National Research University Higher School of Economics

as a manuscript

Alexey Alexandrovich Naumov

Non-asymptotic analysis of high-dimensional random objects and applications in machine learning

DISSERTATION SUMMARY
for the purpose of obtaining academic degree
Doctor of Science in Computer Science

Moscow - 2022

The dissertation was prepared at the National Research University Higher School of Economics.

Academic Advisors:

Vladimir Vasilievich Ulyanov, Doctor of Science, Professor, Lomonosov Moscow State University

Alexander Nikolaevich Tikhomirov, Doctor of Science, Professor, Komi Center of Science, Ural Branch, Russian Academy of Sciences

Contents

1 Introduction

Many theoretical and applied problems in mathematics, computer science, engineering are naturally related to the study of high-dimensional random objects, such as random matrices, graphs, processes, algorithms, etc. At first sight, these different objects have quite little in common. Each has its own ideas, mathematical approaches, and methods. Even the probabilistic nature and structure may be different. But there are some basic probabilistic principles that appear in the study of the above objects in high-dimensional spaces. These general principles usually take the form of non-asymptotic probability inequalities. The term non-asymptotic here means that we are not dealing with limit theorems as in many probabilistic results, but with explicit estimates that can be either dimension-free or contain a dependence on a dimension parameter.

In this dissertation we will look at three topics, which the author has dealt with over the last five years:

- Bootstrap method and Bayesian inference;
- Linear stochastic approximation (LSA) and Temporal difference (TD) algorithms;
- Markov chain Monte-Carlo algorithms and variance minimization.

Theoretical analysis of these algorithms requires to develop new non-asymptotic inequalities for linear and non-linear statistics of random objects which could be of independent interest. We will briefly discuss the content of the thesis.

In chapter 1 we study the problem of Gaussian comparison, i.e. one has to evaluate how the probability of a ball under a Gaussian measure is affected, if the mean and the covariance operators of this Gaussian measure are slightly changed. We present particular examples motivating our results when such "large ball probability" problem naturally arises in probability and statistics, including bootstrap validation, Bayesian inference, high-dimensional CLT. We derive sharp bounds for the Kolmogorov distance between the probabilities of two Gaussian elements to hit a ball in a Hilbert space. The key property of these bounds is that they are dimension-free and depend on the nuclear (Schatten-one) norm of the difference between the covariance operators of the elements. We also state a tight dimension free anti-concentration bound for a squared norm of a Gaussian element in Hilbert space which refines the well known results on the density of a chi-squared distribution

In chapter 2 we study the exponential stability of random matrix products driven by independent identically distributed (i.i.d.) noise or a general (possibly unbounded) state space Markov chain. Exponential stability plays a crucial role in the analysis of linear stochastic approximation (LSA) algorithms. This family of methods arises in many machine learning tasks and used to obtain approximate solutions of a linear system $\bar{\mathbf{A}}\theta = \bar{\mathbf{b}}$ for which $\bar{\mathbf{A}}$ and $\bar{\mathbf{b}}$ can only be accessed through random estimates $\{(\mathbf{A}(Z_n), \mathbf{b}(Z_n))\}_{n\in\mathbb{N}}$, where $\mathbf{A} : \mathsf{Z} \to \mathbb{R}^{d\times d}$, $\mathbf{b} : \mathsf{Z} \to \mathbb{R}^d$ are measurable functions and $(Z_k)_{k\in\mathbb{N}}$ is either an i.i.d. sequence with distribution π satisfying $\mathbb{E}[\mathbf{A}(Z_1)] = \bar{\mathbf{A}}$ and $\mathbb{E}[\mathbf{b}(Z_1)] = \bar{\mathbf{b}}$, or a Markov chain, taking values in a general state-space Z with unique invariant distribution π and $\lim_{n\to+\infty}\mathbb{E}[\mathbf{A}(Z_n)] = \bar{\mathbf{A}}, \lim_{n\to+\infty}\mathbb{E}[\mathbf{b}(Z_n)] = \bar{\mathbf{b}}$. As an application we provide non-asymptotic bounds for LSA and TD algorithms.

In chapter 3 we propose a novel and practical variance reduction approach for additive functionals of dependent sequences. This approach combines the use of control variates with the minimisation of an empirical variance estimate. We analysed finite sample properties of the proposed method and derive finite-time bounds of the excess asymptotic variance to zero. We applied this methodology to Stochastic Gradient MCMC (SGMCMC) methods for Bayesian inference on large data sets and combine it with existing variance reduction methods for SGMCMC. The crucial role in the theoretical analysis play novel concentration inequalities for quadratic forms of Markov chain.

Object and goals of the dissertation The goal of the dissertation is twofold. The first goal is to obtain non-asymptotic inequalities for high-dimensional random objects which could be of independent interest. In particular, we develop Gaussian comparison and anti-concentration inequalities, concentration for quadratic forms of Markov chains, moment bounds for products of random matrices driven by i.i.d. or Markovian noise. The second goal is to apply obtained result for theoretical analysis of machine learning algorithms. We study particular problems of Bayesian inference and bootstrap method, variance reduction methods for MCMC, convergence of LSA and RL algorithms.

The obtained results

1. We derived tight non-asymptotic bounds for the Kolmogorov distance between the probabilities of two Gaussian elements to hit a ball in a Hilbert space. We also established an anti-concentration bound for the squared norm of a non-centered Gaussian element in a Hilbert space.

2. We offered a bootstrap procedure for building sharp confidence sets for the true spectral projector of covariance matrix from the given data. We proved validity of the proposed procedure for Gaussian samples with an explicit error bound for the error of bootstrap approximation.

3. We study the exponential stability of random matrix products driven by a general (possibly unbounded) state space Markov chain, provided that (i) the underlying Markov chain satisfies a super-Lyapunov drift condition, (ii) the growth of the matrix-valued functions is controlled by an appropriately defined function Using this result, we give finite-time p-th moment bounds for constant and decreasing stepsize linear stochastic approximation schemes with Markovian noise on general state space and for TD algorithms with linear function approximation.

4. We provided a non-asymptotic analysis of linear stochastic approximation (LSA) algorithms with fixed stepsize and driven by i.i.d. noise. Our analysis is based on new results regarding moments and high probability bounds for products of matrices which are shown to be tight. We derive high probability bounds on the performance of LSA under weaker conditions than previous works. However, in contrast, we establish polynomial concentration bounds with order depending on the stepsize. We show that our conclusions cannot be improved without additional assumptions on the sequence of random matrices $\{\mathbf{A}(Z_n) : n \in \mathbb{N}^*\}$, and in particular that no Gaussian or exponential high probability bounds can hold. Finally, we pay a particular attention to establishing bounds with sharp order with respect to the number of iterations and the stepsize.

5. We proposed a novel and practical variance reduction approach for additive functionals of dependent sequences. We analysed finite sample properties of the proposed method and derive finite-time bounds of the excess asymptotic variance to zero. We applied this methodology to Stochastic Gradient MCMC (SGMCMC) methods.

Author's contribution Includes the development of the listed above methods and algorithms, theoretical analysis and applications to bootstrap and Bayesian inference, MCMC and TD learning.

As a result of the work on this dissertation, 10 papers were published:

First-tier publications

1. Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, and Vladimir Ulyanov. Large ball probabilities, Gaussian comparison and anti-concentration. Bernoulli, 25, 4A (2019), pp. 2538–2563, WOS Q2 / Scopus Q1 (main co-author; the author of this thesis proved Gaussian comparison and anti-concentration inequalities).

2. Alexey Naumov, Vladimir Spokoiny, Yury Tavyrikov, and Vladimir Ulyanov. Nonasymptotic Estimates for the Closeness of Gaussian Measures on Balls. Doklady Mathematics, 98, 2 (2018), pp. 490—493, Scopus Q2 (main co-author; the author of this thesis proved Gaussian comparison and anti-concentration inequalities).

3. Alexey Naumov, Vladimir Spokoiny, and Vladimir Ulyanov. Bootstrap confidence sets for spectral projectors of sample covariance. Probability Theory and Related Fields, 174, 3–4 (2019), pp. 1091–1132, WOS/Scopus Q1 (main co-author; the author of this thesis proved bootstrap validity).

4. Alexey Naumov, Vladimir Spokoiny, and Vladimir Ulyanov. Confidence Sets for Spectral Projectors of Covariance Matrices. Doklady Mathematics, 98, 2 (2018), pp. 511—514, Scopus Q2 (main co-author; the author of this thesis proved bootstrap validity).

5. Denis Belomestny, Leonid Iosipoi, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Variance reduction for Markov chains with application to MCMC. Statistics and Computing, 30, 4 (2020), pp. 973–997, WOS/Scopus Q1.

6. 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 (2021), pp. 507–535, WOS Q2 / Scopus Q1 (main co-author; the author of this thesis proved concentration inequalities for quadratic forms of Markov chains).

7. Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite Time Analysis of Linear Two-timescale Stochastic Approximation with Markovian Noise. Proceedings of Thirty Third Conference on Learning Theory. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2144–2203, CORE A*.

8. Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Hoi-To Wai. On the Stability of Random Matrix Product with Markovian Noise: Application to Linear Stochastic Approximation and TD Learning. Proceedings of Thirty Fourth Conference on Learning Theory. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, 2021, pp. 1711–1752, CORE A* (main co-author; the author of this thesis proved main result on the stability of product of random matrices).

9. Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Hoi-To Wai, and Kevin Scaman. Tight High Probability Bounds for Linear Stochastic Approximation with Fixed Stepsize. Advances in Neural Information Processing

Systems 34 (NeurIPS 2021). Ed. by M. Ranzato and A. Beygelzimer and K. Nguyen and P.S. Liang and J.W. Vaughan and Y. Dauphin. CORE A*.

10. Friedrich Götze, Alexey Naumov, and Alexander Tikhomirov. Distribution of linear statistics of singular values of the product of random matrices. Bernoulli, 23, 4B (2017), pp. 3067–3113, WOS Q2 / Scopus Q1 (main co-author; the author of this thesis proved central limit theorem for linear statistics of singular values of random matrix products).

Reports at conferences and seminars

1. Seminar "Modern Methods in Applied Stochastics and Nonparametric Statistics", Berlin, Germany, 06.01.16, "Distribution of Linear Statistics of Singular Values of the Product of Random Matrices";
2. Seminar "Modern Methods in Applied Stochastics and Nonparametric Statistics", Berlin, Germany, 18.01.17, "Bootstrap confidence sets for spectral projectors of sample covariance ";
3. ProbabLY ON Random matrices, Lyon, France, 3.04.2017–7.04.2017, "Estimation of a spectral projector";
4. The 2nd Chinese-Russian Seminar on Asymptotic Methods in Probability Theory and Mathematical Statistics & The 10th Probability Limit Theory and Statistic Large Sample Theory Seminar, Changchun, China, 24.09.2017–26.09.2017, "Large ball probabilities in statistical inference";
5. 12th International Vilnius Conference on Probability Theory and Mathematical Statistics 2018 IMS Annual Meeting on Probability and Statistics, Vilnius, Lithuania, 02.07.2018–06.07.2018, "Bootstrap confidence sets for spectral projectors of sample covariance";
6. Warm-Up Week - Bielefeld Stochastic Summer, Bielefeld, Germany, 20.08.2018–23.08.2018, "Random matrices and high-dimensional inference";
7. Seminar "Séminaire de Statistique", Paris, France, 13.05.19, "Gaussian approximation for maxima of large number of quadratic forms of high-dimensional random vectors";
8. Workshop SDEs/SPDEs: Theory, Numerics and their interplay with Data Science, Crete, Greece, 26.06.2019–30.06.2019, "Variance reduction for dependent sequences via empirical variance minimization with applications";
9. Mathematical Methods of Statistics, Lumini, France, 16.12.2019–20.12.2019, "Variance reduction for dependent sequences with application to MCMC";
10. **Conference on Computational Learning Theory (COLT2020)**, Graz, Austria, 09.07.2020–12.07.2020, "Finite Time Analysis of Linear Two-timescale Stochastic Approximation with Markovian Noise";
11. Seminar "Modern Methods in Applied Stochastics and Nonparametric Statistics", Berlin, Germany, 06.01.21, "Finite time analysis of linear two-timescale stochastic approximation with Markovian noise ";
12. Summer school Information, control and optimization, Voronovo, Russia, 10.06.2021–17.06.2021, "Stochastic approximation and Reinfrocement learning";

13. **Conference on Computational Learning Theory (COLT2021)**, Boulder, Colorado, USA, 15.08.2021–19.08.2021, "On the Stability of Random Matrix Product with Markovian Noise: Application to Linear Stochastic Approximation and TD Learning".

14. **Advances in Neural Information Processing Systems 34 (NeurIPS2021)**, 7.12.2021–10.12.2021, "Tight High Probability Bounds for Linear Stochastic Approximation with Fixed Stepsize".

2 Notations

This section gathers the general notations that are used throughout the thesis. Some additional notations may arise in individual chapters.

Let (Z, d) be a complete separable metric space with sigma-algebra $\mathcal{Z}$. Fix a measurable function $V : \mathsf{Z} \to [1, \infty)$. Let $\mathrm{P} : \mathsf{Z} \times \mathcal{Z} \to \mathbb{R}_+$ be a Markov kernel. Let $m \in \mathbb{N}^*$, ν a probability on $\mathcal{Z}$ and ϵ. A set $\mathsf{C} \in \mathcal{Z}$ is said to be $(m, \epsilon\nu)$-*small* for P if for all $z \in \mathsf{C}$ and $\mathsf{A} \in \mathcal{Z}$, $\mathrm{P}^m(z, \mathsf{A}) \geq \epsilon\nu(\mathsf{A})$. A set $\mathsf{A} \in \mathcal{Z}$ is said to be *accessible* if for all $z \in \mathsf{Z}$, there exists $m \in \mathbb{N}^*$ such that $\mathrm{P}(z, \mathsf{A}) > 0$.

We denote by P and Q symmetric positive definite matrices.

Table 1: Table of notation use throughout the paper

Notation	Meaning
$\Vert g \Vert_V$, $g : \mathsf{Z} \to \mathbb{R}$	$\Vert g \Vert_V = \sup_{z \in \mathsf{Z}} \lvert g(z) \rvert / V(z)$
L_∞^V	set of all measurable functions $g : \mathsf{Z} \to \mathbb{R}$ satisfying $\Vert g \Vert_V < \infty$
$\mathrm{P}V(z), z \in \mathsf{Z}$	$\mathrm{P}V(z) = \int_{\mathsf{Z}} V(z')\mathrm{P}(z, \mathrm{d}z')$
$\mathbb{M}_1(\mathsf{Z})$	set of probability measures on $(\mathsf{Z}, \mathcal{Z})$
$\Vert \mu \Vert_V$, $\mu \in \mathbb{M}_1(\mathsf{Z})$	$\Vert \mu \Vert_V = \sup_{f : \Vert f \Vert_V \leq 1} \int_{\mathsf{Z}} f(z)\mu(\mathrm{d}z)$
$\mathbb{S}_p(\mathsf{Z}, \mathsf{d}), p \geq 1$	$\mathbb{S}_p(\mathsf{Z}, \mathsf{d}) := \{\lambda \in \mathbb{M}_1(\mathsf{Z}) : \int_{\mathsf{Z}} \mathsf{d}^p(x, y)\lambda(\mathrm{d}y) < \infty \text{ for all } x \in \mathsf{Z}\}$
$\Pi(\lambda, \nu)$, $\lambda, \nu \in \mathbb{M}_1(\mathsf{Z})$	coupling set, i.e. $\xi \in \Pi(\lambda, \nu)$ is the measure on $\mathsf{Z} \times \mathsf{Z}$ satisfying for all $A \in \mathcal{B}(\mathsf{Z})$, $\xi(A, \mathsf{Z}) = \lambda(A)$ and $\xi(\mathsf{Z}, A) = \nu(A)$
$W_p^{\mathsf{d}}(\lambda, \nu)$, $p \geq 1$ and $\lambda, \nu \in \mathbb{S}_p(\mathsf{Z}, \mathsf{d})$	$W_p^{\mathsf{d}}(\lambda, \nu) := \inf_{\Pi(\lambda, \nu)} \{\int_{\mathsf{Z} \times \mathsf{Z}} \mathsf{d}^p(x, y)\, \xi(\mathrm{d}x, \mathrm{d}y)\}^{1/p}$
$\mathrm{KL}(\lambda \vert \nu)$, $\lambda, \nu \in \mathbb{M}_1(\mathsf{X})$	Kullback-Leibler divergence of λ with respect to ν, i.e., $\mathrm{KL}(\lambda \vert \nu) = \int \log(\mathrm{d}\lambda/\mathrm{d}\nu)\mathrm{d}\lambda$ if $\lambda \ll \nu$ and $\mathrm{KL}(\lambda \vert \nu) = \infty$ otherwise
$\Vert h \Vert_{\mathsf{Lip}}$, $h : \mathsf{Z} \to \mathbb{R}$	$\Vert h \Vert_{\mathsf{Lip}} := \sup_{x \neq y \in \mathsf{Z}} \{\lvert h(y) - h(x) \rvert / \mathsf{d}(x, y)\}$
$\lvert h \rvert_\infty$, $h : \mathsf{Z} \to \mathbb{R}$	$\lvert h \rvert_\infty = \sup_{z \in \mathsf{Z}} \lvert h(z) \rvert$
$\mathrm{Lip}_{\mathsf{d}}(L)$	class of Lipschitz functions with $\Vert h \Vert_{\mathsf{Lip}} \leq L$
$\mathrm{Lip}_{b,\mathsf{d}}(L, B)$	class of bounded Lipschitz functions with $\Vert h \Vert_{\mathsf{Lip}} \leq L$ and $\lvert h \rvert_\infty \leq B$
$a \lesssim b$ ($a \gtrsim b$)	there exists some absolute constant C such that $a \leq Cb$ ($a \geq Cb$ resp.)
$a \asymp b$	there exist c, C such that $c\, a \leq b \leq C\, a$
I_d	d-dimensional identity matrix
$\Vert x \Vert_Q$	$\Vert x \Vert_Q = \{x^\top Q x\}^{1/2}$ (note that $\Vert x \Vert = \Vert x \Vert_{\mathrm{I}_d}$)
κ_{Q}	the condition number of Q, i.e. $\kappa_{\mathsf{Q}} = \lambda_{\mathsf{min}}^{-1}(Q)\lambda_{\mathsf{max}}(Q)$
$\Vert \mathbf{A} \Vert_Q$	$\Vert \mathbf{A} \Vert_Q = \max_{\Vert x \Vert_Q = 1} \Vert \mathbf{A}x \Vert_Q$ (again note that $\Vert \mathbf{A} \Vert = \Vert \mathbf{A} \Vert_{\mathrm{I}_d}$)
$\Vert \mathbf{A} \Vert_{P,Q}$	$\Vert \mathbf{A} \Vert_{P,Q} = \max_{\Vert x \Vert_P = 1} \Vert \mathbf{A}x \Vert_Q$
$\Vert \mathbf{A} \Vert_p$, $p \geq 1$	the Schatten p-norm, i.e. $\Vert \mathbf{A} \Vert_p = \{\sum_{\ell=1}^d \sigma_\ell^p(\mathbf{A})\}^{1/p}$
$\Vert \mathbf{X} \Vert_{p,q}$, $p, q \geq 1$	$\Vert \mathbf{X} \Vert_{p,q} = \{\mathbb{E}[\Vert \mathbf{X} \Vert_p^q]\}^{1/q}$
$\mathbb{S}^{d-1}$	$\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \Vert x \Vert = 1\}$
$X \in \mathrm{SG}(\sigma^2)$	Sub-Gaussian random variable X with variance factor σ^2, i.e. for all $\lambda \in \mathbb{R}$, $\log \mathbb{E}[\mathrm{e}^{\lambda X}] \leq \lambda^2\sigma^2/2$

3 Large-ball probabilities and applications to bootstrap and Bayesian inference

3.1 Gaussian comparison and anti-concentration inequalities

The results of this subsection are published in [48], [81], [83] and [82].

In many statistical and probabilistic applications one faces the problem of Gaussian comparison, i.e. one has to evaluate how the probability of a ball under a Gaussian measure is affected, if the mean and the covariance operators of this Gaussian measure are slightly changed. Below we present particular examples motivating our results when such "large ball probability" problem naturally arises, including bootstrap validation, Bayesian inference, high-dimensional CLT. This chapter presents sharp bounds for the Kolmogorov distance between the probabilities of two Gaussian elements to hit a ball in a Hilbert space. The key property of these bounds is that they are dimension-free and depend on the nuclear (Schatten-one) norm of the difference between the covariance operators of the elements. We also state a tight dimension free anti-concentration bound for a squared norm of a Gaussian element in Hilbert space which refines the well known results on the density of a chi-squared distribution; see Theorem 3.7.

Section 3.2 presents some application examples where the "large ball probability" issue naturally arises and explains how the new bounds of this paper can be used to improve the existing results. The key observation behind the improvement is that in all mentioned examples we only need to know the properties of Gaussian measures on a class of balls. It means, in particular, that we would like to compare two Gaussian measures on the class of balls instead on the class of all measurable sets. The latter can be upperbounded by general Pinsker's inequality via the Kullback–Leibler divergence. In case of Gaussian measures this divergence can be expressed explicitly in terms of parameters of the underlying measures, see e.g. [101]. However, the obtained bound involves the inverse of the covariance operators of the considered Gaussian measures. In particularly, small eigenvalues have the largest impact which is contra-intuitive if a probability of a ball is considered. Our bounds only involve the operator and Frobenius norms of the related covariance operators and apply even in Hilbert space setup.

The proofs of the present optimal results are based in particular on Theorem 3.6 below. This theorem gives sharp upper bounds for a probability density function $p_\xi(x, \mathsf{a})$ of $\|\xi - \mathsf{a}\|^2$, where ξ is a Gaussian element with zero mean in a Hilbert space $\mathbb{H}$ with norm $\|\cdot\|$ and $\mathsf{a} \in \mathbb{H}$. It is well known that $p_\xi(x, \mathsf{a})$ can be considered as a density function of a weighted sum of non-central χ^2 distributions. An explicit but cumbersome representation for $p_\xi(x, \mathsf{a})$ in finite dimensional space $\mathbb{H}$ is available (see e.g. Section 18 in [60]). However, it involves some special characteristics of the related Gaussian measure which makes it hard to use in specific situations. Our result from Theorem 3.6 is much more transparent and provide sharp uniform upper bound on the underlying density.

One can even get two-sided bounds for $p_\xi(x, \mathsf{a})$ but under additional conditions, see e.g. [18]. Asymptotic properties of $p_\xi(x, \mathsf{a})$, small balls probabilities $\mathbb{P}\bigl(\|\xi - a\| \le \varepsilon\bigr)$, or large deviation bounds $\mathbb{P}\bigl(\|\xi\| \ge 1/\varepsilon\bigr)$ for small ε can be found e.g. in [19], [71], [72], [73] and [113].

3.1.1 Main results

Let $\mathbb{H}$ be a real separable Hilbert space with a scalar product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$. If dimension of $\mathbb{H}$ is finite and equals p, we shall write $\mathbb{R}^p$ instead of $\mathbb{H}$. Let Σ_ξ be a covariance operator of an arbitrary Gaussian random element in $\mathbb{H}$. By $\{\lambda_{k\xi}\}_{k\ge 1}$ we denote the set of its eigenvalues arranged in the non-increasing order, i.e. $\lambda_{1\xi} \ge \lambda_{2\xi} \ge \ldots$, and let $\lambda_\xi := \operatorname{diag}(\lambda_{j\xi})_{j=1}^{\infty}$. Note that $\sum_{j=1}^{\infty} \lambda_{j\xi} < \infty$. Introduce the following quantities

$$\Lambda_{k\xi}^2 := \sum_{j=k}^{\infty} \lambda_{j\xi}^2, \quad k = 1, 2,$$

and

$$\kappa(\Sigma_\xi) := \begin{cases} \Lambda_\xi^{-1}, & \text{if } 3\lambda_{1,\xi}^2 \le \Lambda_{1\xi}^2, \\ (\lambda_{1\xi}\Lambda_{2\xi})^{-1/2}, & \text{if } 3\lambda_{1\xi}^2 > \Lambda_{1\xi}^2,\ 3\lambda_{2\xi}^2 \le \Lambda_{2\xi}^2, \\ (\lambda_{1\xi}\lambda_{2\xi})^{-1/2}, & \text{if } 3\lambda_{1\xi}^2 > \Lambda_{1\xi}^2,\ 3\lambda_{2\xi}^2 > \Lambda_{2\xi}^2. \end{cases} \tag{3.1}$$

It is easy to see that $\|\Sigma_\xi\|_{\mathsf{Fr}} = \Lambda_{1\xi}$. Moreover, it is straightforward to check that

$$\frac{0.9}{(\Lambda_{1\xi}\Lambda_{2\xi})^{1/2}} \le \kappa(\Sigma_\xi) \le \frac{1.8}{(\Lambda_{1\xi}\Lambda_{2\xi})^{1/2}}. \tag{3.2}$$

Hence, $\kappa(\Sigma_\xi) \asymp (\Lambda_{1\xi}\Lambda_{2\xi})^{-1/2}$ and therefore equivalent results can be formulated in terms of any of the quantities introduced. The following theorem is the main result of this section.

Theorem 3.1. *Let ξ and η be Gaussian elements in $\mathbb{H}$ with zero mean and covariance operators Σ_ξ and Σ_η respectively. For any $\mathsf{a} \in \mathbb{H}$*

$$\begin{aligned} &\sup_{x>0} |\mathbb{P}(\|\xi - \mathsf{a}\| \le x) - \mathbb{P}(\|\eta\| \le x)| \\ &\lesssim \Bigl\{\kappa(\Sigma_\xi) + \kappa(\Sigma_\eta)\Bigr\}\Bigl(\|\lambda_\xi - \lambda_\eta\|_1 + \|\mathsf{a}\|^2\Bigr). \end{aligned} \tag{3.3}$$

We see that the obtained bounds are expressed in terms of the specific characteristics of the matrices Σ_ξ and Σ_η such as their operator and the Frobenius norms rather than the dimension p. Another nice feature of the obtained bounds is that they do not involve the inverse of Σ_ξ or Σ_η. In other words, small or vanishing eigenvalues of Σ_ξ or Σ_η do not affect the obtained bounds in the contrary to the Pinsker bound. Similarly, only the squared norm $\|\mathsf{a}\|^2$ of the shift a shows up in the results, while the Pinsker bound involves $\|\Sigma_\xi^{-1/2}\mathsf{a}\|$ which can be very large or infinite if Σ_ξ is not well conditioned.

Let us consider $\kappa(\Sigma_\xi)$ in the first factor on the r.h.s of (3.3): $\kappa(\Sigma_\xi) + \kappa(\Sigma_\eta)$. The representation (3.1) mimics well the three typical situations: in the "large-dimensional case" with three or more significant eigenvalues $\lambda_{j\xi}$, one can take $\kappa(\Sigma_\xi) = \|\Sigma_\xi\|_{\mathsf{Fr}}^{-1} = \lambda_{1\xi}^{-1}$. In the "two dimensional" case, when the sum $\Lambda_{k\xi}^2$ is of the order $\lambda_{k\xi}^2$ for $k = 1, 2$, we have that $\kappa(\Sigma_\xi)$ behaves as the product $(\lambda_{1\xi}\lambda_{2\xi})^{-1/2}$. In the intermediate case of a spike model with one large eigenvalue $\lambda_{1\xi}$ and many small eigenvalues $\lambda_{j\xi}, j \ge 2$, we have that $\kappa(\Sigma_\xi)$ behaves as $(\lambda_{1\xi}\Lambda_{2\xi})^{-1/2}$.

As it was mentioned earlier (see (3.2)), the result of Theorem 3.1 may be equivalently formulated in a "unified" way in terms of $(\Lambda_{1\xi}\Lambda_{2\xi})^{-1/2}$ and $(\Lambda_{1\eta}\Lambda_{2\eta})^{-1/2}$. Moreover, we specify the bound (3.3) in the "high-dimensional" case, $3\|\Sigma_\xi\|^2 \le \|\Sigma_\xi\|_{\mathsf{Fr}}^2, 3\|\Sigma_\eta\|^2 \le \|\Sigma_\eta\|_{\mathsf{Fr}}^2$, which means at least three significantly positive eigenvalues of the matrices Σ_ξ and Σ_η. In this case $\Lambda_{2\xi}^2 \ge 2\Lambda_{1\xi}^2/3, \Lambda_{2\eta}^2 \ge 2\Lambda_{1\eta}^2/3$ and we get the following corollary.

Corollary 3.2. *Let ξ and η be Gaussian elements in $\mathbb{H}$ with zero mean and covariance operators Σ_ξ and Σ_η respectively. Then for any $\mathsf{a} \in \mathbb{H}$*

$$\sup_{x>0} |\mathbb{P}(\|\xi - \mathsf{a}\| \le x) - \mathbb{P}(\|\eta\| \le x)| \lesssim \Bigl(\frac{1}{(\Lambda_{1\xi}\Lambda_{2\xi})^{1/2}} + \frac{1}{(\Lambda_{1\eta}\Lambda_{2\eta})^{1/2}} \Bigr) \Bigl(\|\lambda_\xi - \lambda_\eta\|_1 + \|\mathsf{a}\|^2 \Bigr).$$

Moreover, assume that

$$3\|\Sigma_\xi\|^2 \le \|\Sigma_\xi\|_{\mathsf{Fr}}^2 \quad \textit{and} \quad 3\|\Sigma_\eta\|^2 \le \|\Sigma_\eta\|_{\mathsf{Fr}}^2 .$$

Then for any $\mathsf{a} \in \mathbb{H}$

$$\sup_{x>0} |\mathbb{P}(\|\xi - \mathsf{a}\| \le x) - \mathbb{P}(\|\eta\| \le x)| \lesssim \Bigl(\frac{1}{\|\Sigma_\xi\|_{\mathsf{Fr}}} + \frac{1}{\|\Sigma_\eta\|_{\mathsf{Fr}}} \Bigr) \Bigl(\|\lambda_\xi - \lambda_\eta\|_1 + \|\mathsf{a}\|^2 \Bigr).$$

We complement the result of Theorem 3.1 and Corollary 3.2 with several additional remarks. The first remark is that by the Weilandt–Hoffman inequality, $\|\lambda_\xi - \lambda_\eta\|_1 \le \|\Sigma_\xi - \Sigma_\eta\|_1$, see e.g. [77]. This yields the bound in terms of the nuclear norm of the difference $\Sigma_\xi - \Sigma_\eta$, which may be more useful in a number of applications.

Corollary 3.3. *Under conditions of Theorem 3.1 we have*

$$\sup_{x>0} \bigl|\mathbb{P}\bigl(\|\xi - \mathsf{a}\| \le x) - \mathbb{P}(\|\eta\| \le x)\bigr| \lesssim \Bigl\{ \kappa(\Sigma_\xi) + \kappa(\Sigma_\eta) \Bigr\} \Bigl(\|\Sigma_\xi - \Sigma_\eta\|_1 + \|\mathsf{a}\|^2 \Bigr).$$

Since the right-hand-side of (3.3) does not change if we exchange ξ and η, Theorem 3.1 and its Corollaries hold for the balls with the same shift a. In particular, the following corollary is true.

Corollary 3.4. *Under conditions of Theorem 3.1 we have*

$$\sup_{x>0} \Bigl|\mathbb{P}(\|\xi - \mathsf{a}\| \le x) - \mathbb{P}(\|\eta - \mathsf{a}\| \le x)\Bigr| \lesssim \Bigl\{ \kappa(\Sigma_\xi) + \kappa(\Sigma_\eta) \Bigr\} \Bigl(\|\lambda_\xi - \lambda_\eta\|_1 + \|\mathsf{a}\|^2 \Bigr).$$

The result of Theorem 3.1 may be also rewritten in terms of the operator norm

$$\|\Sigma_\xi^{-1/2} \Sigma_\eta \Sigma_\xi^{-1/2} - \mathrm{I}\,\|.$$

Indeed, using the inequality $\|\mathbf{AB}\|_1 \le \|\mathbf{A}\|_1 \|\mathbf{B}\|$ we immediately obtain the following corollary.

Corollary 3.5. *Under conditions of Theorem 3.1 we have*

$$\sup_{x>0} |\mathbb{P}(\|\xi - \mathsf{a}\| \le x) - \mathbb{P}(\|\eta\| \le x)| \lesssim \Bigl\{ \kappa(\Sigma_\xi) + \kappa(\Sigma_\eta) \Bigr\} \Bigl(\operatorname{Tr}\bigl(\Sigma_\xi\bigr) \, \|\Sigma_\xi^{-1/2} \Sigma_\eta \Sigma_\xi^{-1/2} - \mathrm{I}\,\| + \|\mathsf{a}\|^2 \Bigr).$$

We now discuss the origin of the value $\kappa(\Sigma_\xi)$ which appears in the main theorem and its corollaries. Analysing the proof of Theorem 3.1 one may find out that it is necessary to get an upper bound for a probability density function (p.d.f.) $p_\xi(x)$ (resp. $p_\eta(x)$) of $\|\xi\|^2$ (resp. $\|\eta\|^2$) and the more general p.d.f. $p_\xi(x, \mathsf{a})$ of $\|\xi - \mathsf{a}\|^2$ for all $\mathsf{a} \in \mathbb{H}$. The same arguments remain true for $p_\eta(x)$. The following theorem provides uniform bounds.

Theorem 3.6. *Let ξ be a Gaussian element in $\mathbb{H}$ with zero mean and covariance operator Σ_ξ. Then it holds for any a that*

$$\sup_{x \geq 0} p_\xi(x, \mathsf{a}) \lesssim \kappa(\Sigma_\xi) \tag{3.4}$$

with $\kappa(\Sigma_\xi)$ from (3.1)*. In particular, $\kappa(\Sigma_\xi) \lesssim (\Lambda_{1\xi}\Lambda_{2\xi})^{-1/2}$.*

Since $\xi \overset{\mathsf{d}}{=} \sum_{j=1}^{\infty} \sqrt{\lambda_{j\xi}} Z_j \mathbf{e}_{j\xi}$, we obtain that $\|\xi\|^2 \overset{\mathsf{d}}{=} \sum_{j=1}^{\infty} \lambda_{j\xi} Z_j^2$. Here and in what follows $\{\mathbf{e}_{j\xi}\ \}_{j=1}^{\infty}$ is the orthonormal basis formed by the eigenvectors of Σ_ξ corresponding to $\{\lambda_{j\xi}\}_{j=1}^{\infty}$. In the case $\mathbb{H} = \mathbb{R}^p$, $\mathsf{a} = 0, \Sigma_\xi \asymp \mathrm{I}$ one has that the distribution of $\|\xi\|^2$ is close to standard χ^2 with p degrees of freedom and

$$\sup_{x \geq 0} p_\xi(x, 0) \asymp p^{-1/2}.$$

Hence, the bound (3.4) gives the right dependence on p because $\kappa(\Sigma_\xi) \asymp p^{-1/2}$. However, a lower bound for $\sup_{x \geq 0} p_\xi(x, \mathsf{a})$ in the general case is still an open question. Another possible extension is a non-uniform upper bound for the p.d.f. of $\|\xi - \mathsf{a}\|^2$. In this direction for any $\lambda > \lambda_{1\xi}$ we can prove that

$$p_\xi(x, \mathsf{a}) \leq \frac{\exp\bigl(-(x^{1/2} - \|\mathsf{a}\|)^2/(2\,\lambda)\bigr)}{2\sqrt{\lambda_{1\xi}\lambda_{2\xi}}} \prod_{j=3}^{\infty} (1 - \lambda_{j\xi}/\lambda)^{-1/2};$$

see [48][Lemma B.1]. It is still an open question whether it is possible to replace the $\lambda_{k\xi}$'s in the denominator by $\Lambda_{k\xi}$, $k = 1, 2$.

A direct corollary of Theorem 3.6 is the following theorem which states for a rather general situation a dimension-free anti-concentration inequality for the squared norm of a Gaussian element ξ. In the "high dimensional situation", this anti-concentration bound only involves the Frobenius norm of Σ_ξ.

Theorem 3.7 (ε-band of the squared norm of a Gaussian element)**.** *Let ξ be a Gaussian element in $\mathbb{H}$ with zero mean and a covariance operator Σ_ξ. Then for arbitrary $\varepsilon > 0$, one has*

$$\sup_{x > 0} \mathbb{P}(x < \|\xi - \mathsf{a}\|^2 < x + \varepsilon) \lesssim \kappa(\Sigma_\xi)\,\varepsilon$$

with $\kappa(\Sigma_\xi)$ from (3.1)*. In particular, $\kappa(\Sigma_\xi)$ can be replaced by $(\Lambda_{1\xi}\,\Lambda_{2\xi})^{-1/2}$.*

The lower bounds that justify the structure of estimates in Theorem 3.1 and Theorem 3.7 may be found in [48].

3.2 Application examples

This section collects some examples where the developed results seem to be very useful.

3.2.1 Bootstrap validity for the Maximum Likelihood Estimation (MLE)

Consider an independent sample $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ with a joint distribution $\mathbb{P} = \prod_{i=1,\ldots,n} \mathrm{P}_i$. The parametric maximum likelihood approach assumes that $\mathbb{P}$ belongs to a given parametric family $\bigl(\mathbb{P}_\theta\,, \theta \in \Theta \subseteq \mathbb{R}^p\bigr)$ dominated by a measure μ, that is, $\mathbb{P} = \mathbb{P}_{\theta^*}$ for $\theta^* \in \Theta$. The corresponding log-likelihood function can be written as a sum of marginal log-likelihoods $\ell_i(Y_i, \theta)$:

$$L(\theta) := \log \frac{d\mathbb{P}_\theta}{d\mu}(\mathbf{Y}) = \sum_{i=1}^{n} \ell_i(Y_i, \theta), \qquad \ell_i(Y_i, \theta) = \log \frac{d\mathrm{P}_{i,\theta}}{d\mu_i}(Y_i).$$

The MLE $\tilde{\theta}$ of the true parameter θ^* is defined as the point of maximum of $L(\theta)$:

$$\tilde{\theta} := \operatorname*{argmax}_{\theta \in \Theta} L(\theta), \qquad L(\tilde{\theta}) := \max_{\theta \in \Theta} L(\theta).$$

If the parametric assumption is misspecified, the target θ^* is defined as the best parametric fit:

$$\theta^* := \operatorname*{argmax}_{\theta \in \Theta} \mathbb{E}\, L(\theta).$$

The likelihood based confidence set $E(\zeta)$ for the target parameter θ^* is given by

$$E(\zeta) := \bigl\{\theta \colon L(\tilde{\theta}) - L(\theta) \le \zeta\bigr\}.$$

The value ζ should be selected to ensure the prescribed coverage probability $1 - \alpha$:

$$\mathbb{P}\bigl(\theta^* \notin E(\zeta)\bigr) \le \alpha. \tag{3.5}$$

However, it depends on the unknown measure $\mathbb{P}$. The bootstrap approach is a resampling technique based on the conditional distribution of the reweighted log-likelihood $L^\circ(\theta)$

$$L^\circ(\theta) = \sum_{i=1}^{n} \ell_i(Y_i, \theta) w_i$$

with i.i.d. random weights w_i given the data $\mathbf{Y}$. Below we assume that $w_i \sim \mathcal{N}(1, 1)$. The bootstrap confidence set is defined as

$$E^\circ(\zeta) := \bigl\{\theta \colon \sup_{\theta' \in \Theta} L^\circ(\theta') - L^\circ(\theta) \le \zeta\bigr\}.$$

The bootstrap distribution is perfectly known and the bootstrap quantile ζ° is defined by the condition

$$\mathbb{P}^\circ\bigl(\tilde{\theta} \notin E^\circ(\zeta^\circ)\bigr) = \mathbb{P}^\circ\Bigl(\sup_{\theta \in \Theta} L^\circ(\theta) - L^\circ(\tilde{\theta}) > \zeta^\circ\Bigr) = \alpha.$$

The bootstrap approach suggests to use ζ° in place of ζ to ensure (3.5) in an asymptotic sense. Bootstrap consistency means that for n large

$$\mathbb{P}\bigl(\theta^* \notin E(\zeta^\circ)\bigr) = \mathbb{P}\bigl(L(\tilde{\theta}) - L(\theta^*) > \zeta^\circ\bigr) \approx \alpha;$$

see e.g. [101]. A proof of this result is quite involved. The key steps are the following two approximations:

$$\sup_{\theta \in \Theta} L(\theta) - L(\theta^*) \approx \frac{1}{2}\bigl\|\xi + \mathsf{a}\bigr\|^2, \tag{3.6}$$

$$\sup_{\theta \in \Theta} L^\circ(\theta) - L^\circ(\tilde{\theta}) \approx \frac{1}{2}\bigl\|\xi^\circ\bigr\|^2,$$

where ξ is a Gaussian vector with the variance Σ given by

$$\Sigma := D^{-1} \operatorname{Var}\bigl[\nabla L(\theta^{*})\bigr] D^{-1}, \qquad D^{2} = -\nabla^{2} \mathbb{E} L(\theta^{*}),$$

while ξ° is conditionally (given $\mathbf{Y}$) Gaussian w.r.t. the bootstrap measure $\mathbb{P}^{\circ}$ with the covariance Σ° given by

$$\Sigma^{\circ} := D^{-1} \left(\sum_{i=1}^{n} \nabla \ell_i(Y_i, \theta^{*}) \bigl\{ \nabla \ell_i(Y_i, \theta^{*}) \bigr\}^{\mathsf{T}} \right) D^{-1}.$$

The vector a in (3.6) is the so called modeling bias and it vanishes if the parametric assumption $\mathbb{P} = \mathbb{P}_{\theta^{*}}$ is precisely fulfilled. The matrix Bernstein inequality ensures that Σ° is close to Σ in the operator norm for n large; see e.g. [107]. This yields bootstrap validity under the true parametric assumption in a weak sense. However, for quantifying the quality of the bootstrap approximation one has to measure the distance between two high dimensional Gaussian distributions $\mathcal{N}(\mathsf{a}, \Sigma)$ and $\mathcal{N}(0, \Sigma^{\circ})$. The recent paper [101] used the approach based on the Pinsker inequality which gives a bound in the total variation distance $\| \cdot \|_{\mathsf{TV}}$ via the Kullback-Leibler divergence between these two measures. A related bound involves the Frobenius norm $\| \cdot \|_{\mathsf{Fr}}$ of the matrix $\Sigma^{-1/2} \Sigma^{\circ} \Sigma^{-1/2} - \mathrm{I}_p$ and the norm of the vector $\beta := \Sigma^{-1/2} \mathsf{a}$:

$$\bigl\| \mathcal{N}(\mathsf{a}, \Sigma) - \mathcal{N}(0, \Sigma^{\circ}) \bigr\|_{\mathsf{TV}} \le \frac{1}{2} \Bigl(\bigl\| \Sigma^{-1/2} \Sigma^{\circ} \Sigma^{-1/2} - \mathrm{I}_p \bigr\|_{\mathsf{Fr}} + \bigl\| \Sigma^{-1/2} \mathsf{a} \bigr\| \Bigr); \tag{3.7}$$

see e.g. [101]. However, if we limit ourselves to the centered balls then these bounds can be significantly improved. Namely, by the main result of Theorem 3.1 and Corollary 3.2 below, we get under some technical conditions

$$\Bigl| \mathbb{P}\Bigl(\bigl\| \xi + \mathsf{a} \bigr\|^{2} > 2\zeta^{\circ} \Bigr) - \alpha \Bigr| \le \frac{\mathsf{C}}{\|\Sigma\|_{\mathsf{Fr}}} \Bigl(\|\Sigma - \Sigma^{\circ}\|_{1} + \|\mathsf{a}\|^{2} \Bigr). \tag{3.8}$$

The "small modeling bias" condition on a from [101] means that the value $\|\Sigma^{-1/2}\mathsf{a}\|$ is small and it ensures that a possible model misspecification does not destroy the validity of the bootstrap. Comparison of (3.8) with (3.7) reveals a number of benefits of (3.8). First, the "shift" term is proportional to the squared norm of the vector a, while the bound (3.7) depends on the norm of $\Sigma^{-1/2}\mathsf{a}$, i.e. on the whole spectrum of Σ. Normalization by $\Sigma^{-1/2}$ can significantly inflate the vector a in directions where the eigenvalues of Σ are small. In the contrary, the bound (3.8) only involves the squared norm $\|\mathsf{a}\|^{2}$ and the Frobenius norm of Σ, and the improvement from $\bigl\|\Sigma^{-1/2}\mathsf{a}\bigr\|$ to $\|\mathsf{a}\|^{2} / \|\Sigma\|_{\mathsf{Fr}}$ can be enormous if some eigenvalues of Σ nearly vanish. Further, the Frobenius norm $\bigl\|\Sigma^{-1/2}\Sigma^{\circ}\Sigma^{-1/2} - \mathrm{I}_p\bigr\|_{\mathsf{Fr}}$ can be much larger than the ratio $\bigl\|\Sigma - \Sigma^{\circ}\bigr\|_{1} / \|\Sigma\|_{\mathsf{Fr}}$ by the same reasons.

3.2.2 Prior impact in linear Gaussian modeling

Consider a linear regression model

$$Y_i = \Psi_i^{\mathsf{T}} \theta + \varepsilon_i$$

The assumption of homogeneous Gaussian errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^{2})$ yields the log-likelihood

$$L(\theta) = -\frac{1}{2\sigma^{2}} \sum_{i=1}^{n} (Y_i - \Psi_i^{\mathsf{T}} \theta)^{2} + R = -\frac{1}{2\sigma^{2}} \bigl\| \mathbf{Y} - \Psi^{\mathsf{T}} \theta \bigr\|^{2} + R,$$

where the term R does not depend on θ. A Gaussian prior $\pi = \pi_G = \mathcal{N}\bigl(0, G^{-2}\bigr)$ results in the posterior

$$\vartheta_G \,\big|\, \mathbf{Y} \propto \exp\biggl(L(\theta) - \frac{1}{2}\|G\theta\|^2 \biggr) \propto \exp\biggl(-\frac{1}{2\sigma^2}\bigl\|\mathbf{Y} - \Psi^{\mathsf{T}}\theta\bigr\|^2 - \frac{1}{2}\|G\theta\|^2 \biggr).$$

We shall represent the quantity $L_G(\theta) := L(\theta) - \frac{1}{2}\|G\theta\|^2$ in the form

$$L_G(\theta) \, = \, L_G(\breve{\theta}_G) - \frac{1}{2}\bigl\|D_G(\theta - \breve{\theta}_G)\bigr\|^2,$$

where

$$\begin{aligned} \breve{\theta}_G &:= \bigl(\Psi\Psi^{\mathsf{T}} + \sigma^2 G^2\bigr)^{-1}\Psi\mathbf{Y}, \\ D_G^2 &:= \sigma^{-2}\Psi\Psi^{\mathsf{T}} + G^2. \end{aligned}$$

In particular, it implies that the posterior distribution $\mathbb{P}(\vartheta_G \,\big|\, \mathbf{Y})$ of ϑ_G given $\mathbf{Y}$ is $\mathcal{N}(\breve{\theta}_G, D_G^{-2})$. A contraction property is a kind of concentration of the posterior on the elliptic set

$$E_G(\mathsf{r}) \, = \, \bigl\{\theta \colon \|W(\theta - \breve{\theta}_G)\| \le \mathsf{r}\bigr\},$$

where W is a given linear mapping from $\mathbb{R}^p$. The desirable credibility property manifests the prescribed conditional probability of $\vartheta_G \in E(\mathsf{r}_G)$ given $\mathbf{Y}$ with r_G defined for a given α by

$$\mathbb{P}\Bigl(\bigl\|W\bigl(\vartheta_G - \breve{\theta}_G\bigr)\bigr\| \ge \mathsf{r}_G \,\big|\, \mathbf{Y}\Bigr) \, = \, \alpha. \tag{3.9}$$

Under the posterior measure $\vartheta_G \sim \mathcal{N}(\breve{\theta}_G, D_G^{-2})$, this bound reads as

$$\mathbb{P}\bigl(\|\xi_G\| \ge \mathsf{r}_G\bigr) \, = \, \alpha \tag{3.10}$$

with a zero mean normal vector $\xi_G \sim \mathcal{N}(0, \Sigma_G)$ for $\Sigma_G = W D_G^{-2} W^{\mathsf{T}}$. The question of a prior impact can be stated as follows: whether the obtained credible set significantly depends on the prior covariance G. Consider another prior $\pi_1 = \mathcal{N}(0, G_1^{-2})$ with the covariance matrix G_1^{-2}. The corresponding posterior ϑ_{G_1} is again normal but now with parameters $\breve{\theta}_{G_1} = \bigl(\Psi\Psi^{\mathsf{T}} + \sigma^2 G_1^2\bigr)^{-1}\Psi\mathbf{Y}$ and $D_{G_1}^2 = \sigma^{-2}\Psi\Psi^{\mathsf{T}} + G_1^2$. We aim at checking the posterior probability of the credible set $E_G(\mathsf{r}_G)$:

$$\mathbb{P}\Bigl(\bigl\|W\bigl(\vartheta_{G_1} - \breve{\theta}_G\bigr)\bigr\| \ge \mathsf{r}_G \,\big|\, \mathbf{Y}\Bigr).$$

Clearly this probability can be written as

$$\mathbb{P}\Bigl(\bigl\|\xi_{G_1} + \mathsf{a}\bigr\| \ge \mathsf{r}_G\Bigr)$$

with $\xi_{G_1} \sim \mathcal{N}(0, \Sigma_{G_1})$ for $\Sigma_{G_1} = W D_{G_1}^{-2} W^{\mathsf{T}}$ and

$$\mathsf{a} \, := \, W\bigl(\breve{\theta}_{G_1} - \breve{\theta}_G\bigr).$$

Therefore,

$$\Bigl|\mathbb{P}\Bigl(\bigl\|W\bigl(\vartheta_{G_1} - \breve{\theta}_G\bigr)\bigr\| \ge \mathsf{r}_G \,\big|\, \mathbf{Y}\Bigr) - \alpha\Bigr| \, \le \, \sup_{\mathsf{r}>0}\Bigl|\mathbb{P}\Bigl(\bigl\|\xi_{G_1} + \mathsf{a}\bigr\| \ge \mathsf{r}\Bigr) - \mathbb{P}\Bigl(\bigl\|\xi_G\bigr\| \ge \mathsf{r}\Bigr)\Bigr|\,.$$

Again, the Pinsker inequality allows to upperbound the total variation distance between the Gaussian measures $\mathcal{N}(0, \Sigma_G)$ and $\mathcal{N}(\mathsf{a}, \Sigma_{G_1})$, however the answer is given via the Kullback-Leibler distance between these two measures:

$$\bigl\| \mathcal{N}(0, \Sigma_G) - \mathcal{N}(\mathsf{a}, \Sigma_{G_1}) \bigr\|_{\mathsf{TV}} \le \mathsf{C} \Bigl(\bigl\| \Sigma_G^{-1/2} \Sigma_{G_1} \Sigma_G^{-1/2} - \mathrm{I}_p \bigr\|_{\mathsf{Fr}} + \bigl\| \Sigma_{G_1}^{-1/2} \mathsf{a} \bigr\| \Bigr); \tag{3.11}$$

see e.g. [86]. Results of this paper allow to significantly improve this bound. In particular, only the nuclear norm $\bigl\| \Sigma_G - \Sigma_{G_1} \bigr\|_1$, the norm of the vector a and the Frobenius norm of Σ_G are involved. If $G^2 \ge G_1^2$, then $\Sigma_G \le \Sigma_{G_1}$ and

$$\bigl\| \Sigma_G - \Sigma_{G_1} \bigr\|_1 = \operatorname{Tr} \Sigma_{G_1} - \operatorname{Tr} \Sigma_G$$

and thus, by the main result of Theorem 3.1 and Corollary 3.2 below, it holds under some technical conditions

$$\Bigl| \mathbb{P}\Bigl(\bigl\| W\bigl(\vartheta_{G_1} - \breve{\theta}_G \bigr) \bigr\| \ge \mathsf{r}_G \bigm| \mathbf{Y} \Bigr) - \alpha \Bigr| \le \frac{\mathsf{C}\bigl(\operatorname{Tr} \Sigma_{G_1} - \operatorname{Tr} \Sigma_G + \|\mathsf{a}\|^2 \bigr)}{\|\Sigma_G\|_{\mathsf{Fr}}} .$$

This new bound significantly outperforms (3.11); see the discussion of the previous paragraph.

3.2.3 Nonparametric Bayes approach

One of the central question in the *nonparametric Bayes* approach is whether one can use the corresponding credible set as a *frequentist confidence set* for the true underlying mean $\mathbb{E}\,\mathbf{Y} = f = \Psi^{\mathsf{T}} \theta^*$. Here we consider the model $\mathbf{Y} = f + \varepsilon = \Psi^{\mathsf{T}} \theta + \varepsilon$ in $\mathbb{R}^n$ with a homogeneous Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_n)$ and a Gaussian prior $\mathcal{N}(0, G^{-2})$ on θ. The credible set $E_G(\mathsf{r})$ for ϑ_G yields the credible set $\mathcal{E}_G(\mathsf{r})$ for the corresponding response $f = \Psi^{\mathsf{T}} \theta$:

$$\mathcal{E}_G(\mathsf{r}) = \bigl\{ f = \Psi^{\mathsf{T}} \theta \colon \| \mathbf{A}\, \Psi^{\mathsf{T}} (\theta - \breve{\theta}_G) \| \le \mathsf{r} \bigr\},$$

with some linear mapping $\mathbf{A}$. The radius $\mathsf{r} = \mathsf{r}_G$ is fixed to ensure the prescribed credibility $1 - \alpha$ for the corresponding set $\mathcal{E}_G(\mathsf{r}_\alpha)$ due to (3.9) or (3.10) with $W = \mathbf{A} \Psi^{\mathsf{T}}$ and $\Sigma_G = \mathbf{A} \Psi^{\mathsf{T}} D_G^{-2} \Psi \mathbf{A}^{\mathsf{T}} = \sigma^2 \mathbf{A} \Pi_G \mathbf{A}^{\mathsf{T}}$, with $\Pi_G = \Psi^{\mathsf{T}} \bigl(\Psi \Psi^{\mathsf{T}} + \sigma^2 G^2 \bigr)^{-1} \Psi$. The frequentist coverage probability of the true response f is given by

$$\mathbb{P}\bigl(f \in \mathcal{E}_G(\mathsf{r}) \bigr) = \mathbb{P}\bigl(\| \mathbf{A} (f - \Psi^{\mathsf{T}} \breve{\theta}_G) \| \le \mathsf{r} \bigr) = \mathbb{P}\bigl(\| \mathbf{A}\, \Psi^{\mathsf{T}} (\theta^* - \breve{\theta}_G) \| \le \mathsf{r} \bigr).$$

The aim is to show that the the latter is close to $1 - \alpha$. For the posterior mean $\breve{\theta}_G = \bigl(\Psi \Psi^{\mathsf{T}} + \sigma^2 G^2 \bigr)^{-1} \Psi \mathbf{Y}$, it holds

$$\mathbb{E}\bigl[\mathbf{A} \bigl(f - \Psi^{\mathsf{T}} \breve{\theta}_G \bigr) \bigr] = \mathbf{A} \bigl(\mathrm{I} - \Pi_G \bigr) f =: \mathsf{a}.$$

Further,

$$\Sigma := \operatorname{Var}\bigl\{ \mathbf{A} \bigl(f - \Psi^{\mathsf{T}} \breve{\theta}_G \bigr) \bigr\} = \operatorname{Var}\bigl\{ \mathbf{A} \Pi_G \, \varepsilon \bigr\} = \sigma^2 \mathbf{A} \Pi_G^2 \mathbf{A}^{\mathsf{T}}$$

and hence, the vector $\mathbf{A} \bigl(f - \Psi^{\mathsf{T}} \breve{\theta}_G \bigr)$ is under $\mathbb{P}$ normal with mean $\mathsf{a} = \mathbf{A} \bigl(\mathrm{I} - \Pi_G \bigr) f$ and variance $\Sigma = \sigma^2 \mathbf{A} \Pi_G^2 \mathbf{A}^{\mathsf{T}}$. Therefore,

$$\mathbb{P}\bigl(f \in \mathcal{E}_G(\mathsf{r}) \bigr) = \mathbb{P}\bigl(\bigl\| \mathsf{a} + \xi \bigr\| \le \mathsf{r} \bigr).$$

Here $\xi \sim \mathcal{N}(0,\Sigma)$. So, it suffices to compare two probabilities

$$\mathbb{P}\bigl(\|\mathsf{a}+\xi\| \le \mathsf{r}\bigr) \;\text{ vs }\; \mathbb{P}\bigl(\|\xi_G\| \le \mathsf{r}\bigr)$$

for all $\mathsf{r} \ge 0$. Existing results cover only very special cases; see e.g. [62, 20, 86, 23, 24, 6] and references therein. Most of the mentioned results are of asymptotic nature and do not quantify the accuracy of the coverage probability. The results of this paper enable to study this accuracy in a straightforward way. Note first that the covariance operators $\Sigma = \sigma^2 \mathbf{A}\Pi_G^2 \mathbf{A}^\top$ and $\Sigma_G = \sigma^2 \mathbf{A}\Pi_G \mathbf{A}^\top$ satisfy $\Sigma \le \Sigma_G$. This yields that

$$\|\Sigma_G - \Sigma\|_1 = \operatorname{Tr}\Sigma_G - \operatorname{Tr}\Sigma .$$

Theorem 3.1 and Corollary 3.2 allow to evaluate under some technical conditions the coverage probability of the credibility set

$$\bigl|\mathbb{P}\bigl(f \not\in \mathcal{E}_G(\mathsf{r}_G)\bigr) - \alpha\bigr| \le \frac{\mathsf{C}\bigl(\operatorname{Tr}\Sigma_G - \operatorname{Tr}\Sigma + \|\mathsf{a}\|^2\bigr)}{\|\Sigma\|_{\mathsf{Fr}}} .$$

The right hand-side of this bound can be easily evaluated. The value $\|\mathsf{a}\| = \mathbf{A}\bigl(\mathrm{I} - \Pi_G\bigr)f$ is small under usual smoothness assumptions on f. The difference

$$\operatorname{Tr}\Sigma_G - \operatorname{Tr}\Sigma = \sigma^2 \operatorname{Tr}\bigl\{\mathbf{A}(\Pi_G - \Pi_G^2)\mathbf{A}^\top\bigr\}$$

is small under standard condition on the design Ψ and on the spectrum of G^2; see e.g. [100].

3.2.4 Central Limit Theorem in finite- and infinite-dimensional spaces

Another motivation for the current paper comes from the limit theorem in high-dimensional spaces for convex sets, in particular, for non-centred balls. Applications of smoothing inequalities require to evaluate the probability of hitting the vicinity of a convex set, see e.g. [13], [12]. This question is closely related to the anti-concentration inequalities considered below in Theorem 3.7. Recently, significant interest was shown in understanding of the anti-concentration phenomenon for weighted sums of random variables, particularly, in random matrix and number theory. We refer the interested reader to [96], [50].

Let $Y_1, \dots, Y_n$ be i.i.d. random vectors in $\mathbb{R}^p$. Assume that all these vectors have zero mean and the covariance operator Σ. Let X be a Gaussian random vector in $\mathbb{R}^p$ with zero mean and the same covariance operator Σ. We are interested to bound

$$\delta(\mathbf{C}) = \sup_{A \in \mathbf{C}} \left| \mathbb{P}\left(\frac{Y_1 + \dots + Y_n}{\sqrt{n}} \in A\right) - \mathbb{P}(X \in A)\right| \tag{3.12}$$

for some class $\mathbf{C}$ of Borel sets. It is worth emphasizing that the probabilities of hitting the vicinities of a set $A \in \mathbf{C}$, play the crucial role in the form of the bound for $\delta(\mathbf{C})$. Assume the class $\mathbf{C}$ satisfies the following two conditions:

(i) Class $\mathbf{C}$ is invariant under affine symmetric transformations, that is, $\mathbf{D}A + \mathsf{a} \in \mathbf{C}$ if $A \in \mathbf{C}, \mathsf{a} \in \mathbb{R}^p$ and $\mathbf{D} : \mathbb{R}^p \to \mathbb{R}^p$ is a linear symmetric invertible operator.

(ii) Class $\mathbf{C}$ is invariant under taking ε-neighborhoods for all $\varepsilon > 0$. More precisely, $A^\varepsilon, A^{-\varepsilon} \in \mathbf{C}$ if $A \in \mathbf{C}$, where

$$A^\varepsilon = \{x \in \mathbb{R}^p : \rho_A(x) \le \varepsilon\} \;\text{ and }\; A^{-\varepsilon} = \{x \in A : B_\varepsilon(x) \subset A\},$$

with $\rho_A(x) = \inf_{y\in A} |x-y|$ as the distance between $A \subset \mathbb{R}^p$ and $x \in \mathbb{R}^p$, and $B_\varepsilon(x) = \{y \in \mathbb{R}^p : |x-y| \le \varepsilon\}$.

Let X_0 be a Gaussian random vector in $\mathbb{R}^p$ with zero mean and the identity covariance operator I. Assume that the class $\mathbf{C}$ in (3.12) is such that for all $A \in \mathbf{C}$ and $\varepsilon > 0$

$$\mathbb{P}(X_0 \in A^\varepsilon \backslash A) \le a_p\, \varepsilon, \quad \mathbb{P}(X_0 \in A \backslash A^{-\varepsilon}) \le a_p\, \varepsilon, \tag{3.13}$$

where $a_p = a_p(\mathbf{C})$ is the so called isoperimetric constant of $\mathbf{C}$, e.g. taking $\mathbf{C}$ as the class of all convex sets in $\mathbb{R}^p$ we get $a_p \le 4\, p^{1/4}$; see [5].

It is known (see [12][Theorem 1.2]) that if $\mathbf{C}$ satisfies conditions (i), (ii) and (3.13) then for some absolute constant C one has

$$\delta(\mathbf{C}) \le C\,(1 + a_p)\; \mathbb{E}\,|Y_1|^3/\sqrt{n}. \tag{3.14}$$

Therefore, the inequalities (3.13), i.e. knowledge of a_p, play the crucial role in the form of the bound (3.14).

We have a similar situation in infinite-dimensional spaces. Though contrary to the finite dimensional case even if $\mathbf{C}$ is a rather small class of "good" subsets, e.g. the class of all balls, the convergence of $\mathbb{P}\Big((Y_1 + \cdots + Y_n)/\sqrt{n} \in A\Big)$ to $\mathbb{P}\big(X \in A\big)$ for each $A \in \mathbf{C}$, implied by the central limit theorem, can not be uniform in $A \in \mathbf{C}$; see e.g. [98][pp. 69–70]. However, the convergence becomes uniform for a class of all balls with center at some fixed point, say a. Such classes naturally appear in various statistical problems; see e.g. [90] or our previous application examples. Thus, similar to the inequalities (3.13) we need to get sharp bounds for the probability $\mathbb{P}(x < \|X - \mathsf{a}\|^2 < x + \varepsilon)$ for the Gaussian element X in a Hilbert space $\mathbb{H}$. Due to our Theorem 3.7 below, it holds under some technical conditions that

$$\mathbb{P}\big(x < \|X - \mathsf{a}\|^2 < x + \varepsilon\big) \le \frac{\mathtt{C}\,\varepsilon}{\|\Sigma\|_{\mathsf{Fr}}}$$

for an absolute constant $\mathtt{C}$.

3.2.5 Bootstrap confidence sets for spectral projectors of sample covariance

Let $X, X_1, \ldots, X_n$ be independent identically distributed (i.i.d.) random vectors taking values in $\mathbb{R}^p$ with mean zero and $\mathbb{E}\,\|X\|^2 < \infty$. Denote by $\boldsymbol{\Sigma}$ its $p \times p$ symmetric covariance matrix defined as $\boldsymbol{\Sigma} := \mathbb{E}(XX^\mathsf{T})$. We also consider the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ of the observations $X_1, \ldots, X_n$ defined as the average of $X_j X_j^\mathsf{T}$: with $\mathbf{X} := [X_1, \ldots, X_n] \in \mathbb{R}^{p\times n}$,

$$\widehat{\boldsymbol{\Sigma}} := \frac{1}{n}\sum_{j=1}^{n} X_j X_j^\mathsf{T} = \frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}.$$

In statistical applications, the true covariance matrix $\boldsymbol{\Sigma}$ is typically unknown and one often uses the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ as its estimator. The accuracy $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|$ of estimation of $\boldsymbol{\Sigma}$ by $\widehat{\boldsymbol{\Sigma}}$, in particular, for p much larger than n, has been actively studied in the literature. We refer to [107] for an overview of the recent results based on the matrix Bernstein inequality; see also [110]. A bound in term of the effective rank $\mathtt{r}(\boldsymbol{\Sigma}) := \mathrm{Tr}(\boldsymbol{\Sigma})/\|\boldsymbol{\Sigma}\|$ can be found in [65] and [55]. This or similar bounds on the spectral norm $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|$ can be effectively applied to relate the eigenvalues of $\boldsymbol{\Sigma}$ and of $\widehat{\boldsymbol{\Sigma}}$ under the spectral gap condition. This paper focuses on a slightly different problem of recovering the spectral projectors on the eigen-subspaces of $\mathbf{\Sigma}$ for few significantly positive eigenvalues. Such tasks naturally arise in many dimensionality reduction techniques for large p. In particular, the famous principal component analysis (PCA) projects the vector X onto the subspace spanned by the eigenvectors for the first principal eigenvalues. Surprisingly, the problem of recovering the spectral projectors (eigenvectors or eigen-subspaces) of $\mathbf{\Sigma}$ from the sample $X_1, \ldots, X_n$ for significantly positive spectral values is much less studied than the problem of recovering the covariance matrix $\mathbf{\Sigma}$. Recently [66] established sharp non-asymptotic bounds on the Frobenius distance $\|\mathbf{P}_r - \widehat{\mathbf{P}}_r\|_2$ between the spectral projectors $\mathbf{P}_r$ and its empirical counterparts $\widehat{\mathbf{P}}_r$ for the rth eigenvalue, as well as its asymptotic behaviour for large samples. This enables to build some asymptotic confidence sets for the target projector $\mathbf{P}_r$ as a proper elliptic vicinity of $\widehat{\mathbf{P}}_r$. However, it is well known that such asymptotic results apply only for really large samples due to a slow convergence of the normalized U-statistics to the limiting normal law.

To formulate the main result we need to introduce additional notations. Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p$ be the eigenvalues of $\mathbf{\Sigma}$ and $\mathbf{u}_j, j = 1, \ldots, p$, be the corresponding orthonormal eigenvectors. Matrix $\mathbf{\Sigma}$ has the following spectral decomposition

$$\mathbf{\Sigma} = \sum_{j=1}^{p} \sigma_j \mathbf{u}_j \mathbf{u}_j^{\mathsf{T}}. \tag{3.15}$$

Let $\mu_1 > \mu_2 > \ldots > \mu_q > 0$ with some $1 \leq q \leq p$, be strictly distinct eigenvalues of $\mathbf{\Sigma}$ and $\mathbf{P}_r, r = 1, \ldots, q$, be the corresponding spectral projectors (orthogonal projectors in $\mathbb{R}^p$). Denote $m_r := \operatorname{Rank}(\mathbf{P}_r)$. We may rewrite (3.15) in terms of distinct eigenvalues and corresponding spectral projectors, namely

$$\mathbf{\Sigma} = \sum_{r=1}^{q} \mu_r \mathbf{P}_r.$$

Denote by $\Delta_r := \{j \colon \sigma_j = \mu_r\}$. Then $|\Delta_r| = m_r$. Define $g_r := \mu_r - \mu_{r+1} > 0$ for $r \geq 1$. Let $\overline{g}_r := \min(g_{r-1}, g_r)$ for $r \geq 2$ and $\overline{g}_1 := g_1$. The quantity $\overline{g}_r$ is the r-th spectral gap of the eigenvalue μ_r.

Consider now the sample covariance matrix $\widehat{\mathbf{\Sigma}}$. Similarly to (3.15), it can be represented as

$$\widehat{\mathbf{\Sigma}} = \sum_{j=1}^{p} \widehat{\sigma}_j \widehat{\mathbf{u}}_j \widehat{\mathbf{u}}_j^{\mathsf{T}},$$

where $\widehat{\sigma}_1 \geq \widehat{\sigma}_2 \geq \ldots \geq \widehat{\sigma}_p, \widehat{\mathbf{u}}_1, \ldots, \widehat{\mathbf{u}}_p$ are the eigenvalues and the corresponding eigenvectors of $\widehat{\mathbf{\Sigma}}$. Following [66] we may define clusters of eigenvalues $\widehat{\sigma}_j, j \in \Delta_r$. Let $\widehat{\mathbf{E}} := \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$. One can show that

$$\inf_{j \notin \Delta_r} |\widehat{\sigma}_j - \mu_r| \geq \overline{g}_r - \|\widehat{\mathbf{E}}\|, \quad \sup_{j \in \Delta_r} |\widehat{\sigma}_j - \mu_r| \leq \|\widehat{\mathbf{E}}\|.$$

Assume that $\|\widehat{\mathbf{E}}\| \leq \overline{g}_r/2$. Then all $\widehat{\sigma}_j, j \in \Delta_r$ may be covered by an interval

$$(\mu_r - \|\widehat{\mathbf{E}}\|, \mu_r + \|\widehat{\mathbf{E}}\|) \subset (\mu_r - \overline{g}_r/2, \mu_r + \overline{g}_r/2).$$

The rest of the eigenvalues of $\widehat{\mathbf{\Sigma}}$ are outside of the interval

$$\Big(\mu_r - (\overline{g}_r - \|\widehat{\mathbf{E}}\|), \mu_r + (\overline{g}_r - \|\widehat{\mathbf{E}}\|)\Big) \supset [\mu_r - \overline{g}_r/2, \mu_r + \overline{g}_r/2].$$

Let $\|\widehat{\mathbf{E}}\| < \frac{1}{4}\min_{1\le s\le r}\overline{g}_s =: \overline{\delta}_r$. The set $\{\widehat{\sigma}_j, j \in \cup_{s=1}^r \Delta_s\}$ consists of r clusters, the diameter of each cluster being strictly smaller than $2\overline{\delta}_r$ and the distance between any two clusters being larger than $2\overline{\delta}_r$. We denote by $\widehat{\mathbf{P}}_r$ the projector on subspace spanned by the direct sum of $\widehat{\mathbf{u}}_j, j \in \Delta_r$.

It follows from [66][Lemma 5] that $\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\|_2^2$ has nearly weighted χ^2 distribution; see also [83][Theorem 4]. Therefore, after centering and standardization, it can be approximated by the standard normal distribution under some conditions on the spectrum of $\mathbf{\Sigma}$:

$$\mathcal{L}\left(\frac{\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\|_2^2 - \mathbb{E}\,\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\|_2^2}{\mathrm{Var}^{1/2}(\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\|_2^2)}\right) \approx \mathcal{N}(0,1), \tag{3.16}$$

see [66][Theorem 6]. This allows to build an asymptotic elliptic confidence set for $\mathbf{P}_r$ in the form

$$\left\{\mathbf{P}_r \colon \frac{\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\|_2^2 - \mathbb{E}\,\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\|_2^2}{\mathrm{Var}^{1/2}(\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\|_2^2)} \le z_\alpha\right\},$$

where z_α is a proper quantile of the standard normal law. However, there are at least two drawbacks of this approach. First, weak approximation in (3.16) can be very poor in some cases, especially if the effective rank of $\mathbf{\Sigma}$ is not large. Second, this construction requires to know or to estimate the values $\mathbb{E}\,\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\|_2^2$ and $\mathrm{Var}(\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\|_2^2)$ which depend on the unknown covariance operator $\mathbf{\Sigma}$. A partial solution of this problem is discussed in [66]. It involves splitting the sample into three subsamples, and pilot estimation of the mean and the variance of $\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\|_2^2$. The approach only applies in some special cases, in particular, if the covariance matrix has a nearly spike structure.

In this chapter we propose a bootstrap procedure which 1) does not rely on the asymptotic distribution of the error $\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\|_2^2$; does not require to know the moments of $\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\|_2^2$; does not involve any data splitting; provides an explicit error bound for the bootstrap approximation in the case when sample comes from the Gaussian distribution. The procedure is based on the resampling idea which allows to estimate directly the quantiles

$$\gamma_\alpha := \inf\left\{\gamma > 0 \colon \mathbb{P}\Big(n\|\widehat{\mathbf{P}}_r - \mathbf{P}_r\|_2^2 > \gamma\Big) \le \alpha\right\} \tag{3.17}$$

without estimating the covariance matrix $\mathbf{\Sigma}$.

Bootstrap methods belong nowadays to most popular ways for measuring the significance of a test or for building a confidence set. The existing theory based on the high order expansions of the related statistics states the bootstrap validity for various parametric methods. However, an extension to a non-classical situation with a limited sample size and/or high parameter dimension meets serious problems. We refer to series of works [101], [29] which validate a bootstrap procedure for a test based on the maximum of huge number of statistics. Here we make a further step in understanding the range of applicability of a weighted bootstrap method in constructing a finite sample confidence set for a spectral projector. A proof of bootstrap validity in this setup is a challenging task. The spectral projector is a non-linear and non-regular function of the covariance matrix, which itself is a quadratic function of the underlying multivariate distribution

We introduce the following weighted version of $\widehat{\mathbf{\Sigma}}$:

$$\mathbf{\Sigma}^\circ := \frac{1}{n}\sum_{i=1}^n w_i X_i X_i^\mathsf{T},$$

where $w_1, \ldots, w_n$ are i.i.d. random variables, independent of $\mathbf{X} = (X_1, \ldots, X_n)$, with $\mathbb{E}\, w_1 = 1$, $\operatorname{Var} w_1 = 1$. A typical example used in this section is to apply i.i.d. Gaussian weights $w_i \sim \mathcal{N}(1,1)$. Denote by $\mathbb{P}^\circ(\cdot) := \mathbb{P}(\cdot \,\big|\, \mathbf{X})$ and $\mathbb{E}^\circ$ the corresponding conditional probability and expectation. It is obvious that

$$\mathbb{E}^\circ\, \mathbf{\Sigma}^\circ = \widehat{\mathbf{\Sigma}}. \tag{3.18}$$

In what follows we will often refer to "$\mathbf{X}$–world" and "bootstrap world". In the $\mathbf{X}$–world the sample $\mathbf{X}$ is random opposite to the bootstrap world, where $\mathbf{X}$ is fixed, but $w_1, \ldots, w_n$ are random. Then, equation (3.18) implies that in the bootstrap world we know precisely the expectation of $\mathbf{\Sigma}^\circ$ opposite to the $\mathbf{X}$–world, where $\mathbf{\Sigma}$ is unknown. Similarly to (3.15) we may write

$$\mathbf{\Sigma}^\circ = \sum_{j=1}^{p} \sigma_j^\circ \mathbf{u}_j^\circ \mathbf{u}_j^{\circ\mathsf{T}}.$$

Let us denote by $\mathbf{P}_r^\circ$ a projector on the subspace spanned by the direct sum of $\mathbf{u}_j^\circ, j \in \Delta_r$. For a given α we define the quantile γ_α° as

$$\gamma_\alpha^\circ := \min\Big\{\gamma > 0 \colon \mathbb{P}^\circ\Big(n\|\mathbf{P}_r^\circ - \widehat{\mathbf{P}}_r\|_2^2 > \gamma\Big) \le \alpha\Big\}. \tag{3.19}$$

Note that this value γ_α° is defined w.r.t. the bootstrap measure, therefore, it depends on the data $\mathbf{X}$. This bootstrap critical value γ_α° is applied in the $\mathbf{X}$–world to build the confidence set

$$\mathcal{E}(\alpha) := \big\{\mathbf{P} \colon n\|\mathbf{P} - \widehat{\mathbf{P}}_r\|_2^2 \le \gamma_\alpha^\circ\big\}.$$

The main result given below justifies this construction and evaluate the coverage probability of the true projector $\mathbf{P}_r$ by this set. It states that

$$\mathbb{P}\big(\mathbf{P}_r \notin \mathcal{E}(\alpha)\big) = \mathbb{P}\big(n\|\mathbf{P}_r - \widehat{\mathbf{P}}_r\|_2^2 > \gamma_\alpha^\circ\big) \approx \alpha.$$

To formulate the main result of this section we introduce additional notation. Define the following block-matrix

$$\Gamma_r := \begin{pmatrix} \Gamma_{r1} & \mathbf{O} & \ldots & \mathbf{O} \\ \mathbf{O} & \Gamma_{r2} & \mathbf{O}\ldots & \mathbf{O} \\ \ldots & & & \\ \mathbf{O} & \ldots & \mathbf{O} & \Gamma_{rq} \end{pmatrix}, \tag{3.20}$$

where $\Gamma_{rs}, s \neq r$ are diagonal matrices of order $m_r m_s \times m_r m_s$ with values $2\mu_r\mu_s/(\mu_r - \mu_s)^2$ on the main diagonal. Let $\lambda_1(\Gamma_r) \ge \lambda_2(\Gamma_r) \ge \ldots$ be the eigenvalues of Γ_r. The available bounds on the distance between the covariance matrix and its empirical counterpart claim that the eigenvalues of $\mathbf{\Sigma}$ can be recovered with accuracy $O(1/\sqrt{n})$; see e.g. [107], [110], [65], [55]. Therefore, the part of the spectrum of $\mathbf{\Sigma}$ below a threshold of order $O(1/\sqrt{n})$ cannot be estimated. The same applies to the matrix Γ_r. Introduce the corresponding value $\mathfrak{m}$:

$$\lambda_{\mathfrak{m}}(\Gamma_r) \ge \operatorname{Tr}\Gamma_r\left(\sqrt{\frac{\log n}{n}} + \sqrt{\frac{\log p}{n}}\right) > \lambda_{\mathfrak{m}+1}(\Gamma_r). \tag{3.21}$$

Denote by $\Pi_{\mathfrak{m}}$ a projector on the subspace spanned by the eigenvectors of Γ_r corresponding to its largest $\mathfrak{m}$ eigenvalues. Now we state our main result of this section.

Theorem 3.8. *Let observations $X, X_1, \ldots, X_n$ be i.i.d. Gaussian random vectors in $\mathbb{R}^p$ with $\mathbb{E}\, X = 0$ and $\mathbb{E}\, X X^\mathsf{T} = \mathbf{\Sigma}$. Let γ_α° be defined by (3.19) for any $\alpha : 0 < \alpha < 1$, with i.i.d. Gaussian random weights $w_i \sim \mathcal{N}(1,1)$ for $i = 1, \ldots, n$. Then the following bound is fulfilled*

$$\left| \alpha - \mathbb{P}\Bigl(n \| \widehat{\mathbf{P}}_r - \mathbf{P}_r \|_2^2 > \gamma_\alpha^\circ \Bigr) \right| \lesssim \diamondsuit,$$

where

$$\diamondsuit := \frac{\mathfrak{m} \operatorname{Tr} \Gamma_r}{\sqrt{\lambda_1(\Gamma_r)\lambda_2(\Gamma_r)}} \left(\sqrt{\frac{\log n}{n}} + \sqrt{\frac{\log p}{n}} \right) + \frac{\operatorname{Tr}(\mathbf{I} - \Pi_{\mathfrak{m}})\Gamma_r}{\sqrt{\lambda_1(\Gamma_r)\lambda_2(\Gamma_r)}}$$
$$+ \frac{m_r \operatorname{Tr}^3 \mathbf{\Sigma}}{\overline{g}_r^3 \sqrt{\lambda_1(\Gamma_r)\lambda_2(\Gamma_r)}} \left(\sqrt{\frac{\log^3 n}{n}} + \sqrt{\frac{\log^3 p}{n}} \right)$$

and $\mathfrak{m}$ is defined by (3.21).

The proof of Theorem 3.8 may be found in [83]. We outline its main steps. We may show (see [83][Section 4.2])

$$\mathbf{X}\text{–world:} \qquad \mathcal{L}\bigl(n \| \widehat{\mathbf{P}}_r - \mathbf{P}_r \|_2^2 \bigr) \approx \mathcal{L}\bigl(\| \xi \|^2 \bigr), \quad \xi \sim \mathcal{N}(0, \Gamma_r),$$

where Γ_r defined in (3.20). Further, in [83][Section 4.3] we demonstrate that the similar relation holds in the bootstrap world, namely

$$\text{Bootstrap world:} \quad \mathcal{L}\bigl(n \| \mathbf{P}_r^\circ - \widehat{\mathbf{P}}_r \|_2^2 \bigr) \approx \mathcal{L}\bigl(\| \xi^\circ \|^2 \bigr), \quad \xi^\circ \sim \mathcal{N}(0, \Gamma_r^\circ),$$

where Γ_r° is defined in [83][Eq. 24]. To compare ξ and ξ° we apply Gaussian comparison inequality, Theorem 3.1.

Although an analytic expression for the value γ_α° is not available, one can evaluate it from numerical simulations by generating a large number M of independent samples $\{w_1, \ldots, w_n\}$ and computing from them the empirical distribution function of $n \| \mathbf{P}_r^\circ - \widehat{\mathbf{P}}_r \|_2^2$. In fact, standard arguments, see e.g. [99][Section 5.1], in combination with [83][Theorem 5] suggest that the accuracy of Monte-Carlo approximation is of order $M^{-1/2}$. Theorem 3.8 justifies the use of this value γ_α° in place of γ_α defined in (3.17) provided that the error $\diamondsuit$ is sufficiently small

In the conclusion of this section we illustrate the performance of the bootstrap procedure on an artificial example.

Example 1. First we describe our setup. Let n be a sample size. We consider the different values of n, namely $n = 100, 300, 500, 1000, 2000, 3000$. Let $X_1, \ldots, X_n$ have the normal distribution in $\mathbb{R}^p$, with zero mean and covariance matrix $\mathbf{\Sigma}$. The value of p and the choice of $\mathbf{\Sigma}$ will be described below. The distribution of $n \| \widehat{\mathbf{P}}_1 - \mathbf{P}_1 \|_2^2$ is evaluated by using $M = 3000$ Monte-Carlo samples from the normal distribution with zero mean and covariance $\mathbf{\Sigma}$. The bootstrap distribution for a given realization X is evaluated by $M = 3000$ Monte-Carlo samples of bootstrap weights $\{w_1, \ldots, w_n\}$. Since this distribution is random and depends on X, we finally use the median from 50 realizations of X for each quantile. We consider the following parameters: $p = 500$, $\mu_1 = 36, \mu_2 = 30, \mu_3 = 25, \mu_4 = 19$ and all other eigenvalues $\mu_s, s = 5, \ldots, 500$ are uniformly distributed in $[1, 5]$. Here we get $\overline{g}_1 = 6$ and $\mathsf{r}(\mathbf{\Sigma}) = 51.79$. Figure 1 shows the corresponding PP-plots for the empirical distribution of $n \| \widehat{\mathbf{P}}_1 - \mathbf{P}_1 \|_2^2$ against its bootstrap counterpart. Table 2 shows the coverage probabilities of the quantiles estimated using the bootstrap.

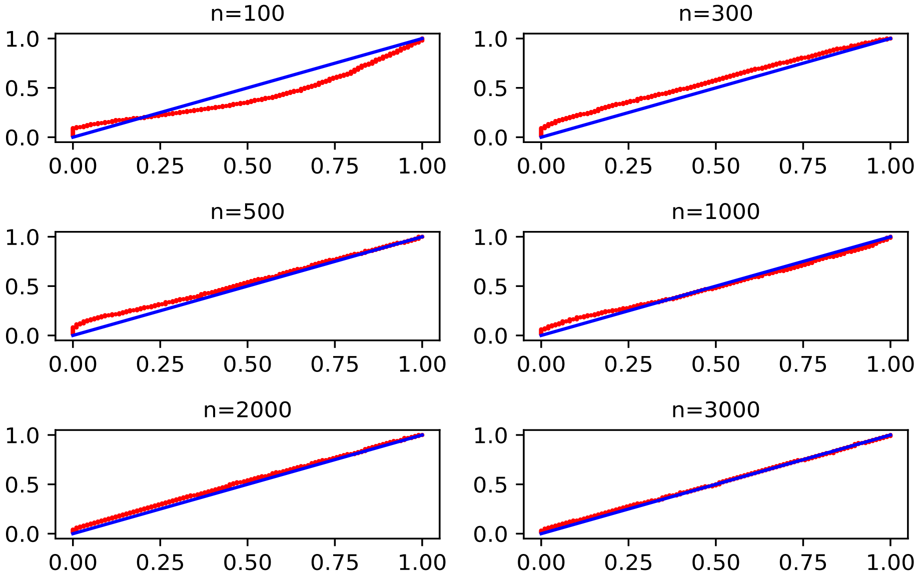

Figure 1: PP-plot of the bootstrap procedure for Example 1.

Table 2: Coverage probabilities for Example 1. For each n the first line corresponds to the median value of the coverage probability and the second line corresponds to the interquartile range.

	Confidence levels					
n	0.99	0.95	0.90	0.85	0.80	0.75
100	0.997	0.986	0.954	0.924	0.889	0.850
	0.004	0.026	0.052	0.074	0.091	0.104
300	0.992	0.937	0.873	0.812	0.754	0.692
	0.026	0.093	0.165	0.207	0.236	0.271
500	0,988	0.962	0.902	0.846	0.788	0.623
	0.054	0.139	0.227	0.264	0.323	0.174
1000	0.992	0.974	0.943	0.890	0.841	0.783
	0.021	0.062	0.114	0.066	0.153	0.170
2000	0.988	0.954	0.891	0.843	0.795	0.741
	0.021	0.059	0.081	0.098	0.126	0.142
3000	0.994	0.961	0.908	0.864	0.815	0.763
	0.016	0.053	0.073	0.081	0.092	0.101

4 On the Stability of Random Matrix Product: Application to Linear Stochastic Approximation and TD Learning

This chapter is concerned with the linear stochastic approximation (LSA) algorithm for solving the linear system $\bar{\mathbf{A}}\theta = \bar{\mathbf{b}}$ with the unique solution $\theta^\star$. In particular, we consider the LSA scheme based on the observations $\{(\mathbf{A}(Z_n), \mathbf{b}(Z_n))\}_{n\in\mathbb{N}}$, where $\mathbf{A} : \mathsf{Z} \to \mathbb{R}^{d\times d}$, $\mathbf{b} : \mathsf{Z} \to \mathbb{R}^d$ are measurable functions and $(Z_k)_{k\in\mathbb{N}}$ is

1. either an i.i.d. sequence with distribution π satisfying
$$\mathbb{E}[\mathbf{A}(Z_1)] = \bar{\mathbf{A}} \text{ and } \mathbb{E}[\mathbf{b}(Z_1)] = \bar{\mathbf{b}}\,. \tag{4.1}$$
2. or a Markov chain, taking values in a general state-space Z with unique invariant distribution π and $\lim_{n\to+\infty} \mathbb{E}[\mathbf{A}(Z_n)] = \bar{\mathbf{A}}, \lim_{n\to+\infty} \mathbb{E}[\mathbf{b}(Z_n)] = \bar{\mathbf{b}}$.

With a sequence of stepsizes $\{\alpha_n\}_{n\in\mathbb{N}}, \alpha_n > 0$ the LSA algorithm consists in the sequences of estimates $\{\theta_n\}_{n\in\mathbb{N}}$, defined by:

$$\theta_n = \theta_{n-1} - \alpha_n\{\mathbf{A}(Z_n)\theta_{n-1} - \mathbf{b}(Z_n)\}\,, \tag{4.2}$$

with the deterministic initialization θ_0. The LSA recursion (4.2) encompasses a wide range of algorithms. LSA is central to the analysis of identification algorithms and control of linear systems. Early results have focused on these two applications and studied both the asymptotic behaviour of the sequence $(\theta_n)_{n\in\mathbb{N}}$ and the tracking error; see [43, 51, 53, 74] and the references therein.

LSA is also a cornerstone in the analysis of linear value-function estimation (LVE) that are popular in reinforcement learning [103, 15]. Seminal works on this topic [15, 108, 14] established conditions for asymptotic convergence. Finite-time bound for LVE (and more generally LSA) has attracted a renewed interest. In the case when $(Z_i)_{i\in\mathbb{N}^*}$ is an i.i.d. sequence, [69, 30] have investigated mean-squared error bounds for LSA. Recent developments [16, 102, 27] have considered the setting that $(Z_i)_{i\in\mathbb{N}^*}$ is a Markov chain, and provided finite-time analysis. On a related subject, [54, 112, 37, 63] considered linear two-timescale stochastic approximation that involves coupled LSA recursions.

4.1 LSA driven by general state space Markov chain

The results of this subsection are published in [42].

Most of the existing results on LSA are limited by strong conditions such as (i) uniform geometric ergodicity (UGE) on the Markov chain and/or (ii) uniformly bounded $\mathbf{A}, \mathbf{b}$, i.e. $\sup_{z\in\mathsf{Z}}\{\|\mathbf{A}(z)\|+\|\mathbf{b}(z)\|\} < +\infty$. These conditions are restrictive since the UGE condition typically requires the state space to be finite or compact and do not extend to general (unbounded) state space. This is of course a limitation because many applications involve general unbounded state space; see e.g. [74] and [15, p. 305].

In this chapter, we aim to provide high-order moment bounds on the LSA with Markovian noise. Our results are applicable under the relaxed conditions: (i) $(Z_i)_{i\in\mathbb{N}^*}$ is a Markov chain on a general (possibly unbounded) state-space satisfying a super-Lyapunov drift condition, and (ii) for some constant $C \geq 0$, for any $z \in \mathsf{Z}$, $\|\mathbf{A}(z)\| \leq C\mathrm{W}_1(z)$, $\|\mathbf{b}(z)\| \leq C\mathrm{W}_2(z)$, with $\mathrm{W}_1, \mathrm{W}_2 : \mathbb{R}_+ \to [1, +\infty)$ deduced from the drift condition in (i). They are strictly weaker than the conditions required in previously reported works. In particular, $\mathbf{A}, \mathbf{b}$ can be potentially unbounded.

For $m, n \in \mathbb{N}$, $m < n$ and $z_{m+1:n} = (z_{m+1}, \ldots, z_n) \in \mathsf{Z}^{n-m}$, we define

$$\Gamma_{m+1:n}(z_{m+1:n}) = \textstyle\prod_{i=m+1}^{n}\{\mathrm{I}_d - \alpha_i\mathbf{A}(z_i)\}\,.$$

A key property used for deriving our bounds is an exponential stability result on the matrix product above, $\Gamma_{m+1:n}(Z_{m+1:n})$, for $m, n \in \mathbb{N}$, $m < n$. To motivate why this is relevant to LSA, suppose that the Markov chain $(Z_n)_{n\in\mathbb{N}^*}$ is ergodic so that, for all $z \in \mathsf{Z}$, the following limits $\bar{\mathbf{A}} = \lim_{n\to\infty} \mathbb{E}_z[\mathbf{A}(Z_n)]$, $\bar{\mathbf{b}} = \lim_{n\to\infty} \mathbb{E}_z[\mathbf{b}(Z_n)]$ exist. Assume in addition that the limiting matrix $-\bar{\mathbf{A}}$ is *Hurwitz*, i.e. the real parts of its eigenvalues are strictly negative, and denote by $\theta^\star$ the unique solution of the linear system $\bar{\mathbf{A}}\theta^\star = \bar{\mathbf{b}}$. The n-th error vector $\tilde{\theta}_n = \theta_n - \theta^\star$ may be expressed, for all $n \in \mathbb{N}$, by

$$\tilde{\theta}_n = \textstyle\sum_{j=1}^n \alpha_j \Gamma_{j+1:n}(Z_{j+1:n})\bar{\varepsilon}(Z_j) + \Gamma_{1:n}(Z_{1:n})\tilde{\theta}_0 \,, \tag{4.3}$$

where $\bar{\varepsilon}(Z_j) = \mathbf{b}(Z_j) - \bar{\mathbf{b}} - \{\mathbf{A}(Z_j) - \bar{\mathbf{A}}\}\theta^\star$. Obtaining a bound on p-th moments for $\{\|\tilde{\theta}_n\|\}_{n\in\mathbb{N}}$ naturally requires that the sequence of random matrices $\{\mathbf{A}(Z_i)\}_{i\in\mathbb{N}^*}$ to be (V, q)-*exponentially stable*. Recall that for $q \geq 1$ and a function $\mathrm{V} : \mathsf{Z} \to [1, \infty)$, $\{\mathbf{A}(Z_i)\}_{i\in\mathbb{N}^*}$ is said to be (V, q)-exponentially stable if there exists $\mathsf{a}_q, \mathrm{C}_q > 0$ and $\alpha_{\infty,q} < \infty$ such that, for any sequence of positive step sizes $(\alpha_i)_{i\in\mathbb{N}^*}$ satisfying $\sup_{i\in\mathbb{N}^*} \alpha_i \leq \alpha_{\infty,q}$, $z \in \mathsf{Z}$, $m, n \in \mathbb{N}$, $m < n$,

$$\mathbb{E}_z[\|\Gamma_{m+1:n}(Z_{m+1:n})\|^q] \leq \mathrm{C}_q \exp\left(-\mathsf{a}_q \textstyle\sum_{i=m+1}^n \alpha_i\right) \mathrm{V}(z) \,. \tag{4.4}$$

Intuitively, (V, q)-exponential stability means that the q-th moment of the product of random matrices $\Gamma_{m+1:n}(Z_{m+1:n})$ behaves similarly to that of the product of *deterministic* matrices $G_{m+1:n} = \prod_{i=m+1}^n (\mathrm{I}_d - \alpha_i \bar{\mathbf{A}})$, under the assumption that $-\bar{\mathbf{A}}$ is Hurwitz.

Fix $p, q, r \in \mathbb{N}^*$ such that $p^{-1} = q^{-1} + r^{-1}$. Assume that the sequence $\{\mathbf{A}(Z_i)\}_{i\in\mathbb{N}^*}$ is (V, q)-exponentially stable for some $q > 1$, the r-th moments of the noise term $\|\bar{\varepsilon}(Z_n)\|$ and initialization error $\tilde{\theta}_0$ are bounded. Using (4.3), we can readily derive bounds for the p-th moment, $\mathbb{E}_z^{1/p}[\|\tilde{\theta}_n\|^p]$ by applying the Hölder's inequality. Note that the r-th moment bound for the "noise" terms may follow from classical Lyapunov drift conditions, which is implied by super-Lyapunov drift conditions.

In this section:

- We establish (V, q)-exponential stability of the sequence of matrices $\{\mathbf{A}(Z_k)\}_{k\in\mathbb{N}^*}$, and provide explicit expression for constants appearing in (4.4); see Theorem 4.2. Compared to the prior works, our result can be applied to the settings where the function $\mathbf{A}(\cdot)$ is unbounded, not symmetric and $(Z_k)_{k\in\mathbb{N}^*}$ is a Markov chain on a general (unbounded) state-space not constrained to be uniformly geometrically ergodic. A discussion of how our results relax the restrictive conditions in previously reported works is given after the statement of Theorem 4.2.

- We provide finite-time bound and first-order expansion for the p-th moment of the error $(\tilde{\theta}_n)_{n\in\mathbb{N}^*}$ for LSA recursion (4.3). More precisely, we show that $\mathbb{E}_z^{1/p}[\|\tilde{\theta}_n\|^p] = \mathcal{O}(\alpha_n^{1/2})\, \mathrm{V}_p(z)$ both for constant $\alpha_n \equiv \alpha$ (where α is sufficiently small) or nonincreasing stepsizes under weak additional conditions including $\alpha_n = C/(n + n_0)^{\mathrm{t}}$, for any $\mathrm{t} \in (0, 1]$; see Theorem 4.3. From our analysis on the LSA error $\tilde{\theta}_n$, we identify a leading term, denoted $J_n^{(0)}$, which is a weighted additive linear functional of the error process $(\bar{\varepsilon}(Z_n))_{n\in\mathbb{N}^*}$. Furthermore, the leading term $J_n^{(0)}$ and its remainder $H_n^{(0)} = \tilde{\theta}_n - J_n^{(0)}$ admit a separation of scales. For example, when $\alpha_n = C/(n + n_0)$, the leading term has a p-th moment bound of $\mathcal{O}(n^{-1/2})\, \mathrm{V}_p(z)$, and the remainder has a p-th moment bound of $\mathcal{O}(n^{-1}\log(n))\, \mathrm{V}_p(z)$; see Theorem 4.4.

- Finally, we apply our results to TD-learning for LVE. We give sufficient conditions for a Markov Reward Process on general (unbounded) state space (with unbounded reward and feature functions) to satisfy the assumptions of Theorem 4.3 and Theorem 4.4. Therefore, the convergence bounds we derive hold for these algorithms.

4.1.1 Main Results

Consider a Markov chain $(Z_k)_{k\in\mathbb{N}}$ with Markov kernel P. We assume without loss of generality that $(Z_k)_{k\in\mathbb{N}}$ is the canonical process corresponding to P on $(\mathsf{Z}^{\mathbb{N}}, \mathcal{Z}^{\otimes\mathbb{N}})$. We denote by $\mathbb{P}_\mu$ and $\mathbb{E}_\mu$ the corresponding probability distribution and expectation with initial distribution μ. By construction, for any $\mathsf{A} \in \mathcal{Z}$, $\mathbb{P}_\mu\left(Z_k \in \mathsf{A} \,|\, Z_{k-1}\right) = \mathrm{P}(Z_{k-1}, \mathsf{A})$, $\mathbb{P}_\mu$-a.s. In the case $\mu = \delta_z$, $z \in \mathsf{Z}$, $\mathbb{P}_\mu$ and $\mathbb{E}_\mu$ are denoted by $\mathbb{P}_z$ and $\mathbb{E}_z$. In addition, throughout this paper, we assume

UE 1. *The Markov kernel* $\mathrm{P} : \mathsf{Z} \times \mathcal{Z} \to \mathbb{R}_+$ *is irreducible and aperiodic. There exist* $c > 0, \mathrm{b} > 0, \delta \in (1/2, 1]$, $R_0 \geq 0$, *and* $V : \mathsf{Z} \to [\mathrm{e}, \infty)$ *such that by setting* $W = \log V$, $\mathsf{C}_0 = \{z : W(z) \leq R_0\}$, $\mathsf{C}_0^{\complement} = \{z : W(z) > R_0\}$, *we have*

$$\mathrm{P}V(z) \leq \exp[-cW^\delta(z)]V(z)1_{\{\mathsf{C}_0^{\complement}\}}(z) + \mathrm{b}\,1_{\{\mathsf{C}_0\}}(z)\,. \tag{4.5}$$

In addition, for any $R \geq 1$, *the level sets* $\{z : W(z) \leq R\}$ *are* $(m_R, \varepsilon_R\nu)$*-small for* P, *with* $m_R \in \mathbb{N}^*$, $\varepsilon_R \in (0, 1]$ *and* ν *being a probability measure on* $(\mathsf{Z}, \mathcal{Z})$.

Since $(\mathsf{Z}, \mathcal{Z})$ is a general state-space, irreducibility here means that the Markov kernel P admits an accessible small set; see [38, Chapter 9]. The drift condition (4.5) in UE 1 is referred to as a multiplicative or super-Lyapunov drift condition and plays a key role in studying the large deviations of additive functionals of Markov chains; see [109]. Eq. (4.5) implies the classical Foster-Lyapunov drift condition, $\mathrm{P}V(z) \leq \lambda V(z) + \mathrm{b}\,\mathbf{b}1_{\mathsf{C}_0}(z)$ with

$$\lambda = \exp(-c \inf\nolimits_{\mathsf{C}_0^{\complement}} W^\delta) \leq \exp(-c) < 1\,. \tag{4.6}$$

It follows from [38, Theorem 15.2.4] that under UE 1 the Markov kernel P is V-uniformly geometrically ergodic and admits a unique stationary distribution π, i.e. there exists $\rho \in (0, 1)$ and $\mathrm{B}_V < \infty$ such that for each $z \in \mathsf{Z}$ and $n \in \mathbb{N}$,

$$\|\mathrm{P}^n(z, \cdot) - \pi\|_V \leq \mathrm{B}_V \rho^n V(z)\,. \tag{4.7}$$

UE 1 is a special case of condition **(DV3)** in [68, 67] which plays a key role in multiplicative regularity of Markov chains. A key consequence of UE 1 is a bound for products (see [42][Lemma 10] and [67, Theorem 1.2]): for any $z \in \mathsf{Z}$, $n \in \mathbb{N}$, and non-increasing sequence $(\alpha_i)_{i\in\mathbb{N}^*} \subset [0, 1]$, we get

$$\mathbb{E}_z[\exp\{c\textstyle\sum_{k=0}^{n-1}\alpha_k W^\delta(Z_k)\}] \leq \exp\{\tilde{\mathrm{b}}\textstyle\sum_{k=0}^{n-1}\alpha_k\}\exp\{\alpha_1 W(z)\}\,,$$

where $\tilde{\mathrm{b}} = \log \mathrm{b} + \sup_{r\geq \mathrm{e}}\{cr^\delta - r\}$ and c is defined in (4.5). UE 1 is satisfied with $\delta = 1$ for Gaussian linear vector auto-regressive process and also non-linear auto-regressive process under exponential moment condition for innovation process, see e.g. [89].

We also impose some constraints on $\mathbf{A}$. For $\varepsilon \in (0, 1)$ consider the following assumptions

A 1 (ε)**.** *Given* $\varepsilon \in (0, 1)$ *there exists* $\mathrm{C}_A > 0$ *such that for any* $1 \leq i, j \leq d$, *the* (i, j)*-th element of* $\mathbf{A}$ *satisfies* $\|[\mathbf{A}]_{i,j}\|_{W^\beta} \leq \mathrm{C}_A$, *where* $\beta < \min(2\delta - 1, \delta/(1 + \varepsilon))$ *and* δ *is given in UE 1.*

A 2. *The square matrix* $-\bar{\mathbf{A}} = -\mathbb{E}_\pi[\mathbf{A}(Z_0)]$ *is Hurwitz.*

A 1(ε), A 2 are standard conditions on the parameter matrices in LSA. Meanwhile, A 2 guarantees the existence of a unique solution $\theta^\star$ to $\bar{\mathbf{A}}\theta = \bar{\mathbf{b}}$. It is a sufficient and necessary condition for the solution of the ordinary differential equation $\dot{\theta}_t = -\bar{\mathbf{A}}\theta_t$ to converge exponentially to $\theta^\star$ [59, Lemma 4.1.2]. The same kind of result holds for the discrete system $\theta^{\mathsf{d}}_{n+1} - \theta^{\mathsf{d}}_n = -\alpha\bar{\mathbf{A}}\theta^{\mathsf{d}}_n$.

Proposition 4.1 (See [88][Lemma 9.1, p. 140)**.** *] Assume that $-\bar{\mathbf{A}}$ is a Hurwitz matrix. Then there exists a unique positive definite matrix Q satisfying the Lyapunov equation $\bar{\mathbf{A}}^\top Q + Q\bar{\mathbf{A}} = \mathrm{I}$. In addition, setting*

$$a = \|Q\|^{-1}/2\,, \quad \text{and} \quad \alpha_\infty = (1/2)\|\bar{\mathbf{A}}\|_Q^{-2}\|Q\|^{-1}\,, \tag{4.8}$$

then for any $\alpha \in [0, \alpha_\infty]$, we get $\|\mathrm{I} - \alpha\bar{\mathbf{A}}\|_Q^2 \leq 1 - a\alpha$. If in addition $\alpha \leq \|Q\|^2$ then $1 - a\alpha \geq 1/2$.

The above proposition implies that the discrete system converges exponentially as $\|\theta^{\mathsf{d}}_{n+1}\| \leq \sqrt{\kappa_{\mathsf{Q}}}(1 - a\alpha)^{n/2}\|\theta^{\mathsf{d}}_0\|$ for $\alpha \in (0, \alpha_\infty)$.

Our aim is to establish (V, q)-exponential stability of the sequence $\{\mathbf{A}(Z_k)\}_{k\in\mathbb{N}^*}$. The following example illustrates that, even if the function $\mathbf{A}(z)$ is bounded, for the matrix product to be exponentially stable, it is necessary for the Markov chain $(Z_k)_{k\in\mathbb{N}}$ to be geometrically ergodic.

The following theorem establishes the (V, p)-exponential stability of the sequence $\{\mathbf{A}(Z_k)\}_{k\in\mathbb{N}^*}$. For ease of notation, we simply denote $\Gamma_{m+1:n} = \Gamma_{m+1:n}(Z_{m+1:n})$.

Theorem 4.2. *Assume UE 1, A 1(ε) and A 2. Then for any $p \geq 1$, there exists $\alpha_{\infty,p} > 0$, given in [42][Eq. 87], such that for any non-increasing sequence $(\alpha_k)_{k\in\mathbb{N}^*}$ satisfying $\alpha_1 \in (0, \alpha_{\infty,p})$, $z_0 \in \mathsf{Z}$ and $m, n \in \mathbb{N}$, $m < n$, it holds*

$$\mathbb{E}^{1/p}_{z_0}[\|\Gamma_{m+1:n}\|^p] \leq \mathrm{C}_{\mathsf{st},p}\,\mathrm{e}^{-(a/4)\sum_{\ell=m+1}^n \alpha_\ell} V^{1/2p}(z_0)\,, \tag{4.9}$$

where a, $\mathrm{C}_{\mathsf{st},p}$, and h are defined in (4.8), *[42][Eq. 89], and [42][Eq. 86], respectively.*

The theorem shows that provided $(\alpha_k)_{k\in\mathbb{N}^*}$ satisfies $\sum_{k\in\mathbb{N}^*}\alpha_k = +\infty$,

$$\mathbb{E}^{1/p}_z[\|\Gamma_{m+1:n}\|^p] \to 0$$

as $(n - m) \to \infty$ for any $p \geq 1$. Specifically, it has a similar convergence rate as the deterministic matrix product $\|G_{m+1:n}\| = \|\prod_{i=m+1}^n(\mathrm{I}_d - \alpha_i\bar{\mathbf{A}})\| \lesssim \mathrm{e}^{-a\sum_{\ell=m+1}^n \alpha_\ell}$.

Theorem 4.2 generalizes previously reported works. [51, 52] used a slightly different definitions allowing to consider non-Markovian processes satisfying more general mixing conditions (like ϕ- or β-mixing). As we will see later, when specialized to Markov chains, the results we obtain significantly improve those reported in these works. [89] established (V, q)-exponential stability for general state-space Markov chain under a super-Lyapunov drift condition (similar to UE 1). However, the results in [89] assume constant stepsize and $\bar{\mathbf{A}}(z)$ being symmetric and non-negative definite for any $z \in \mathsf{Z}$. Non-negative definiteness plays a key role in the arguments: in such case, for any $z \in \mathsf{Z}$, the spectral norm $\|\,\mathrm{I}_d - \alpha\mathbf{A}(z)\| \leq 1$ provided that $\|\mathbf{A}(z)\| \leq \alpha^{-1}$ for $\alpha > 0$ which is no longer true for general matrix-valued function $\mathbf{A}(z)$. Similar results, also under the condition that $\mathbf{A}(z)$ is symmetric for any $z \in \mathsf{Z}$, were obtained by [35] based on perturbation theory for linear operators in Banach space and spectral theory. However, the bounds provided in [35] are only qualitative and it is difficult to make these results quantitative because they are based on perturbation arguments of linear operators in Banach spaces. The restrictions imposed on these prior works have limited their applications to more general algorithms, in particular to most RL algorithms. As we will see below, the application to linear value-function estimation in temporal difference learning involve non-symmetric matrix function $\mathbf{A}$. In contrast, our result (cf. Theorem 4.2) can be applied to the setting where for some $z \in \mathsf{Z}$, $\mathbf{A}(z)$ is not necessary non-negative symmetric but only Hurwitz.

Notice that the case of uniformly geometric ergodic Markov chain is covered by UE 1. In this case the set Z is small and drift function V can be chosen to be constant. Together with the assumption of bounded $\mathbf{A}(\cdot)$, the exponential stability of product of random matrices has been implicitly established in [102, 37, 63, 27]. In particular, their results on LSA can be applied on the recursion $y_0 = y$, $y_{n+1} = \{\mathrm{I}_d - \alpha_{n+1}\mathbf{A}(Z_{n+1})\}y_n$, $n \in \mathbb{N}$. Through studying the decomposition:

$$y_{n+1} = \{\mathrm{I}_d - \alpha_{n+1}\bar{\mathbf{A}}\}y_n - \alpha_{n+1}(\mathbf{A}(Z_{n+1}) - \bar{\mathbf{A}})y_n, \ \forall\ n \in \mathbb{N}, \tag{4.10}$$

they derived bounds on $\mathbb{E}_{z_0}[\|y_{n+1}\|^p] = \mathbb{E}_{z_0}[\|\Gamma_{1:n+1}y\|^p]$. However, generalizing this approach for other classes of Markov chains (e.g., UE 1) or unbounded function appears to be impossible.

4.1.2 Application to Linear Stochastic Approximation

This section illustrates how to apply Theorem 4.2 to analyze LSA schemes with Markovian noise. First, we state the assumptions on $\mathbf{b}(\cdot)$ and step sizes which can be either constant or diminishing. For $\mathsf{K} \in \mathbb{N}^*$, consider the following assumption:

A 3 (K). *There exists* $\mathrm{C}_{b,\mathsf{K}} > 0$ *such that* $\max_{1\le \ell\le d}\|\mathbf{b}_\ell\|_{V^{1/\mathsf{K}}} \le \mathrm{C}_{b,\mathsf{K}}$*, where* $\mathbf{b}_\ell$ *is the* ℓ*-th component of* $\mathbf{b}$.

A 4. *There exists a constant* $0 < \mathrm{c}_\alpha \le a/16$ *such that for* $k \in \mathbb{N}$, $\alpha_k/\alpha_{k+1} \le 1 + \alpha_{k+1}\,\mathrm{c}_\alpha$.

It is easy to check that A 4 is satisfied by diminishing step sizes $\alpha_n = \mathrm{C}_a(n + n_0)^{-\mathrm{t}}$, $\mathrm{t} \in (0, 1]$ and constant step sizes.

Theorem 4.3. *Let* $\mathsf{K} \ge 8$. *Assume UE 1, A 1(*ε*), A 2 and A 3(*K*). For any* $2 \le p \le K/4$*, there exists* $\alpha_{\infty,p}^{(0)}$ *defined in [42][Eq. 25] such that for any non-increasing sequence* $(\alpha_k)_{k\in\mathbb{N}^*}$ *satisfying* $\alpha_1 \in (0, \alpha_{\infty,p}^{(0)})$ *and A 4,* $z \in \mathsf{Z}$*, and* $n \in \mathbb{N}$*, it holds*

$$\mathbb{E}_z^{1/p}[\|\tilde{\theta}_n\|^p] \le \mathrm{M}_0\,\mathrm{C}_{\mathsf{st},2p}\,\mathrm{e}^{-(a/4)\sum_{\ell=1}^{n}\alpha_\ell}V^{1/(4p)}(z) + (\mathrm{C}_{\mathsf{J},p}^{(0)} + \mathrm{C}_{\mathsf{H},p}^{(0)})\sqrt{\alpha_n}V^{2/\mathsf{K}+1/(4p)}(z), \tag{4.11}$$

where $\mathrm{M}_0 = \mathbb{E}_z^{1/(2p)}[\|\tilde{\theta}_0\|^{2p}]$ *and* $\mathrm{C}_{\mathsf{J},p}^{(0)}, \mathrm{C}_{\mathsf{H},p}^{(0)}$ *are defined in [42][Eq. 32], [42][Eq. 34], respectively.*

Most often, the distribution of the initial value $\tilde{\theta}_0$ does not depend on the initial value of the Markov chain z. In this case $\mathbb{E}_z^{1/(2p)}[\|\tilde{\theta}_0\|^{2p}]$ is a constant. With a sufficiently small step size, Theorem 4.3 shows that the L_p norm of error vector converges under UE 1 for the Markov chain. Compared to [102], we consider relaxed conditions on the Markov chain and allow for diminishing step sizes in the LSA.

We show that the finite-time L_p error bound can be derived through applying the stability of random matrix product (see Theorem 4.2). We recall that the error vector $\tilde{\theta}_{n+1} = \theta_{n+1} - \theta^\star$ may be expressed as

$$\tilde{\theta}_{n+1} = \Gamma_{1:n+1}\tilde{\theta}_0 + \textstyle\sum_{j=1}^{n+1}\alpha_j\Gamma_{j+1:n+1}\bar{\varepsilon}(Z_j) \equiv \tilde{\theta}_{n+1}^{(\mathsf{tr})} + \tilde{\theta}_{n+1}^{(\mathsf{fl})}\,. \tag{4.12}$$

Using the Hölder's inequality and Theorem 4.2, the transient term $\tilde{\theta}_{n+1}^{(\mathsf{tr})}$ can be bounded as follows

$$\mathbb{E}_z^{1/p}[\left\|\tilde{\theta}_{n+1}^{(\mathsf{tr})}\right\|^p] \le \mathbb{E}_z^{1/(2p)}[\|\Gamma_{1:n+1}\|^{2p}]\mathbb{E}_z^{1/(2p)}[\|\tilde{\theta}_0\|^{2p}] \le \mathrm{M}_0\,\mathrm{C}_{\mathsf{st},2p}\,\mathrm{e}^{-(a/4)\sum_{\ell=1}^{n+1}\alpha_\ell}V^{1/(4p)}(z).$$

As for the fluctuation term $\tilde{\theta}^{(\mathsf{fl})}_{n+1}$, it can be verified that $\tilde{\theta}^{(\mathsf{fl})}_{n+1} = J^{(0)}_{n+1} + H^{(0)}_{n+1}$, where the latter terms are defined by the following pair of recursions:

$$\begin{aligned} J^{(0)}_{n+1} &= (\mathrm{I}_d - \alpha_{n+1} A)\, J^{(0)}_n + \alpha_{n+1} \bar{\varepsilon}(Z_{n+1}), & J^{(0)}_0 &= 0, \\ H^{(0)}_{n+1} &= (\mathrm{I}_d - \alpha_{n+1} \mathbf{A}(Z_{n+1}))\, H^{(0)}_n - \alpha_{n+1} \widetilde{\mathbf{A}}(Z_{n+1}) J^{(0)}_n, & H^{(0)}_0 &= 0, \end{aligned} \tag{4.13}$$

and $\widetilde{\mathbf{A}}(z) = \mathbf{A}(z) - A$. Furthermore, we observe that

$$J^{(0)}_{n+1} = \textstyle\sum_{j=1}^{n+1} \alpha_j G_{j+1:n+1} \bar{\varepsilon}(Z_j), \quad H^{(0)}_{n+1} = -\textstyle\sum_{j=1}^{n+1} \alpha_j \Gamma_{j+1:n+1} \widetilde{\mathbf{A}}(Z_j) J^{(0)}_{j-1} \,. \tag{4.14}$$

Rosenthal's inequality for Markov chains implies that

$$\mathbb{E}^{1/p}_z[\big\| J^{(0)}_{n+1} \big\|^p] \leq \mathrm{C}^{(0)}_{\mathsf{J},p} \sqrt{\alpha_{n+1}} V^{1/\mathsf{K}}(z)\,, \tag{4.15}$$

To analyze $H^{(0)}_{n+1}$, from (4.14) we apply the Hölder's inequality twice to get

$$\mathbb{E}^{1/p}_z[\|H^{(0)}_{n+1}\|^p] \leq \textstyle\sum_{j=1}^{n+1} \alpha_j \mathbb{E}^{1/(2p)}_z[\|\Gamma_{j+1:n+1}\|^{2p}]\, \mathbb{E}^{1/(4p)}_z[\|\widetilde{\mathbf{A}}(Z_j)\|^{4p}]\, \mathbb{E}^{1/(4p)}_z[\|J^{(0)}_{j-1}\|^{4p}].$$

Finally, applying (4.15) we get

$$\mathbb{E}^{1/p}_z[\|H^{(0)}_{n+1}\|^p] \leq \mathrm{C}^{(0)}_{\mathsf{H},p} \sqrt{\alpha_{n+1}} V^{2/\mathsf{K}+1/(4p)}(z). \tag{4.16}$$

Refining the error bound $\mathbb{E}^{1/p}_z[\|\tilde{\theta}^{(\mathsf{fl})}_n\|^p]$ It is possible to obtain a bound on $\mathbb{E}^{1/p}_z[\|H^{(0)}_n\|^p]$ tighter than $\mathcal{O}(\sqrt{\alpha_n})$ obtained in (4.16). This establishes in particular that $J^{(0)}_n$ is the leading term in the decomposition of the fluctuation term $\tilde{\theta}^{(\mathsf{fl})}_{n+1} = J^{(0)}_{n+1} + H^{(0)}_{n+1}$. To this end, we rely on an extra decomposition step similar to (4.13). We may further decompose the error term $H^{(0)}_n$ as $H^{(0)}_n = J^{(1)}_n + H^{(1)}_n$ such that

$$\begin{aligned} J^{(1)}_{n+1} &= (\mathrm{I}_d - \alpha_{n+1} A) J^{(1)}_n - \alpha_{n+1} \widetilde{\mathbf{A}}(Z_{n+1}) J^{(0)}_n, & J^{(1)}_0 &= 0, \\ H^{(1)}_{n+1} &= (\mathrm{I}_d - \alpha_{n+1} \mathbf{A}(Z_{n+1})) H^{(1)}_n - \alpha_{n+1} \widetilde{\mathbf{A}}(Z_{n+1}) J^{(1)}_n, & H^{(1)}_0 &= 0, \end{aligned} \tag{4.17}$$

where $J^{(0)}_n$ is defined in (4.13). For diminishing step sizes, here we should strengthen the previous assumption A 4 as:

A 5. *We have $\mathcal{A}_0 < \infty$, where $\mathcal{A}_n = \sum_{\ell=n}^{\infty} \alpha_\ell^2$. There exists a constant $0 < \mathrm{c}_\alpha \leq a/32$ such that for $k \in \mathbb{N}$, $\alpha_k/\alpha_{k+1} \leq 1 + \alpha_{k+1} \mathrm{c}_\alpha$ and $\alpha_k/\mathcal{A}_{k+1} \leq (2/3)\, \mathrm{c}_\alpha$.*

It is easy to check that A 5 is satisfied by diminishing step sizes $\alpha_n = \mathrm{C}_a (n + n_0)^{-\mathrm{t}}$, $\mathrm{t} \in (\frac{1}{2}, 1]$.

Theorem 4.4. *Let $\mathsf{K} \geq 32$ and assume UE 1, A 1(ε), A 2, and A 3(K). For any $2 \leq p \leq \mathsf{K}/16$ and any non-increasing sequence $(\alpha_k)_{k \in \mathbb{N}}$ satisfying $\alpha_0 \in (0, \alpha^{(1)}_{\infty,p})$ such that $\alpha_k \equiv \alpha$ or A 5 holds. For any $z \in \mathsf{Z}$, $n \in \mathbb{N}$, it holds*

$$\mathbb{E}^{1/p}_z[\|H^{(0)}_n\|^p] \leq V^{3/\mathsf{K}+9/(16p)}(z) \begin{cases} \mathrm{C}^{(\mathsf{f})}_p\, \alpha \sqrt{\log(1/\alpha)}, & \textit{if } \alpha_n \equiv \alpha, \\ \mathrm{C}^{(\mathsf{d})}_p \sqrt{\alpha_n \mathcal{A}_n \log(1/\alpha_n)}, & \textit{if under A 5,} \end{cases} \tag{4.18}$$

where $\alpha^{(1)}_{\infty,p}$, $\mathrm{C}^{(\mathsf{f})}_p, \mathrm{C}^{(\mathsf{d})}_p$ are given in [42][Eq. 92], [42][Eq. 94], respectively.

The theorem shows that the previous bound of $\mathbb{E}^{1/p}_z[\|H^{(0)}_n\|^p] = \mathcal{O}(\sqrt{\alpha_n})$ can be improved to $\mathcal{O}(\sqrt{\alpha_n \mathcal{A}_n \log(1/\alpha_n)})$. Take for example a diminishing step size as $\alpha_n = \mathrm{C}_a (n + n_0)^{-1}$, our result shows that the fluctuation term admits a *clear separation of scales* as

$$\tilde{\theta}^{(\mathsf{fl})}_n = J^{(0)}_n + H^{(0)}_n \text{ with } \mathbb{E}^{1/p}_z[\|J^{(0)}_n\|^p] = \mathcal{O}(n^{-1/2}),\ \mathbb{E}^{1/p}_z[\|H^{(0)}_n\|^p] = \mathcal{O}(n^{-1}\sqrt{\log n}).$$

4.1.3 Temporal Difference Learning Algorithms

Following the notation from [104, Chapter 12], we consider a discounted Markov Reward Process (MRP) denoted by the tuple $(\mathsf{X}, \mathrm{Q}, \mathrm{R}, \gamma)$, where Q is the state transition kernel defined on a general state space $(\mathsf{X}, \mathcal{X})$. We do not assume that X is finite and countable, the only requirement being that $\mathcal{X}$ is countably generated: we may assume for example that $\mathsf{X} = \mathbb{R}^d$. For any given state $x \in \mathsf{X}$, the scalar $\mathrm{R}(x)$ represents the reward of being at the state x. The reward function is possibly unbounded. Finally, $\gamma \in (0,1)$ is the discount factor. The value function $V^\star : \mathsf{X} \to \mathbb{R}$ is defined as the expected discounted reward $V^\star(x) = \mathbb{E}_x[\sum_{k=0}^\infty \gamma^k \mathrm{R}(X_k)]$.

Let $d \in \mathbb{N}^*$, we associate with every state $x \in \mathsf{X}$ a *feature vector* $\psi(x) \in \mathbb{R}^d$ and approximate $V^\star(x)$ by a linear combination $V_\theta(x) = \psi(x)^\top \theta$ (see [108, 104]). Temporal difference learning algorithms may be expressed as

$$\theta_{k+1} = \theta_k + \alpha_{k+1} \varphi_k \{\mathrm{R}(X_k) + \gamma \psi(X_{k+1})^\top \theta_k - \psi(X_k)^\top \theta_k\}, \tag{4.19}$$

where $\{\varphi_k\}_{k\in\mathbb{N}}$ is a sequence of eligibility vectors. For the TD(0) algorithm, $\varphi_k = \psi(X_k)$. For the TD(λ) algorithm, $\varphi_k = (\lambda\gamma)\varphi_{k-1} + \psi(X_k)$. Note that for TD($\lambda$), (4.19) corresponds to (4.2) with the extended Markov chain $Z_k = (X_k, X_{k+1}, \varphi_k)$ and $\bar{\mathbf{A}}(Z_k) = -\varphi_k(\psi(X_k)^\top - \gamma\psi(X_{k+1})^\top)$, $b(Z_k) = \varphi_k \mathrm{R}(X_k)$. [102] were able to study TD(λ) while that $(Z_k)_{k\in\mathbb{N}^*}$ is not necessary uniformly ergodic. Indeed, a core argument in their application is the use of [15, Lemma 6.7] which implies that if Z is a finite state space and $(X_k)_{k\in\mathbb{N}}$ is uniformly ergodic, then $\|\mathbb{E}_z[\bar{\mathbf{A}}(Z_k)] - \bar{\mathbf{A}}\| \leq C\rho^k$ and $\|\mathbb{E}_z[b(Z_k)] - \bar{\mathbf{b}}\| \leq C\rho^k$, for any $z \in \mathsf{Z}$, $k \in \mathbb{N}^*$ and for some $C \geq 0$, $\rho \in (0,1)$. This is precisely the condition considered by [102] to derive their bounds. Obviously [15, Lemma 6.7] does not extend to general (unbounded) state space.

As a replacement, to verify our assumption UE 1, we consider here a τ-truncated version of the eligibility trace

$$\varphi_k = \phi_\tau(X_{k-\tau+1:k}) \quad \text{where} \quad \phi_\tau(x_{0:\tau-1}) = \textstyle\sum_{s=0}^{\tau-1} (\lambda\gamma)^s \psi(x_{\tau-1-s})\,. \tag{4.20}$$

TD(0) algorithm is a special case of (4.20) with $\tau = 1$ and we recover the TD(λ) algorithm by letting $\tau \to \infty$. The recursion (4.19) with eligibility vector defined in (4.20) is a special case of (4.2). To see this, we define $Z_k = [X_{k-\tau}, \ldots, X_k]^\top$ and observe that (4.19) can be obtained by using in (4.2) the following matrix/vector, for $z = [x_0, \ldots, x_\tau]^\top = x_{0:\tau} \in \mathsf{X}^{\tau+1}$,

$$\mathbf{A}(z) = \phi_\tau(x_{0:\tau-1})\{\psi(x_{\tau-1}) - \gamma\psi(x_\tau)\}^\top, \quad \mathbf{b}(z) = \phi_\tau(x_{0:\tau-1})\mathrm{R}(x_{\tau-1})\,. \tag{4.21}$$

Note that compared to [102], we do not consider TD(λ) but (4.20). Consider the following assumptions.

M 1. *The Markov kernel* $\mathrm{Q} : \mathsf{X} \times \mathcal{X} \to \mathbb{R}_+$ *is irreducible and strongly aperiodic. There exist* $c > 0, \mathrm{b} > 0, \delta \in (1/2, 1]$, $R_0 \geq 0$, *and* $\tilde{V} : \mathsf{X} \to [\mathrm{e}, \infty)$ *such that by setting* $\tilde{W} = \log \tilde{V}$, $\mathsf{C}_0 = \{x : \tilde{W}(x) \leq R_0\}$, $\mathsf{C}_0^\complement = \{x : \tilde{W}(x) > R_0\}$, *we have*

$$\mathrm{Q}\tilde{V}(x) \leq \exp[-c\tilde{W}^\delta(x)]\tilde{V}(x)\mathbf{b}\mathbf{1}_{\mathsf{C}_0^\complement}(x) + \mathrm{b}\,\mathbf{b}\mathbf{1}_{\mathsf{C}_0}(x)\,, \tag{4.22}$$

in addition, for any $R \geq 1$, *the level sets* $\{x : \tilde{W}(x) \leq R\}$ *are* $(1, \varepsilon_R \nu)$*-small for* Q, *with* $\varepsilon_R \in (0,1]$ *and* ν *being a probability measure on* $(\mathsf{X}, \mathcal{X})$.

It follows from [38, Theorem 15.2.4] that the Markov kernel Q admits a unique stationary distribution π_0.

M 2. $\pi_0(\psi\psi^\top)$ *is positive definite.*

In the following, we show that under M 1, M 2, the TD(λ) algorithm with truncated eligibility trace (4.19) satisfies the assumptions in Section 4.1.2. In this case, the state-space is set to be $\mathsf{Z} = \mathsf{X}^{\tau+1}$ and the Markov kernel P is given, for any $z = x_{0:\tau} \in \mathsf{X}^{\tau+1}$, by

$$\mathrm{P}(x_{0:\tau}; \mathrm{d}x'_{0:\tau}) = \prod_{\ell=1}^{\tau} \delta_{x_\ell}(\mathrm{d}x'_{\ell-1}) \mathrm{Q}(x_\tau, \mathrm{d}x'_\tau) \,, \tag{4.23}$$

where δ_x denotes the Dirac measure at $x \in \mathsf{X}$.

1. It follows from [42][Lemma 35] that P is irreducible, aperiodic and has a unique invariant distribution $\pi(\mathrm{d}x_{0:\tau}) = \pi_0(\mathrm{d}x_0) \prod_{\ell=1}^{\tau} \mathrm{Q}(x_{\ell-1}, \mathrm{d}x_\ell)$. By [42][Lemma 36], the super-Lyapunov drift condition (4.5) is satisfied with

   $$V(x_{0:\tau}) = \exp\left(c_0 \sum_{i=0}^{\tau-1} (i+1) \tilde{W}^{\delta}(x_i) + \tilde{W}(x_\tau)\right) ,$$

   where c_0 is defined in [42][Eq. 121]. Hence, UE 1 is verified.

2. Let $\|\psi(x)\| \leq \mathrm{C}_\psi \, W^{\beta/2}(x)$ and for $\mathsf{K} \geq 1$, $|\mathrm{R}(x)| \leq \mathrm{C}_{\mathrm{R},\mathsf{K}} \, V^{1/2\mathsf{K}}(x)$, where $\mathrm{C}_\psi, \mathrm{C}_{\mathrm{R},\mathsf{K}} > 0$ are some constants. Then A 1(ε) and A 3(K) are satisfied with

   $$\bar{\mathrm{C}}_A = (1+\gamma) \, \mathrm{C}_\psi^2 \, /(1-\lambda\gamma), \quad \bar{\mathrm{C}}_{b,\mathsf{K}} = \mathrm{C}_{\mathrm{R},\mathsf{K}} \, \mathrm{C}_\psi (\beta\mathsf{K}/\mathrm{e})^{\beta/2}/(1-\lambda\gamma). \tag{4.24}$$

3. Eq. (4.21) implies

   $$A = \sum_{\ell=0}^{\tau-1} \mathbb{E}_{\pi_0}[\psi(X_{\tau-1-\ell})\{\psi(X_{\tau-1}) - \gamma\psi(X_\tau)\}^\top] \,.$$

   Assumption A 2 follows from [42][Lemma 33].

Collecting the above results shows that the assumptions required by Theorem 4.3 are satisfied, thereby proving that the L_p error of TD(λ) algorithm (4.19) (with truncated eligibility trace) converges according to the rate specified in (4.11).

4.2 Tight High Probability Bounds for Linear Stochastic Approximation with Fixed Stepsize

The results of this subsection are published in [41].

In this section we consider the case i.i.d. noise $(Z_k)_{k\in\mathbb{N}}$ and fixed stepsize $\alpha_n \equiv \alpha$. To simplify notations we write $\mathbf{A}_n = \mathbf{A}(Z_n), \mathbf{b}_n = \mathbf{b}(Z_n)$. We obtain new results regarding moments and high probability bounds for products of matrices which are shown to be tight. These results clarify the stability result (4.4). We derive high probability bounds on the performance of LSA under weaker conditions on the sequence $\{(\mathbf{A}_n, \mathbf{b}_n) : n \in \mathbb{N}^*\}$ than previous works. However, in contrast, we establish polynomial concentration bounds with order depending on the stepsize. We show that our conclusions cannot be improved without additional assumptions on the sequence of random matrices $\{\mathbf{A}_n : n \in \mathbb{N}^*\}$, and in particular that no Gaussian or exponential high probability bounds can hold. Finally, we pay a particular attention to establishing bounds with sharp order with respect to the number of iterations and the stepsize and whose leading terms contain the covariance matrices appearing in the central limit theorems.

We require the following main assumption in this section:

A 6. $\{(\mathbf{A}_n, \mathbf{b}_n)\}_{n\in\mathbb{N}^*}$ *is an i.i.d. sequence satisfying the following conditions.*

1. $\mathbb{E}[\mathbf{b}_1] = \bar{\mathbf{b}}$ *and there exists* $\bar{\mathrm{C}}_b > 0$ *such that, for any* $u \in \mathbb{S}^{d-1}$, $u^\top(\mathbf{b}_1 - \bar{\mathbf{b}}) \in \mathrm{SG}(\bar{\mathrm{C}}_b^2)$.

2. *There exists* $\bar{\mathrm{C}}_A > 0$ *such that* $\|\mathbf{A}_1\| \leq \bar{\mathrm{C}}_A$ *almost surely.*

3. The matrix $-\bar{\mathbf{A}} = -\mathbb{E}[\mathbf{A}_1]$ is Hurwitz, i.e. for any eigenvalue λ of $\bar{\mathbf{A}}$, $\mathrm{Re}(\lambda) > 0$.

Both conditions A 6-1, 2 are standard in analysis of LSA, e.g., in [30, 102, 76]. For example, the assumption on the sub-Gaussianity of $\mathbf{b}_1$ is used in [30] and is relaxed from [102], the almost sure boundedness of $\mathbf{A}_1$ is also used in [30, 102].

The aim of this section is to derive high probability bounds on $u^\top\{\theta_n - \theta^\star\}$ for any $n \in \mathbb{N}$, $u \in \mathbb{S}^{d-1}$.

Below, we present a counterexample to show that under only A 6, if $\alpha > 0$ is fixed, then there exists $\bar{p} > 0$ such that $\lim_{n\to+\infty} \mathbb{E}[\|\theta_n - \theta^\star\|^p] = +\infty$ for $p \geq \bar{p}$. As a corollary, it is impossible to obtain any exponential high probability bounds for $\{\|\theta_n - \theta^\star\| : n \in \mathbb{N}\}$.

Example 2. Consider (4.2) with $d = 1$ taking $\mathbf{b}_n = 0$ for any $n \in \mathbb{N}^*$ and for $\{\mathbf{A}_n : n \in \mathbb{N}^*\}$ an i.i.d. sequence of *biased* Rademacher r.v.s with parameter $q_A \in (1/2, 1)$:

$$\mathbf{A}_n = \begin{cases} 1 & \text{with probability } q_A\,, \\ -1 & \text{with probability } 1 - q_A\,. \end{cases} \tag{4.25}$$

This choice is associated with $\theta^\star = 0$ and corresponds to the recursion: $\theta_n = \prod_{k=1}^n (1 - \alpha\mathbf{A}_k)\theta_0$, for some $\theta_0 \neq 0$. For any $p \geq 1$ and $\alpha \in (0, 1)$, we have by definition,

$$\mathbb{E}\left[|\theta_n|^p\right] = \{q_A(1-\alpha)^p + (1-q_A)(1+\alpha)^p\}^n \, |\theta_0|^p \;.$$

Using the lower bounds $(1-\alpha)^p \geq 1 - \alpha p$ and $(1+\alpha)^p \geq 1 + \alpha p + p(p-1)\alpha^2/2$, we get for any $p \geq 1$ and $\alpha \in (0, 1)$,

$$\mathbb{E}\left[|\theta_n|^p\right] \geq \{1 - p\alpha[(2q_A - 1) - (p-1)\alpha(1-q_A)/2]\}^n \, |\theta_0|^p \;.$$

If $\alpha \in (0, 1)$ is fixed, then for any $p > \bar{p}_{q,\alpha} = 1 + 2(2q_A - 1)/[\alpha(1-q_A)]$, we have $\lim_{n\to+\infty} \mathbb{E}\left[|\theta_n|^p\right] = +\infty$. On the other hand, if $\alpha \in (0, 2(2q_A - 1)/(1-q_A))$, then $\lim_{n\to+\infty} \mathbb{E}[\theta_n^2] = 0$. Therefore $\{\theta_n,\, n \in \mathbb{N}\}$ converges in distribution to the Dirac measure at 0 which corresponds to the unique stationary distribution of this sequence as a Markov chain. In such a case, this distribution admit p moments for any $p \geq 0$.

However, this result is specific to this particular case and does not hold if only A 6 holds. Consider $\{\theta_n,\, n \in \mathbb{N}\}$ defined by (4.2) with $\{\mathbf{A}_n,\, n \in \mathbb{N}^*\}$ given in (4.25) and $\{\mathbf{b}_n,\, n \in \mathbb{N}^*\}$ be an i.i.d. sequence of zero-mean Gaussian random variables with unit variance independent of $\{\mathbf{A}_n,\, n \in \mathbb{N}^*\}$. We show in [41][Appendix B.2] that there exists $\alpha_{2,\infty}$ such that for any $\alpha \in (0, \alpha_{2,\infty}]$, the Markov chain $\{\theta_n,\, n \in \mathbb{N}\}$ admits a unique invariant distribution π_α for any $\alpha > 0$. Further, for any $\alpha \in (0, \alpha_{2,\infty}]$ there exists $p_\alpha \geq 1$ such that $\int_{\mathbb{R}} |\theta|^p \, \mathrm{d}\pi_\alpha(\theta) = +\infty$ for any $p \geq p_\alpha$.

It is, however, possible to obtain any p-th moment uniform bound for $\{\|\theta_n - \theta^\star\| : n \in \mathbb{N}\}$ by strengthening A 6-3 to:

A 7. *There exist $\tilde{a} \in (0, 1)$, $\tilde{\alpha}_\infty > 0$ and a positive definite d-dimensional matrix $\tilde{Q}$ such that almost surely, for any $\alpha \in (0, \tilde{\alpha}_\infty]$, $\|\mathrm{I} - \alpha\mathbf{A}_1\|_{\tilde{Q}} < 1 - \tilde{a}\alpha$.*

Conditions similar to A 7 are considered in [25] for the analysis of SA schemes with decreasing stepsize. For example, A 7 holds in the case of regularized linear regression. We take $\mathbf{A}_1 = \lambda\,\mathrm{I} + \mathbf{a}_1\mathbf{a}_1^\top$, for some $\lambda > 0$ and under the assumption that $\|\mathbf{a}_1\|$ is bounded almost surely. The LSA recursion (4.2) approximates the solution to $(\lambda\,\mathrm{I} + \mathbb{E}[\mathbf{a}_1\mathbf{a}_1^\top])\theta = \bar{\mathbf{b}}$, which admits a unique solution.

On the other hand, examples where A7 does not hold are common. For instance, we may consider TD(0) learning with linear function approximation. For a Markov Reward Process with X as the state space, $\mathrm{P} : \mathsf{X} \times \mathcal{X} \to [0, 1]$ as the transition probability, $\mathrm{R} : \mathsf{X} \to \mathbb{R}$ as the reward function, and $\gamma \in (0, 1)$ as a discount factor, TD(0) learning is described as in (4.2) with

$$\mathbf{A}_n = \phi(x_n)\{\phi(x_n) - \gamma\phi(x'_n)\}^\top, \quad \mathbf{b}_n = \mathrm{R}(x_n)\phi(x_n)\,, \tag{4.26}$$

where $\phi : \mathsf{X} \to \mathbb{R}^d$ is a feature map. A typical setting is when x_n is drawn from the stationary distribution of P and $x'_n \sim \mathrm{P}(x_n, \cdot)$. It is easy to verify A 6 provided that $\|\phi(x)\|$, $\mathrm{R}(x)$ are bounded for all $x \in \mathsf{X}$ [108]. However, A 7 is violated as $\mathbf{A}_n$ is only rank-one.

Our next endeavor is to establish moment estimates on the product below:

$$\Gamma^{(\alpha)}_{m:n} = \textstyle\prod_{i=m}^{n}(\mathrm{I} - \alpha\mathbf{A}_i)\,, \quad m, n \in \mathbb{N}^*, \quad m \leq n\,. \tag{4.27}$$

Here for $A_1, \ldots, A_N$, d-dimensional matrices we denote $\prod_{\ell=i}^{j} A_\ell = A_j \ldots A_i$ if $i \leq j$ and with the convention $\prod_{\ell=i}^{j} A_\ell = \mathrm{I}_d$ if $i > j$. We also define its expected value as $G^{(\alpha)}_{m:n} = \mathbb{E}[\Gamma^{(\alpha)}_{m:n}] = (\mathrm{I} - \alpha\bar{\mathbf{A}})^{n-m+1}$.

4.2.1 Moment and High-probability Bounds for Products of Random Matrices

Recall from Proposition 4.1 that the norm of the expected value $G^{(\alpha)}_{1:n} = \mathbb{E}[\Gamma^{(\alpha)}_{1:n}]$ decays exponentially with n as $\left\|G^{(\alpha)}_{1:n}\right\| \leq \sqrt{\kappa_{\mathsf{Q}}}(1 - \alpha a)^{n/2}$. We expect a similar phenomenon for the moment bound of $\|\Gamma^{(\alpha)}_{1:n}\|$. Precisely, in this section, we show that if p is fixed, then there exists $\alpha_{p,\infty} > 0$ such that for any $\alpha \in (0, \alpha_{p,\infty}]$, the p-th moment of $\Gamma^{(\alpha)}_{m:n}$ decays exponentially with $n - m$.

We present the main technical result on the product of general random matrices as follows, whose proof is based on the framework introduced in [58].

Proposition 4.5. *Let $\{\mathbf{Y}_\ell,\ \ell \in \mathbb{N}\}$ be an independent sequence and P be a positive definite matrix. Assume that for each $\ell \in \mathbb{N}$ there exist $m_\ell \in (0, 1)$ and $\sigma_\ell > 0$ such that $\|\mathbb{E}[\mathbf{Y}_\ell]\|^2_P \leq 1 - m_\ell$ and $\|\mathbf{Y}_\ell - \mathbb{E}[\mathbf{Y}_\ell]\|_P \leq \sigma_\ell$ almost surely. Define $\mathbf{Z}_n = \prod_{\ell=0}^{n} \mathbf{Y}_\ell = \mathbf{Y}_n\mathbf{Z}_{n-1}$, for $n \geq 1$ and starting from $\mathbf{Z}_0$. Then, for any $2 \leq q \leq p$ and $n \geq 1$,*

$$\|\mathbf{Z}_n\|^2_{p,q} \leq \kappa_P \prod_{\ell=1}^{n}(1 - m_\ell + (p-1)\sigma_\ell^2) \left\|P^{1/2}\mathbf{Z}_0 P^{-1/2}\right\|^2_{p,q}\,, \tag{4.28}$$

where we recall that $\kappa_P = \lambda^{-1}_{\mathsf{min}}(P)\lambda_{\mathsf{max}}(P)$.

In order to bound $\Gamma^{(\alpha)}_{1:n}$ using Proposition 4.5, we identify the latter with $\mathbf{Y}_\ell = \mathrm{I} - \alpha\mathbf{A}_\ell, \ell \geq 1$, $\mathbf{Y}_0 = \mathrm{I}$. As $-\bar{\mathbf{A}}$ is Hurwitz, applying Proposition 4.1 yields $\|\mathbb{E}[\mathbf{Y}_\ell]\|^2_Q = \left\|\mathrm{I} - \alpha\bar{\mathbf{A}}\right\|^2_Q \leq 1 - a\alpha$. Further, A 6-2 ensures that almost surely,

$$\|\mathbf{Y}_\ell - \mathbb{E}[\mathbf{Y}_\ell]\|_Q = \alpha\left\|\mathbf{A}_\ell - \bar{\mathbf{A}}\right\|_Q \leq 2\alpha\sqrt{\kappa_{\mathsf{Q}}}\bar{\mathrm{C}}_A = b_Q\alpha\,.$$

Therefore, (4.28) holds with $m_\ell = a\alpha$ and $\sigma_\ell = b_Q\alpha$. As $\|\mathrm{I}\|_p = d^{1/p}$, we obtain the following corollary.

Corollary 4.6. *Assume A 6-2-3. Then, for any $\alpha \in [0, \alpha_\infty]$, $2 \leq q \leq p$, and $n \in \mathbb{N}$,*

$$\mathbb{E}^{1/q}\left[\|\Gamma^{(\alpha)}_{1:n}\|^q\right] \leq \left\|\Gamma^{(\alpha)}_{1:n}\right\|_{p,q} \leq \sqrt{\kappa_{\mathsf{Q}}}d^{1/p}(1 - a\alpha + (p-1)b_Q^2\alpha^2)^{n/2}\,, \tag{4.29}$$

where α_∞ was defined in (4.8), and $b_Q = 2\sqrt{\kappa_{\mathsf{Q}}}\bar{\mathrm{C}}_A$.

Note that Corollary 4.6 shows $\sup_{n\in\mathbb{N}} \mathbb{E}[\|\Gamma_{1:n}^{(\alpha)}\|^p] < +\infty$ for any $\alpha \in (0, \alpha_{p,\infty}]$, where

$$\alpha_{p,\infty} = \alpha_\infty \wedge a/(2b_Q^2(p-1))\,. \tag{4.30}$$

This kind of condition relating the choice of α with the required order p is necessary as illustrated in Example 2. Corollary 4.6 further leads to the high-probability bound:

Corollary 4.7. *Assume A 6-2-3. Then, for any $\alpha \in (0, \alpha_\infty)$ where α_∞ was defined in* (4.8)*, $\delta \in (0,1)$ and $n \in \mathbb{N}$, with probability at least $1-\delta$,*

$$\|\Gamma_{1:n}^{(\alpha)}\| \leq \sqrt{\kappa_{\mathsf{Q}}} \exp\left[-(an\alpha - \alpha^2 b_Q^2 n)/2 + b_Q\alpha\sqrt{2n\log(d/\delta)}\right]\,.$$

The result in Corollary 4.7 is tight with respect to δ. See example in [41] that continues Example 2.

We conclude the section with a complementary result of Corollary 4.6 that does not require A 6-2:

Proposition 4.8. *Assume A 6-3, $\|\mathbf{A}_1 - \bar{\mathbf{A}}\| \in \mathrm{SG}(\bar{\mathrm{C}}_A')$ for some $\bar{\mathrm{C}}_A' > 0$. Then, for any $\alpha \in (0, \alpha_\infty)$ where α_∞ was defined in* (4.8)*, $2 \leq q \leq p$, and $n \in \mathbb{N}$,*

$$\mathbb{E}^{1/q}\left[\|\Gamma_{1:n}^{(\alpha)}\|^q\right] \leq \left\|\Gamma_{1:n}^{(\alpha)}\right\|_{p,q} \leq \sqrt{\kappa_{\mathsf{Q}}} d^{1/p}(1 - a\alpha + q(p-1)(b_Q')^2\alpha^2)^{n/2}\,, \tag{4.31}$$

where $b_Q' = 2\sqrt{\kappa_{\mathsf{Q}}}\bar{\mathrm{C}}_A'$.

Finite-time High-probability Bounds for LSA Relying on the results established in Section 4.2.1 and the decomposition (similar to (4.12))

$$\theta_n - \theta^\star = \tilde{\theta}_n^{(\mathsf{tr})} + \tilde{\theta}_n^{(\mathsf{fl})}\,, \quad \tilde{\theta}_n^{(\mathsf{tr})} = \Gamma_{1:n}^{(\alpha)}\{\theta_0 - \theta^\star\}\,, \quad \tilde{\theta}_n^{(\mathsf{fl})} = \alpha \textstyle\sum_{j=1}^n \Gamma_{j+1:n}^{(\alpha)} \bar{\varepsilon}_j\,. \tag{4.32}$$

we derive high probability bounds on $u^\top\{\theta_n - \theta^\star\}$ for any $n \in \mathbb{N}$ and $u \in \mathbb{S}^{d-1}$, where $\{\theta_n,\, n \in \mathbb{N}\}$ is defined in (4.2). We begin our study with the transient term $\tilde{\theta}_n^{(\mathsf{tr})}$. Observe that

Proposition 4.9. *Assume A 6 and let $p_0 \geq 2$. Then, for any $n \in \mathbb{N}^*$, $\alpha \in (0, \alpha_{p_0,\infty})$, where $\alpha_{p_0,\infty}$ is defined in* (4.30)*, $u \in \mathbb{S}^{d-1}$ and $\delta \in (0,1)$ it holds with probability at least $1-\delta$ that*

$$|u^\top \Gamma_{1:n}^{(\alpha)}(\theta_0 - \theta^\star)| \leq \sqrt{\kappa_{\mathsf{Q}}} d^{1/p_0}(1 - a\alpha/4)^n \,\|\theta_0 - \theta^\star\|\, \delta^{-1/p_0}\,,$$

where a was defined in (4.8)*.*

Proposition 4.9 only provides a polynomial high probability bound with respect to δ. This is due to the fact that only polynomial moments of $\|\Gamma_{1:n}^{(\alpha)}\|$ up to a maximal order are uniformly bounded in the number of iterations n.

We now turn to the fluctuation term $\tilde{\theta}_n^{(\mathsf{fl})}$ defined in (4.32). Note that under A 6, the sequence $\{\bar{\varepsilon}_n,\, n \in \mathbb{N}\}$ is i.i.d.. From this observation and following [42], we consider the decomposition

$$\tilde{\theta}_n^{(\mathsf{fl})} = \alpha \sum_{j=1}^n \Gamma_{j+1:n}^{(\alpha)} \bar{\varepsilon}_j = J_n^{(\alpha,0)} + H_n^{(\alpha,0)}\,, \tag{4.33}$$

where $\{(J_n^{(\alpha,0)}, H_n^{(\alpha,0)}) \,:\, n \in \mathbb{N}\}$ are defined by induction for $n \geq 0$ as:

$$\begin{aligned} J_{n+1}^{(\alpha,0)} &= \left(\mathrm{I} - \alpha\bar{\mathbf{A}}\right) J_n^{(\alpha,0)} + \alpha\bar{\varepsilon}_{n+1}\,, & J_0^{(\alpha,0)} &= 0\,, \\ H_{n+1}^{(\alpha,0)} &= \left(\mathrm{I} - \alpha\mathbf{A}_n\right) H_n^{(\alpha,0)} - \alpha(\mathbf{A}_{n+1} - \bar{\mathbf{A}}) J_n^{(\alpha,0)}\,, & H_0^{(\alpha,0)} &= 0\,. \end{aligned} \tag{4.34}$$

The latter recurrence can be written as

$$J_n^{(\alpha,0)} = \alpha \sum_{j=1}^{n} G_{j+1:n}^{(\alpha)} \bar{\varepsilon}_j \,, \quad H_n^{(\alpha,0)} = -\alpha \sum_{j=1}^{n} \Gamma_{j+1:n}^{(\alpha)} (\mathbf{A}_j - \bar{\mathbf{A}}) J_{j-1}^{(\alpha,0)} \,.$$

Note that $J_n^{(\alpha,0)}$ is a linear statistics of the random variables $\{\bar{\varepsilon}_j,\, j \in \{1, \ldots, n\}\}$ which are centered and i.i.d. under A 6. Next, we show that $J_n^{(\alpha,0)}$ is the leading term as the stepsize $\alpha \downarrow 0$. Denote for any $n \in \mathbb{N}^*$ and $\alpha > 0$, the covariance matrix of $J_n^{(\alpha,0)}$ as

$$\boldsymbol{\Sigma}_n^\alpha = \operatorname{Cov}(J_n^{(\alpha,0)}) \,. \tag{4.35}$$

We obtain the following statement:

Proposition 4.10. *Assume A 6. Then for any $n \in \mathbb{N}^*$, $\alpha \in (0, \alpha_\infty]$, where α_∞ is defined in* (4.8)*, $u \in \mathbb{S}^{d-1}$ and $\delta \in (0,1)$, it holds with probability at least $1-\delta$,*

$$\left| u^\top J_n^{(\alpha,0)} \right| < \mathsf{D}_1 \sqrt{\{u^\top \boldsymbol{\Sigma}_n^\alpha u\} \log(2/\delta)} + \alpha \sqrt{1 + \log(1/(a\alpha))}\, \mathsf{D}_2 \log^{3/2}(2/\delta) \,, \tag{4.36}$$

where $\mathsf{D}_1 = 60\sqrt{3}\mathrm{e}^{4/3}$ and D_2 is defined in [41][Eq. 49].

We analyze further the covariance associated with $J_n^{(\alpha,0)}$ and its dependence with respect to n and α. First, note that for any $\alpha \in (0, \alpha_{2,\infty}]$, $\{\boldsymbol{\Sigma}_n^\alpha,\, n \in \mathbb{N}^*\}$ converges to $\alpha \boldsymbol{\Sigma}^\alpha$ as $n \to \infty$ where $\boldsymbol{\Sigma}^\alpha = \alpha \sum_{k=0}^{\infty} G_{1:k} \boldsymbol{\Sigma}_\varepsilon G_{1:k}^\top$ is the unique solution of the Ricatti equation

$$\bar{A} \boldsymbol{\Sigma}^\alpha + \boldsymbol{\Sigma}^\alpha \bar{A}^\top - \alpha \bar{A} \boldsymbol{\Sigma}^\alpha \bar{A}^\top = \boldsymbol{\Sigma}_\varepsilon \,, \quad \text{with} \quad \boldsymbol{\Sigma}_\varepsilon = \mathbb{E}[\varepsilon_1 \varepsilon_1^\top] \,. \tag{4.37}$$

Notice that we focus on the cases where $\boldsymbol{\Sigma}_\varepsilon$ is full-rank. Using Proposition 4.1, we obtain that for any $n \geq 0$,

$$\| \boldsymbol{\Sigma}_n^\alpha - \alpha \boldsymbol{\Sigma}^\alpha \| \leq \alpha^2 \sum_{k>n} \|G_{1:k}\|^2 \, \|\boldsymbol{\Sigma}_\varepsilon\| \leq \alpha a^{-1} \kappa_{\mathsf{Q}} \, \|\boldsymbol{\Sigma}_\varepsilon\| \, (1 - \alpha a)^n \,. \tag{4.38}$$

We now give an expansion of $\boldsymbol{\Sigma}^\alpha$ with respect to α. It is well-known that as $\alpha \downarrow 0$, $\boldsymbol{\Sigma}^\alpha$ converges to $\boldsymbol{\Sigma}$, the unique solution of the Lyapunov equation (see [88, Lemma 9.1])

$$\bar{\mathbf{A}} \boldsymbol{\Sigma} + \boldsymbol{\Sigma} \bar{\mathbf{A}}^\top = \boldsymbol{\Sigma}_\varepsilon \,. \tag{4.39}$$

Our next result states the convergence of $\boldsymbol{\Sigma}^\alpha$ to $\boldsymbol{\Sigma}$ is of the order of the stepsize α.

Proposition 4.11. *Assume that A 6-3 holds. Then, for any $\alpha \in (0, \alpha_\infty]$, where α_∞ is defined in* (4.8)*,*

$$\| \boldsymbol{\Sigma}^\alpha - \boldsymbol{\Sigma} \| \, [Q] \leq \alpha a^{-1} \| \bar{\mathbf{A}} \boldsymbol{\Sigma} \bar{\mathbf{A}}^\top \|_Q \,,$$

where $\boldsymbol{\Sigma}^\alpha$ and $\boldsymbol{\Sigma}$ are defined in (4.37) *and* (4.39) *respectively and a is given in* (4.8)*.*

The last step in bounding $\tilde{\theta}_n^{(\mathsf{fl})}$ is to consider $H_n^{(\alpha,0)}$. We proceed similarly to (4.34) and consider the decomposition $H_n^{(\alpha,0)} = J_n^{(\alpha,1)} + H_n^{(\alpha,1)}$, where $\{(J_n^{(\alpha,1)}, H_n^{(\alpha,1)}) : n \in \mathbb{N}\}$ are defined by induction for $n \geq 0$ as:

$$\begin{aligned} J_{n+1}^{(\alpha,1)} &= (\mathrm{I} - \alpha \bar{\mathbf{A}}) J_n^{(\alpha,1)} - \alpha (\mathbf{A}_{n+1} - \bar{\mathbf{A}}) J_n^{(\alpha,0)}, & J_0^{(\alpha,1)} &= 0 \,, \\ H_{n+1}^{(\alpha,1)} &= (\mathrm{I} - \alpha \mathbf{A}_{n+1}) H_n^{(\alpha,1)} - \alpha (\mathbf{A}_{n+1} - \bar{\mathbf{A}}) J_n^{(\alpha,1)}, & H_0^{(\alpha,1)} &= 0 \,. \end{aligned} \tag{4.40}$$

In our next result we bound each term of this decomposition separately.

Proposition 4.12. *Assume A 6 and let $p_0 \geq 2$. Then, for any $n \in \mathbb{N}$, $\alpha \in (0, \alpha_{p_0,\infty})$, where $\alpha_{p_0,\infty}$ is defined in* (4.30)*, $u \in \mathbb{S}^{d-1}$ and $\delta \in (0, 1/2)$, with probability at least $1 - 2\delta$, it holds*

$$\left|u^\top J_n^{(\alpha,1)}\right| < \mathrm{e}\mathsf{D}_3 \alpha \log^2(1/\delta)\,, \quad \left|u^\top H_n^{(\alpha,1)}\right| < \mathsf{D}_4 \alpha p_0^2 \delta^{-1/p_0}\,, \tag{4.41}$$

where D_3 and D_4 are given in [41][Eq. 57 and 60].

Now we are ready to combine the previous bounds and to state the main result of this section.

Theorem 4.13. *Assume A 6 and let $p_0 \geq 2$. Then, for any $n \in \mathbb{N}$, $\alpha \in (0, \alpha_{p_0,\infty})$, where $\alpha_{p_0,\infty}$ is defined in* (4.30)*, $u \in \mathbb{S}^{d-1}$ and $\delta \in (0, 1/4)$, with probability at least $1 - 4\delta$, it holds*

$$\alpha^{-1/2}|u^\top(\theta_n - \theta^\star)| < \mathsf{D}_1 \sqrt{\{u^\top \boldsymbol{\Sigma}^\alpha u\} \log(2/\delta)} + \alpha^{1/2} q^{(1)}(\alpha, \delta) + (1 - a\alpha/4)^n \Delta^{(1)}(\alpha, \delta)\,, \tag{4.42}$$

where $\boldsymbol{\Sigma}^\alpha$ is the unique solution of (4.37)*, $\mathsf{D}_1 = 60\sqrt{3}\mathrm{e}^{4/3}$, a is defined in* (4.8)*,*

$$\begin{aligned} q^{(1)}(\alpha, \delta) &= \big(\mathrm{e}\mathsf{D}_3 \log^2(1/\delta) + \sqrt{1 + \log(1/a\alpha)}\mathsf{D}_2 \log^{3/2}(2/\delta)\big) + \mathsf{D}_4 p_0^2 \delta^{-1/p_0}\,, \\ \Delta^{(1)}(\alpha, \delta) &= \mathsf{D}_1 \sqrt{a^{-1}\kappa_{\mathsf{Q}} \|\boldsymbol{\Sigma}_\varepsilon\| \log(2/\delta)} + \sqrt{\kappa_{\mathsf{Q}}} d^{1/p_0} \|\theta_0 - \theta^\star\| \alpha^{-1/2} \delta^{-1/p_0}\,, \end{aligned} \tag{4.43}$$

where κ_{Q} and $\boldsymbol{\Sigma}_\varepsilon$ are defined in (4.8) *and* (4.37) *respectively.*

We now discuss the high probability bound (4.42). First, the term $\Delta^{(1)}(\alpha, \delta)$, and in particular the initial condition vanishes exponentially fast in the number of iterations n. In addition, $q^{(1)}(\alpha, \delta)$ and $\Delta^{(1)}(\alpha, \delta)$ are of order δ^{-1/p_0} as $\delta \to 0$ and therefore (4.42) provides polynomial high probability bounds on LSA. However, this conclusion is expected as illustrated in Example 2. Finally, the discussion of (4.42) with respect to α is postponed to the next section.

Under A 7 we can provide a better bound for $H_n^{(\alpha,1)}$.

Proposition 4.14. *Assume A 6 and A 7. Then, for any $n \in \mathbb{N}$, $\alpha \in (0, \alpha_\infty \wedge \tilde{\alpha}_\infty)$, where α_∞ is defined in* (4.8)*, $u \in \mathbb{S}^{d-1}$ and $\delta \in (0, 1/2)$, with probability at least $1 - 2\delta$, it holds*

$$\left|u^\top J_n^{(\alpha,1)}\right| < \mathrm{e}\mathsf{D}_3 \alpha \log^2(1/\delta)\,, \quad \left|u^\top H_n^{(\alpha,1)}\right| < \mathrm{e}\mathsf{D}_5 \alpha \log^2(1/\delta)\,, \tag{4.44}$$

where D_3 and D_5 are given in [41][Eq. 57 and 61].

As a result, we can establish exponential high probability bounds with respect to δ.

Theorem 4.15. *Assume A 6 and A 7. Then, for any $n \in \mathbb{N}$, $\alpha \in (0, \alpha_\infty \wedge \tilde{\alpha}_\infty)$, $u \in \mathbb{S}^{d-1}$ and $\delta \in (0, 1/4)$, with probability at least $1 - 4\delta$, it holds*

$$\alpha^{-1/2}|u^\top(\theta_n - \theta^\star)| < \mathsf{D}_1 \sqrt{\{u^\top \boldsymbol{\Sigma}^\alpha u\} \log(2/\delta)} + \alpha^{1/2} q^{(2)}(\alpha, \delta) + (1 - \alpha\tilde{a})^{n/2} \Delta^{(2)}(\alpha, \delta)\,,$$

where $\mathsf{D}_1 = 60\sqrt{3}\mathrm{e}^{4/3}$, $\boldsymbol{\Sigma}^\alpha$ is solution of (4.37)*,*

$$\begin{aligned} q^{(2)}(\alpha, \delta) &= \mathrm{e}(\mathsf{D}_3 + \mathsf{D}_5) \log^2(1/\delta) + \sqrt{1 + \log(1/\tilde{a}\alpha)}\mathsf{D}_2 \log^{3/2}(2/\delta)\,, \\ \Delta^{(2)}(\alpha, \delta) &= \mathsf{D}_1 \sqrt{\tilde{a}^{-1}\kappa_{\tilde{Q}} \|\boldsymbol{\Sigma}_\varepsilon\| \log(2/\delta)} + \kappa_{\tilde{Q}}^{1/2} \|\theta_0 - \theta^\star\| \, \alpha^{-1/2}\,, \end{aligned} \tag{4.45}$$

where $\boldsymbol{\Sigma}_\varepsilon$ is defined in (4.37)*.*

Optimality of the derived bounds with respect to α: analysis of $(\theta_n)_{n\in\mathbb{N}}$ as a Markov chain In this section, we study the sequence $\{\theta_n,\, n \in \mathbb{N}\}$ defined in (4.2) as a Markov chain. This perspective will allow us to show that the bounds that we derived in Theorem 4.13 are near-Berstein high probability bounds with respect to the stepsize α. Denote by R_α the Markov kernel associated with $\{\theta_n \,:\, n \in \mathbb{N}\}$. First, we show that if α is small enough then R_α is geometrically ergodic with respect to the Wasserstein distance of order 2 denoted by W_2 and give a representation of its stationary distribution as an infinite sum.

Theorem 4.16. *Assume A 6. Then, for any $\alpha \in (0, \alpha_{2,\infty})$, where $\alpha_{2,\infty}$ is defined in* (4.30)*, R_α admits a unique stationary distribution $\pi_\alpha \in \mathcal{P}_2(\mathbb{R}^d)$ and for any $n \in \mathbb{N}$,*

$$W_2^2(\delta_\theta R_\alpha^n, \pi_\alpha) \le \sqrt{\kappa_{\mathsf{Q}} d (1 - a\alpha/2)^n} \int_{\mathbb{R}^d} \left\| \tilde{\theta} - \theta \right\|^2 \mathrm{d}\pi_\alpha(\tilde{\theta}) \,. \tag{4.46}$$

Further, if $\{(\mathbf{A}_k, \mathbf{b}_k) \,:\, k \in \mathbb{N}_-\}$ is any sequence of i.i.d. random variables with the same distribution as $(\mathbf{A}_1, \mathbf{b}_1)$, then the following limit exists almost surely and in L^2 and has distribution π_α:

$$\theta_\infty^{(\alpha)} = \lim_{n\to-\infty} \theta_n^{(\alpha,\leftarrow)} \,, \quad \theta_n^{(\alpha,\leftarrow)} = \alpha \sum_{k=n}^{1} \Gamma_{k:0} \mathbf{b}_{k-1} \,, \quad \Gamma_{k:0} = \prod_{i=k}^{0} (\mathrm{I}_d - \alpha \mathbf{A}_i) \,. \tag{4.47}$$

Based on Theorem 4.13, we easily get concentration bounds for the family of distributions $\{\pi_\alpha \,:\, \alpha \in (0, \alpha_{2,\infty})\}$ around $\theta^\star$.

Theorem 4.17. *Assume A 6 and let $p_0 \ge 2$. Then, for any $\alpha \in (0, \alpha_{p_0,\infty})$, where $\alpha_{p_0,\infty}$ is defined in* (4.30)*, $u \in \mathbb{S}^{d-1}$ and $\delta \in (0, 1/4)$, with probability at least $1 - 4\delta$, it holds*

$$\alpha^{-1/2} |u^\top(\theta_\infty^{(\alpha)} - \theta^\star)| < \mathsf{D}_1 \sqrt{\{u^\top \mathbf{\Sigma} u\} \log(2/\delta)} + \alpha^{1/2}[a^{-1/2} \|\bar{\mathbf{A}} \mathbf{\Sigma} \bar{\mathbf{A}}^\top\|_Q^{1/2} + q^{(1)}(\alpha, \delta)] \,, \tag{4.48}$$

where $\mathbf{\Sigma}$ is the unique solution of (4.39)*, $\mathsf{D}_1 = 60\sqrt{3}\mathrm{e}^{4/3}$, a is defined in* (4.8)*, and $q^{(1)}(\alpha, \delta)$ in* (4.43)*.*

Our results is only polynomial in δ and we cannot expect improving this dependency as illustrated in Example 2 for fixed α. The leading term in (4.48) as $\alpha \downarrow 0$ is $\sqrt{\mathsf{D}_1 \{u^\top \mathbf{\Sigma} u\}}$. In our next result, we establish a central limit theorem for the family $(\theta_\infty^{(\alpha)})_{\alpha\in(0,\alpha_{2,\infty}]}$ where $\mathbf{\Sigma}$ plays the role of the asymptotic covariance matrix. As a result, (4.48) is a Bernstein-type high probability bound with respect to α and therefore (4.48) is sharp. Define for any $\alpha \in (0, \alpha_{2,\infty}]$,

$$\tilde{\theta}_\infty^{(\alpha)} = \alpha^{-1/2} \{\theta_\infty^{(\alpha)} - \theta^\star\} \,. \tag{4.49}$$

Theorem 4.18. *Assume A 6. Then, the family $\{\tilde{\theta}_\infty^{(\alpha)} \,:\, \alpha \in (0, \alpha_{2,\infty}]\}$ converges in law as $\alpha \downarrow 0$ to a zero-mean Gaussian random variable with covariance matrix $\mathbf{\Sigma}$ defined by* (4.39)*.*

Note that this result was established in [87, Theorem 1] for general stochastic approximation schemes but under stronger conditions on the sequence $\{\varepsilon_n,\, n \in \mathbb{N}^*\}$. In particular, it is assumed that the distribution of ε_1 admits a density with respect to the Lebesgue measure. We relax this condition and provide a new proof for this result.

5 Variance reduction in MCMC algorithms

The results of this subsection are published in [8] and [11].

Variance reduction aims at reducing the stochastic error of a Monte Carlo estimate; see [92], [95], [47], and [46] for a an introduction to this field. Recently one witnessed a revival of interest in variance reduction techniques for dependent sequences with applications to Bayesian inference and reinforcement learning among others; see, for instance, [79], [61], [33], [26], [2], and references therein.

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

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

where $(X_k)_{k=0}^{N-1}$ is an i.i.d. sample from π. Using the central limit theorem, one can construct an asymptotically valid confidence interval for the value $\pi(f)$ of the form $\pi_N(f) \pm \mathsf{q}\, N^{-1/2}(\mathrm{Var}_\pi(f))^{1/2}$, where q is a quantile of a normal distribution, and $\mathrm{Var}_\pi(f) = \int_{\mathsf{X}} \{f(x) - \pi(f)\}^2 \pi(\mathrm{d}x)$. A general way to reduce the variance $\mathrm{Var}_\pi(f)$ is to select another function g in a set $\mathcal{G}$ such that $\pi(g) = 0$ and $\mathrm{Var}_\pi(f-g) \ll \mathrm{Var}_\pi(f)$. Such a function g is called a *control variate* (CV). A natural approach to learn $g \in \mathcal{G}$ is to minimize the empirical variance

$$D_n(f-g) = (n-1)^{-1} \sum\nolimits_{k=0}^{n-1} \bigl(f(X_k) - g(X_k) - \pi_n(f-g)\bigr)^2, \tag{5.1}$$

constructed using a new independent learning sample $(X_k)_{k=0}^{n-1}$. This leads to the Empirical Variance Minimisation (EVM) method recently studied in [7] and [10]. In many problems of interest, drawing an i.i.d. sample from π is not an option, yet it is possible to obtain a non-stationary dependent sequence $(X_k)_{k=0}^{\infty}$ whose marginal distribution converges to π. This situation is typical in Bayesian statistics, where π represents a posterior distribution and $(X_k)_{k=0}^{\infty}$ is sampled using Markov chain Monte Carlo (MCMC) methods. Under appropriate conditions, the central limit theorem also holds and therefore, it is possible to construct the asymptotic confidence interval for $\pi(f)$ of the form

$$\left[\pi_N(f) - \mathsf{q}\sqrt{\frac{V_\infty(f)}{N}}, \pi_N(f) + \mathsf{q}\sqrt{\frac{V_\infty(f)}{N}}\,\right], \tag{5.2}$$

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

$$V_\infty(f) := \lim_{N\to\infty} N \cdot \mathbb{E}\Bigl[\bigl(\pi_N(f) - \pi(f)\bigr)^2\Bigr]. \tag{5.3}$$

A sensible approach is to select a control variate $g \in \mathcal{G}$ by minimizing an estimate for the asymptotic variance $V_\infty(f-g)$. When the spectral estimate of $V_\infty(f-g)$ is used, this leads to the Empirical Spectral Variance Minimization (ESVM); see [9].

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

$$\mathrm{d}Y_t = -\nabla U(Y_t)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t\,, \tag{5.4}$$

where $(W_t)_{t\geq 0}$ is the standard Brownian motion. Note that ∇U does not depend on the normalizing constant of π which is typically unknown in Bayesian inference.

Under some technical conditions, the distribution of Y_t converges to π as $t \to \infty$, see [93]. The gradient-based MCMC algorithms are based on a time-discretized version of (5.4). In the Bayesian setting, a computational bottleneck of these algorithms is that the complexity of the gradient ∇U evaluation scales proportionally to the number of observations (sample size) K which can be very time consuming in the "big data" limit. To alleviate this problem, [111] proposed to replace the "full" gradient ∇U by a stochastic gradient estimate based on sums over random *minibatches*. This algorithm, Stochastic Gradient Langevin Dynamics (SGLD), has emerged as a key MCMC algorithm in Bayesian inference for large scale datasets. The analysis of SGLD and its finite sample performance has attracted a wealth of contributions; see, for example, [75], [106], [80], [32], and the references therein. These works show that the use of stochastic gradient comes at a price: while the resulting estimate of the gradient is still unbiased, its variance might annihilate the computational advantages of SGLD [32]. Several proposals have been made to reduce the variance of the stochastic gradient estimate of the "full" gradient, inspired by several methods, proposed for incremental stochastic optimization; see [94], [61], and [33]. [39] has investigated the properties of the Stochastic Average Gradient (SAGA) and Stochastic Variance Reduced Gradient (SVRG) estimators for Langevin dynamics. These results have been later completed and sharpened by [32], [26], [22]. Other variance reduction approaches include various subsampling schemes and constructing alternative estimates for the gradient (see, for instance, [2] and [114]).

This chapter is organized as follows. In Section 5.1, we analyze the ESVM approach for general dependent sequences. In particular, the ESVM method is described in Section 5.1.1. In Section 5.1.2, we study the theoretical properties of the ESVM method for asymptotically stationary dependent sequences. Here we provide a bound for the excess risk $V_\infty(f - \widehat{g}_n) - \inf_{g \in \mathcal{G}} V_\infty(f-g)$, where a control variate $\widehat{g}_n \in \mathcal{G}$ is chosen by minimization of the spectral variance V_n based on $(X_k)_{k=0}^{n-1}$, that is, $\widehat{g}_n \in \operatorname{argmin} V_n(f-g)$. The precise definition of V_n will be given in Section 5.1.1. In Section 5.2, we apply these results to Markov chains which are uniformly geometrically ergodic in Wasserstein distance. While Section 5.2.1 is devoted to the (undajusted) Langevin Dynamics, in Section 5.2.2 we use the ESVM approach for variance reduction in SGLD-type algorithms. We show that in both cases, the excess variance can be bounded, with high probability and up to logarithmic factors, as

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

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

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

for some constant $C > 0$. Note that these intervals can be much tighter than ones in (5.2), provided that n is large and $\inf_{g \in \mathcal{G}} V_\infty(f-g)$ is small. The latter condition is satisfied if the class $\mathcal{G}$ is rich enough. In Section 5.3, we illustrate performance of the proposed variance reduction method on various benchmark problems.

5.1 Empirical Spectral Variance Minimization

5.1.1 Method

Let $(\Omega, \mathfrak{F}, (\mathfrak{F}_k)_{k \geq 0}, \mathbb{P}^\circ)$ be a filtered probability space and $(X_k)_{k=0}^\infty$ be a random process adapted to the filtration $(\mathfrak{F}_k)_{k \geq 0}$ and taking values in X. Let $f : \mathsf{X} \to \mathbb{R}$ be a function such that $\pi(f^2) < \infty$ and $\mathbb{E}[f^2(X_k)] < \infty$ for all $k \in \mathbb{N}$. Let also $\mathcal{G}$ be a set of control variates, that is, functions $g \in \mathcal{G}$ satisfying $\pi(g^2) < \infty$, $\pi(g) = 0$, and $\mathbb{E}[g^2(X_k)] < \infty$ for all $k \in \mathbb{N}$. Particular examples of classes $\mathcal{G}$ are given below in Section 5.2. Denote the class of functions $h = f - g$ for $g \in \mathcal{G}$ by $\mathcal{H}$,

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

CS1. *For any $h \in \mathcal{H}$, there exists a symmetric, summable, and positive semidefinite sequence $(\rho^{(h)}(\ell))_{\ell \in \mathbb{Z}}$ satisfying*

1) $\rho^{(h)}(0) = \mathrm{Var}_\pi(h)$,
2) *for any $\ell \in \mathbb{N}_0$ and constant $R > 0$ independent of h and ℓ,*
$$\sum\nolimits_{k \in \mathbb{N}_0} \left| \mathbb{E}\big[\tilde{h}(X_k)\tilde{h}(X_{k+\ell})\big] - \rho^{(h)}(\ell) \right| \le R,$$
3) $\lim\limits_{\ell \to \infty} \sum\nolimits_{k \in \mathbb{N}_0} \left| \mathbb{E}\big[\tilde{h}(X_k)\tilde{h}(X_{k+\ell})\big] - \rho^{(h)}(\ell) \right| = 0.$

Proposition 5.1. *Assume that the condition CS 1 holds. Then, for all $h \in \mathcal{H}$, the asymptotic variance $V_\infty(h)$ defined in* (5.3) *exists and can be represented as*

$$V_\infty(h) = \sum\nolimits_{\ell \in \mathbb{Z}} \rho^{(h)}(\ell). \tag{5.5}$$

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

$$V_n(h) := \sum\nolimits_{|\ell| < b_n} w_n(\ell) \rho_n^{(h)}(\ell), \tag{5.6}$$

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

$$\rho_n^{(h)}(\ell) = \rho_n^{(h)}(-\ell) := n^{-1} \sum\nolimits_{k=0}^{n-\ell-1} \big(h(X_k) - \pi_n(h)\big)\big(h(X_{k+\ell}) - \pi_n(h)\big). \tag{5.7}$$

Here the truncation point b_n is an integer depending on n and the lag window w_n is a kernel of the form $w_n(\ell) = w(\ell/b_n)$, where w is a symmetric non-negative function supported on $[-1, 1]$ such that $\sup_{y \in [0,1]} |w(y)| \le 1$ and $w(y) = 1$ for $y \in [-1/2, 1/2]$. There are several other estimates for the asymptotic variance $V_\infty(h)$; see [44] and the references therein. The ESVM estimator is obtained by

$$\widehat{h}_n \in \operatorname{argmin}_{h \in \mathcal{H}} V_n\big(h\big). \tag{5.8}$$

The ESVM method is summarized 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 functions $g \in \mathcal{G}$.
2. Find $\widehat{g}_n \in \operatorname{argmin}_{g \in \mathcal{G}} V_n(f - g)$, where $V_n$ is computed based on $\mathbf{X}_n$.
Output: $\pi_N\left(f - \widehat{g}_n\right)$ computed based on $\mathbf{X}'_N$.
```

5.1.2 Theoretical analysis

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

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

To obtain a quantitative bound for the asymptotic variance of $\widehat{h}_{n,\varepsilon}$, we need to specify the decay rate of the sequence $(\rho^{(h)}(\ell))_{\ell \in \mathbb{Z}}$ from CS 1.

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

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

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

Theorem 5.2. *Assume that CS 1 and CD 1 hold. Assume additionally that for any $n \in \mathbb{N}$ there exists a decreasing continuous function α_n satisfying*

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

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

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

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

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

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

where $\lesssim$ stands for inequality up to a constant depending on λ, R, D, and ς. In the next section we shall apply Theorem 5.2 to the analysis of the ESVM algorithm for dependent sequences in ULA and SGLD.

5.2 Applications

In general, Theorem 5.2 can be applied to different types of dependent sequences satisfying conditions CS 1 and CD 1. In what follows, we let (X, d) be a complete separable metric space (equipped with its Borel σ-algebra $\mathcal{X}$) and consider P to be a Markov kernel on $(\mathsf{X}, \mathcal{X})$. Let $\Omega = \mathsf{X}^\mathbb{N}$ be the set of X-valued sequences endowed with the σ-field $\mathfrak{F} = \mathcal{X}^\mathbb{N}$, $(X_k)_{k=0}^\infty$ be the coordinate process, and $\mathfrak{F}_k = \sigma(X_\ell, \ell \leq k)$ be the canonical filtration. For every probability measure ξ on $(\mathsf{X}, \mathcal{X})$ there exists a unique probability $\mathbb{P}_\xi$ on $(\mathsf{X}^\mathbb{N}, \mathcal{X}^{\otimes \mathbb{N}})$ such that the coordinate process $(X_k)_{k=0}^\infty$ is a Markov chain with Markov kernel P and initial distribution ξ. We denote by $\mathbb{E}_\xi$ the associated expectation. We focus below on the case where P is W_p^{d}-uniformly ergodic for $p = 1$ or $p = 2$.

W 1 (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 a constant $\Delta_p(P) \in [0, 1)$ such that*

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

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

$$W_p^{\mathsf{d}}(\xi \mathrm{P}^n, \pi) \leq \Delta_p^n(\mathrm{P}) W_p^{\mathsf{d}}(\xi, \pi) \,, \quad n \in \mathbb{N}. \tag{5.9}$$

If there is no risk of confusion, we denote for simplicity $\Delta_p = \Delta_p(\mathrm{P})$. Let us start with a general result for Markov kernels satisfying W 1 (2). We show below that this assumption implies CS 1 and CD 1 when $\mathcal{H}$ is a subset of Lipschitz functions, and establish an exponential concentration inequality for $V_n(h)$, $h \in \mathcal{H}$. As it was emphasized in [78] and [36], powerful tools for exploring concentration properties of W_2^{d}-ergodic Markov kernels are the transportation cost-information inequalities.

Definition 5.3. *For $p \geq 1$, we say that $\mu \in \mathbb{M}_1(\mathsf{X})$ satisfies L^p-transportation cost-information inequality with constant $\alpha > 0$ if for any $\nu \in \mathbb{M}_1(\mathsf{X})$, $W_p^{\mathsf{d}}(\mu, \nu) \leq \sqrt{2\alpha \, \mathrm{KL}(\nu|\mu)}$. We write briefly $\mu \in T_p(\alpha)$ for this relation.*

L^p-transportation cost-information inequalities are well-studied in the literature, see, for instance, [4] and references therein. The cases $p = 1$ and $p = 2$ are of particular interest. Relations between $T_1(\alpha)$ and concentration inequalities are covered in [70] and [17]. In particular, $T_1(\alpha)$ is known to be equivalent to Gaussian concentration for all Lipschitz functions, see [17]. In turn $T_2(\alpha)$ is a stronger inequality than $T_1(\alpha)$. It was first established for the standard Gaussian measure on $\mathbb{R}^d$ by Talagrand in [105]. Moreover, the celebrated result by Bakry-Emery [3] implies that the measure $\pi(dx) = \mathrm{e}^{-U(x)} \mathrm{d}x$ satisfies $T_2(\alpha)$ if $\nabla^2 U \geq \alpha^{-1} \,\mathrm{I}$, see [4, Chapter 9.6]. We are especially interested in $T_2(\alpha)$, since it is known to be stable under both independent and Markovian tensorisations, see [85] and [36].

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

Proposition 5.4. *Let $\mathcal{H} \subseteq \mathrm{Lip}_{\mathsf{d}}(L)$ and assume that W 1 (2) holds. Then, for any initial distribution $\xi \in \mathbb{S}_2(\mathsf{X}, \mathsf{d})$, CS 1 is satisfied with*

$$\rho^{(h)}(\ell) = \mathbb{E}_\pi\big[\tilde{h}(X_0)\tilde{h}(X_{|\ell|})\big], \quad R = A_1 L^2 (1 - \Delta_2)^{-1} W_2(\xi, \pi), \tag{5.10}$$

where A_1 is a constant given in [11][Eq. A.12], and CD 1 is satisfied with

$$\varsigma = L\sqrt{D}\bigg[\int \{W_2^{\mathsf{d}}(\delta_x, \pi)\}^2 \pi(\mathrm{d}x)\bigg]^{1/2}, \quad \lambda = \Delta_2, \quad D = \sup\nolimits_{h \in \mathcal{H}} \mathrm{Var}_\pi(h). \tag{5.11}$$

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

$$\mathbb{P}_\xi\big(\big|V_n(h) - \mathbb{E}_\xi[V_n(h)]\big| \geq t\big) \leq 2 \exp\left(-\frac{(1 - \Delta_2)^2 n t^2}{c\alpha L^2 b_n^2 \big(D + Rn^{-1} + t\big)}\right), \tag{5.12}$$

where $c > 0$ is an absolute constant.

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

Proposition 5.5. *Let $\mathcal{H} \subset \mathrm{Lip}_{b,\mathsf{d}}(L, B)$ and assume that W1 (1) holds. Then for any initial distribution $\xi \in \mathbb{S}_1(\mathsf{X}, \mathsf{d})$, 1 is satisfied with*

$$\rho^{(h)}(\ell) = \mathbb{E}_\pi\big[\tilde{h}(X_0)\tilde{h}(X_{|\ell|})\big], \quad R = A_2 B(1 - \Delta_1^{1/2})^{-1}, \tag{5.13}$$

where A_2 is a constant given in [11][Eq. A.18], and CD 1 is satisfied with

$$\varsigma = 2LB \int W_1^{\mathsf{d}}(\delta_x, \pi)\pi(\mathrm{d}x), \quad \lambda = \Delta_1, \quad D = \sup_{h \in \mathcal{H}} \mathrm{Var}_\pi(h). \tag{5.14}$$

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

$$\mathbb{P}_\xi\big(\big|V_n(h) - \mathbb{E}_\xi[V_n(h)]\big| \geq t\big) \leq \frac{\mathsf{C}_{\mathsf{R},1}^p B^{2p} b_n^{3p/2} p^p}{n^{p/2} t^p} + \frac{\mathsf{C}_{\mathsf{R},2}^p B^{2p} b_n^{2p} p^{2p}}{n^{p-1} t^p}, \tag{5.15}$$

where constants $\mathsf{C}_{\mathsf{R},1}$ and $\mathsf{C}_{\mathsf{R},2}$ are given in [11][Eq. A.28].

5.2.1 Langevin dynamics

In this case, $\mathsf{X} = \mathbb{R}^d$ and we assume that π has an everywhere positive density w.r.t the Lebesgue measure, i.e., $\pi(\theta) = Z^{-1}\mathrm{e}^{-U(\theta)}$, where $Z = \int \mathrm{e}^{-U(\vartheta)}\mathrm{d}\vartheta$ is the normalization constant. Consider the first-order Euler-Maruyama discretization of the Langevin Dynamics from (5.4),

$$\theta_{k+1} = \theta_k - \gamma \nabla U(\theta_k) + \sqrt{2\gamma}\,\xi_{k+1}, \tag{5.16}$$

where $\gamma > 0$ is a step size and $(\xi_k)_{k=1}^\infty$ is an i.i.d. sequence of the standard Gaussian d-dimensional random vectors. The idea of using (5.16) to approximately sample from π has been advocated by [93] which coin the term Unadjusted Langevin Algorithm (ULA). Consider the following assumption on U.

ULA 1. *The function U is continuously differentiable on $\mathbb{R}^d$ with gradient ∇U satisfying the following two conditions.*

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

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

The Unadjusted Langevin Algorithm has been widely studied under the above assumptions, see, for example, [40] and [31]. As it is known from [40], under ULA 1 the associated Markov kernel, denoted by $\mathrm{P}_\gamma^{(\mathsf{ULA})}$, is W_2^{d}-uniformly ergodic. For completeness, we state below [40, Proposition 3].

Proposition 5.6. *Assume ULA 1 and set $\kappa = 2m_U L_U/(m_U + L_U)$. Then for any step size $\gamma \in (0, 2/(m_U + L_U))$, $\mathrm{P}_\gamma^{(\mathsf{ULA})}$ satisfies W1 (2) with $\mathsf{d}(\vartheta, \vartheta') = \|\vartheta - \vartheta'\|$ and $\Delta_2 = \sqrt{1 - \kappa\gamma}$. Moreover, $\mathrm{P}_\gamma^{(\mathsf{ULA})}$ has a unique invariant measure $\pi_\gamma^{(\mathsf{ULA})}$.*

It is shown in [40, Corollary 7] that, for any step size $\gamma \in (0, 2/(m_U + L_U))$,

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

We define the asymptotic variance as

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

At each iteration of the algorithm, ∇U is computed. Hence it is an appealing option to use this gradient to construct Stein control variates (see, for instance, [1], [79], and [84]), given by

$$g_\phi(\theta) = -\langle \phi(\theta)\,,\, \nabla U(\theta)\rangle + \operatorname{div}\bigl(\phi(\theta)\bigr), \tag{5.17}$$

where $\phi : \mathsf{X} \to \mathbb{R}^d$ is a continuously differentiable Lipschitz function, $\langle \cdot\,,\cdot\rangle$ is the standard scalar product in $\mathbb{R}^d$, and $\operatorname{div}(\phi)$ is the divergence of ϕ. Under rather mild conditions on π and ϕ, it follows from the integration by parts that $\pi(g_\phi) = 0$ (see [79, Propositions 1 and 2]). Note that if $\phi(\theta) \equiv b$, $b \in \mathbb{R}^d$, we get $g_b(\theta) = -\langle b\,,\, \nabla U(\theta)\rangle$. Then for a parametric class $\mathcal{H} = \{f - g_b : \|b\| \le B\}$, assuming that $f \in \operatorname{Lip}_{\mathsf{d}}(L_1)$ and that condition 1 holds, we get $\mathcal{H} \subset \operatorname{Lip}_{\mathsf{d}}(\max(L_1, BL_U))$. For other approaches to construct control variates we refer reader to [57], [34], and [21]. The next result follows now from Theorem 5.2 and Proposition 5.4.

Theorem 5.7. *Let $\mathcal{H} \subset \operatorname{Lip}_{\mathsf{d}}(L)$ and assume that ULA 1 holds. Assume additionally that $\xi \in T_2(\beta)$ for some $\beta > 0$. Fix any $\gamma \in (0, 2/(m_U + L_U))$ and set $b_n = 2\lceil \log(n)/\log(1/\Delta_2)\rceil$ with $\Delta_2 = \sqrt{1-\kappa\gamma}$ and $\kappa = 2m_U L_U/(m_U + L_U)$. Then, for any $\varepsilon > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\begin{aligned}&V_\infty^{(\mathsf{ULA})}(\widehat{h}_{n,\varepsilon}) - \inf\nolimits_{h\in\mathcal{H}} V_\infty^{(\mathsf{ULA})}(h)\\ &\qquad\qquad \lesssim \mathtt{C}_1\,\varepsilon\log(n) + \mathtt{C}_2\sqrt{\frac{\log^2(n)\log(|\mathcal{H}_\varepsilon|/\delta)}{n}} + \mathtt{C}_3\,\frac{\log^2(n)\log\left(|\mathcal{H}_\varepsilon|/\delta\right)}{n},\end{aligned}$$

where

$$\mathtt{C}_1 = \frac{\sqrt{R}+\sqrt{D}}{\kappa\gamma},\ \mathtt{C}_2 = \frac{L\sqrt{(\beta\vee\gamma)(D+R)}}{\kappa^2\gamma^2} + \frac{\sqrt{DR}}{\kappa\gamma},\ \mathtt{C}_3 = \frac{L^2(\beta\vee\gamma)}{\kappa^4\gamma^4} + \frac{R}{\kappa\gamma} + \frac{\varsigma}{\kappa^2\gamma^2}$$

with R, ς from Proposition 5.4 and $D = \sup_{h\in\mathcal{H}} \operatorname{Var}_{\pi_\gamma^{(\mathsf{ULA})}}(h)$.

Corollary 5.8. *Under the assumptions of Theorem 5.7, the following holds.*

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

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

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

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

Here $\lesssim$ stands for inequality up to a constant depending on ρ and other constants from Theorem 5.7. Moreover, if additionally the constant $\pi_\gamma^{(\mathsf{ULA})}(f)$ is in the class $\mathcal{H}$, then $\inf_{h\in\mathcal{H}} V_\infty^{(\mathsf{ULA})}(h) = 0$ and these bounds hold for the asymptotic variance itself.

5.2.2 Extension to the Stochastic Gradient Langevin Dynamics

In this section, we shall consider the situations where the target π is given by the posterior distribution in the Bayesian inference problem, that is, $\pi(\theta) \propto \exp\left(-U(\theta)\right)$, where $U(\theta) = U_0(\theta) + \sum_{i=1}^K U_i(\theta)$ with K being a number of observations. Computing $\nabla U(\theta)$ requires a computational budget that scales linearly with K. Hence it is often impossible to apply procedures based on discretisation of

Langevin Dinamics directly. One possible solution advocated by [111] is to replace $\nabla U(\theta)$ by an unbiased estimate. This gives rise to the SGLD algorithm, where the parameters are updated according to

$$\begin{aligned} \theta_{k+1} &= \theta_k - \gamma G(\theta_k, S_{k+1}) + \sqrt{2\gamma}\,\xi_{k+1}, \\ G(\theta, S) &= \nabla U_0(\theta) + KM^{-1} \sum\nolimits_{i \in S} \nabla U_i(\theta), \end{aligned} \tag{5.18}$$

where each S_{k+1} is a random batch taking values in S_M (here S_M is the set of all subsets S of $\{1, \ldots, K\}$ with $|S| = M$) which is sampled from a uniform distribution over S_M independently of $\mathcal{F}_k$ (here $(\mathcal{F}_k)_{k \geq 0}$ is the filtration generated by $\{(\theta_\ell, S_\ell)\}_{\ell \geq 0}$). Note that $\mathbb{E}[G(\theta_k, S_{k+1}) | \mathcal{F}_k] = \nabla U(\theta_k)$ and therefore $G(\theta_k, S_{k+1})$ is an unbiased estimate of $\nabla U(\theta_k)$. The available variance reduction techniques for SGLD usually replace the stochastic gradient in (5.18) with more sophisticated estimates which preserve unbiasedness but have lower variance.

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

$$G_{\mathrm{FP}}(\theta, S) = \nabla U_0(\theta) + KM^{-1} \sum\nolimits_{i \in S} \big(\nabla U_i(\theta) - \nabla U_i(\hat{\theta})\big) + \sum\nolimits_{i=1}^{K} \nabla U_i(\hat{\theta}). \tag{5.19}$$

The SGLD-FP algorithm is obtained by plugging this approximation into (5.18).

More sophisticated variance reduction methods typically use reference values $(g_k^i)_{i=1}^K$ of the gradient $(\nabla U_i)_{i=1}^K$ from previous iterates (and not only the last iterate); as a result, constructed sequence $(\theta_k)_{k=0}^\infty$ is often not Markovian. One particular example is SAGA-LD method, adapted from [94, 33]. If $i \in S_k$, the reference value is updated, that is, $g_{k+1}^i = \nabla U_i(\theta_k)$. Otherwise, the reference value is simply propagated, that is, $g_{k+1}^i = g_k^i$. One then considers the following gradient estimator

$$G_{\mathrm{SAGA}}^k(\theta, S) = \nabla U_0(\theta) + KM^{-1} \sum\nolimits_{i \in S} \big(\nabla U_i(\theta) - g_k^i\big) + g_k\,, \quad g_k = \sum\nolimits_{i=1}^{K} g_k^i\,. \tag{5.20}$$

The recursion is initialized with $g_0^i = \nabla U_i(\theta_0)$, $i \in \{1, \ldots, K\}$, and $g_0 = \sum_{i=1}^K g_0^i$. Finally, the gradient is computed according to (5.20) and plugged into (5.18).

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

SGLD1. *The function $U(\theta) = U_0(\theta) + \sum_{i=1}^K U_i(\theta)$ satisfies the following conditions.*

1) Lipschitz gradient: for any $i \in \{0, \ldots, K\}$, U_i is continuously differentiable on $\mathbb{R}^d$ with $\widetilde{L}_U$-Lipschitz gradient;

2) Convexity: for any $i \in \{0, \ldots, K\}$, U_i is convex;

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

Note that using Stein control variates with SGLD-based sampling procedure (5.18) eliminates benefits of using $G(\theta, S)$ instead of exact gradient $\nabla U(\theta)$. Following [45], we replace ∇U by its stochastic counterpart. More precisely, for k-th iteration of SGLD algorithm, we consider the control variates of the form

$$g_\phi(\theta, S) = -\langle \phi(\theta)\,,\, G(\theta, S) \rangle + \operatorname{div}\big(\phi(\theta)\big). \tag{5.21}$$

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

Proposition 5.9. *Assume SGLD 1. Then for any step size* $\gamma \in \big(0, \widetilde{L}_U^{-1}(K+1)^{-1}\big)$, $\overline{\mathrm{P}}$ *satisfies W 1 (2) with* $\Delta_2 = \sqrt{1 - \gamma m_U}$ *and* $\mathsf{d}(x, x') = \|\vartheta - \vartheta'\| + \mathbf{b}1_{S \neq S'}$ *for any* $x = (\vartheta, S)$ *and* $x' = (\vartheta', S')$. *Moreover,* $\overline{\mathrm{P}}$ *has a unique invariant measure* $\overline{\pi} = \pi_\gamma^{(\mathsf{SGLD})} \otimes \Upsilon_M$.

Similarly to Langevin Dynamics, we define

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

Theorem 5.10. *Let* $\mathcal{H} \subseteq \mathrm{Lip}_{b,\mathsf{d}}(L, B)$ *and assume that 1 holds. Fix any* $\gamma \in \big(0, \widetilde{L}_U^{-1}(K+1)^{-1}\big)$ *and set* $b_n = 2\lceil \log(n)/\log(1/\Delta_1)\rceil$ *with* $\Delta_1 = \sqrt{1 - \gamma m_U}$. *Then, for any* $\varepsilon > 0$ *and* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$,

$$\begin{aligned} &V_\infty^{(\mathsf{SGLD})}(\widehat{h}_{n,\varepsilon}) - \inf\nolimits_{h \in \mathcal{H}} V_\infty^{(\mathsf{SGLD})}(h) \\ &\lesssim \mathtt{C}_4\, \varepsilon \log(n) + \mathtt{C}_5 \sqrt{\frac{\log^5(n)}{n} \left(\frac{|\mathcal{H}_\varepsilon|}{\delta}\right)^{1/\log(n)}} + \mathtt{C}_6\, \frac{\log n}{n}, \end{aligned}$$

where

$$\mathtt{C}_4 = \frac{\sqrt{R} + \sqrt{D}}{m_U \gamma},\ \mathtt{C}_5 = \frac{B^2 R_1(L, \xi)}{(m_U \gamma)^2} + \frac{B^2 R_2(L, \xi)}{(m_U \gamma)^{4 + 2/\log n}} + \frac{\sqrt{RD}}{m_U \gamma},\ \mathtt{C}_6 = \frac{D(m_U \gamma) + \varsigma}{(m_U \gamma)^2}$$

with R, ς *from Proposition 5.5,* $D = \sup_{h \in \mathcal{H}} \mathrm{Var}_{\pi_\gamma^{(\mathsf{SGLD})}}(h)$, *and constants* $R_1(L, \xi)$, $R_2(L, \xi)$ *which can be tracked from [11][Eq. A.27].*

Proof. By Proposition 5.9, 1-2 holds with $\Delta_2 = \sqrt{1 - \gamma m_U}$, and, by Lyapunov inequality, W1 (1) also holds with $\Delta_1 = \Delta_2$. Hence, the second part of Proposition 5.5 can be applied with $p = \log n$. The remaining part follows from Theorem 5.2 with computation of the inverse function in the right-hand side of (5.15). □

Corollary 5.11. *Under the assumptions of Theorem 5.10, 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* $1 - 1/n$,

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

where $\lesssim$ *stands for inequality up to a constant depending on* ρ *and other constants from Theorem 5.10. Moreover, if additionally* $\overline{\pi}(f) \in \mathcal{H}$, *then* $\inf_{h \in \mathcal{H}} V_\infty^{(\mathsf{SGLD})}(h) = 0$ *and these bounds hold for the asymptotic variance itself.*

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

5.3 Experiments

In this section, we numerically compare the following two methods to choose control variates: Empirical Variance Minimisation (EVM) method, where a control variate is determined by minimizing the marginal variance, see (5.1), and Empirical Spectral Variance Minimisation (ESVM) method, where a control variate is determined by minimizing the spectral variance, see (5.6). Implementation is available at `https://github.com/svsamsonov/vr_sg_mcmc`.

5.3.1 Toy example

We first consider a multimodal distribution in $\mathbb{R}^2$ from [91]. Namely, let $\pi(x_1, x_2) = Z^{-1} \mathrm{e}^{-U(x_1,x_2)}$, where Z is the normalization constant and

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

We choose $M = 1$ and $\mu = \sigma = 3$; the respective density profile is presented in Figure 2. Our aim is to estimate $\pi(f)$ with $f(x_1, x_2) = x_1 + x_2$ using ULA. The parametric class g_φ in (5.17) is generated by $\varphi(x) = \sum_{k=1}^{p} \beta_k \psi_k(x)$, where $\psi_k = e^{-\|x-\mu_k\|^2/2\sigma_\psi^2}$ with all μ_k regularly spaced in $[-3, 3] \times [-3, 3]$ and $\sigma_\psi = 2$. Boxplots displaing variation of 100 estimates for EVM and ESVM are presented in the same Figure 2. Furthermore, we compute sample autocovariance functions for a trajectory with and without adding ESVM and EVM control variates. The results reflect a spectacular decrease in high-order autocovariance for ESVM, see Figure 2. Note that EVM aims at minimizing only the lag-zero autocovariance, that is why the autocovariance function for ESVM-adjusted trajectory decreases much faster.

Experiment	n_{burn}	n_{test}	γ	batch size
Toy example, Section 5.3.1	10^3	10^4	0.1	-
Gaussian Mixture, Section 5.3.2	10^4	10^5	0.01	10

Table 2: Experiment model parameters

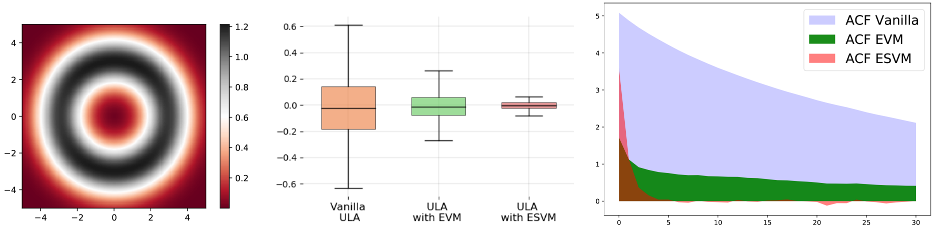

Figure 2: Toy example from Section 5.3.1. From left to right: (1) density profile, (2) boxplots displaing variation of 100 estimates for vanilla ULA, ULA with EVM, and ULA with ESVM, (3) sample autocovariance functions a trajectory with and without ESVM and EVM.

5.3.2 Gaussian mixture model

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

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

The density of the posterior distribution over μ is given in Figure 3. It has 2 modes roughly corresponding to $\mu = 1$ and $\mu = -1$. To generate data from this posterior distribution and estimate posterior mean, we use SGLD. The parametric class g_φ in (5.21) is generated by $\varphi(x) = \beta_0 x^2 + \beta_1 x + \beta_2$. Boxplots displaing variation of 100 estimates for EVM and ESVM and respective sample autocovariance functions are also presented in Figure 3. Note that the increase in lag-zero autocovariance for ESVM is explained by the additional randomness in (5.21). On contrary, EVM favors far too small coefficients to overcome this additional randomness, which leads to poor variance reduction.

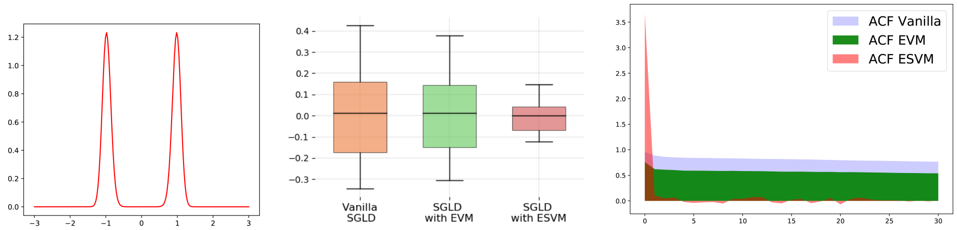

Figure 3: Gaussian mixture model from Section 5.3.2. From left to right: (1) density of the posterior distribution, (2) boxplots displaing variation of 100 estimates for vanilla SGLD, SGLD with EVM, and SGLD with ESVM, (3) sample autocovariance functions for a trajectory with and without ESVM and EVM.

5.3.3 Bayesian logistic regression

The probability of the i-th output $y_i \in \{-1, 1\}$, $i = 1, \ldots, K$, is given by $\mathrm{p}(y_i|\mathbf{x}_i, \theta) = (1 + \mathrm{e}^{-y_i \langle \theta\,,\,\mathbf{x}_i \rangle})^{-1}$, where $\mathbf{x}_i$ is a $d \times 1$ vector of predictors and θ is the vector of unknown regression coefficients. We complete the Bayesian model by considering the Zellner g-prior $\mathcal{N}_d(0, g(\mathbf{X}^\top\mathbf{X})^{-1})$ for θ where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ is an $K \times d$ design matrix, see [56, Section 2]. Normalizing the covariates, for $\tilde{\mathbf{x}}_i = (\mathbf{X}^\top\mathbf{X})^{-1/2}\mathbf{x}_i$ and $\tilde{\theta} = (\mathbf{X}^\top\mathbf{X})^{1/2}\theta$, we get $\langle \theta\,,\,\mathbf{x}_i \rangle = \langle \tilde{\theta}\,,\,\tilde{\mathbf{x}}_i \rangle$, under the Zellner g-prior, $\tilde{\theta} \sim \mathcal{N}_d(0, g\mathbf{I}_d)$.

We analyse the performance of EVM and ESVM methods on two datasets from the UCI repository. The first dataset, EEG, contains $K = 14\,980$ observations in dimension $d = 15$, the second dataset, SUSY, has $K = 500\,000$ observations in dimension $d = 19$. The data is first split into a training set $\mathcal{T}_N^{\mathrm{train}} = \{(y_i, \mathbf{x}_i)\}_{i=1}^K$ and a test set $\mathcal{T}_K^{\mathrm{test}} = \{(y_i', \mathbf{x}_i')\}_{i=1}^K$ by randomly picking $K = 100$ test points from the data. We use the SGLD-FP and SAGA-LD algorithms to approximately sample from the posterior distribution $\mathrm{p}(\tilde{\theta}|\mathcal{T}_N^{\mathrm{train}})$. Given a sample $(\tilde{\theta}_k)_{k=0}^{n-1}$, we can estimate the predictive distribution for a fixed test point $(y', \mathbf{x}')$, that is, $\mathrm{p}(y'|\mathbf{x}') = \int_{\mathbb{R}^d} \mathrm{p}(y'|\mathbf{x}', \tilde{\theta})\,\mathrm{p}(\tilde{\theta}|\mathcal{T}_N^{\mathrm{train}})\,\mathrm{d}\tilde{\theta}$, by computing the ergodic mean $n^{-1}\sum_{k=0}^{n-1} f(\tilde{\theta}_k)$ for $f(\tilde{\theta}) = \mathrm{p}(y'|\mathbf{x}', \tilde{\theta})$. To get rid of randomness caused by the random choice of a test point, we estimate the average predictive distribution for the whole test set $\mathcal{T}_K^{\mathrm{test}}$ by computing the ergodic mean for the function $f(\tilde{\theta}) = K^{-1}\sum_{i=1}^K \mathrm{p}(y_i'|\mathbf{x}_i', \tilde{\theta})$. Boxplots for the estimation of average predictive distribution are shown in Figure 4. Note that ESVM leads to a significant variance reduction for both SGLD-FP and SAGA-LD.

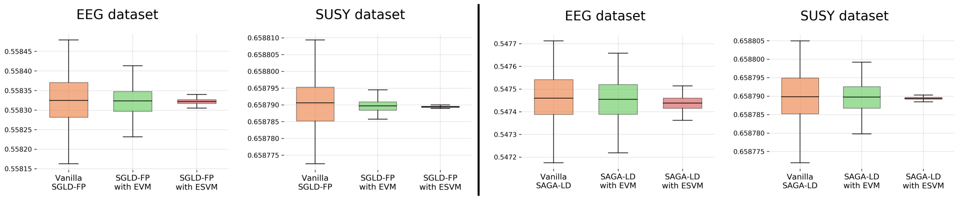

Figure 4: Bayesian logistic regression for EEG and SUSY datasets from Section 5.3.3. Boxplots displaing variation of 100 estimates of average predictive distribution for (1) left panel: vanilla SGLD-FP, SGLD-FP with EVM, and SGLD-FP with ESVM, (2) right panel: vanilla SAGA-LD, SAGA-LD with EVM, and SAGA-LD with ESVM.

Further, for the EEG dataset we plot in Figure 5 a part of the trajectory $f(\tilde{\theta}_m) = K^{-1}\sum_{i=1}^K \mathrm{p}(y_i'|\mathbf{x}_i', \tilde{\theta}_m)$ for 500 consecutive sample values $\tilde{\theta}_m$ with and without adding the ESVM control variate. These trajectories are accompanied by the sample autocovariance functions for vanilla and variance-reduced samples for both EVM and ESVM. Again, since EVM aims at minimizing only lag-zero autocovariance, the decrease in autocovariance function for this method is smaller than for ESVM. We also report in Figure 6 how autocovariance functions change with batch sizes. Note that for small batch sizes ESVM still manages to remove correlations, while EVM almost fails. At the same time, increasing the batch size leads to similar results for EVM and ESVM.

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

Table 4: Experimental hyperparameters

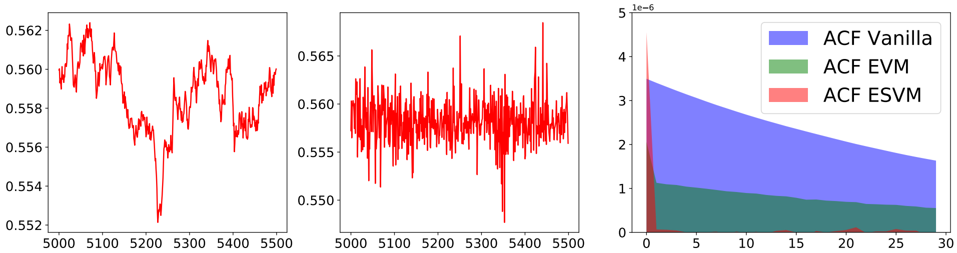

Figure 5: Bayesian logistic regression for the EEG dataset from Section 5.3.3. From left to right: (1) part of a trajectory without ESVM, (2) part of a trajectory with ESVM, (3) sample autocovariance functions for a trajectory with andwithout ESVM and EVM.

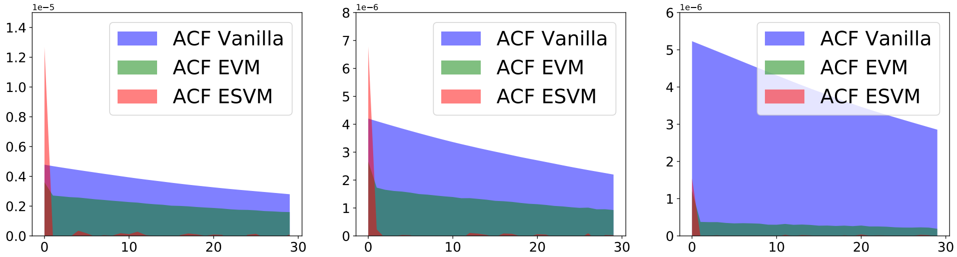

Figure 6: Bayesian logistic regression for the EEG dataset from Section 5.3.3. Comparison of sample autocovariance for different batch sizes. From left to right: batch size 5, 15, 150 respectively.

5.3.4 Bayesian Probabilistic Matrix Factorization

A typical problem in Recommendation Systems is to predict user's rating for a particular item given other user's ratings of this item and how a given user evaluated other items. A common approach to this problem is Probabilistic Matrix Factorization via Bayesian inference, see [97]. Namely, we are interested in approximating matrix $R \in \mathbb{R}^{M\times N}$, where M is a number of users, N is a number of rated items, and $R_{i,j}$ stands for rating assigned by i-th user to j-th item. Due to natural limitations (user is unlikely to rate all possible items), we observe only a some small subset of elements of R and want to predict ratings of the hidden part. In Probabilistic Matrix Factorization, we aim at representing R as a product $R = U^{\mathsf{T}}V + C$, where $U \in \mathbb{R}^{D\times M}$, $V \in \mathbb{R}^{D\times N}$, and $C \in \mathbb{R}^{M\times N}$ being a matrix of biases with elements $C_{i,j} = a_i + b_j$, $a \in \mathbb{R}^M$, $b \in \mathbb{R}^N$. In the subsequent experiments we assume that rank parameter $D = 10$ is fixed. The naive solution would be to find

$$U, V, a, b = \operatorname{argmin}_{U,V,a,b} \sum\nolimits_{(i,j)\in I_{\text{train}}} \big(R_{i,j} - \langle U_i\,,\, V_j\rangle - a_i - b_j\big)^2,$$

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

$$\begin{aligned} U, V, a, b = \operatorname{argmin}_{U,V,a,b} \sum\nolimits_{(i,j)\in I_{\text{train}}} \big(R_{i,j} - \langle U_i\,,\, V_j\rangle - a_i - b_j\big)^2 \\ + \lambda_U \|U\|^2 + \lambda_V \|V\|^2 + \lambda_a \|a\|^2 + \lambda_b \|b\|^2, \end{aligned}$$

but it requires careful tuning of penalisation coefficients $\lambda_U, \lambda_V, \lambda_a, \lambda_b$. We thus would benefit a lot from Bayesian approach for tuning weights; this was pointed out in [97]. We follow a slightly simplified formulation proposed by [28], that is, we consider

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

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

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

2. Sample new λ from $p(\lambda_U, \lambda_V, \lambda_a, \lambda_b|U, V, a, b)$ using the Gibbs sampler.

The experiments are performed on the Movielens dataset $ml{-}100k$ (link to dataset). We apply our control variates procedure as a postprocessing step following [2]. The functional of interest is the mean squared error over the test subsample, $f(U, V, a, b) = \sum_{(i,j)\in I_{\text{test}}} (R_{i,j} - \langle U_i\,,\, V_j\rangle - a_i - b_j)^2$. Since the dimension of parameter space is very high, first-order control variates are the only option among Stein's control variates. Parts of SGLD- and SGLD-FP-based trajectories before and after using control variates, and confidence intervals for estimation of f are presented in Figure 7.

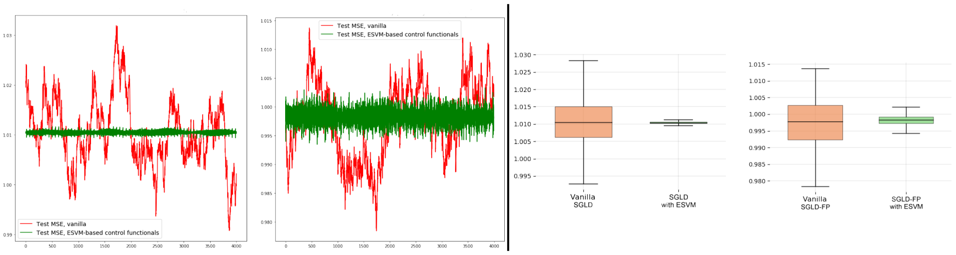

Figure 7: Bayesian Probabilistic Matrix Factorization from Section 5.3.4. Left Panel: test MSE trajectory for SGLD (left) and SGLD-FP (right) with and without ESVM. Right Panel: confidence intervals for test MSE trajectory for SGLD (left) and SGLD-FP (right).

6 Conclusion

This thesis is based on published papers [48, 81, 83, 82, 8, 11, 64, 42, 41, 49].

Let us list the main results that are obtained in this thesis and submitted for defense.

1. Tight non-asymptotic bounds for the Kolmogorov distance between the probabilities of two Gaussian elements to hit a ball in a Hilbert space.
2. Anti-concentration bound for the squared norm of a non-centered Gaussian element in a Hilbert space.
3. Bootstrap procedure for building sharp confidence sets for the true spectral projector of covariance matrix from the given data.
4. Exponential stability of random matrix products driven by i.i.d. sequence or a general (possibly unbounded) state space Markov chain.
5. Finite-time p-th moment bounds for constant and decreasing stepsize linear stochastic approximation schemes with i.i.d. or Markovian noise on general state space.
6. Novel and practical variance reduction approach for additive functionals of dependent sequences.

7 Acknowledgements

The dissertation was prepared within the framework of the HSE University Basic Research Program.

References

[1] Roland Assaraf and Michel Caffarel. “Zero-variance principle for Monte Carlo algorithms”. In: *Phys. Rev. Lett.* 83.23 (1999), pp. 4682–4685.

[2] Jack Baker, Paul Fearnhead, Emily B Fox, and Christopher Nemeth. “Control variates for stochastic gradient MCMC”. In: *Statistics and Computing* 29.3 (2019), pp. 599–615.

[3] Dominique Bakry and Michel Émery. “Diffusions hypercontractives”. In: *Séminaire de probabilités de Strasbourg* 19 (1985), pp. 177–206.

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

[5] Keith Ball. “The reverse isoperimetric problem for Gaussian measure”. In: *Discrete Comput. Geom.* 10.4 (1993), pp. 411–420. ISSN: 0179-5376. DOI: 10.1007/BF02573986. URL: https://doi.org/10.1007/BF02573986.

[6] Eduard Belitser. “On coverage and local radial rates of credible sets”. In: *Ann. Statist.* 45.3 (2017), pp. 1124–1151. ISSN: 0090-5364. DOI: 10.1214/16-AOS1477. URL: https://doi.org/10.1214/16-AOS1477.

[7] D Belomestny, L Iosipoi, and N Zhivotovskiy. “Variance reduction via empirical variance minimization: convergence and complexity”. In: *arXiv preprint, arXiv:1712.04667* (2017).

[8] D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, and S. Samsonov. "Variance reduction for Markov chains with application to MCMC". In: *Stat. Comput.* 30.4 (2020), pp. 973–997. ISSN: 0960-3174. DOI: 10.1007/s11222-020-09931-z. URL: https://doi.org/10.1007/s11222-020-09931-z.

[9] D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, and S. Samsonov. "Variance reduction for Markov chains with application to MCMC". In: *Statistics and Computing* 30.4 (2020), pp. 973–997.

[10] D. V. Belomestny, L. S. Iosipoi, and N. K. Zhivotovskiy. "Variance Reduction in Monte Carlo Estimators via Empirical Variance Minimization". In: *Doklady Mathematics* 98.2 (2018), pp. 494–497.

[11] Denis Belomestny, Leonid Iosipoi, Eric Moulines, Alexey Naumov, and Sergey Samsonov. "Variance Reduction for Dependent Sequences with Applications to Stochastic Gradient MCMC". In: *SIAM/ASA J. Uncertain. Quantif.* 9.2 (2021), pp. 507–535. DOI: 10.1137/19M1301199. URL: https://doi.org/10.1137/19M1301199.

[12] V. Bentkus. "A Lyapunov type bound in $\mathbf{R}^d$". In: *Teor. Veroyatn. Primen.* 49.2 (2004), pp. 400–410. ISSN: 0040-361X. DOI: 10.1137/S0040585X97981123. URL: https://doi.org/10.1137/S0040585X97981123.

[13] V. Bentkus. "On the dependence of the Berry-Esseen bound on dimension". In: *J. Statist. Plann. Inference* 113.2 (2003), pp. 385–402. ISSN: 0378-3758. DOI: 10.1016/S0378-3758(02)00094-0. URL: https://doi.org/10.1016/S0378-3758(02)00094-0.

[14] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations.* Vol. 22. Springer Science & Business Media, 2012.

[15] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming.* Belmont, MA: Athena Scientific, 1996.

[16] J. Bhandari, D. Russo, and R. Singal. "A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation". In: *Conference On Learning Theory.* 2018, pp. 1691–1692.

[17] S. G. Bobkov and F. Götze. "Exponential Integrability and Transportation Cost Related to Logarithmic Sobolev Inequalities". English (US). In: *Journal of Functional Analysis* 163.1 (Apr. 1999), pp. 1–28. ISSN: 0022-1236.

[18] Sergey G. Bobkov, Alexey A. Naumov, and Vladimir V. Ulyanov. "Two-sided inequalities for the density function's maximum of weighted sum of chi-square variables". In: *arXiv e-prints*, arXiv:2012.10747 (Dec. 2020), arXiv:2012.10747. arXiv: 2012.10747 [math.PR].

[19] Vladimir I. Bogachev. *Gaussian measures.* Vol. 62. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 1998, pp. xii+433. ISBN: 0-8218-1054-5. DOI: 10.1090/surv/062. URL: https://doi.org/10.1090/surv/062.

[20] Dominique Bontemps. "Bernstein-von Mises theorems for Gaussian regression with increasing number of regressors". In: *Ann. Statist.* 39.5 (2011), pp. 2557–2584. ISSN: 0090-5364. DOI: 10.1214/11-AOS912. URL: https://doi.org/10.1214/11-AOS912.

[21] Nicolas Brosse, Alain Durmus, Sean Meyn, and Eric Moulines. "Diffusion approximations and control variates for MCMC". In: *arXiv preprint, arXiv:1808.01665* (2018).

[22] Nicolas Brosse, Alain Durmus, and Eric Moulines. “The promises and pitfalls of Stochastic Gradient Langevin Dynamics”. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS’18. 2018, pp. 8278–8288.

[23] Ismaël Castillo. “A semiparametric Bernstein–von Mises theorem for Gaussian process priors”. In: *Probab. Theory Related Fields* 152.1-2 (2012), pp. 53–99. ISSN: 0178-8051. DOI: 10.1007/s00440-010-0316-5. URL: https://doi.org/10.1007/s00440-010-0316-5.

[24] Ismaël Castillo and Richard Nickl. “Nonparametric Bernstein-von Mises theorems in Gaussian white noise”. In: *Ann. Statist.* 41.4 (2013), pp. 1999–2028. ISSN: 0090-5364. DOI: 10.1214/13-AOS1133. URL: https://doi.org/10.1214/13-AOS1133.

[25] Siddharth Chandak and Vivek S. Borkar. “Concentration of Contractive Stochastic Approximation and Reinforcement Learning”. In: *arXiv:2106.14308* (2021).

[26] 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”. In: *Proceedings of Machine Learning Research* 80 (2018).

[27] S. Chen, A. Devraj, A. Busic, and S. Meyn. “Explicit mean-square error bounds for monte-carlo and linear stochastic approximation”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 4173–4183.

[28] Tianqi Chen, Emily B. Fox, and Carlos Guestrin. “Stochastic Gradient Hamiltonian Monte Carlo”. In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML’14. 2014, pp. 1683–1691.

[29] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. “Central limit theorems and bootstrap in high dimensions”. In: *Ann. Probab.* 45.4 (2017), pp. 2309–2352. ISSN: 0091-1798. DOI: 10.1214/16-AOP1113. URL: https://doi.org/10.1214/16-AOP1113.

[30] Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. “Finite sample analyses for TD(0) with function approximation”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[31] Arnak Dalalyan. “Theoretical guarantees for approximate sampling from smooth and log-concave densities”. In: *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 79.3 (2017), pp. 651–676.

[32] Arnak S. Dalalyan and Avetik G. Karagulyan. “User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient”. In: *Stoch. Proc. Appl.* 129.12 (2019), pp. 5278–5311.

[33] 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*. 2014, pp. 1646–1654.

[34] Petros Dellaportas and Ioannis Kontoyiannis. “Control variates for estimation based on reversible Markov chain Monte Carlo samplers”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.1 (2012), pp. 133–161.

[35] B. Delyon and A. Yuditsky. “On Small Perturbations of Stable Markov Operators: Unbounded Case”. In: *Theory Probab. Appl.* 43.4 (1999), pp. 577–587.

[36] H. Djellout, A. Guillin, and L. Wu. “Transportation cost-information inequalities and applications to random dynamical systems and diffusions”. In: *Ann. Probab.* 32.3B (2004), pp. 2702–2732. ISSN: 0091-1798.

[37] Thinh T Doan. “Finite-Time Analysis and Restarting Scheme for Linear Two-Time-Scale Stochastic Approximation”. In: *arXiv preprint arXiv:1912.10583* (2019).

[38] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, Cham, 2018, pp. xviii+757. ISBN: 978-3-319-97703-4; 978-3-319-97704-1.

[39] Kumar Avinava Dubey, Sashank J Reddi, Sinead A Williamson, Barnabas Poczos, Alexander J Smola, and Eric P Xing. “Variance Reduction in Stochastic Gradient Langevin Dynamics”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 1154–1162.

[40] Alain Durmus and Eric Moulines. “High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm”. In: *Bernoulli* 25.4A (Nov. 2019), pp. 2854–2882.

[41] Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, and Hoi-To Wai. “Tight High Probability Bounds for Linear Stochastic Approximation with Fixed Stepsize”. In: *NeurIPS*. 2021.

[42] Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Hoi-To Wai. “On the Stability of Random Matrix Product with Markovian Noise: Application to Linear Stochastic Approximation and TD Learning”. In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Ed. by Mikhail Belkin and Samory Kpotufe. Vol. 134. Proceedings of Machine Learning Research. PMLR, 15–19 Aug 2021, pp. 1711–1752. URL: https://proceedings.mlr.press/v134/durmus21a.html.

[43] E. Eweda and O. Macchi. “Quadratic mean and almost-sure convergence of unbounded stochastic approximation algorithms with correlated observations”. In: *Ann. Inst. H. Poincaré Sect. B (N.S.)* 19.3 (1983), pp. 235–255. ISSN: 0020-2347.

[44] James M. Flegal and Galin L. Jones. “Batch means and spectral variance estimators in Markov chain Monte Carlo”. In: *Ann. Statist.* 38.2 (Apr. 2010), pp. 1034–1070.

[45] Nial Friel, Antonietta Mira, and Chris. J. Oates. “Exploiting Multi-Core Architectures for Reduced-Variance Estimation with Intractable Likelihoods”. In: *Bayesian Analysis* 11.1 (2015), pp. 215–245.

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

[47] Emmanuel Gobet. *Monte-Carlo Methods and Stochastic Processes*. CRC Press, Boca Raton, FL, 2016, pp. xxv+309. ISBN: 978-1-4987-4622-9.

[48] Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, and Vladimir Ulyanov. “Large ball probabilities, Gaussian comparison and anti-concentration”. In: *Bernoulli* 25.4A (2019), pp. 2538–2563. ISSN: 1350-7265. DOI: 10.3150/18-BEJ1062. URL: https://doi.org/10.3150/18-BEJ1062.

[49] Friedrich Götze, Alexey Naumov, and Alexander Tikhomirov. "Distribution of linear statistics of singular values of the product of random matrices". In: *Bernoulli* 23.4B (2017), pp. 3067–3113. ISSN: 1350-7265. DOI: 10.3150/16-BEJ837. URL: https://doi.org/10.3150/16-BEJ837.

[50] Friedrich Götze and Andrei Yu. Zaitsev. "New applications of Arak's inequalities to the Littlewood-Offord problem". In: *Eur. J. Math.* 4.2 (2018), pp. 639–663. ISSN: 2199-675X. DOI: 10.1007/s40879-018-0215-3. URL: https://doi.org/10.1007/s40879-018-0215-3.

[51] L. Guo. "Stability of recursive stochastic tracking algorithms". In: *SIAM Journal on Control and Optimization* 32.5 (1994), pp. 1195–1225.

[52] L. Guo and L. Ljung. "Exponential stability of general tracking algorithms". In: *IEEE Transactions on Automatic Control* 40.8 (1995), pp. 1376–1387.

[53] L. Guo and L. Ljung. "Performance analysis of general tracking algorithms". In: *IEEE Transactions on Automatic Control* 40.8 (1995), pp. 1388–1402.

[54] H. Gupta, R Srikant, and L. Ying. "Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning". In: *Advances in Neural Information Processing Systems*. 2019, pp. 4706–4715.

[55] Ramon van Handel. "Structured random matrices". In: *Convexity and concentration*. Vol. 161. IMA Vol. Math. Appl. Springer, New York, 2017, pp. 107–156.

[56] Timothy E Hanson, Adam J Branscum, Wesley O Johnson, et al. "Informative *g*-Priors for Logistic Regression". In: *Bayesian Analysis* 9.3 (2014), pp. 597–612.

[57] Shane G Henderson. "Variance reduction via an approximating Markov process". PhD thesis. Stanford University, 1997.

[58] De Huang, Jonathan Niles-Weed, Joel A Tropp, and Rachel Ward. "Matrix Concentration for Products". In: *arXiv preprint arXiv:2003.05437* (2020).

[59] B. Jacob and H. Zwart. *Linear Port-Hamiltonian Systems on Infinite-dimensional Spaces*. Operator Theory: Advances and Applications 223. 10.1007/978-3-0348-0399-1. Springer, 2012. ISBN: 978-3-0348-0398-4. DOI: 10.1007/978-3-0348-0399-1.

[60] Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous univariate distributions. Vol. 1*. Second. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1994, pp. xxii+756. ISBN: 0-471-58495-9.

[61] Rie Johnson and Tong Zhang. "Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction". In: *Advances in Neural Information Processing Systems*. 2013, pp. 315–323.

[62] Iain M. Johnstone. "High dimensional Bernstein–von Mises: simple examples". In: *Borrowing strength: theory powering applications—a Festschrift for Lawrence D. Brown*. Vol. 6. Inst. Math. Stat. (IMS) Collect. Inst. Math. Statist., Beachwood, OH, 2010, pp. 87–98.

[63] M. Kaledin, E. Moulines, A. Naumov, V. Tadic, and Hoi-To Wai. "Finite time analysis of linear two-timescale stochastic approximation with Markovian noise". In: *Conference On Learning Theory*. 2020.

[64] Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. “Finite Time Analysis of Linear Two-timescale Stochastic Approximation with Markovian Noise”. In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2144–2203. URL: http://proceedings.mlr.press/v125/kaledin20a.html.

[65] Vladimir Koltchinskii and Karim Lounici. “Concentration inequalities and moment bounds for sample covariance operators”. In: *Bernoulli* 23.1 (2017), pp. 110–133. ISSN: 1350-7265. DOI: 10.3150/15-BEJ730. URL: https://doi.org/10.3150/15-BEJ730.

[66] Vladimir Koltchinskii and Karim Lounici. “Normal approximation and concentration of spectral projectors of sample covariance”. In: *Ann. Statist.* 45.1 (2017), pp. 121–157. ISSN: 0090-5364. DOI: 10.1214/16-AOS1437. URL: https://doi.org/10.1214/16-AOS1437.

[67] I. Kontoyiannis and S. Meyn. “Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes”. In: *Electronic Journal of Probability* 10 (2005), pp. 61–123.

[68] I. Kontoyiannis and S. P. Meyn. “Spectral theory and limit theorems for geometrically ergodic Markov processes”. In: *Ann. Appl. Probab.* 13.1 (2003), pp. 304–362. ISSN: 1050-5164. DOI: 10.1214/aoap/1042765670. URL: https://doi.org/10.1214/aoap/1042765670.

[69] C. Lakshminarayanan and C. Szepesvari. “Linear stochastic approximation: How far does constant step-size and iterate averaging go?” In: *International Conference on Artificial Intelligence and Statistics*. 2018, pp. 1347–1355.

[70] M. Ledoux. *The Concentration of Measure Phenomenon*. Vol. 89. AMS Surveys and Monographs, 2001.

[71] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Isoperimetry and processes, Reprint of the 1991 edition. Springer-Verlag, Berlin, 2011, pp. xii+480. ISBN: 978-3-642-20211-7.

[72] W. V. Li and Q.-M. Shao. “Gaussian processes: inequalities, small ball probabilities and applications”. In: *Stochastic processes: theory and methods*. Vol. 19. Handbook of Statist. North-Holland, Amsterdam, 2001, pp. 533–597. DOI: 10.1016/S0169-7161(01)19019-X. URL: https://doi.org/10.1016/S0169-7161(01)19019-X.

[73] Mikhail Lifshits. *Lectures on Gaussian processes*. SpringerBriefs in Mathematics. Springer, Heidelberg, 2012, pp. x+121. ISBN: 978-3-642-24938-9; 978-3-642-24939-6. DOI: 10.1007/978-3-642-24939-6. URL: https://doi.org/10.1007/978-3-642-24939-6.

[74] Lennart Ljung. “Recursive identification algorithms”. In: *Circuits, Systems and Signal Processing* 21.1 (2002), pp. 57–68.

[75] Yi-An Ma, Tianqi Chen, and Emily Fox. “A Complete Recipe for Stochastic Gradient MCMC”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 2917–2925.

[76] Odile Macchi and Eweda Eweda. “Second-order convergence analysis of stochastic adaptive linear filtering”. In: *IEEE Transactions on Automatic Control* 28.1 (1983), pp. 76–85.

[77] A. S. Markus. “Eigenvalues and singular values of the sum and product of linear operators”. In: *Uspehi Mat. Nauk* 19.4 (118) (1964), pp. 93–123. ISSN: 0042-1316.

[78] K. Marton. "Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration". In: *Ann. Probab.* 24.2 (Apr. 1996), pp. 857–866.

[79] Antonietta Mira, Reza Solgi, and Daniele Imparato. "Zero variance Markov chain Monte Carlo for Bayesian estimators". In: *Statistics and Computing* 23.5 (2013), pp. 653–662.

[80] Tigran Nagapetyan, Andrew B Duncan, Leonard Hasenclever, Sebastian J Vollmer, Lukasz Szpruch, and Konstantinos Zygalakis. "The True Cost of Stochastic Gradient Langevin Dynamics". In: *arXiv preprint, arXiv:1706.02692* (2017).

[81] A. A. Naumov, V. G. Spokoiny, Yu. E. Tavyrikov, and V. V. Ulyanov. "Nonasymptotic Estimates for the Closeness of Gaussian Measures on Balls". In: *Doklady Mathematics* 98.2 (2018), pp. 490–493. DOI: 10.1134/S1064562418060248. URL: https://doi.org/10.1134/S1064562418060248.

[82] A. A. Naumov, V. G. Spokoiny, and V. V. Ulyanov. "Confidence Sets for Spectral Projectors of Covariance Matrices". In: *Doklady Mathematics* 98.2 (2018), pp. 511–514. DOI: 10.1134/S1064562418060285. URL: https://doi.org/10.1134/S1064562418060285.

[83] Alexey Naumov, Vladimir Spokoiny, and Vladimir Ulyanov. "Bootstrap confidence sets for spectral projectors of sample covariance". In: *Probab. Theory Related Fields* 174.3-4 (2019), pp. 1091–1132. ISSN: 0178-8051. DOI: 10.1007/s00440-018-0877-2. URL: https://doi.org/10.1007/s00440-018-0877-2.

[84] Chris J Oates, Mark Girolami, and Nicolas Chopin. "Control functionals for Monte Carlo integration". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.3 (2017), pp. 695–718.

[85] F. Otto and C. Villani. "Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality". In: *Journal of Functional Analysis* 173.2 (2000), pp. 361–400.

[86] Maxim Panov and Vladimir Spokoiny. "Finite sample Bernstein–von Mises theorem for semiparametric problems". In: *Bayesian Anal.* 10.3 (2015), pp. 665–710. ISSN: 1936-0975. DOI: 10.1214/14-BA926. URL: https://doi.org/10.1214/14-BA926.

[87] G. Pflug. "Stochastic Minimization with Constant Step-Size: Asymptotic Laws". In: *SIAM Journal on Control and Optimization* 24.4 (1986), pp. 655–666. DOI: 10.1137/0324039. eprint: https://doi.org/10.1137/0324039. URL: https://doi.org/10.1137/0324039.

[88] A. S. Poznyak. *Advanced Mathematical Tools for Automatic Control Engineers: Deterministic Techniques.* Oxford: Elsevier, 2008.

[89] P. Priouret and A. Veretenikov. "A remark on the stability of the LMS tracking algorithm". In: *Stochastic analysis and applications* 16.1 (1998), pp. 119–129.

[90] Yuri V. Prokhorov and Vladimir V. Ulyanov. "Some approximation problems in statistics and probability". In: *Limit theorems in probability, statistics and number theory.* Vol. 42. Springer Proc. Math. Stat. Springer, Heidelberg, 2013, pp. 235–249. DOI: 10.1007/978-3-642-36068-8_11. URL: https://doi.org/10.1007/978-3-642-36068-8_11.

[91] Danilo Jimenez Rezende and Shakir Mohamed. "Variational Inference with Normalizing Flows". In: *arXiv preprint arXiv:1505.05770* (2015).

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

[93] G. O. Roberts and R. L. Tweedie. "Exponential convergence of Langevin distributions and their discrete approximations". In: *Bernoulli* 2.4 (1996), pp. 341–363. ISSN: 1350-7265.

[94] Nicolas L. Roux, Mark Schmidt, and Francis R. Bach. "A Stochastic Gradient Method with an Exponential Convergence Rate for Finite Training Sets". In: *Advances in Neural Information Processing Systems 25.* 2012, pp. 2663–2671.

[95] Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo Method.* Vol. 10. John Wiley & Sons, 2016.

[96] Mark Rudelson and Roman Vershynin. "The Littlewood-Offord problem and invertibility of random matrices". In: *Adv. Math.* 218.2 (2008), pp. 600–633. ISSN: 0001-8708. DOI: `10.1016/j.aim.2008.01.010`. URL: `https://doi.org/10.1016/j.aim.2008.01.010`.

[97] R. Salakhutdinov and A. Mnih. "Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo". In: *Proceedings of the 25th International Conference on Machine Learning (ICML-08).* 2008, pp. 880–887.

[98] Vjačeslav V. Sazonov. *Normal approximation—some recent advances.* Vol. 879. Lecture Notes in Mathematics. Springer-Verlag, Berlin-New York, 1981, pp. vii+105. ISBN: 3-540-10863-7.

[99] Jun Shao. *Mathematical statistics.* Second. Springer Texts in Statistics. Springer-Verlag, New York, 2003, pp. xvi+591. ISBN: 0-387-95382-5. DOI: `10.1007/b97553`. URL: `https://doi.org/10.1007/b97553`.

[100] Vladimir Spokoiny. "Penalized maximum likelihood estimation and effective dimension". In: *Ann. Inst. Henri Poincaré Probab. Stat.* 53.1 (2017), pp. 389–429. ISSN: 0246-0203. DOI: `10.1214/15-AIHP720`. URL: `https://doi.org/10.1214/15-AIHP720`.

[101] Vladimir Spokoiny and Mayya Zhilova. "Bootstrap confidence sets under model misspecification". In: *Ann. Statist.* 43.6 (2015), pp. 2653–2675. ISSN: 0090-5364. DOI: `10.1214/15-AOS1355`. URL: `https://doi.org/10.1214/15-AOS1355`.

[102] R. Srikant and Lei Ying. "Finite-Time Error Bounds For Linear Stochastic Approximation and TD Learning". In: *Conference on Learning Theory.* 2019.

[103] Richard S. Sutton. "Learning to predict by the methods of temporal differences". In: *Machine Learning* 3.1 (Aug. 1988), pp. 9–44. ISSN: 1573-0565. DOI: `10.1007/BF00115009`.

[104] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* Second. The MIT Press, 2018.

[105] M. Talagrand. "Transportation cost for Gaussian and other product measures". In: *Geom. Funct. Anal.* 6.3 (1996), pp. 587–600. ISSN: 1016-443X.

[106] Yee Whye Teh, Alexandre H Thiery, and Sebastian J Vollmer. "Consistency and Fluctuations for Stochastic Gradient Langevin Dynamics". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 193–225.

[107] Joel A. Tropp. "User-friendly tail bounds for sums of random matrices". In: *Found. Comput. Math.* 12.4 (2012), pp. 389–434. ISSN: 1615-3375. DOI: `10.1007/s10208-011-9099-z`. URL: `https://doi.org/10.1007/s10208-011-9099-z`.

[108] J. N. Tsitsiklis and B. Van Roy. "An analysis of temporal-difference learning with function approximation". In: *IEEE Transactions on Automatic Control* 42.5 (May 1997), pp. 674–690. ISSN: 2334-3303. DOI: 10.1109/9.580874.

[109] S. Varadhan. *Large deviations and applications.* SIAM, 1984.

[110] Roman Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *Compressed sensing.* Cambridge Univ. Press, Cambridge, 2012, pp. 210–268.

[111] M. Welling and Y. W. Teh. "Bayesian Learning via Stochastic Gradient Langevin Dynamics". In: *Proceedings of the 28th International Conference on Machine Learning (ICML-11).* 2011, pp. 681–688.

[112] Tengyu Xu, Shaofeng Zou, and Yingbin Liang. "Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples". In: *Advances in Neural Information Processing Systems.* 2019, pp. 10633–10643.

[113] Vadim Yurinsky. *Sums and Gaussian vectors.* Vol. 1617. Lecture Notes in Mathematics. Springer-Verlag, Berlin, 1995, pp. xii+305. ISBN: 3-540-60311-5. DOI: 10.1007/BFb0092599. URL: https://doi.org/10.1007/BFb0092599.

[114] Difan Zou, Pan Xu, and Quanquan Gu. "Subsampled Stochastic Variance-Reduced Gradient Langevin Dynamics". In: *International Conference on Uncertainty in Artificial Intelligence.* 2018.