in Kantorovich-Wasserstein semi-metric.

Chapter 2

Variance reduction for dependent sequences with applications to Stochastic Gradient MCMC

2.1. Introduction and problem statement

In this chapter of the thesis we propose and analyze a novel and practical variance reduction approach for additive functionals of dependent sequences. Key technical element of the analysis are the Rosenthal type inequalities similar to the ones derived in Chapter 1. This chapter is based on the results published in [39].

In the following, we outline the variance reduction setting for Monte Carlo methods. The primary objective of these methods is to compute the integral $\pi(f) = \int_{\mathsf{X}} f(x)\pi(\mathrm{d}x)$ w.r.t. a probability measure π for some integrable function $f : \mathsf{X} \mapsto \mathbb{R}$, defined on $(\mathsf{X}, \mathcal{X})$. Typically in practice, the space $\mathsf{X} \subseteq \mathbb{R}^d$ and a measure π admits a density w.r.t. the Lebesgue measure on $\mathbb{R}^d$. For simplicity, we also denote this density by π. In this case, the considered problem reduces to computing the integral

$$\int_{\mathsf{X}} f(x)\pi(x)\,\mathrm{d}x \tag{2.1}$$

It is known that for $\mathsf{X} = [0,1]^d$, approximating the integral (2.1) with a given accuracy using deterministic algorithms requires exponential (in the problem dimension d) number of function evaluations, see [62]. Therefore, as d increases, deterministic methods quickly become impractical. An alternative to quadrature formulas is given by the stochastic methods, relying on the Monte Carlo estimates and its modifications [26]. Note that $\pi(f) = \mathsf{E}_\pi[f(X)]$, where the random variable X is distributed according to π. Hence, due to the law of large numbers, a consistent estimate of $\pi(f)$ is given by

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

where $(X_k)_{k=0}^{N-1}$ are i.i.d. random variables with distribution π. Moreover, if $\pi(f^2) < \infty$, an asymptotic confidence interval for $\pi(f)$ with confidence level $1-\alpha$ can be constructed using the central limit theorem:

$$\left[\pi_N(f) - \mathsf{q}_{1-\alpha/2}\sqrt{\tfrac{\mathrm{Var}_\pi(f)}{N}}, \pi_N(f) + \mathsf{q}_{1-\alpha/2}\sqrt{\tfrac{\mathrm{Var}_\pi(f)}{N}}\right], \tag{2.2}$$

where $\mathsf{q}_{1-\alpha/2}$ is the respective quantile of the standard normal distribution. Hence, more accurate estimate for $\pi(f)$ can be achieved either by increasing N or by reducing the variance $\mathrm{Var}_\pi(f)$. One of the widely used methods for the latter purpose is the *control variates method*, see [23], [24].

The essence of the control variates method is to construct an easily computable random variable Y (a control variate), such that $\mathsf{E}[Y] = 0$, $\mathsf{E}[Y^2] < \infty$, and the variance of $f(X) - Y$ is small (recall that the random variable X is distributed according to π). Typically in practice [27], control variates of the form $Y = g(X)$ are used, where g satisfies the condition $\pi(g) = 0$. The practical application of the control variates method in this setting usually consists of two stages. First, we select a class of control variates $\mathcal{G} = \{g : \pi(g) = 0\}$. Then, based on i.i.d. variables $(X_k)_{k=0}^{n-1}$ with distribution π and some optimization problem, one selects a particular function

$\widehat{g}_n \in \mathcal{G}$. After that, for a new sample $(X'_k)_{k=0}^{N-1}$ from π, independent of $(X_k)_{k=0}^{n-1}$, we construct the estimate

$$\pi_N(f - \widehat{g}_n) = N^{-1} \sum\nolimits_{k=0}^{N-1} \{f(X'_k) - \widehat{g}_n(X'_k)\}\,. \tag{2.3}$$

Note that $\pi(f - \widehat{g}_n) = \pi(f)$, and a successful choice of $\widehat{g}_n$ can provide more accurate asymptotic confidence intervals for $\pi(f)$ of the form

$$\big[\pi_N(f - \widehat{g}_n) - \mathsf{q}_{1-\alpha/2} \sqrt{\tfrac{\mathrm{Var}_\pi(f-\widehat{g}_n)}{N}}, \pi_N(f - \widehat{g}_n) + \mathsf{q}_{1-\alpha/2} \sqrt{\tfrac{\mathrm{Var}_\pi(f-\widehat{g}_n)}{N}}\big]\,. \tag{2.4}$$

The potential benefit from the control variates method depends on how the length of the asymptotic confidence interval in the modified experiment (2.4) scales compared to that in (2.2).

2.2. Control variates for dependent observations

Generating i.i.d. random variables with distribution π is often either impossible or computationally inefficient [7]. This situation is typical for $\mathsf{X} = \mathbb{R}^d$ when the dimension d is high. In such cases, it is often easier to construct a sequence of dependent random variables $(X_k)_{k=0}^{\infty}$, with distribution of X_k converging to π as k increases. Assuming that the central limit theorem holds, the asymptotic confidence interval for $\pi(f)$ in such an experiment can be written as

$$\Big[\pi_N(f) - \mathsf{q}_{1-\alpha/2} \sqrt{\tfrac{V_\infty(f)}{N}}, \pi_N(f) + \mathsf{q}_{1-\alpha/2} \sqrt{\tfrac{V_\infty(f)}{N}}\Big], \tag{2.5}$$

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

$$V_\infty(f) := \lim_{N\to\infty} N \cdot \mathsf{E}[\big(\pi_N(f) - \pi(f)\big)^2]\,. \tag{2.6}$$

Hence, a natural objective of variance reduction methods in the outlined setting of dependent sequences is to design experiment that achieve a lower asymptotic variance $V_\infty(\cdot)$.

Dependent sequences $(X_k)_{k=0}^{\infty}$ are most commonly constructed using Markov Chain Monte Carlo (MCMC) algorithms. Such algorithms construct $(X_k)_{k=0}^{\infty}$ as a Markov chain with a unique invariant distribution π, and then estimate $\pi(f)$ based on the ergodic average $\pi_N(f)$. The latter estimates $\pi_N(f)$ may have large variance due to correlations between the neighboring elements of the chain. A notably popular group of MCMC algorithms are based on Langevin dynamics, see [36, 37, 63]. Probably the most popular of them is the SGLD [33] algorithm, see Section 2.6. However, there are modifications of the SGLD algorithm which produces non-Markovian dependent sequences $(X_k)_{k=0}^{\infty}$, for example the SAGA method outlined in [29].

An important question related to the control variates method is the choice of criterion used to select the best control variate. In case of independent observations, the criterion is usually based on the least squares method [27], [64] or the empirical variance [65]. The latter approach leads to the Empirical Variance Minimization (EVM) method, which is extensively studied in [65]. In case of dependent observations $(X_k)_{k=0}^{\infty}$, the situation is more complicated, because $V_\infty(\cdot)$ in (2.6) is difficult to estimate. For $(X_k)_{k=0}^{\infty}$ being a Markov chain, recent papers [66–68] use the least-squares based method, which aims to minimize the marginal variance $\mathrm{Var}_\pi(\cdot)$ instead of V_∞. Another approach from [32] utilizes the technique of minimizing the estimate of V_∞, similar to

the one considered in the current work. However, the authors in [32] consider Markov chains with a kernel that is either V-geometrically ergodic or satisfies the L^p-transportation inequality (see Definition 3.1 in [39]), which is rather restrictive.

2.3. Contributions

The main contributions of this chapter are listed below:

- We propose an extension of the ESVM variance reduction method, proposed in [32], for dependent random sequences satisfying the covariance stationarity assumption (see **(CS)** in Section 2.4). We also provide high-probability bounds for the excess asymptotic variance $V_\infty(f-\widehat{g}_n) - \inf_{g\in\mathcal{G}} V_\infty(f-g)$, where $\widehat{g}_n$ is constructed using the suggested algorithm;
- We derive concentration inequalities for quadratic forms of functions of Markov chains satisfying the geometric ergodicity condition in the Kantorovich-Wasserstein metric $\mathbf{W}_{\mathsf{d},1}$. These results are applied to the iterates of the SGLD algorithm (see Section 2.6). In particular, we show that for the parametric class of control variates $\mathcal{G}$, it holds

$$V_\infty(f-\widehat{g}_n) - \inf_{g\in\mathcal{G}} V_\infty(f-g) \lesssim n^{-1/2}\log^{5/2}(n)\,,$$

where n is the number of observations $X_0,\dots,X_{n-1}$ used for estimating $\widehat{g}_n$.

2.4. Empirical Spectral Variance Minimization

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 $\mathcal{G}$ 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 $k\in\mathbb{N}$. Examples of classes $\mathcal{G}$ include, for example, the class of Stein control variates (2.10). Denote the class of functions $\mathcal{H} := \{f-g:\ g\in\mathcal{G}\}$. Recall that $\bar{h}=h-\pi(h)$ for $h\in\mathcal{H}$. We impose the following condition:

(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 conditions below:*

- $\rho^{(h)}(0) = \mathrm{Var}_\pi(h)$;
- *There exist constant $R>0$ independent of h and ℓ, such that for any $\ell\in\mathbb{N}_0$,*

$$\sum\nolimits_{k\in\mathbb{N}_0}\left|\mathsf{E}\left[\bar{h}(X_k)\bar{h}(X_{k+\ell})\right] - \rho^{(h)}(\ell)\right| \leq R\,,$$

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

The condition **(CS)** guarantees the existence of $V_\infty(h)$ defined in (2.6) for any function $h\in\mathcal{H}$, see [39, Proposition 2.1]. Further, this variance can be expressed as:

$$V_\infty(h) = \sum_{\ell\in\mathbb{Z}}\rho^{(h)}(\ell). \tag{2.7}$$

Since the closed-form computation of $V_\infty(h)$ is typically impossible, the suggested variance reduction algorithm should rely on its empirical counterpart, $V_n(h)$. Among existing consistent

estimators of the asymptotic variance (see [69] for review) we use the *spectral variance estimator*:

$$V_n(h) = \sum_{|\ell|<b_n} w_n(\ell)\rho_n^{(h)}(|\ell|)\,, \quad \rho_n^{(h)}(|\ell|) = n^{-1}\sum_{k=0}^{n-\ell-1}\big(h(X_k)-\pi_n(h)\big)\big(h(X_{k+\ell})-\pi_n(h)\big)\,. \tag{2.8}$$

Here b_n is an integer truncation level (typically b_n increase with n), and weights $w_n(\ell)$ are given by $w_n(\ell) = w(\ell/b_n)$ for a symmetric non-negative function w, such that $\sup_{y\in[0,1]}|w(y)| \le 1$ and $w(y) = 1$ for $y \in [-1/2, 1/2]$.

Now we state the version of the general ESVM algorithm for dependent sequences satisfying **(CS)**. Based on the spectral estimator $V_n(h)$ defined in (2.8), we select the control variate $\widehat{g}_n$ (equivalently, $\widehat{h}_n = f - \widehat{g}_n$) as a minimizer

$$\widehat{h}_n \in \arg\min_{h\in\mathcal{H}} V_n\big(h\big). \tag{2.9}$$

We summarize the ESVM method in Algorithm 1.

Algorithm 1 Empirical Spectral Variance Minimization (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 $g \in \mathcal{G}$.

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

Constructing a control variate. If π is known at least up to a normalizing constant, it is possible to construct control variates depending only on the gradient $\nabla \log\pi$ using the Stein operator, as suggested in [70, 71], see also [27, 64, 72]. The latter operator gives rise to the popular class of *Stein control variates*:

$$g_\phi(\theta) = \langle\phi(\theta), \nabla\log\pi(\theta)\rangle + \mathrm{div}\big(\phi(\theta)\big), \tag{2.10}$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is a smooth function, and $\mathrm{div}(\phi)$ is the divergence of ϕ. Then under appropriate tail conditions on $\log\pi(\theta)$, one can ensure that $\pi(g_\phi) = 0$, see [72].

Theoretical analysis. To simplify the optimization in (2.9), we consider the minimization over an ε-net within $\mathcal{H}$. Formally, assuming that $\mathcal{H}$ 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 $\mathcal{H}_\varepsilon \subset \mathcal{H}$ such that for any $h \in \mathcal{H}$ there exists $h_\varepsilon \in \mathcal{H}_\varepsilon$ with the distance between h an h_ε in $L^2(\pi)$-norm being less than or equal to ε. Consider now the minimization over $\mathcal{H}_\varepsilon$:

$$\widehat{h}_{n,\varepsilon} \in \arg\min_{h\in\mathcal{H}_\varepsilon} V_n(h)\,. \tag{2.11}$$

Since $\mathcal{H}_\varepsilon$ is finite, the optimization problem (2.11) is tractable. We demonstrate that the asymptotic variance $V_\infty(\widehat{h}_{n,\varepsilon})$ closely approximates $\inf_{h\in\mathcal{H}} V_\infty(h)$, the smallest asymptotic variance over $\mathcal{H}$. For this purpose, we impose an assumption on the decay rate of $(\rho^{(h)}(\ell))_{\ell\in\mathbb{Z}}$ from **(CS)**:

(CD). *There exist* $\varsigma > 0$ *and* $\lambda \in [0,1)$ *such that, for any* $h \in \mathcal{H}$ *and* $\ell \in \mathbb{N}_0$, $\left|\rho^{(h)}(\ell)\right| \le \varsigma\lambda^\ell$.

Theorem 6. *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*

$$\sup_{h\in\mathcal{H}} \mathsf{P}\Big(\big|V_n(h) - \mathsf{E}[V_n(h)]\big| > t\Big) \le \alpha_n(t), \quad t > 0. \tag{2.12}$$

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

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

where α_n^{-1} *is an inverse function for* α_n*,* $D = \sup_{h\in\mathcal{H}} \mathrm{Var}_\pi(h)$*, and the first asymptotic variance is conditional on* $X_0, \dots, X_{n-1}$ *used to construct* $\widehat{h}_{n,\varepsilon}$ *in* (2.11).

Theorem 6 generalizes results previously obtained in [65] for independent sequences and in [32] for Markov chains which are geometrically ergodic in V-total variation norm or Kantorovich-Wasserstein metric $\mathbf{W}_{\mathsf{d},2}$. This theorem applies to any dependent sequence that allows to verify **(CS)** and **(CD)**, and study the concentration properties of the quadratic form $V_n(h)$ in (2.12) for $h \in \mathcal{H}$. Below we study these properties for $(X_k)_{k\ge0}$ being a Markov chain satisfying the uniform ergodicity property w.r.t. the Kantorovich-Wasserstein metric $\mathbf{W}_{\mathsf{d},1}$.

2.5. Applications to the Markov kernels, geometrically ergodic in the Kantorovich-Wasserstein distance

In what follows, we assume that $(X_k)_{k=0}^\infty$ is a Markov chain on a complete separable metric space (X, d), and let P be the corresponding Markov kernel on $(\mathsf{X}, \mathcal{X})$. We focus on the case where P is $\mathbf{W}_{\mathsf{d},p}$-uniformly ergodic for some $p \ge 1$, that is:

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

$$\sup_{(x,x')\in\mathsf{X}^2,\, x\ne x'} \frac{\mathbf{W}_{\mathsf{d},p}(\delta_x\mathrm{P}, \delta_{x'}\mathrm{P})}{\mathsf{d}(x,x')} = \Delta_p\,.$$

The statement of [41, 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, for any probability measure ξ with finite p-th moment,

$$\mathbf{W}_{\mathsf{d},p}(\xi\mathrm{P}^n, \pi) \le \Delta_p^n \mathbf{W}_{\mathsf{d},p}(\xi, \pi)\,, \quad n \in \mathbb{N}. \tag{2.13}$$

The setting of Markov kernels satisfying **(WE)**-2 has been partially addressed before in [32]. However, in such a setting one need to rely on rather restrictive additional properties of the Markov kernel $\mathrm{P}(x,\cdot)$, see the related discussion in [39, Proposition 3.2]. Below we focus on a more general setting of **(WE)**-1. In this particular setting **(CS)** and **(CD)** can be verified for $\mathcal{H}$ being a subset of bounded Lipschitz functions.

Proposition 2. *Let $\mathcal{H} \subset \mathrm{Lip}_{b,\mathsf{d}}(L,\mathsf{B})$ and assume that **(WE)**-1 holds. Then for any initial distribution $\xi \in \mathbb{S}_1(\mathsf{X},\mathsf{d})$, assumptions **(CS)** and **(CD)** are satisfied with*

$$\rho^{(h)}(\ell) = \mathsf{E}_\pi\big[\bar{h}(X_0)\bar{h}(X_{|\ell|})\big], \quad \lambda = \Delta_1, \tag{2.14}$$

and constants R and ς are given in [39, Proposition 3.3]. Moreover, for any $p \in \mathbb{N}$,

$$\mathsf{P}_\xi\big(\big|V_n(h) - \mathsf{E}_\xi[V_n(h)]\big| \geq t\big) \leq \frac{\bar{C}^p_{\mathsf{R},1} B^{2p} b_n^{3p/2} p^p}{n^{p/2} t^p} + \frac{\bar{C}^p_{\mathsf{R},2} B^{2p} b_n^{2p} p^{2p}}{n^{p-1} t^p}, \tag{2.15}$$

where $\bar{C}_{\mathsf{R},1}$ and $\bar{C}_{\mathsf{R},2}$ are constants given in [39, eq.(A.28)].

Proof of Proposition 2 is based on a suitable version of Rosenthal inequality adapted from [45]. Note that $V_n(h)$ is a quadratic form in $h(X_0),\dots,h(X_{n-1})$. Studying the concentration properties of such objects is already challenging in the setting of independent observations [73]. Recent results for the setting of Markov chains, unfortunately, covers only the uniformly geometrically ergodic setting [74] and [75], which is more restrictive as compared to the setting of the current chapter. Moreover, the latter condition fails to cover the algorithms studied in Section 2.6.

2.6. Applications to Langevin-based MCMC algorithms

In this section we consider a setting when the distribution of interest π is a probability measure on $\mathbb{R}^d$ which admits density w.r.t. the Lebesgue measure. We also denote this density by π. Further, assume that there is a function $U(\theta) : \mathbb{R}^d \to \mathbb{R}$, such that $\tilde{C} = \int_{\mathbb{R}^d} e^{-U(\theta)} \mathrm{d}\theta < \infty$, and $\pi(\theta) = e^{-U(\theta)}/\tilde{C}$ for $\theta \in \mathbb{R}^d$. We assume that U is known, but not the normalization constant $\tilde{C}$. Then for approximate sampling from π one can use methods based on the Langevin diffusion

$$\mathrm{d}Y_t = -\nabla U(Y_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t\,, \tag{2.16}$$

with $(W_t)_{t\geq 0}$ being a d-dimensional Wiener process. Under appropriate conditions, (2.16) admits a unique strong solution [36]. Moreover, the distribution of Y_t converges to π at the exponential rate, see e.g. [63]. With the Euler scheme used to discretize (2.16), we obtain the Unadjusted Langevin Algorithm (ULA). Given a step size $\gamma > 0$ and an i.i.d. sequence of standard Gaussian vectors $(\xi_k)_{k\geq 1}$, iterates of the ULA are written as a recurrence

$$\theta_{k+1} = \theta_k - \gamma\nabla U(\theta_k) + \sqrt{2\gamma}\,\xi_{k+1}\,. \tag{2.17}$$

The ULA algorithm and its theoretical properties attracted lot of attention [36, 37, 76]. Note that computing the gradient ∇U may be computationally expensive, if $U(\theta) = U_0(\theta) + \sum_{i=1}^K U_i(\theta)$ and the number of terms K is large. In such a setting, following [33], we can use the Stochastic Gradient Langevin Dynamics (SGLD) algorithm:

$$\theta_{k+1} = \theta_k - \gamma G(\theta_k, S_{k+1}) + \sqrt{2\gamma}\,\xi_{k+1}\,, \quad G(\theta,S) = \nabla U_0(\theta) + KM^{-1}\sum\nolimits_{i\in S}\nabla U_i(\theta)\,. \tag{2.18}$$

Random variable $S_{k+1} \in \mathsf{S}_M$ is called a mini-batch. Here S_M is the set of all subsets S of $\{1,\dots,K\}$ with $|S| = M$. Note that S_{k+1} is chosen independently of $\mathcal{F}_k = \sigma(\{(\theta_\ell, S_\ell)\}_{0\leq\ell\leq k})$. For theoretical

analysis of SGLD we impose the following assumptions on U:

(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 $U(\theta') \geq U(\theta) + \langle \nabla U(\theta), \theta' - \theta \rangle + (m_U/2)\|\theta' - \theta\|^2$.*

These assumptions are classical for the analysis of SGLD, see e.g. [77, 78]. Denote by $\mathrm{P}_{\mathsf{SGLD}}$ the transition kernel of SGLD and let Υ_M be a uniform distribution over S_M. Set $\overline{\mathrm{P}} := \mathrm{P}_{\mathsf{SGLD}} \otimes \Upsilon_M$ and note that the underlying chain now has a form $X_k = (\theta_k, S_{k+1})$. It is shown in [39, Proposition 3.7] that $\overline{\mathrm{P}}$ satisfies the assumption **(WE)**-1 with $\Delta_1 = \sqrt{1 - \gamma m_U}$, and admits a unique invariant measure. Define the corresponding asymptotic variance of SGLD iterates as $V_\infty^{(\mathsf{SGLD})}(\cdot)$. Then we obtain the following result:

Theorem 7. *Let $\mathcal{H} \subseteq \mathrm{Lip}_{b,\mathsf{d}}(L, \mathsf{B})$ 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 $1 - \delta$,*

$$V_\infty^{(\mathsf{SGLD})}(\widehat{h}_{n,\varepsilon}) - \inf_{h \in \mathcal{H}} V_\infty^{(\mathsf{SGLD})}(h) \lesssim \bar{C}_4\, \varepsilon \log(n) + \bar{C}_5 \sqrt{\frac{\log^5(n)}{n}} \left(\frac{|\mathcal{H}_\varepsilon|}{\delta} \right)^{1/\log(n)} + \bar{C}_6\, \frac{\log n}{n}, \tag{2.19}$$

where $\bar{C}_4$, $\bar{C}_5$, and $\bar{C}_6$ are constants provided in [39, Theorem 3.8], and the first asymptotic variance is conditional on $X_0, \dots, X_{n-1}$ used to construct $\widehat{h}_{n,\varepsilon}$ in (2.11).

Corollary 1. *Under the assumptions of Theorem 7, if class $\mathcal{H}$ is parametric, that is, $|\mathcal{H}_\varepsilon| \leq C_\rho \varepsilon^{-\rho}$ for all $\varepsilon \in (0,1)$ and some constants $C_\rho, \rho > 0$. Then 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 7.

If the class $\mathcal{H}$ is constructed using the Stein control variates, we can ensure the inclusion $\mathcal{H} \subseteq \mathrm{Lip}_{b,\mathsf{d}}(L, \mathsf{B})$ by taking smooth and compactly supported functions ϕ. As a result of Corollary 1, the asymptotic confidence intervals constructed by our method for the SGLD algorithm take the form

$$\pi_N(\widehat{h}_{n,\varepsilon}) \pm \mathfrak{q}_{1-\alpha/2} \sqrt{\frac{\inf_{h \in \mathcal{H}} V_\infty(h) + C n^{-1/2}}{N}} \tag{2.20}$$

for some constant $C > 0$. Assuming that the number of test observations $N = n$ (see Algorithm 1), the ESVM allows to obtain the length of the asymptotic confidence interval of order $n^{-3/4}$, provided that the class $\mathcal{H}$ is chosen so that $\inf_{h \in \mathcal{H}} V_\infty(h)$ is sufficiently small. This result is interesting to compare with one previously obtained in [65]. The result of (2.20) is comparable with the "slow rates" reported in [65] in the case of independent observations.

Numerical demonstrations of the proposed methodology applied to the ULA and SGLD algorithms and their modifications can be founded in [39].

Chapter 3

Variance reduction with martingale representations

3.1. Introduction and problem setup

The particular emphasis of Chapter 2 focuses on the setting of variance reduction with an analytically known target distribution, π. This setup perfectly suits the setting of control variates, see [32, 70, 71, 79, 80]. One of the most common tools in such a scenario is the Stein control variates [81], [71], as described in Section 2.4. The main problem with this approach is that it requires the direct access to π and $\nabla \log \pi$, which is not always possible. However, it turns out that if π is not known analytically and Stein control variates cannot be directly applied, it is still possible to suggest alternative constructions of control variates, which is the focus of this chapter. The results of this chapter are published in [40]

Similar to Section 2.1, we aim at computing $\pi(f) := \int_{\mathbb{R}^d} f(x)\pi(\mathrm{d}x)$, where $f : \mathbb{R}^d \to \mathbb{R}$ is a function in $L^2(\pi)$ and π has a smooth and everywhere positive density w.r.t the Lebesgue measure. By a slight abuse of notation, we use the same letter π for the probability measure and its density with respect to the Lebesgue measure. We further aim to enhance the estimate of $\pi(f)$, which writes as $\pi_n^x(f) = \frac{1}{n}\sum_{p=1}^n f(X_p^x)$, where $(X_p^x)_{p\in\mathbb{N}_0}$ is a Markov chain, satisfying a recurrence relation

$$X_p^x = \Phi(X_{p-1}^x, \xi_p), \quad p = 1, 2, \ldots, \quad X_0 = x \tag{3.1}$$

for some i.i.d. random vectors $\xi_p \in \mathbb{R}^m$ with distribution P_ξ and some Borel-measurable function $\Phi\colon \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$. We use an upper index x for X_p^x to highlight its dependence on the initial condition, which we assume to be $X_0 = x$. In fact, this is quite general class of Markov chains (see [41, Theorem 1.3.6]), which covers such MCMC algorithms as the Unadjusted and Metropolis-Adjusted Langevin Algorithms (see [35] or [41, Chapter 2]). Details are provided in [40, Example 2.1-2.2]. We further assume that we have access to a complete orthonormal system in $\mathrm{L}^2(\mathbb{R}^m, P_\xi)$, further denoted by $(\phi_k)_{k\geq 0}$. Since the dependence on the initial condition is explicit in π_n^x and X_p^x, we simply write E and Var instead of E_x and Var_x when applicable.

3.2. Contributions.

The main contributions of this chapter are the following:

- We propose a generic variance reduction method for additive functionals of Markov chains satisfying the recurrence representation (3.1). Compared to Stein control variates techniques, the knowledge of the stationary distribution is not required.
- We provide a non-asymptotic analysis of our algorithm applied to the normal noise model, which covers as a special example the Langevin dynamics.

3.3. Martingale representation

We first prove a general discrete-time martingale representation for Markov chains of type (3.1), which is used later to construct an efficient variance reduction algorithm. Let $(\phi_k)_{k\in\mathbb{Z}_+}$ be

a complete orthonormal system in $\mathrm{L}^2(\mathbb{R}^m, P_\xi)$ with $\phi_0 \equiv 1$. In particular, we have

$$\mathsf{E}[\phi_i(\xi)\phi_j(\xi)] = \delta_{ij}, \quad i,j \in \mathbb{N}$$

with $\xi \sim P_\xi$. Notice that this implies that the random variables $\phi_k(\xi)$, $k \geq 1$, are centered. Let $(\xi_p)_{p\in\mathbb{N}}$ be i.i.d. $m-$dimensional random vectors with distribution P_ξ. We denote via $(\mathcal{G}_p)_{p\in\mathbb{N}_0}$ the filtration generated by $(\xi_p)_{p\in\mathbb{N}}$ with $\mathcal{G}_0 = \mathrm{triv}$. Then we obtain the following expansion result:

Theorem 8. *For any $q \in \mathbb{N}$, $j < q$, bounded measurable function f, and $x \in \mathbb{R}^d$, it holds in $\mathrm{L}^2\big(\mathbb{R}^{mq}, P_\xi^{\otimes q}\big)$, that*

$$f(X_q^x) = \mathsf{E}\left[f(X_q^x)\middle|\,\mathcal{G}_j\right] + \sum\nolimits_{k=1}^{\infty}\sum\nolimits_{l=j+1}^{q} \bar{a}_{q-l+1,k}(X_{l-1}^x)\phi_k\left(\xi_l\right) \tag{3.2}$$

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

$$\bar{a}_{r,k}(y) = \mathsf{E}\left[f(X_r^y)\phi_k\left(\xi_1\right)\right] \quad r,k \in \mathbb{N}. \tag{3.3}$$

The coefficients $\bar{a}_{r,k}$ in (3.3) can be alternatively written as

$$\bar{a}_{r,k}(x) = \mathsf{E}\left[\phi_k\left(\xi\right) Q_{r-1}\left(\Phi(x,\xi)\right)\right] \quad \textit{with } Q_r(y) = \mathsf{E}\left[f(X_r^y)\right], \quad r \in \mathbb{N}. \tag{3.4}$$

Discussion. Setting

$$\pi_n^x(f) = n^{-1}\sum\nolimits_{p=1}^{n} f(X_p^x)\,,$$

we obtain the following exact representation for any bounded measurable function f:

$$\pi_n^x(f) = \tfrac{1}{n}\sum\nolimits_{q=1}^{n}\mathsf{E}[f(X_q^x)] + \tfrac{1}{n}\sum\nolimits_{k=1}^{\infty} M_{n,k}^x\,, \text{ with } M_{n,k}^x = \sum\nolimits_{l=1}^{n}\sum\nolimits_{r=1}^{n-l+1}\bar{a}_{r,k}(X_{l-1}^x)\phi_k(\xi_l)\,. \tag{3.5}$$

Construction of a control variate. We now show how the representation (3.5) can be used to construct variance-reduced estimates for $\pi(f)$. The first natural candidate implied by (3.5) is

$$\pi_n^{(x,K)}(f) = \pi_n^x(f) - n^{-1}\sum\nolimits_{k=1}^{K} M_{n,k}^x\,, \tag{3.6}$$

where K is a truncation parameter. However, its computational complexity scales *quadratically* with the number of observations n. In order to overcome the problem we set the second truncation level n_0 - the maximal number of estimated coefficients $\bar{a}_{r,k}, r \in \{1,\ldots,n_0\}$. Corresponding estimator writes as

$$\pi_{n,n_0}^{(x,K)}(f) = \pi_n^x(f) - n^{-1}\sum_{k=1}^{K} M_{n,k,n_0}^x\,, \quad M_{n,k,n_0}^x = \sum_{l=1}^{n}\sum_{r=1}^{\{n-\ell+1\}\wedge n_0}\bar{a}_{r,k}(X_{l-1}^x)\phi_k(\xi_l)\,. \tag{3.7}$$

It remains to define an estimator of $\bar{a}_{r,k}$. Towards this aim we first approximate $Q_r(\cdot)$ by the functions of the form $Q_{r,\boldsymbol{\beta}}(y) = \sum_{b=1}^{b_0}\boldsymbol{\beta}_b\psi_b(y)$ with some basis functions $\{\psi_b\}_{b=1}^{b_0}$ and $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^{b_0}$. Vector $\boldsymbol{\beta}$ is estimated via the least-squares approach, that is, for $r \in \{0,\ldots,n_0-1\}$, we find

$$\hat{\boldsymbol{\beta}}_r \in \operatorname*{arg\,min}_{\beta\in\mathbb{R}_0^b}\sum\nolimits_{s=1}^{n-r}\left|f(X_{r+s}^x) - Q_{r,\boldsymbol{\beta}}(X_s^x)\right|^2\,, \tag{3.8}$$

and then compute the estimates $\widehat{a}_{r,k}$ of the functions $\bar{a}_{r,k}$ according to the formulas

$$\widehat{a}_{r+1,k}(y) = \int \phi_k(z) Q_{\hat{\boldsymbol{\beta}}_r,r}(\Phi(y,z)) P_\xi(\mathrm{d}z) \tag{3.9}$$

where Φ is defined in (3.1). The estimator obtained by plugging (3.9) into (3.7) is referred to as the MAD-CV (MArtingale Decomposition Control Variate) estimator. The resulting estimate

$$\widehat{\pi}^{(x,K)}_{n,n_0}(f) = \pi^x_n(f) - n^{-1} \sum_{k=1}^{K} \widehat{M}^x_{n,k,n_0}\,, \quad \widehat{M}^x_{n,k,n_0} = \sum_{l=1}^{n} \sum_{r=1}^{(n-l+1)\wedge n_0} \widehat{a}_{r,k}(X^x_{l-1})\phi_k(\xi_l) \tag{3.10}$$

remains unbiased for $\pi(f)$ (if computed on a new trajectory independent of regression data). We summarize the suggested algorithm in Algorithm 2 above.

Algorithm 2 Martingale decomposition control variate (MAD-CV)

Input: Independent sequences $\mathbf{X}_N = (X^x_k)_{k=0}^{N-1}$, and $\tilde{\mathbf{X}}_n = (\tilde{X}^x_k)_{k=0}^{n-1}$, satisfying the recurrence (3.1); truncation point n_0.

1. Solve the r-step ahead prediction problem for $\hat{\boldsymbol{\beta}}_r$ in (3.8) based on $\mathbf{X}_N$;

2. Compute the estimates $\widehat{a}_{r,k}$ according to $\widehat{a}_{r+1,k}(y) = \int \phi_k(z) Q_{\hat{\boldsymbol{\beta}}_r,r}(\Phi(y,z)) P_\xi(\mathrm{d}z)$

Output: MAD-CV estimator $\widehat{\pi}^{(x,K)}_{n,n_0}(f)$ for $\pi(f)$, computed for a new trajectory $\tilde{\mathbf{X}}_n$.

3.4. Gaussian noise model

We analyze the MAD-CV algorithm for the Markov chains $(X^x_p)_{p\geq 0}$ driven by a normal noise:

$$X^x_p = \Phi(X^x_{p-1}, Z_p), \quad Z_p \sim \mathcal{N}(0, \mathrm{I}_d), \quad p = 1, 2, \ldots, \quad X^x_0 = x\,. \tag{3.11}$$

For a multi-index $\mathbf{k} = (k_i) \in \mathbb{N}_0^d$, we denote by $\mathbf{H}_\mathbf{k}(x)$ the normalized Hermite polynomial on $\mathbb{R}^d$, that is, $\mathbf{H}_\mathbf{k}(x) := \prod_{i=1}^d H_{k_i}(x_i)$, $x = (x_i) \in \mathbb{R}^d$ with $H_{k_i}(\cdot)$ being a univariate Hermite polynomial of degree k_i. In this case we specify the estimator (3.6) as

$$\pi^{(x,K)}_n(f) = \pi^x_n(f) - n^{-1} \sum_{0<\|\mathbf{k}\|\leq K} \sum_{l=1}^{n} \sum_{r=1}^{n-l+1} \bar{a}_{r,k}(X^x_{l-1}) \mathbf{H}_\mathbf{k}(Z_l)\,. \tag{3.12}$$

We aim to apply the MAD-CV algorithm for estimating expectations under the stationary distribution of ergodic diffusion processes. Let $b(x) : \mathbb{R}^d \to \mathbb{R}^d$ be a drift function, $(\mathsf{W}_t)_{t\geq 0}$ be a $d-$dimensional Wiener process and assume that the SDE

$$\mathrm{d}\mathsf{X}^x_t = -b(\mathsf{X}^x_t)\,\mathrm{d}t + \mathrm{d}\mathsf{W}_t,\, \mathsf{X}_0 = x \tag{3.13}$$

admits a unique strong solution $(\mathsf{X}^x_t)_{t\geq 0}$ for any $x \in \mathbb{R}^d$. Consider the Euler-Maruyama discretization of the SDE (3.13), i.e. the homogeneous Markov chain $(X^x_k)_{k\geq 0}$, starting from $X^x_0 = x \in \mathbb{R}^d$ and defined by the following recursion: for any $k \in \mathbb{N}$,

$$X^x_{k+1} = X^x_k - \gamma b(X^x_k) + \sqrt{\gamma} Z_{k+1}, \tag{3.14}$$

where $\gamma > 0$ is a stepsize and $(Z_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. $d-$dimensional standard normal vectors. Note that the recurrence (3.14) is a particular case of the general scheme (3.11) with $\Phi(x,z) = x - \gamma b(x) + \sqrt{\gamma} z$. We impose some technical conditions on the drift function b, following [82], namely,

A 4. *There exist a constant $L > 0$, such that $\|b(x) - b(y)\| \le L\|x - y\|$ for any $x, y \in \mathbb{R}^d$.*

A 5. *There exist a constant $m > 0$, such that $\langle b(x) - b(y), x - y\rangle \ge m\|x - y\|^2$ for any $x, y \in \mathbb{R}^d$.*

Under the assumptions **A** 4 and **A** 5, one can obtain the following bound on the variance of additive functionals of the Markov chains of the form (3.14):

Theorem 9. *Let $(X_k^x)_{k\ge 0}$ be a Markov chain given by the recurrence (3.14), and assume that **A** 4 and **A** 5 hold, and let $f : \mathbb{R}^d \to \mathbb{R}$ be a $K \times d$ times continuously differentiable function for some $K \in \mathbb{N}$. Assume in addition that there exist constants C_f and C_b, such that for any $x \in \mathbb{R}^d$, any multi-index $\mathbf{k} \in \mathbb{N}_0^d$ with $0 < \|\mathbf{k}\| \le K$, and any $u \in \{1, \dots, d\}$,*

$$|f^{(\mathbf{k})}(x)| \le C_f, \quad |b_u^{(\mathbf{k})}(x)| \le C_b\,.$$

Then, for $0 < \gamma < \min(1/C_b, m/\mathrm{L}^2)$ and any $n \in \mathbb{N}$,

$$\mathrm{Var}\left[\pi_n^{(x,K)}(f)\right] \lesssim \tfrac{\gamma^{K-2}}{n}.$$

Moreover, with the truncation point $n_0(\gamma) = \lceil K \log \gamma^{-1}/(2m\gamma)\rceil$, variance of the truncated estimate $\pi_{n,n_0(\gamma)}^{(x,K)}(f)$ can be bounded as

$$\mathrm{Var}\left[\pi_{n,n_0(\gamma)}^{(x,K)}(f)\right] \lesssim \tfrac{\gamma^{K-2}}{n},$$

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

In order to prove Theorem 9 we first establish the rate at which the coefficients $\bar{a}_{r,k}$ decrease with the growth of r. Then we relate $\mathrm{Var}\left[\pi_n^{(x,K)}(f)\right]$ with $\bar{a}_{r,k}(\cdot)$ based on an appropriate version of the Gaussian Poincare inequality [2]. Note that under conditions of Theorem 9, the variance of the estimate $\pi_n^{(x,K)}(f)$ for the discretized diffusion (3.14) satisfies

$$\mathrm{Var}\left[\pi_n^{(x,K)}(f)\right] \lesssim \tfrac{\gamma^{K-2}}{n}\,.$$

At the same time, the variance of the standard Monte Carlo estimate $\pi_n^x(f)$ is of order $1/(n\gamma)$ and this order can not be improved in general. Thus, for $K \ge 2$ and γ small enough we have a clear variance reduction effect.

Remark 1. *In the particular case of the Unadjusted Langevin Algorithm (2.17), assumptions of the Theorem 9 can be verified for the smooth and strongly convex potential U, that is, for $U \in C^2(\mathbb{R}^d)$ satisfying*

$$m_U\|x\|^2 \le \langle \nabla^2 U(y)x, x\rangle \le M_U\|x\|^2$$

for some $m_U > 0$, $M_U > 0$, and any $x, y \in \mathbb{R}^d$.

3.5. Numerical experiments

We compare the variance reduction achieved by MAD-CV against plain MCMC estimates based on the ULA algorithm (see (2.17)). Consider the target density π, which is a mixture of a two d-dimensional standard normal distributions

$$\pi(x) = \frac{1}{2\sqrt{(2\pi)^d}} \left(\mathrm{e}^{-(1/2)\|x-\mu\|^2} + \mathrm{e}^{-(1/2)\|x+\mu\|^2} \right) . \tag{3.15}$$

We fix $d = 2$, $\mu = (0.5, 0.5)$, and estimate $\pi(f)$ with $f(x) = x_1 + x_2$ and $f(x) = x_1^2 + x_2^2$. Using the ULA with constant step size $\gamma = 0.2$, we sample a training trajectory of length 5×10^4 with the starting point $X_0 = (1, 1)$. Then we solve the least squares problems (3.8) with the class of regressors $\{x_1, x_2, x_1^2, x_1x_2, x_2^2\}$ for the different choices of truncation point $n_0 \in [2, 20]$. We finally estimate the cost-to-variance ratio as follows

$$\mathcal{R}(f, K, n, n_0) = \frac{\mathrm{cost}\{\pi_n^x(f)\} \, \mathrm{Var}[\pi_n^x(f)]}{\mathrm{cost}\{\pi_{n,n_0}^{(x,K)}\} \, \mathrm{Var}[\pi_{n,n_0}^{(x,K)}(f)]} . \tag{3.16}$$

Note that $\mathcal{R}(f, K, n, n_0) > 1$ indicates that the variance reduction procedure is more cost-efficient as compared to the simple increase of the trajectory length n. We estimate the approximate value of $\mathcal{R}(f, K, n, n_0)$ based on 100 independent trajectories, each of length $n = 5 \times 10^4$. Here we set

$$\mathrm{cost}\{\pi_{n,n_0}^{(x,K)}(f)\} = \mathrm{cost}\{\pi_n^x(f)\} \times n_0 \times t(K) ,$$

where $t(K)$ is the number of evaluated coefficients $\hat{a}_{r,\mathbf{k}}(x)$. Variance reduction costs for different truncation points n_0 are summarized in Figure 3.1. We refer the reader to [40] for additional numerical examples.

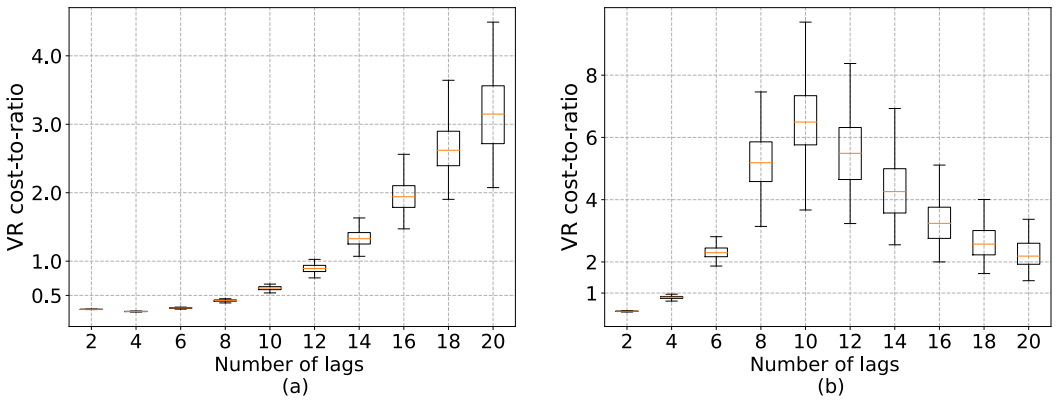

Figure 3.1. Cost-to-variance ratios (3.16) as functions of the truncation level n_0 for the mixture (3.15) of two-dimensional Gaussian distributions. Subfigure (a) : $f(x) = x_1 + x_2$, subfigure (b) : $f(x) = x_1^2 + x_2^2$.

Conclusion

1. In Chapter 1 we obtain new counterparts of Rosenthal and Bernstein inequalities for additive functionals of ergodic Markov chains that converge to the stationary distribution exponentially either in V-total variation norm or in the Kantorovich-Wasserstein semi-metric. The proof method we employ is based on the cumulant expansion techniques and the connections between cumulants and centered moments established through the Leonov-Shiryaev formula.

2. In Chapter 2, we propose an extension of the control variates method for variance reduction to the case of dependent random sequences that satisfy the covariance stationarity assumption. We derive concentration inequalities for quadratic forms of functions of Markov chains satisfying the geometric ergodicity condition in the Kantorovich-Wasserstein metric $\mathbf{W}_{\mathsf{d},1}$ and apply these results to MCMC algorithms based the Stochastic Gradient Langevin Dynamics.

3. In Chapter 3 we propose a novel variance reduction approach for additive functionals of Markov chains based on a discrete-time martingale representation. We study the variance reduction achieved by our method in a special setting of the normal noise model, covering the Unadjusted Langevin Algorithm (ULA). Our theoretical analysis is based on the Poincare inequality for Gaussian random vectors.

Bibliography

1. M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89. AMS Surveys and Monographs, 2001.
2. S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.
3. Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
4. Ron Dorfman and Kfir Yehuda Levy. Adapting to mixing time in stochastic optimization with Markovian data. In *International Conference on Machine Learning*, pages 5429–5446. PMLR, 2022.
5. Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.
6. G. O. Roberts and J. S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probab. Surv.*, 1:20–71, 2004.
7. Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50:5–43, 2003.
8. Bernard Bercu, Bernard Delyon, and Emmanuel Rio. *Concentration inequalities for sums and martingales*. Springer, 2015.
9. Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Number 47 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
10. Florence Merlevède, Magda Peligrad, and Emmanuel Rio. A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151(3-4):435–474, 2011.
11. Emmanuel Rio et al. *Asymptotic theory of weakly dependent random processes*, volume 80. Springer, 2017.
12. Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.
13. Błażej Miasojedow. Hoeffding's inequalities for geometrically ergodic markov chains on general state space. *Statistics & Probability Letters*, 87:115–120, 2014.
14. Radosław Adamczak and Witold Bednorz. Exponential concentration inequalities for additive functionals of Markov chains. *ESAIM: Probability and Statistics*, 19:440–481, 2015.
15. Jianqing Fan, Bai Jiang, and Qiang Sun. Hoeffding's inequality for general markov chains and its applications to statistical learning. *The Journal of Machine Learning Research*, 22(1):6185–6219, 2021.
16. Alexandros G. Dimakis, Soummya Kar, José M. F. Moura, Michael G. Rabbat, and Anna Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010.
17. Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with

convergence rate o(1/n). In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

18. Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR, 2018.
19. Stéphan JM Clémençon. Moment and probability inequalities for sums of bounded additive functionals of regular Markov chains via the Nummelin splitting technique. *Statistics & probability letters*, 55(3):227–238, 2001.
20. Michał Lemańczyk. General Bernstein-like inequality for additive functionals of Markov chains. *Journal of Theoretical Probability*, 34(3):1426–1454, 2021.
21. R. Bentkus and R. Rudzkis. Exponential estimates for the distribution of random variables. *Litovsk. Mat. Sb.*, 20(1):15–30, 216, 1980.
22. V. P. Leonov and A. N. Sirjaev. On a method of semi-invariants. *Theor. Probability Appl.*, 4:319–329, 1959.
23. Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo Method*, volume 10. John Wiley & Sons, 2016.
24. Emmanuel Gobet. *Monte-Carlo Methods and Stochastic Processes*. CRC Press, Boca Raton, FL, 2016.
25. Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, New York, 1999.
26. Paul Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53. Springer Science & Business Media, 2013.
27. Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for monte carlo integration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):695–718, 2017.
28. Rie Johnson and Tong Zhang. Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.
29. Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A Fast Incremental Gradient Method with Support for Non-Strongly Convex Composite Objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.
30. Niladri S Chatterji, Nicolas Flammarion, Yi-An Ma, Peter L Bartlett, and Michael I Jordan. On the Theory of Variance Reduction for Stochastic Gradient Monte Carlo. *Proceedings of Machine Learning Research*, 80, 2018.
31. Jack Baker, Paul Fearnhead, Emily B Fox, and Christopher Nemeth. Control variates for stochastic gradient MCMC. *Statistics and Computing*, 29(3):599–615, 2019.
32. D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, and S. Samsonov. Variance reduction for Markov chains with application to MCMC. *Statistics and Computing*, 30(4):973–997, 2020.
33. 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)*, pages 681–688, 2011.

34. Denis Belomestny, Stefan Häfner, and Mikhail Urusov. Variance reduction for discretised diffusions via regression. *Journal of Mathematical Analysis and Applications*, 458:393–418, 2018.
35. K. L. Mengersen and R. L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101–121, 02 1996.
36. Arnak Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 79(3):651–676, 2017.
37. A. Durmus and É. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27(3):1551–1587, 2017.
38. Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Probability and moment inequalities for additive functionals of geometrically ergodic Markov chains. *Journal of Theoretical Probability*, pages 1–50, 2024.
39. Denis Belomestny, Leonid Iosipoi, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Variance reduction for dependent sequences with applications to stochastic gradient MCMC. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):507–535, 2021.
40. Denis Belomestny, Eric Moulines, and Sergey Samsonov. Variance reduction for additive functionals of Markov chains via martingale representations. *Statistics and Computing*, 32(1):16, 2022.
41. R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, 2018.
42. L. Saulis and V. A. Statulevičius. *Limit theorems for large deviations*, volume 73 of *Mathematics and its Applications (Soviet Series)*. Kluwer Academic Publishers Group, Dordrecht, 1991. Translated and revised from the 1989 Russian original.
43. Iosif Pinelis. Optimum Bounds for the Distributions of Martingales in Banach Spaces. *The Annals of Probability*, 22(4):1679 – 1706, 1994.
44. Haskell P. Rosenthal. On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel J. Math.*, 8:273–303, 1970.
45. Paul Doukhan and Michael H Neumann. Probability and moment inequalities for sums of weakly dependent random variables, with applications. *Stochastic Processes and their Applications*, 117(7):878–903, 2007.
46. Hermann Thorisson. On maximal and distributional coupling. *The Annals of Probability*, pages 873–876, 1986.
47. Peter W Glynn and Dirk Ormoneit. Hoeffding's inequality for uniformly ergodic markov chains. *Statistics & probability letters*, 56(2):143–146, 2002.
48. Katalin Marton. A measure concentration inequality for contracting Markov chains. *Geometric & Functional Analysis GAFA*, 6(3):556–571, 1996.
49. Jérôme Dedecker, Sébastien Gouëzel, et al. Subgaussian concentration inequalities for geometrically ergodic Markov chains. *Electronic Communications in Probability*, 20, 2015.
50. A. Joulin and Y. Ollivier. Curvature, concentration and error estimates for Markov chain Monte Carlo. *The Annals of Probability*, 38(6):2418 – 2442, 2010.
51. Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral

methods. *Electronic Journal of Probability*, 20(none):1 – 32, 2015.

52. J. Fan, B. Jiang, and Q. Sun. Hoeffding's lemma for Markov chains and its applications to statistical learning. *arXiv preprint arXiv:1802.00211*, 2018.
53. J. Fan, B. Jiang, and Q. Sun. Bernstein's inequality for general Markov chains. *arXiv preprint arXiv:1805.10721*, 2018.
54. Ioannis Kontoyiannis and Sean P Meyn. Geometric ergodicity and the spectral gap of non-reversible Markov chains. *Probability Theory and Related Fields*, 154(1-2):327–339, 2012.
55. Patrice Bertail and Stéphan Clémençon. Sharp bounds for the tails of functionals of Markov chains. *Theory of Probability & Its Applications*, 54(3):505–515, 2010.
56. Gabriela Ciołek and Patrice Bertail. New Bernstein and Hoeffding type inequalities for regenerative Markov chains. *Latin American journal of probability and mathematical statistics*, 16:1–19, 02 2019.
57. Krishna B Athreya and Peter Ney. A new approach to the limit theory of recurrent Markov chains. *Transactions of the American Mathematical Society*, 245:493–501, 1978.
58. E. Nummelin. A splitting technique for harris recurrent Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 43:309–318, 1978.
59. Paul Doukhan and Sana Louhichi. A new weak dependence condition and applications to moment inequalities. *Stochastic Process. Appl.*, 84(2):313–342, 1999.
60. M. Hairer, J.C. Mattingly, and M. Scheutzow. Asymptotic coupling and a general form of Harris' theorem with applications to stochastic delay equations. *Probability theory and related fields*, 149(1-2):223–259, 2011.
61. M. Hairer, A.M. Stuart, and S.J. Vollmer. Spectral gaps for Metropolis-Hastings algorithms in infinite dimensions. *Ann. Appl. Probab.*, 24:2455–290, 2014.
62. Nikolai Sergeevich Bakhvalov. On the optimality of linear methods for operator approximation in convex classes of functions. *USSR Computational Mathematics and Mathematical Physics*, 11(4):244–249, 1971.
63. G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.
64. Leah F South, Chris J Oates, Antonietta Mira, and Christopher Drovandi. Regularized zero-variance control variates. *Bayesian Analysis*, 1(1):1–24, 2022.
65. D Belomestny, L Iosipoi, and N Zhivotovskiy. Variance reduction via empirical variance minimization: convergence and complexity. *arXiv preprint, arXiv:1712.04667*, 2017.
66. Zhuo Sun, Chris J Oates, and François-Xavier Briol. Meta-learning control variates: Variance reduction with limited data. In *Uncertainty in Artificial Intelligence*, pages 2047–2057. PMLR, 2023.
67. L F South, T Karvonen, C Nemeth, M Girolami, and C J Oates. Semi-exact control functionals from Sard's method. *Biometrika*, 109(2):351–367, 09 2021.
68. Leah F South, Marina Riabiz, Onur Teymur, and Chris J Oates. Postprocessing of mcmc. *Annual Review of Statistics and Its Application*, 9:529–555, 2022.
69. James M. Flegal and Galin L. Jones. Batch means and spectral variance estimators in Markov chain monte carlo. *Ann. Statist.*, 38(2):1034–1070, 04 2010.
70. Roland Assaraf and Michel Caffarel. Zero-variance principle for Monte Carlo algorithms.

Physical review letters, 83(23):4682, 1999.

71. Antonietta Mira, Reza Solgi, and Daniele Imparato. Zero variance Markov chain Monte Carlo for Bayesian estimators. *Statistics and Computing*, 23(5):653–662, 2013.
72. Chris J. Oates, Jon Cockayne, François-Xavier Briol, and Mark Girolami. Convergence rates for a class of estimators based on Stein's method. *Bernoulli*, 25(2):1141 – 1159, 2019.
73. Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.
74. Quentin Duchemin, Yohann De Castro, and Claire Lacour. Concentration inequality for u-statistics of order two for uniformly ergodic markov chains. *Bernoulli*, 29(2):929–956, 2023.
75. Quentin Duchemin, Yohann De Castro, and Claire Lacour. Three rates of convergence or separation via u-statistics in a dependent framework. *Journal of Machine Learning Research*, 23(201):1–59, 2022.
76. Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm. *Bernoulli*, 25(4A):2854–2882, 11 2019.
77. Yi-An Ma, Tianqi Chen, and Emily Fox. A Complete Recipe for Stochastic Gradient MCMC. In *Advances in Neural Information Processing Systems*, pages 2917–2925, 2015.
78. Arnak S. Dalalyan and Avetik G. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stoch. Proc. Appl.*, 129(12):5278–5311, 2019.
79. Shane G Henderson. *Variance reduction via an approximating Markov process*. PhD thesis, Stanford University, 1997.
80. P. Dellaportas and I. Kontoyiannis. Control variates for estimation based on reversible Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1), 2012.
81. Roland Assaraf and Michel Caffarel. Zero-variance principle for Monte Carlo algorithms. *Phys. Rev. Lett.*, 83(23):4682–4685, 1999.
82. Valentin De Bortoli and Alain Durmus. Convergence of diffusions and their discretizations: from continuous to discrete processes and back. *arXiv preprint arXiv:1904.09808*, 2019.