On statistical inference for selected multivariate stochastic processes of diffusion type

PhD Dissertation

Jan Niklas Dexheimer

Department of Mathematics

Aarhus University

May 30, 2022

On statistical inference for selected multivariate stochastic processes of diffusion type

PhD Dissertation by
Jan Niklas Dexheimer

Department of Mathematics, Aarhus University
Ny Munkegade 118, 8000 Aarhus C, Denmark

Supervised by
Associate Professor Claudia Strauch, Aarhus University
Professor Sören Christensen, Christian-Albrechts-Universität zu Kiel

Submitted to Graduate School of Natural Sciences, Aarhus, May 30, 2022

AARHUS
UNIVERSITY
DEPARTMENT OF MATHEMATICS

Contents

Preface

This dissertation concludes my PhD studies which started in May 2019 at the School of Business Informatics and Mathematics, University of Mannheim and continued at the Department of Mathematics, Aarhus University from June 2020 to May 2022 under supervision of Associate Professor Claudia Strauch (main supervisor) from Aarhus University and Professor Sören Christensen (co-supervisor) from Christian-Albrechts-Universität zu Kiel.
The main body of this dissertation consists of an introductory section and the three following self-contained papers:

Paper A Adaptive invariant density estimation for continuous-time mixing Markov processes under sup-norm risk.
To appear in *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*.

Paper B Estimating the characteristics of stochastic damping Hamiltonian systems from continuous observations.
Revision submitted to *Stochastic Processes and their Applications*.
Preprint available at arXiv (arXiv:2109.13190v3).

Paper C On Lasso and Slope drift estimators for Lévy-driven Ornstein–Uhlenbeck processes.
Submitted to *Bernoulli*.
Preprint available at arXiv (arXiv:2205.07813v1).

The papers all correspond to their submitted versions, besides changes in layout, numbering, typesetting and the correction of some minor typing errors. Excerpts of earlier versions of Paper A and Paper B were contained in the progress report for my qualifying examination in May 2021, however both have seen significant changes since then. Paper C was written after my qualifying examination and thus was not part of the progress report. I have contributed extensively to all three papers, both in the research and writing phase.
The introductory chapter of this dissertation is meant to acquaint the reader with diffusion processes and the generic chaining device, since both are central for the global concept, respectively the main arguments of this work. Furthermore summaries of the framework, goals, main results and methodology of each paper are given, together with a comparison to relevant current research.
Probabilists and statisticians are cautious about using the term *"surely"* and usually resort to the safer formulation *"almost surely"*. However, this dissertation would *surely* not have been possible without the great support I received during the last three years.
Firstly, I have to thank all the colleagues I met, both in Mannheim and Aarhus. Be it the golf tournaments, the kaffeslaberas or simple lunch breaks; moments like these just made my PhD studies as delightful as they were.
At this time I especially want to point out Lukas Trottner. Not only for taking care of me during my darkest days on the island of Samos, but also for introducing me to the highest levels of pedantism regarding LaTeX. Without him this dissertation would certainly not look like this (and I probably would not have even cared). Also our daily Faxe Kondi, Cherry Coke or Pepsi Max

Lime tastings – with all the wisdom and gossip that unfolded during them – were an integral part of my PhD studies.
Furthermore I want to thank my co-supervisor Sören Christensen for his hospitality during my stay at the Christian-Albrechts-Universität zu Kiel and his support especially in the last weeks of my PhD studies.
To my main supervisor Claudia Strauch I can only express my deepest gratitude for being a continuous source of wisdom and inspiration during the last three years. Her patience with my (especially during lockdown) unusual working hours, confused questions and sometimes chaotic writing is still unbelievable to me. Not only was she a great support in all professional matters but also always cared for my general well-being. Thank you Claudia, this dissertation would not exist without you!
I am also grateful to all my friends from Germany, which I have sadly not seen enough during the last two years; not only because of moving to a different country but also due to a certain global pandemic taking place. Our trips to the stadium of Mainz 05 were a welcome distraction from the sometimes overwhelming world of mathematics. I specifically want to highlight Christoph *"Härtner"* Hartmann for our Zoom calls during lockdown and our GeoGuessr sessions which kept my head clear enough for research in these confusing times.
I also have to thank my brothers Henning and Nils, their wives Sabrina and Carina and my nieces and nephew Anna, Paul and Lisbet for their support during the last years. Even though they might not have even noticed it, being with them was immensely important for loosening my thoughts and helped me a lot in writing this dissertation.
Lastly, I have to mention my parents, Bernd and Cornelia, for their incredible support and assistance throughout my whole life. They enabled me to start studying at the University of Mannheim, always encouraged me to pursue my academic goals and have been a constant source of motivation. Additionally they have not only helped me to move to Denmark (during a global pandemic!), but also within Denmark, which I am still very thankful for. Furthermore, they endured all my bad moods during lockdown when research did not go the way I planned and were very understanding. I can not and do not want to imagine where and what I would be without them, which is why this dissertation is dedicated to them.

Niklas Dexheimer
Aarhus, May 2022

Summary

In the field of statistics for stochastic processes many works have been focussed on classical Itō diffusions, i.e. processes driven by a stochastic differential equation with Brownian noise. In particular a lot of results, e.g. on invariant density or drift estimation, have been discovered in the scalar setting, which cannot easily be transferred to a multivariate context. The goal of this dissertation is to obtain statistical results for more general and in particular multivariate processes beyond the standard continuous diffusion framework.

In paper A nonparametric adaptive invariant density estimation for general multivariate ergodic Markov processes through a kernel density estimator is investigated. The analysis is carried out under minimal assumptions on the behaviour of the underlying process, in particular on its transition densities, and it is shown that both classical diffusion processes and jump diffusions satisfy these assumptions. A particular result of this work is that the same rate of convergence of the estimator, measured in sup-norm loss, is achieved as for classical diffusions. Furthermore the estimation procedure is also extended to an adaptive estimator.

Paper B focusses on nonparametric invariant density and drift estimation for so-called stochastic damping Hamiltonian systems, also known as kinetic diffusions. The main difference of this class of processes compared to classical diffusions is the degeneracy of the diffusion coefficient, which has several implications on the analysis and behaviour of the process. Firstly it does not permit usage of classical tools in statistical inference for standard diffusions and secondly the process possesses a particular variance structure. This variance structure is then also reflected in the rate of convergence of the invariant density estimator, which is highly non-standard and specific to this class of processes. On the other hand, for drift estimation, which has also been carried out adaptively, the classical nonparametric rate of convergence is achieved. In both of these results the error is again measured in the sup-norm.

In paper C, estimation of the drift coefficient of a Lévy-driven Ornstein–Uhlenbeck process via Lasso and Slope estimators is investigated. In contrast to the abovementioned works, this subject falls into the field of parametric statistics. The results of this paper are threefold; firstly the analysis is generalized from continuous Ornstein–Uhlenbeck processes to Lévy-driven ones, secondly it is obtained that the tuning parameters can be chosen independently of the confidence level in the main result and lastly the minimax optimal rate of covergence is achieved up to numerical constants. The last two results so far had not been available, in particular not even for classical Gaussian Ornstein–Uhlenbeck processes.

Resumé

Inden for statistik for stokastiske processer har der hovedsageligt været fokuseret på klassiske Itō diffusioner i litteraturen, dvs. stokastiske differentialligninger med et Brownsk støj led. Mange resultater, f.eks. om invariant tæthed eller drift estimering, er udviklet i et endimensional set-up som ikke let kan overføres til flere dimensioner. Målet med denne afhandling er at opnå statistiske resultater for mere generelle og især multivariate processer der udvider det standard kontinuerte diffusions set-up.

I Artikel A undersøges en ikke-parametrisk adaptiv invariant tæthed estimering for generelle multivariate ergodiske Markov processer ved hjælp af en kerne tætheds estimator. Analysen er udført under minimale antagelser om adfærden af den underliggende proces, især på dens overgangstætheder, og det er vist, at både klassiske diffusion processer og diffusioner med spring opfylder disse antagelser. Dette arbejde viser specielt, at den samme konvergenshastighed for estimatoren, målt i sup-normen, opnås som for klassiske diffusioner. Endvidere er estimeringsproceduren også udvidet til en adaptiv estimator.

Artikel B fokuserer på ikke-parametrisk invariant tæthed og drift estimering for såkaldte stokastiske dæmpende Hamilton-systemer, også kendt som kinetiske diffusioner. Den største forskel på denne klasse af processer sammenlignet med klassiske diffusioner er at de har degenererede diffusionskoefficienter, som har flere implikationer på analysen og opførslen af processen. For det første tillader det ikke brugen af klassiske værktøjer til statistisk inferens for diffusioner, og for det andet har processen en særlig variansstruktur. Denne variansstruktur afspejles også i konvergenshastigheden af den invariante tæthed, som er meget ikke-standard for denne klasse af processer. På den anden side, for drift estimering, som også er blevet udført adaptivt, opnås den klassiske ikke-parametriske konvergenshastighed. I begge disse resultater måles fejlen igen i sup-normen.

I Artikel C undersøges estimering af drift koefficienten for en Lévy-drevet Ornstein–Uhlenbeck proces via Lasso- og Slope-estimatorer. I modsætning til de ovennævnte resultater falder dette emne ind under parametrisk statistik. Resultaterne fra denne artikel er tredelte; for det første generaliseres analysen af kontinuerte Ornstein–Uhlenbeck processer til Lévy-drevne, for det andet opnås det, at tuning-parametrene kan vælges uafhængigt af konfidensniveauet i hovedresultatet, og til sidst opnås den minimax optimale konvergenshastighed op til numeriske konstanter. De sidste to resultater har hidtil ikke været tilgængelige; ikke engang for klassiske Gaussiske Ornstein–Uhlenbeck processer.

Introduction

1

The purpose of this chapter is to introduce the main results and topics of this dissertation. As the title indicates, the common denominator of all contained works is statistical inference for stochastic processes, which somewhat lie in the vicinity of classical Itō diffusions, but still differ from this class up to some extent.
Consequently, in section 1.1 we introduce Itō diffusions and review some of their basic properties, providing the reader with enough information to separate the investigated processes from diffusions.
Continuing from that, Section 1.2 describes Michel Talagrand's generic chaining device; a useful mathematical tool, which will be heavily employed throughout Papers A, B and C.
The subsequent sections all follow the same structure for presenting the different works forming the main content of this dissertation. As these sections only serve as brief introductions, some technical details are left out for the sake of conciseness. Firstly, the goals and frameworks of the different articles are given, followed by the main results. After this the methodology, i.e. the different proof techniques and other related arguments, will be summarised in a nonrigorous way. Each section then concludes with a brief comparison of the different articles to related and current research.

1.1 Diffusion processes

There is not by any means a general definition of the term "*diffusion*" in mathematics (not even in probability theory or statistics), and in fact some of the definitions are surprisingly different. In this work a $\mathbb{R}^d$-valued stochastic process $\boldsymbol{X} = (X_t)$, will always be referred to as (Itō) diffusion, respectively diffusion process, if it satisfies a stochastic differential equation (SDE) of the form

$$\begin{aligned} \mathrm{d}X_t &= b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \quad t \geq 0 \\ X_0 &= x \in \mathbb{R}^d, \end{aligned} \tag{1.1}$$

where $b\colon \mathbb{R}^d \to \mathbb{R}^d$, $\sigma\colon \mathbb{R}^d \to \mathbb{R}^{d\times d}$, and $(W_t)_{t\geq 0}$ is a $\mathbb{R}^d$-valued Brownian motion, respectively Wiener process. The function b is often referred to as the *drift coefficient* of $\boldsymbol{X}$, σ as the *dispersion coefficient* and $\sigma\sigma^\top$ as *diffusion coefficient*.
These processes have been studied extensively ever since Kiyoshi Itō introduced the now famous Itō integral in [24] and by now various results regarding existence, stability or path behaviour of diffusions exist (see e.g. [19, 25, 26, 34, 36, 37]). Furthermore stochastic integration and SDEs have been generalised substantially from Wiener processes as integrators to general semimartingales (see e.g. [2, 35]).
As diffusions are not only of pure mathematical and theoretical interest but are also used for modelling purposes in various sciences, such as biology, physics, finance and medicine (see e.g. [2, 20, 26]), the statistical analysis of diffusions has been developing rapidly throughout the last years.

1.1.1 *Basic properties of diffusion processes*

In this section we list some of the basic properties of diffusion processes. If the drift and dispersion coefficient are Lipschitz continuous it is known that $\boldsymbol{X}$ is a *Markov* process and even possesses the *strong* Markov property. Its infinitesimal generator $\mathcal{A}$ with domain $\mathcal{D}(\mathcal{A})$ is then given by

$$\mathcal{A}f = \sum_{i=1}^{d} b_i \frac{\partial f}{\partial x_i} + \sum_{i,j=1}^{d} (\sigma\sigma^\top)_{i,j} \frac{\partial^2 f}{\partial x_i \partial x_j}, \quad f \in \mathcal{D}(\mathcal{A}),$$

and furthermore the space of $\mathcal{C}^2(\mathbb{R}^d)$ functions with compact support $\mathcal{C}_0^2(\mathbb{R}^d)$ is a subset of $\mathcal{D}(\mathcal{A})$ (see [34, Chapter 7]).
One important class of diffusion processes are *reversible* diffusions, for which a lot of advanced results are available due to their properties. An example of a *reversible* diffusion is the unique weak solution to the SDE

$$\mathrm{d}X_t = -\nabla V(X_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad t \geq 0, \quad X_0 = \widetilde{X},$$

where $\widetilde{X}$ is independent of the Wiener process $(W_t)_{t\geq 0}$, and $V(x) \in \mathcal{C}^2(\mathbb{R}^d)$ is a so-called *potential* function. Assuming that $\boldsymbol{X}$ is ergodic with invariant distribution μ_V it can be shown that the generator of $\boldsymbol{X}$ is self-adjoint (see [32]) and thus the Markov semigroup $(P_t)_{t\geq 0}$ is symmetric with respect to μ_V, respectively μ_V is reversible for $(P_t)_{t\geq 0}$, i.e.

$$\forall t \geq 0, f, g \in L^2(\mu_V): \int f P_t g \,\mathrm{d}\mu_V = \int g P_t f \,\mathrm{d}\mu_V,$$

hence justifying the name *reversible* diffusion. The symmetry of the Markov semigroup then enables the usage of the theory of functional inequalities for the infinitesimal generator of $\boldsymbol{X}$, such as Poincaré or Sobolev inequalities (a comprehensive account for this is given in [4]), which, if fulfilled, have several implications on $\boldsymbol{X}$, e.g. bounds on the transition densities or its ergodic behaviour. Another useful property of ergodic reversible diffusions is that the density of the invariant distribution (also called invariant density) is known explicitly and depends solely on the potential V. More specifically it holds

$$\mathrm{d}\mu_V = c_\mu \exp(-2V)\,\mathrm{d}\lambda,$$

where $c_\mu > 0$ is the normalizing constant (see e.g. [3, 15, 41]). Thus there is a one to one correspondence between potential and invariant density and also for different regularity conditions concerning the two.

1.1.2 *Examples of statistics for diffusion processes*

Statistics for diffusion processes have been investigated eagerly in the past decades and by now many results are available, especially for reversible and scalar diffusions. A thorough overview of statistics for scalar diffusions can be found in [28], however since many estimators in the one-dimensional setting are based on the concept of *local time*, which is not available in higher dimensions, treating multivariate diffusion processes requires different methods.
In the following we list some examples for research on statistics for diffusion processes which are relevant for the results of this dissertation. One of the first works on nonparametric invariant

density and drift estimation for ergodic reversible multivariate diffusions is given in [15], where a kernel density and Nadaraya–Watson type estimator are employed for estimating said functions based on a continouous record of observations. Surprisingly, it is shown that for $d \geq 3$ the rate of convergence of the invariant density estimator with suitably chosen bandwidth in pointwise L^2 risk is given by

$$T^{-\frac{\beta+1}{2\beta+d}},$$

where $\beta+1$ is the assumed Hölder regularity (for details see Paper A or Paper B) of the invariant density and $T > 0$ the observation time. This outperforms the classical nonparametric rate of convergence for density estimation based on $n \in \mathbb{N}$ independent and identically distributed random variables given as

$$n^{-\frac{\beta}{2\beta+d}},$$

where β is the Hölder regularity of the density of observations. It may be confusing to the reader, why the Hölder regularity is given in one case as $\beta+1$ and in the other case as β but this is justified by the fact that assuming β Hölder regularity for the drift in fact implies $\beta+1$ Hölder regularity for the invariant density of a reversible diffusion due to the one to one correspondence between drift and invariant density mentioned in the preceeding section.
The main reason for the accelerated rate of convergence is an improved variance bound for the kernel density estimator, which in itself follows from assuming a bound on the transition densities of the form

$$\forall t > 0, \|x-y\|^2 < t \colon p_t(x,y) \lesssim t^{-\frac{d}{2}} + t^{3d/2},$$

and a spectral gap inequality, i.e.

$$\exists \kappa > 0 \colon \mu((P_t f - \mu(f))^2) = \mathrm{Var}_\mu(P_t f) \leq \mathrm{e}^{-2\kappa t}\mu(f^2), \quad \forall f \in L^2(\mu), t > 0,$$

where μ is the invariant measure and we denote $\mu(f) = \int f \,\mathrm{d}\mu$. These assumptions are reasonable for *reversible* diffusions as they follow by certain assumptions on the drift coefficient through functional inequalities for the generator. In [41] the investigation of invariant density estimation for ergodic reversible multivariate diffusion processes through a kernel density estimator was extended by measuring the estimation error in the global sup-norm risk. It is shown that the rate of convergence improves similarly as in [15] compared to the classical nonparametric rate of convergence and in particular that the kernel density estimator is minimax optimal over a class of diffusions fulfilling a Poincaré and Nash inequality. The analysis is also carried out adaptively, thus not requiring knowledge of the Hölder regularity of the invariant density.
Two examples of recent results on parametric statistics for diffusion processes are given in [13, 21], where estimation of the drift coefficient $\mathbf{A}_0 \in \mathbb{R}^{d\times d}$ of an Ornstein–Uhlenbeck process

$$\mathrm{d}X_t = -\mathbf{A}_0 X_t \,\mathrm{d}t + \mathrm{d}W_t, \quad t \geq 0,$$

based on a continuous record of observations is investigated under the assumption that $\mathbf{A}_0$ is sparse. For this Lasso and Dantzig (in [13]), respectively Lasso and adaptive Lasso estimators (in [21]) are employed and it is shown that these satisfy bounds close to the minimax optimal rate of estimation

$$\sqrt{\frac{s\log(d^2/s)}{T}},$$

with high probability, where s denotes the sparsity of $\mathbf{A}_0$. Concerning the Lasso estimator, the analysis in both papers is comparable to the methods for the Lasso estimator in the setting of sparse linear regression, as e.g. in [7]. On the one hand it is based on the so-called restricted eigenvalue condition and on the other hand it requires a sufficiently tight deviation inequality for the stochastic error, which in the investigated setting is given as an Itō integral. A peculiar property shown for this model is that the restricted eigenvalue condition does not have to be assumed as for sparse linear regression but indeed follows directly as soon as $\boldsymbol{X}$ is ergodic.

1.2 GENERIC CHAINING

Generic chaining, introduced by Michel Talagrand, is a method for finding *uniform* bounds for a stochastic process $\boldsymbol{X} = (X_t)_{t\in T}$ over the index set T, both in probability and expectation. In particular T does not have to be equal to $[0,\infty)$ or another subset of $\mathbb{R}$ but can also be a space of functions, matrices or other objects. A comprehensive overview about this subject can be found in [42], and a more concise approach is contained in [44, Chapter 8].

1.2.1 *Main results*

One of the first results concerning uniform bounds for stochastic processes is known as *Dudley's inequality,* named after Richard Dudley (see [18]). However, before stating it, we need to introduce the notion of a *covering number* of a set. Let (T,d) be a pseudometric space. Then for each $\varepsilon > 0$ the ε covering number of T with respect to the pseudometric d, denoted as $\mathcal{N}(T,d,\varepsilon)$, is the minimal amount of open balls of radius $\varepsilon > 0$ with respect to d needed to cover the set T. Now Dudley's inequality states that there exists a universal constant $c > 0$, such that for a centered *Gaussian* process $(X_t)_{t\in T}$ with associated pseudometric

$$d_X(s,t) \coloneqq \sqrt{\mathbb{E}[|X_s - X_t|^2]}, \quad s,t \in T, \tag{1.2}$$

it holds

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le c\int_0^\infty \sqrt{\log \mathcal{N}(T,d_X,\varepsilon)}\,\mathrm{d}\varepsilon,$$

where the right-hand side is often referred to as *Dudley's entropy integral*. This result can be improved to hold for stochastic processes with subgaussian increments and mean zero (see Theorem 8.1.3 in [44]), i.e. for which there exist a constant $c > 0$ and a pseudometric d_X, such that

$$\|X_t - X_s\|_{\psi_2} \le c d_X(t,s), \quad \forall s,t \in T, \tag{1.3}$$

holds, where for $\alpha > 0$, $\|\cdot\|_{\psi_\alpha}$ is defined as follows for a random variable X

$$\|X\|_{\psi_\alpha} \coloneqq \inf\{t > 0 : \mathbb{E}[\exp((X/t)^\alpha)] \le 2\}.$$

Despite its usefulness, it can be shown that Dudley's inequality produces suboptimal bounds in some cases (see e.g. Exercise 8.1.12 in [44]). This is where Talagrand's *generic chaining* device and γ_α functional come into play which we will introduce now. A sequence $(T_n)_{n\in\mathbb{N}_0}$ of subsets of T is called *admissible* if

$$|T_0| = 1, \quad |T_n| \le 2^{2^n}, \quad \forall n \in \mathbb{N}.$$

Then for $\alpha > 0$ Talagrand's γ_α functional of the pseudometric space (T, d) is defined as

$$\gamma_\alpha(T, d) := \inf_{T_n} \sup_{t \in T} \sum_{n=0}^{\infty} 2^{n/\alpha} d(t, T_n),$$

where the infimum is taken over all admissible sequences and the distance from a point to a set is defined as usual by $d(t, A) := \inf_{a \in A} d(t, a)$. Talagrand's generic chaining bound (see e.g. Theorem 8.5.3 in [44]) for centered subgaussian processes then states for a stochastic process $(X_t)_{t \in T}$, satisfying (1.3)

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \leq c_1 c_2 \gamma_2(T, d_X),$$

where $c_1 > 0$ is a universal constant and c_2 is the constant from (1.3). As it can be shown, that Talagrand's γ_2 functional is (up to a universal constant) always smaller than Dudley's entropy integral (see Exercise 8.5.7 in [44]), the generic chaining bound yields a tighter bound than Dudley's inequality. In fact, by Talagrand's majorizing measures theorem (see e.g. [44, Section 8.6]) there exist constants $c_1, c_2 > 0$, such that for a centered Gaussian process $(X_t)_{t \in T}$

$$c_1 \gamma_2(T, d_X) \leq \mathbb{E}\left[\sup_{t \in T} X_t\right] \leq c_2 \gamma_2(T, d_X),$$

where d_X is defined as in (1.2). Hence the generic chaining bound is in fact optimal for Gaussian processes. This is not only a strong result in itself, but can also be helpful in bounding the γ_2 functional, since finding suitable admissible sequences can be a challenging task.
As the notation might have already indicated, the generic chaining bound can also be generalised to processes fulfilling condition (1.3) with the ψ_2 norm replaced by the ψ_α norm for some $\alpha > 0$. In this case the expected supremum can be bounded by the corresponding γ_α functional. Furthermore, it can also be extended to processes with a mixed tail behaviour, as for example implied by Bernstein's inequality, and to bounds for the p-th moment of the expected supremum (see e.g. [17]).

1.3 Paper A

1.3.1 *Goal*

The primary goal of Paper A, written in collaboration with Claudia Strauch and Lukas Trottner, was to find a general framework for continuous-time Markov processes which leads to similar improvements in the rate of convergence of nonparametric invariant density estimation via kernel density estimators as discovered for reversible diffusions in [15] and [41].

1.3.2 *Framework*

Throughout the whole paper we assume that the investigated stochastic process $\boldsymbol{X} = (X_t)_{t \geq 0}$ is a Borel right Markov process (see [39]) with absolutely continuous marginal laws and a unique invariant distribution μ admitting a density ρ which is also assumed to be the starting distribution of $\boldsymbol{X}$. As the ultimate aim of our investigation is to estimate the invariant density ρ

based on a continuous record of observations of $\boldsymbol{X}$ up to time T, we introduce the nonparametric kernel density estimator

$$\widehat{\rho}_{h,T}(x) = \frac{1}{T}\int_0^T h^{-d}K((x-X_s)/h)\,\mathrm{d}s = \frac{1}{T}\int_0^T K_h(x-X_s)\,\mathrm{d}s, \quad x \in \mathbb{R},$$

where $K\colon \mathbb{R}^d \to \mathbb{R}$ is a smooth, Lipschitz continuous kernel function, such that $\mathrm{supp}(K) \subset [-1/2, 1/2]^d$, and $h = h(T)$ is the so-called bandwidth. The latter being a tuning parameter that usually depends on T which by a slight abuse of notation will often be suppressed.

1.3.3 *Main results*

The theoretical main result of Paper A consists of uniform moment bounds over countable classes of bounded functions for path integrals under the assumption that the underlying process $\boldsymbol{X}$ fulfills the β-mixing property with rate function Ξ, i.e. for all $t \geq 0$

$$\int \|P_t(x,\cdot) - \mu\|_{\mathrm{TV}}\mu(\mathrm{d}x) \leq \Xi(t),$$

where $(P_t)_{t\geq 0}$ denotes the Markov semigroup of $\boldsymbol{X}$ and $\|\cdot\|_{\mathrm{TV}}$ the total variation norm. The bounds are stated in terms of entropy integrals of the function class with respect to the sup-norm distance and a variance distance induced by $\boldsymbol{X}$ of the form

$$d_{\mathbb{G},t}(f,g) = \mathrm{Var}\left(\frac{1}{\sqrt{t}}\int_0^t f(X_s) - g(X_s)\,\mathrm{d}s\right),$$

thus depending not only on the function class but also on the underlying process $\boldsymbol{X}$.
Applying these uniform moment bounds together with the classical decomposition of the sup-norm risk on a compact set $D \subset \mathbb{R}^d$ of the kernel density estimator $\widehat{\rho}_{h,T}$ with respect to the invariant density ρ into a bias part and a stochastic error, i.e.

$$\mathbb{E}\left[\sup_{x\in D}|\widehat{\rho}_{h,T}(x) - \rho(x)|^p\right]^{1/p} \leq \mathbb{E}\left[\sup_{x\in D}|\widehat{\rho}_{h,T}(x) - \mu(\widehat{\rho}_{h,T})|^p\right]^{1/p} + \sup_{x\in D}|\mu(\widehat{\rho}_{h,T}) - \rho(x)|, \tag{1.4}$$

where we denote $\mu(f) \coloneqq \int f\,\mathrm{d}\mu$ for $f \in L^1(\mu)$, it then suffices to bound the entropy integrals, since the bias part can be bounded under classical Hölder regularity assumptions on ρ.
Introducing similar additional assumptions on the transition densities of $\boldsymbol{X}$ and the rate function Ξ as in the reversible diffusion framework leads to suitably tight variance bounds for the kernel density estimator in dimension $d \geq 2$; ultimately implying the same rate of convergence for the kernel density estimator with suitably chosen bandwidth h as in the case of reversible diffusions in [41] under Hölder-regularity assumptions on ρ.
For the scalar case we base our variance analysis on an assumption similar to the Castellana–Leadbetter condition [9] and show that under this assumption the same rate of convergence as for reversible diffusions is also obtainable by choosing the bandwidth correctly. Furthermore, we show that this assumption holds as soon as the process is V-exponentially ergodic with locally bounded V which also implies the desired mixing property.
As the optimal bandwidth choice for $d \geq 3$ depends on the in general unknown regularity of the estimated invariant density, an adaptive estimation scheme is introduced and it is shown that

this achieves the same rate of convergence up to a negligible iterated logarithm factor. For this it is crucial that the uniform moment bounds hold for all p-th moments of the sup-norm risk.
To conclude the paper, we show that the imposed assumptions are satisfied for a large class of diffusion process with and without jumps, thus making it a suitable framework for invariant density estimation of diffusion type processes.

1.3.4 *Methodology*

For proving the uniform moment bounds we apply Talagrand's generic chaining device in the more refined form stated in [17]. To enable its application, the β-mixing assumption is crucial, since it allows us to argue as in [45] and split the path of $\boldsymbol{X}$ up to time T up into $2n_T$ parts of length m_T, and construct two independent collections of random variables $(\widetilde{X}_k^1)_{k\in\{1,\dots,n_T\}}$, $(\widetilde{X}_k^2)_{k\in\{1,\dots,n_T\}}$, such that they are identically distributed as the corresponding piece of the path of $\boldsymbol{X}$, and equal to the corresponding piece with high probability, depending on the rate function Ξ and the block length m_T. Thus we can apply the classical Bernstein inequality for sums of independent and bounded random variables for verifying that the desired concentration property for $\boldsymbol{X}$ holds which enables the use of the generic chaining device and finally leads to bounds stated in terms of entropy integrals together with some additional error terms due to the mixing procedure.
As written above, we are then required to find suitably tight variance bounds, which can be achieved by arguing similarly as in [15], once we assume, that the process is *exponentially* β-mixing, respectively that the Castellana–Leadbetter type condition is in place and that similar transition density bounds as in the reversible diffusion case are satisfied. Combining this with (1.4) leads to the final result on the rate of estimation of the kernel density estimator.
For our adaptive estimation procedure we use a Lepski-type selection criterion as e.g. in [22]. For this it is crucial that the uniform moment bounds hold for any p-th moment, since Markov's inequality then entails an exponential concentration inequality for kernel density estimators with different bandwidths. As constants are hard to control or obtain in the general β-mixing framework we are investigating, we have to introduce an iterated logarithm which will always be larger than the relevant constants for large enough values of T.
For proving that certain jump diffusions actually fall into the proposed framework, we make use of the results in [31],[30] and [12]. To be more precise, [31] entails the exponential β-mixing property for Lévy-driven Ornstein–Uhlenbeck processes under assumptions on the Lévy measure of the background driving Lévy process, and since the marginal distributions of $\boldsymbol{X}$ are known to be infinitely divisible in this case with explicit Lévy triplet, it is also possible to verify the assumption on the transition density by using the inverse Fourier transform. In the other case of more general diffusion and jump coefficients, [12] proves the desired heat kernel bound and [30] entails the exponential β-mixing property through verification of a Lyapunov drift criterion on the extended generator of $\boldsymbol{X}$, which itself implies V-exponential ergodicity with locally bounded V and thus also the Castellana–Leadbetter type condition.

1.3.5 *Comparison to other research*

Comparing the research of this paper to other works is complicated, as we take a different approach than most other works in the field of statistics for stochastic processes. While the more common approach is to introduce an estimator for a certain quantity of interest of a fixed class of stochastic processes and show its rate optimality or even minimax optimality over this class,

we try to find the most general assumptions such that the same convergence rate as for invariant density estimation for reversible diffusions is obtained.
Because of this Paper A is of course comparable to [15] and [41], which were also mentioned in Section 1.1.2. However as the reversibility assumption is in both works central for the main results, but hard to verify in practice, our approach yields a far more general and easier to apply framework.
Another relevant paper to our research is [1], where invariant density estimation for a class of jump diffusions, similar to one of the examples given in Paper A, is investigated under anisotropic regularity assumptions on the invariant density. However, contrary to Paper A this work solely focusses on the less involved L^2 risk of the estimator and also proves a slower rate of convergence in the scalar case than the one obtained in Paper A. Furthermore said paper also introduces an adaptive estimation scheme which contrary to the adaptive estimation scheme in Paper A achieves the same rate of convergence as the optimal choice of bandwidths in the first results of this paper. However, this is slightly misleading as the scheme relies on a certain constant being large enough *without* stating a specific value for it. We encountered the same problem in our adaptive procedure but overcame it by introducing the iterated logarithm, which for large enough values of T will always exceed said constant.

1.4 PAPER B

1.4.1 *Goals and framework*

In Paper B, written jointly with Claudia Strauch, we investigate nonparametric invariant density and drift estimation, based on continuous observations, for a so-called stochastic damping Hamiltonian system $(Z_t)_{t\geq 0} = \mathbf{Z} = (\boldsymbol{X}, \mathbf{Y}) = ((X_t)_{t\geq 0}, (Y_t)_{t\geq 0})$ satisfying the following SDE

$$\begin{aligned} \mathrm{d}X_t &= Y_t \,\mathrm{d}t \\ \mathrm{d}Y_t &= b(X_t, Y_t)\,\mathrm{d}t + \sigma(X_t, Y_t)\,\mathrm{d}W_t, \end{aligned} \tag{1.5}$$

where $\sigma\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$,

$$b\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d, \quad (x, y) \mapsto -(c(x, y)y + \nabla V(x)) =: b(x, y),$$

with $c\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$, $V\colon \mathbb{R}^d \to \mathbb{R}$, and $(W_t)_{t\geq 0}$ being a $\mathbb{R}^d$ valued Wiener process. This can of course be rewritten as

$$\mathrm{d}Z_t = \widetilde{b}(Z_t)\,\mathrm{d}t + \widetilde{\sigma}(Z_t)\,\mathrm{d}\widetilde{W}_t,$$

where

$$\widetilde{b}(Z_t) = (Y_t, b(X_t, Y_t))^\top, \quad \widetilde{\sigma}(Z_t) = \begin{pmatrix} \mathbf{0}_{d\times d} & \mathbf{0}_{d\times d} \\ \mathbf{0}_{d\times d} & \sigma(X_t, Y_t) \end{pmatrix},$$

with $\mathbf{0}_{d\times d}$ denoting the $d \times d$ dimensional zero matrix and $(\widetilde{W}_t)_{t\geq 0}$ being a $\mathbb{R}^{2d}$ valued Wiener process. In the latter form it gets more apparent why stochastic damping Hamiltonian systems are examples for degenerate diffusions, since we see that $\widetilde{\sigma}$ is not invertible. As this also leads to the generator of $\mathbf{Z}$ not being elliptic anymore as for classical diffusion processes, but *hypoelliptic*, these processes are in particular examples for *hypoelliptic diffusions*. Stochastic damping Hamiltonian systems are usually thought of as a model for a particle moving through space. In this case $\boldsymbol{X}$ models the position of the particle and $\mathbf{Y}$ its velocity, and thus c can be interpreted as friction

and V as the physical potential. Because of this interpretation, stochastic damping Hamiltonian systems are also often called *kinetic diffusions*.
Concerning the mathematical framework, we assume throughout the whole paper that $\mathbf{Z}$ is the unique, non-explosive weak solution of (1.5), such that the associated semigroup is strong Feller. Furthermore, we assume as in Paper A that $\mathbf{Z}$ is ergodic with invariant measure μ admitting a density ρ, and that the marginal laws are absolutely continuous. Additionally, we assume a peculiar heat kernel bound, the exponential β-mixing property and stationarity of $\mathbf{Z}$.
For estimating the invariant density ρ, we introduce a similar kernel density estimator as in Paper A. However, due to $\mathbf{Z}$'s special structure, it is natural to replace the isotropic Hölder regularity conditions by anisotropic ones. Thus, the kernel density estimator in Paper B is given as

$$\widehat{\rho}_{h_1,h_2,T}(x,y) = \frac{1}{T(h_1h_2)^d}\int_0^T K((x-X_s)/h_1,(y-Y_s)/h_2)\,\mathrm{d}s, \tag{1.6}$$

where $K\colon \mathbb{R}^{2d} \to \mathbb{R}$ is a kernel function, satisfying the same assumptions as in Paper A, and h_1, h_2 are different bandwidths. For estimating the drift coefficient b, we employ a Nadaraya–Watson type estimator, for which we first introduce

$$\bar{b}_{j,h_1,h_2,T}(x,y) = \frac{1}{T(h_1h_2)^d}\int_0^T K((x-X_s)/h_1,(y-Y_s)/h_2)\,\mathrm{d}Y_s^j,$$

where $j \in \{1,\dots,d\}$, and K is a similar kernel function as above. As $\bar{b}_{j,\cdot}$ can be thought of as an estimator for $b^j\rho$, natural estimators for b^j are now given as

$$\widehat{b}_{j,h_1,h_2,h_1^{(\rho)},h_2^{(\rho)},T,r_T}(x,y) = \frac{\bar{b}_{j,h_1,h_2,T}(x,y)}{|\widehat{\rho}_{h_1^{(\rho)},h_2^{(\rho)},T}(x,y)| + r_T}, \tag{1.7}$$

$$\widehat{b}_{j,h_1,h_2,h_1^{(\rho)},h_2^{(\rho)},T}(x,y) = \frac{\bar{b}_{j,h_1,h_2,T}(x,y)}{\widehat{\rho}_{h_1^{(\rho)},h_2^{(\rho)},T}(x,y) \vee \rho_\star}, \tag{1.8}$$

where $\rho_\star$ is an a priori lower bound for ρ on the investigated domain, and $r_T > 0$ is chosen such that $r_T \in o(1)$. These definitions in particular ensure the well-definedness of the two estimators, as the denominator is always strictly positive.

1.4.2 *Main results*

The main results of Paper B concern the rate of convergence of the sup-norm risk on a compact domain $D \subset \mathbb{R}^{2d}$ of the invariant density estimator (1.6) and the two drift estimators (1.7), used in a non-adaptive setting, and (1.8), which will be used for an adaptive estimation procedure of the drift.
For (1.6) we obtain highly nonclassical results, which stem mostly from the peculiar behaviour of the transition densities and therefore can be seen as a direct consequence of $\mathbf{Z}$'s degeneracy. In particular, it is shown that the rate of convergence does not only depend on the assumed Hölder regularity of the invariant density and bandwidth choices but also on the domain D.
For both the non-adaptive and the adaptive drift estimators we arrive at the classical nonparametric rate of convergence, thus achieving the same result as for classical, non-degenerate diffusion processes.

1.4.3 *Methodology*

The first step in Paper B consists of showing that there are actually processes fulfilling the imposed assumptions. For this we rely on the results of [11, 27, 46] and show that our assumptions are satisfied as soon as V, c and σ fulfill certain regularity assumptions.
For bounding the rate of convergence of the invariant density estimator we apply the uniform moment bounds from Paper A, which is possible since $\mathbf{Z}$ is exponentially β-mixing. Thus it then suffices to find adequate bounds on the variance of the kernel density estimator. For this we argue similarly to Paper A and use an approach as in [15]. However as the transition density bounds are more complicated and due to the anisotropic setting the proof gets much more involved, similar to [16].
For drift estimation we prove uniform moment bounds over a countable class of bounded functions $\mathcal{G}$ for *stochastic* integrals of the form

$$\frac{1}{\sqrt{T}} \int_0^T g(X_s, Y_s) \, \mathrm{d}Y_s^j, \quad g \in \mathcal{G}, j \in \{1, \dots, d\}.$$

By linearity we can decompose every stochastic integral of this form into a path integral with deterministic integrator, which can be bounded again with the results of Paper A, and an integral with respect to a Wiener process. Thus it suffices to find uniform moment bounds for the latter, which is a continuous martingale. Therefore we can apply Bernstein's inequality for continuous martingales together with another result from Paper A, which allows us to find a concentration inequality for the quadratic variation of said martingale, and ultimately enables us to employ the refined version of Talagrand's generic chaining device in [17] again.
For our adaptive drift estimation approach we rely on a Goldenshluger–Lepski type procedure [23] similar to [29]. For the proofs it is again crucial that the generic chaining approach allows us to find uniform bound for all p-th moments.

1.4.4 *Comparison to other research*

Nonparametric invariant density and drift estimation for stochastic damping Hamiltonian systems were investigated by Cattiaux *et al.* in [10, 11]. As both papers work with discrete observations a lot of emphasis is put on the case, where $\mathbf{Z}$ is only partially observed, i.e. only an observation of the position process $\boldsymbol{X}$ is available. However, as this problem is easily solvable in the continuous observations framework by differentiating, we do not investigate this. The main results of [10, 11] are central limit theorems for the estimators, and in particular there are no results for the rate of convergence.
In [14] adaptive nonparametric invariant density estimation for stochastic damping Hamiltonian systems is investigated in a similar setting as in [11], but only for the scalar case. In this paper, bounds on the rate of convergence of the estimator's L^2 risk are derived which are faster than the classical nonparametric rate of convergence, but do not exhibit the peculiar variance structure of $\mathbf{Z}$ as our results do. Again, a lot of emphasis is put on the problem of partial observations, which is negligible in our setup.
The most comparable work on nonparametric statistics for stochastic damping Hamiltonian systems is [16], where invariant density estimation is investigated based on a continuous record of observations of $\mathbf{Z}$, which is assumed to be bidimensional as in [14]. Bounds for the rate of convergence of the estimator in the pointwise L^2 risk are derived, which are of the same form

as our results on invariant density estimation. However in Paper B we also achieve results for higher dimensions and for the more involved sup-norm risk. Furthermore a minimax lower bound is shown, which up to a logarithmic factor resembles the obtained results, thus indicating rate optimality of the kernel density estimator in this setting.

1.5 Paper C

1.5.1 *Goals and framework*

Paper C, written jointly with Claudia Strauch, pursues three main goals. Firstly, we want to generalise the results obtained for Lasso estimators for the drift parameter of classical Ornstein–Uhlenbeck processes under sparsity constraints in [13, 21] to Lévy-driven Ornstein–Uhlenbeck processes. Secondly, we want to improve the upper bounds in these papers to resemble the minimax optimal rate of convergence for sparse regression, corresponding in the given setting to

$$\sqrt{\frac{s\log(d^2/s)}{T}},$$

with s being the sparsity of the drift parameter, and lastly our analysis should only rely on tuning parameters chosen independently of the confidence level.
More specifically, the investigated setting is as follows. We want to estimate the true drift parameter $\mathbf{A}_0 \in \mathbb{R}^{d\times d}$ of a $\mathbb{R}^d$ valued Lévy-driven Ornstein–Uhlenbeck process $\boldsymbol{X} = (X_t)_{t\geq 0}$, satisfying the following SDE

$$\mathrm{d}X_t = -\mathbf{A}_0\,\mathrm{d}t + \mathrm{d}Z_t, \quad t \geq 0, \tag{1.9}$$

where $\mathbf{Z} = (Z_t)_{t\geq 0}$ is a Lévy-process with characteristic triplet $(b, \mathbf{C} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top, \nu)$, with $b \in \mathbb{R}^d$, $\mathbf{C} \in \mathbb{R}^{d\times d}$ and ν being a Lévy measure on $\mathbb{R}^d$. For this we assume that the real parts of all eigenvalues of $\mathbf{A}_0$ are positive, ν admits a second moment and that a continuous record of observations of $\boldsymbol{X}$ up to time $T > 0$ is available. The assumptions on $\mathbf{A}_0$ and ν in particular imply that $\boldsymbol{X}$ is ergodic with invariant distribution μ (see [31, 38]), which we assume to be the starting distribution, i.e. we assume that $\boldsymbol{X}$ is stationary.
For estimating $\mathbf{A}_0$ we employ the classical Lasso estimator and the Slope estimator introduced in [8], as those are known to achieve the minimax optimal rate of convergence in the setting of sparse linear regression (see [7]). As we do not assume $\boldsymbol{\Sigma}$ to be equal to the identity times some constant as in [7, 13, 21], we have to adjust the definitions of Lasso and Slope estimators and the sparsity assumptions to our setting. Let the negative log-likelihood function be given as

$$\mathcal{L}_T(\mathbf{A}) = -\frac{1}{T}\log\left(\frac{\mathrm{d}\mathbb{P}_T^{\mathbf{A}}}{\mathrm{d}\mathbb{P}_T^{0}}\right), \quad \mathbf{A} \in \mathbb{R}^{d\times d},$$

where $\mathbb{P}_T^{\mathbf{A}}$ is the law of $\boldsymbol{X}$ satisfying (1.9) with drift parameter $\mathbf{A}$ up to time T. Then the Lasso estimator with tuning parameter $\lambda_{\mathrm{L}} > 0$ is defined as

$$\widehat{\mathbf{A}}_{\mathrm{lasso}} \in \operatorname{argmin}_{\mathbf{A}\in\mathbb{R}^{d\times d}}\Big(\mathcal{L}_T(\mathbf{A}) + \lambda_{\mathrm{L}}\|\boldsymbol{\Sigma}^{-1}\mathbf{A}\|_1\Big),$$

and similarly the Slope estimator with tuning parameter $\lambda_{\mathrm{S}} > 0$ is given as

$$\widehat{\mathbf{A}}_{\mathrm{slope}} \in \operatorname{argmin}_{\mathbf{A}\in\mathbb{R}^{d\times d}}\Big(\mathcal{L}_T(\mathbf{A}) + \lambda_{\mathrm{S}}\|\boldsymbol{\Sigma}^{-1}\mathbf{A}\|_*\Big),$$

where

$$\|\mathbf{A}\|_* := \sum_{i=1}^{d^2} \mathrm{vec}(\mathbf{A})_i^{\#} \sqrt{\log(2d^2/i)},$$

with $\mathrm{vec}(\mathbf{A})^{\#}$ denoting a nondecreasing rearrangement of the absolute values of the entries of $\mathbf{A} \in \mathbb{R}^{d\times d}$. Furthermore for our analysis we require $\mathbf{\Sigma}^{-1}\mathbf{A}_0$ to be sparse instead of $\mathbf{A}_0$ as in [13, 21], which also can be justified by the same argument as for our adjusted estimators.
Another crucial assumption for our analysis will concern the existence of a concentration inequality for

$$\frac{1}{T}\int_0^T (u^\top X_s)^2 \,\mathrm{d}s - \int (u^\top x)^2 \mu(\mathrm{d}x), \quad T \geq 0, u \in \mathbb{R}^d \colon \|u\| \leq 1. \tag{1.10}$$

1.5.2 *Main results*

The main results of Paper C consist of achieving all three goals introduced above. More precisely, we show that Lasso and Slope estimators with adequately chosen tuning parameters, independent of the confidence level in the statements in probability, in fact achieve the minimax optimal rate of convergence, both in probability and conditional expectation with respect to a certain event. The choice of tuning parameter for the Lasso estimator then depends on the a priori unknown sparsity of $\mathbf{\Sigma}^{-1}\mathbf{A}_0$, whereas the Slope estimator does not require such knowledge.
To underline the rate optimality of the proposed estimators, we also show a minimax lower bound for general loss functions in the standard Ornstein–Uhlenbeck setting over the set of sparse drift coefficients fulfilling our assumptions on their eigenvalues. This had also not been available before.
Furthermore, we prove that the assumed concentration inequality for (1.10) holds as soon as ν admits a fourth moment.

1.5.3 *Methodology*

For our results on the negative log-likelihood function, in particular its well-definedness, we apply the results of [40] where maximum likelihood estimation for more general diffusion processes with jumps is investigated. For this we have to ensure that the invariant distribution μ admits a second moment, which follows by results of [31, 38] as soon as ν admits a second moment.
As Lasso and Slope estimators are both defined as minimisers of a convex function, we can argue similarly to [7] for obtaining a basic inequality for their errors by classical arguments of convex analysis. This inequality is the starting point of the proofs of Paper C's main results.
As in [13, 21] we then require a restricted eigenvalue type condition to hold with high probability and a suitably tight deviation inequality for the stochastic error term occuring in the mentioned basic inequality. In [21] it was shown through results of [6] that the restricted eigenvalue condition holds true with high probability as soon as a concentration inequality for (1.10) is in place. Arguing similarly we find that this is also true in the Lévy-driven case. Additionally in [13], which investigates classical Ornstein–Uhlenbeck processes, it was shown that the concentration inequality holds as soon as the drift parameter fulfills the imposed assumption on its eigenvalues. As the proof of this result relies on Malliavin calculus and the results of [33], it is not possible to follow the same approach for Lévy-driven Ornstein–Uhlenbeck processes. Nevertheless,

we are able to prove that the concentration inequality for (1.10) holds as soon as ν admits a fourth moment in the Lévy-driven case, by using classical results from martingale theory and an application of Fubini's theorem for stochastic integrals similarly to [5].
Proving the desired deviation inequality for the stochastic error term however is a much more challenging task, as we require an inequality similar to the one obtained in [7], where sparse linear regression is investigated. Since the derivation of said inequality strongly relies on the Gaussianity of the stochastic error term, and the error term in our case is given by an Itō integral, a generalization is not straighforward.
We solve this problem by finding an event, which implies both the restricted eigenvalue condition to hold and the stochastic error term to be subgaussian and showing that this event holds with high probability. Then on this event we can apply Talagrand's generic chaining device and obtain a deviation inequality with respect to Talagrand's γ_2 functional of a certain set. Bounding the γ_2 functional of this set would again be difficult, however by Talagrand's majorizing measures theorem, we can bound it by the expected supremum of a Gaussian process on this set, which is possible through the results of [7] and concludes the proof of the error bounds in probability.
Now as our tuning parameters do not depend on the confidence level of the proven error bounds, we can bound the conditional expectation of the error with respect to the event implying the restricted eigenvalue condition and subgaussianity of the stochastic error term, by bounding the tail integral as in [7].
For the proof of the minimax lower bound we rely on the general scheme for lower bounds in [43], which requires a suitable class of hypotheses. For this we employ a mixture of the techniques applied in [7, 21], and show that these hypotheses are sparse and also only have eigenvalues with positive real part.

1.5.4 *Comparison to other research*

The most natural candidates for a comparison to Paper C are [21] and [13], which were introduced in section 1.1.2. As explained there, both mentioned papers investigate Lasso estimators for the drift of a *classical* Ornstein–Uhlenbeck process, corresponding to the background driving Lévy process $\mathbf{Z}$ in our case having Lévy triplet $(0, \mathbb{I}_{d\times d}, 0)$, i.e. $\mathbf{Z}$ is a standard $\mathbb{R}^d$ valued Wiener process. Comparing the main results of Paper C to the corresponding bounds in [13, 21], we see that our work improves the results of said papers not only in generality but also for the classical Ornstein–Uhlenbeck case by achieving the minimax optimal rate of convergence. For the Lasso estimator there exists the downside of having to know the sparsity of $\mathbf{\Sigma}^{-1}\mathbf{A}_0$ beforehand for the specification of the optimal tuning parameter, however we also introduce the Slope estimator which does not require this knowledge. Furthermore, we also show bounds in conditional expectation, which was not possible with previous results since the tuning parameter depended on the confidence level.
Nevertheless the concentration inequality we obtain for (1.10) is weaker than the one proven in [13], which leads to the time until the derived probability estimates in Paper C hold being larger than in said paper. As explained above the reason for this lies in our more general setting, which makes it hardly possible to apply Malliavin calculus as in [13] or Sobolev inequalites as in [21]. However, as our results only rely on a general concentration inequality, the results of [13] are still applicable for the classical Ornstein–Uhlenbeck setting.
Another part of Paper C we can compare to [21] is the derived minimax lower bound. As [21]

works with assumptions of *row*-sparsity, the proven lower bound also concerns this class. However, the supremum in the statement of the minimax lower bound is taken over *all* row-sparse matrices, hence even the ones which imply the underlying process to not be ergodic. We improve this aspect by showing that the constructed hypotheses all imply $\boldsymbol{X}$ to be ergodic, making the minimax lower bound more meaningful.
Lastly we compare Paper C to the results of [7] where it was shown that Lasso and Slope estimators for sparse linear regression achieve the minimax optimal rate of convergence both in probability and *expectation*. The last result contrasts our findings, which only concern the conditional expectation. However this can be justified by the fact, that the restricted eigenvalue condition in [7] is assumed to hold true, whereas we show that it does in fact hold true with high probability if T is large enough.

References

[1] C. Amorino and A. Gloter. "Invariant density adaptive estimation for ergodic jump-diffusion processes over anisotropic classes". In: *J. Statist. Plann. Inference* 213 (2021), pp. 106–129.

[2] D. Applebaum. *Lévy Processes and Stochastic Calculus*. 2nd ed. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2009.

[3] D. Bakry, P. Cattiaux, and A. Guillin. "Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré". In: *J. Funct. Anal.* 254.3 (2008), pp. 727–759.

[4] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators*. Vol. 348. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Cham, 2014, pp. xx+552.

[5] O. E. Barndorff-Nielsen. "Processes of normal inverse Gaussian type". In: *Finance Stoch.* 2.1 (1997), pp. 41–68.

[6] S. Basu and G. Michailidis. "Regularized estimation in sparse high-dimensional time series models". In: *Ann. Statist.* 43.4 (2015), pp. 1535 –1567.

[7] P. C. Bellec, G. Lecué, and A. B. Tsybakov. "Slope meets Lasso: improved oracle bounds and optimality". In: *Ann. Statist.* 46.6B (2018), pp. 3603–3642.

[8] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. "SLOPE—adaptive variable selection via convex optimization". In: *Ann. Appl. Stat.* 9.3 (2015), pp. 1103–1140.

[9] J. V. Castellana and M. R. Leadbetter. "On smoothed probability density estimation for stationary processes". In: *Stochastic Process. Appl.* 21.2 (1986), pp. 179–193.

[10] P. Cattiaux, J. León, and C. Prieur. "Estimation for Stochastic Damping Hamiltonian Systems under Partial Observation. II. Drift term". In: *ALEA Lat. Am. J. Probab. Math. Stat.* 11 (July 2014).

[11] P. Cattiaux, J. R. León, and C. Prieur. "Estimation for stochastic damping Hamiltonian systems under partial observation. I. Invariant density". In: *Stochastic Process. Appl.* 124.3 (2014), pp. 1236–1260.

[12] Z.-Q. Chen, E. Hu, L. Xie, and X. Zhang. "Heat kernels for non-symmetric diffusion operators with jumps". In: *Journal of Differential Equations* 263.10 (2017), pp. 6576 –6634.

[13] G. Ciołek, D. Marushkevych, and M. Podolskij. "On Dantzig and Lasso estimators of the drift in a high dimensional Ornstein–Uhlenbeck model". In: *Electron. J. Stat.* 14.2 (2020), pp. 4395–4420.

[14] F. Comte, C. Prieur, and A. Samson. "Adaptive estimation for stochastic damping Hamiltonian systems under partial observation". In: *Stochastic Process. Appl.* 127.11 (2017), pp. 3689 –3718.

[15] A. Dalalyan and M. Reiß. "Asymptotic statistical equivalence for scalar ergodic diffusions". In: *Probab. Theory Relat. Fields* 134.2 (2006), pp. 248–282.

[16] S. Delattre, A. Gloter, and N. Yoshida. *Rate of Estimation for the Stationary Distribution of Stochastic Damping Hamiltonian Systems with Continuous Observations*. Preprint. Jan. 28, 2020. arXiv: 2001.10423 [math.ST].

[17] S. Dirksen. "Tail bounds via generic chaining". In: *Electron. J. Probab.* 20 (2015), no. 53, 29.

[18] R. Dudley. "The sizes of compact subsets of Hilbert space and continuity of Gaussian processes". In: *J. Funct. Anal.* 1.3 (1967), pp. 290–330.

[19] R. Durrett. *Stochastic calculus*. Probability and Stochastics Series. A practical introduction. CRC Press, Boca Raton, FL, 1996, pp. x+341.

[20] C. Fuchs. *Inference for diffusion processes – With applications in life sciences*. Springer, Heidelberg, 2013, pp. xx+430.

[21] S. Gaïffas and G. Matulewicz. "Sparse inference of the drift of a high-dimensional Ornstein–Uhlenbeck process". In: *J. Multivariate Anal.* 169 (2019), pp. 1–20.

[22] E. Giné and R. Nickl. "An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation". In: *Probab. Theory Relat. Fields* 143.3-4 (2009), pp. 569–596.

[23] A. Goldenshluger and O. Lepski. "Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality". In: *Ann. Statist.* 39.3 (2011), pp. 1608–1632.

[24] K. Itō. "Stochastic integral". In: *Proc. Imp. Acad. Tokyo* 20 (1944), pp. 519–524.

[25] K. Itō and H. P. McKean Jr. *Diffusion processes and their sample paths*. Die Grundlehren der mathematischen Wissenschaften, Band 125. Second printing, corrected. Springer-Verlag, Berlin-New York, 1974, pp. xv+321.

[26] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus*. Second. Vol. 113. Graduate Texts in Mathematics. Springer-Verlag, New York, 1991, pp. xxiv+470.

[27] V. Konakov, S. Menozzi, and S. Molchanov. "Explicit parametrix and local limit theorems for some degenerate diffusion processes". In: *Ann. Inst. H. Poincaré Probab. Statist.* 46.4 (Nov. 2010), pp. 908–923.

[28] Y. A. Kutoyants. *Statistical Inference for Ergodic Diffusion Processes*. Springer Series in Statistics. New York: Springer, 2004.

[29] O. Lepski. "Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure". In: *Ann. Statist.* 41.2 (2013), pp. 1005–1034.

[30] H. Masuda. "Ergodicity and exponential β-mixing bounds for multidimensional diffusions with jumps". In: *Stochastic Process. Appl.* 117.1 (2007), pp. 35 –56.

[31] H. Masuda. "On multidimensional Ornstein-Uhlenbeck processes driven by a general Lévy process". In: *Bernoulli* 10.1 (2004), pp. 97–120.

[32] E. Nelson. "The adjoint Markoff process". In: *Duke Math. J.* 25 (1958), pp. 671–690.

[33] I. Nourdin and F. G. Viens. "Density formula and concentration inequalities with Malliavin calculus". In: *Electron. J. Probab.* 14 (2009), no. 78, 2287–2309.

[34] B. Øksendal. *Stochastic differential equations*. Sixth. Universitext. An introduction with applications. Springer-Verlag, Berlin, 2003, pp. xxiv+360.

[35] P. E. Protter. *Stochastic integration and differential equations*. Second. Vol. 21. Applications of Mathematics (New York). Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin, 2004, pp. xiv+415.

[36] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes, and martingales. Vol. 1*. Cambridge Mathematical Library. Foundations, Reprint of the second (1994) edition. Cambridge University Press, Cambridge, 2000, pp. xx+386.

[37] L. C. G. Rogers and D. Williams. *Diffusions, Markov processes, and martingales. Vol. 2*. Cambridge Mathematical Library. Itô calculus, Reprint of the second (1994) edition. Cambridge University Press, Cambridge, 2000, pp. xiv+480.

[38] K.-i. Sato and M. Yamazato. "Operator-self-decomposable distributions as limit distributions of processes of Ornstein-Uhlenbeck type". In: *Stochastic Process. Appl.* 17.1 (1984), pp. 73–100.

[39] M. Sharpe. *General theory of Markov processes*. Vol. 133. Pure and Applied Mathematics. Academic Press Inc., Boston, MA, 1988, pp. xii+419.

[40] M. Sørensen. "Likelihood methods for diffusions with jumps". In: *Statistical inference in stochastic processes*. Vol. 6. Probab. Pure Appl. Dekker, New York, 1991, pp. 67–105.

[41] C. Strauch. "Adaptive invariant density estimation for ergodic diffusions over anisotropic classes". In: *Ann. Statist.* 46.6B (2018), pp. 3451 –3480.

[42] M. Talagrand. *Upper and lower bounds for stochastic processes*. Vol. 60. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics. Modern methods and classical problems. Springer, Heidelberg, 2014, pp. xvi+626.

[43] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009, pp. xii+214.

[44] R. Vershynin. *High-dimensional probability*. Vol. 47. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018, pp. xiv+284.

[45] G. Viennet. "Inequalities for absolutely regular sequences: application to density estimation". In: *Probab. Theory Relat. Fields* 107.4 (1997), pp. 467–492.

[46] L. Wu. "Large and moderate deviations and exponential convergence for stochastic damping Hamiltonian systems". In: *Stochastic Process. Appl.* 91.2 (2001), pp. 205 –238.

Adaptive invariant density estimation for continuous-time mixing Markov processes under sup-norm risk

Niklas Dexheimer, Claudia Strauch and Lukas Trottner

A

Abstract

Up to now, the nonparametric analysis of multidimensional continuous-time Markov processes has focussed strongly on specific model choices, mostly related to symmetry of the semigroup. While this approach allows to study the performance of estimators for the characteristics of the process in the minimax sense, it restricts the applicability of results to a rather constrained set of stochastic processes and in particular hardly allows incorporating jump structures. As a consequence, for many models of applied and theoretical interest, no statement can be made about the robustness of typical statistical procedures beyond the beautiful, but limited framework available in the literature. To contribute to the statistical understanding in more general situations, we demonstrate how combining β-mixing assumptions on the process and heat kernel bounds on the transition density representing controls on the long- and short-time transitional behaviour, allow to obtain sup-norm and L^2 kernel invariant density estimation rates that match the well-understood case of reversible multidimensional diffusion processes and are faster than in a sampled discrete data scenario. Moreover, we demonstrate how, up to log-terms, optimal sup-norm *adaptive* invariant density estimation can be achieved within our framework, based on tight uniform moment bounds and deviation inequalities for empirical processes associated to additive functionals of Markov processes. The underlying assumptions are verifiable with classical tools from stability theory of continuous-time Markov processes and PDE techniques, which opens the door to evaluate statistical performance for a vast amount of popular Markov models. We highlight this point by showing how multidimensional jump SDEs with Lévy-driven jump part under different coefficient assumptions can be seamlessly integrated into our framework, thus establishing novel adaptive sup-norm estimation rates for this class of processes.

A.1 Introduction

There exist various probabilistic concepts that permit the investigation of quantitative ergodic properties of Markov processes, providing a number of approaches to analyzing the rate of convergence of the process to equilibrium. Such results actually present precious tools for an adequate statistical modelling of complex systems. Markov models, especially of (jump) diffusion-type, find numerous applications in biology, chemistry, natural resource management, computer vision, Bayesian inference in machine learning, cloud computing and many more [3, 17, 36, 37, 39, 43, 79, 82], and ergodicity can usually be seen as some kind of minimum requirement for the development of a fruitful statistical theory. While the probabilistic picture of quantitative ergodic properties is now quite clear, there are still open questions regarding the statistical implications. With this paper, we want to contribute to closing some gaps concerning adaptive nonparametric invariant density estimation for multivariate Markov processes with no specific structural assumptions on their dynamics.

In contrast to the highly-developed statistical theory for scalar diffusion processes, there are relatively few references for nonparametric or high-dimensional general Markov models. To not let sampling effects obscure the statistical implications, it is natural to base the statistical analysis in this context on a continuous observation scheme (i.e., one assumes that a complete trajectory of the process is available). A substantial point of reference for a thorough statistical analysis of ergodic multivariate diffusion processes is provided by the article [28] where the fundamental question of asymptotic statistical equivalence is investigated. Apart from its principal central statement, the work also nicely demonstrates the implications of probabilistic properties of processes on quantitative statistical results. Specifically, heat kernel bounds and the spectral gap inequality are used to prove tight variance bounds for integral functionals which in turn provide fast convergence rates for the specific problem of invariant density estimation. Similar techniques can be used for the in-depth analysis of other statistical questions such as (adaptive) estimation of the drift vector of an ergodic diffusion (cf. [78], [77]). The results in [28, 77, 78] are developed for diffusion processes with drift of gradient-type and unit diffusion matrix. While in this specific case the reversibility assumption is directly verified, the condition of symmetry of the process presents a significant constraint, in particular for solutions of SDEs with jump noise.

More recently, a Bayesian approach to drift estimation of multivariate diffusion processes is undertaken in [63] and [42]. Whilst [42] work in a reversible setting since their approach relies on placing a Gaussian prior on the potential B of the drift $b = \nabla B$ instead of tackling the drift directly, [63] approach drift estimation for non-reversible diffusions by employing PDE techniques to a penalized likelihood estimator. This opens up an excitingly different viewpoint on the statistical handling of multivariate diffusion processes and in case of [63] avoids the need for reversibility. However, both approaches restrict the setting to assumed periodicity of the drift coefficient. While this assumption (similar to reversibility) can certainly be justified for specific applications, the approach does not yet provide an answer to the question of how to conduct a statistical analysis of multidimensional Markov processes without strong structural constraints on the coefficients. From a different perspective, the very recent contribution [4] yields the remarkable observation that quantitatively similar statistical results as in the reversible diffusion case can also be proven for jump diffusions with Lévy-driven jump part, without the need to rely on a reversible or periodic setting, by focusing on assumptions on the characteristics of the process which guarantee exponential ergodicity as the driving force of the statistical approach.

Another branch of the literature that does not consider specific structural assumptions on the process is based on the so called Castellana–Leadbetter condition or variations thereof [14, 18, 50], which imposes finiteness of the integrated uniform distance between the density of the bivariate law of (X_0, X_t) of a stationary Markov process $\boldsymbol{X}$ with stationary density ρ and the product density $\rho \otimes \rho$. This assumption yields dimension independent parametric estimation rates of the invariant density and is thus not suitable for our goal to extend the dimension dependent minimax optimal estimation rates for continuous diffusion processes to more general classes of multidimensional Markov processes, introduced below.

Throughout, we suppose that $(\boldsymbol{X}, (\mathbb{P}^x)_{x\in\mathbb{R}^d})$ is a non-explosive Borel right Markov process with state space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and semigroup $(P_t)_{t\geq 0}$ defined by

$$P_t(x, B) := \mathbb{P}^x(X_t \in B), \quad x \in \mathbb{R}^d, B \in \mathcal{B}(\mathbb{R}^d),$$

see Definition 8.1 in [74] for an exact characterization. This general class of right-continuous Markov processes includes the more specific class of standard processes, which form the basis

of the classical textbook [12], and even more specifically Feller processes, i.e., càdlàg Markov processes with a strongly continuous semigroup mapping $\mathcal{C}_0(\mathbb{R}^d)$, the space of continuous functions on $\mathbb{R}^d$ vanishing at infinity, onto itself. Under regularity assumptions on the coefficients, the exemplary class of (jump) diffusion processes that we study in detail later on belongs to the class of Feller processes and hence falls into our probabilistic regime. Moreover, due to their natural embedding into potential theory, Borel right Markov processes are the object of stability analysis of continuous-time Markov processes pioneered by Meyn and Tweedie in the 1990s [35, 59, 61, 62], in which the long-time behaviour is quantitatively associated with Lyapunov drift criteria. This approach is central to our approach. Since we are ultimately interested in invariant density estimation, we work in an ergodic setting for $\boldsymbol{X}$ throughout the paper. That is, the following assumption is in place:

($\mathscr{A}0$) The marginal laws of $\boldsymbol{X}$ are absolutely continuous, i.e., for any $t > 0$ and $x \in \mathbb{R}^d$, there exists a measurable function $p_t \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ such that

$$P_t(x, B) = \int_B p_t(x, y)\, \mathrm{d}y, \quad B \in \mathcal{B}(\mathbb{R}^d),$$

and, moreover, $\boldsymbol{X}$ admits a unique absolutely continuous invariant probability measure μ, i.e., there exists a density $\rho \colon \mathbb{R}^d \to \mathbb{R}_+$ such that $\mathrm{d}\mu = \rho\, \mathrm{d}\lambda$ and

$$\mathbb{P}^\mu(X_t \in B) := \int_{\mathbb{R}^d} P_t(x, B)\, \mu(\mathrm{d}x) = \int_{\mathbb{R}^d} \int_B p_t(x, y)\rho(x)\, \mathrm{d}y\, \mathrm{d}x = \int_B \rho(x)\, \mathrm{d}x = \mu(B)$$

for any Borel set B.

We abbreviate $\mathbb{P}^\mu = \mathbb{P}$, $\mathbb{E}^\mu = \mathbb{E}$ and denote $\mu(g) = \int g\, \mathrm{d}\mu$ for $g \in L^1(\mu)$ or $g \geq 0$. Note that in ($\mathscr{A}0$) existence of a density ρ of the invariant distribution μ is not an additional requirement on $\boldsymbol{X}$, but is guaranteed by the Radon–Nikodym theorem thanks to the definition of invariance and the existence of densities for the transition operators.

Turning away from Lyapunov criteria for general ergodic Markov processes, the long-time behaviour of Markovian semigroups is also known to be linked to functional inequalities. The most familiar setting is the L^2 framework with its equivalence between the corresponding Poincaré inequalities and exponential decay of the Markovian semigroup. The relation between both approaches in terms of quantifying ergodic properties of Markov processes is studied in [8].

With a view towards applicability of the statistical results we turn away from the functional inequality approach to stability and focus on mixing conditions verifiable through Lyapunov-type criteria that are applicable for a vast amount of structurally diverse Markov processes. Making the mixing behaviour of the process a cornerstone of the statistical analysis is completely natural when comparing to discrete time theory. For discrete observations it is well-established in the field of weak dependence that different sets of mixing assumptions (e.g., α-mixing or β-mixing) and relaxations thereof can produce variance bounds and deviation inequalities that hold up to analogous results from i.i.d. observations to yield sharp nonparametric estimation results, see [30, 69] for an overview. We provide an answer to the statistically fundamentally interesting question under which conditions on a multivariate continuous-time mixing Markov process drawing inference based on full observations can yield better estimation rates than under partial observations corresponding to a weakly dependent discrete observation sequence.

From a statistical perspective, this extended range of application comes at the price that mixing assumptions are in general not suited to conducting the nonparametric statistical analysis in the ususal minimax sense. The reason for this is that the mixing constants are non-explicit in most cases of interest, rendering the uniform analysis of upper bounds on the statistical error over whole classes of processes impossible. Our focus is therefore on analyzing the sup-norm risk of kernel invariant density estimators of a given single Markov process with the known minimax rates of multivariate reversible diffusion processes serving as benchmark results.

Our particular interest in sup-norm adaptive invariant density estimation is not only rooted in the higher degree of intepretability of such statements compared to the pointwise L^2 risk, but also comes from the observation that certain problems from applied probability can only be handled with statistical tools when sup-norm estimation bounds of a quantity of interest are available. This point is highlighted in [22], where results from this paper are implemented for the development of data-driven stochastic optimal control strategies for the probabilistically quite diverse problems of optimally reflecting underlying diffusions and Lévy processes. Moreover, the general issue of adaptive invariant density estimation is not only interesting for purely intrinsic mathematical reasons, but is also highly relevant for related nonparametric statistical questions in an ergodic setting such as drift estimation for stochastic differential equations via Nadaraya–Watson type path estimators, whose analysis requires sharp invariant density estimation rates. This has been demonstrated for continuous diffusion processes under different risk measures and coefficient assumptions [27, 32, 77, 78], but can potentially also be extended to additional Lévy jump structures. The application of our general kernel invariant density estimation results to Lévy driven SDEs in this paper can therefore serve as basis for future drift estimation investigations of such processes. We also emphasize that the uniform moment bounds for path integrals of Markov processes with general mixing rates that we develop as the fundamental tool for invariant density estimation under exponential mixing can also be utilized for sup-norm risk analysis of invariant density and drift estimators in subexponentially ergodic SDE models—which even in the continuous diffusion case is not well-understood in the literature.

In order to obtain a clear picture and benchmark results that are not distorted by discretization errors, we work under the assumption that a continuous observation of a trajectory $\mathbf{X}^T = (X_t)_{t\in[0,T]}$ of $\boldsymbol{X}$ is available. For the analysis of statistical methods (e.g., for estimating the characteristics of $\boldsymbol{X}$), variance bounds and deviation inequalities are of central importance. Section A.2 focuses on the analysis of the variance of additive functionals of the form $\int_0^t f(X_s)\,\mathrm{d}s$ for the ergodic process $\boldsymbol{X}$. We introduce sets of general assumptions on transition and invariant density which allow to prove tight variance bounds (cf. Propositions A.1 and A.5). Here, we consider an on-diagonal heat kernel bound to regulate the short-time transitional behaviour of the process and either local uniform transition density convergence to the invariant distribution at sufficient speed for any dimension $d \in \mathbb{N}$ or exponential β-mixing in dimension $d \geq 2$ to obtain tight controls on the long-time transitions of the process. The combination of heat kernel bound and local uniform transition density convergence can be interpreted as a localized version of the Castellana–Leadbetter condition that separates the short- and long-time effects and considerably weakens the inherent assumptions on the speed at which the law of X_t approaches a singular distribution as $t \downarrow 0$ in higher dimensions. We give a detailed analysis of this condition. We demonstrate how total variation convergence at sufficient speed implies the local uniform transition density assumption and argue that in case of μ-a.s. exponential ergodicity of the

process, exponential β-mixing and local uniform transition density convergence are essentially equivalent, giving a homogeneous picture of our different sets of assumptions.

In Section A.3 we proceed by showing how the β-mixing property of $\boldsymbol{X}$—which is satisfied for a wide range of Markov processes appearing in applied and theoretical probability theory—is reflected in uniform moment bounds on empirical processes associated to integral functionals of $\boldsymbol{X}$. More precisely, for countable classes $\mathcal{G}$ of bounded measurable functions g, we establish an upper bound on

$$\Big(\mathbb{E}\Big[\sup_{g\in\mathcal{G}}\Big|\frac{1}{T}\int_0^T g(X_s)\,\mathrm{d}s - \int g\,\mathrm{d}\mu\Big|^p\Big]\Big)^{1/p}, \quad p \geq 1,$$

(cf. Theorem A.6) stated in terms of entropy integrals related to $\mathcal{G}$ and the variance of the integral functionals. This result holds for β-mixing Borel right processes on general state spaces without any assumptions on the existence of transition densities, i.e., Assumption (𝒜0) is diminished to stationarity which further increases the applicability of our findings for future investigations. Such moment bounds and associated uniform deviation inequalities are generally the focal point for efficient implementation of adaptive estimation procedures, both for the sup-norm as well as the pointwise and integrated L^2 risk. In our concrete estimation context, we use the uniform moment bounds together with the variance bounds from Section A.2 to establish sharp deviation inequalities for the sup-norm risk of a kernel invariant density estimator that is essential for the adaptive estimation scheme considered in Section A.4 that we describe below.

In presence of additional information on the irregularity of paths provided by the heat-kernel estimate, we establish in Section A.4 that the stationary density of exponentially β-mixing Markov processes can be estimated in any dimension at optimal rates both wrt. sup-norm risk and pointwise L^2 risk—where optimality is understood relative to the benchmark minimax rates known for continuous reversible diffusion processes. We go even further by showing that in dimension $d \geq 3$—where the optimal bandwidth choice depends on the typically unknown degree of Hölder smoothness β—fitting a Lepski type adaptive bandwidth selection scheme proposed in [41] for i.i.d. data to our needs provides optimal estimation rates up to iterated log-factors (see also [52] for an adaptive scheme for anisotropic sup-norm estimation for i.i.d. observations). More precisely, our main result Theorem A.11 shows that, given a kernel estimator $\widehat{\rho}_{h,T}$ for the unknown invariant density ρ with bandwidth choice

$$h \equiv h(T) \sim \begin{cases} \log^2 T/\sqrt{T}, & d = 1, \\ \log T/T^{1/4}, & d = 2, \\ (\log T/T)^{1/(2\beta+d-2)}, & d \geq 3, \end{cases}$$

we have, for any $p \geq 1$ and a bounded open domain D,

$$\mathbb{E}\Big[\sup_{x\in D}\big|\widehat{\rho}_{h,T}(x) - \rho(x)\big|^p\Big]^{1/p} \in \begin{cases} \mathrm{O}\big(\sqrt{\log T/T}\big), & d = 1, \\ \mathrm{O}\big(\log T/\sqrt{T}\big), & d = 2, \\ \mathrm{O}\big((\log T/T)^{\frac{\beta}{2\beta+d-2}}\big), & d \geq 3. \end{cases}$$

Although we consider a nonparametric framework, the question of data-driven estimation only arises in dimension $d \geq 3$. In this case, we suggest to replace the smoothness-dependent bandwidth choice $h(T)$ by the adaptive selector $\widehat{h}_T \equiv \widehat{h}_T^{(k)}$ introduced in (A.16). Then, if the

order of the kernel is sufficiently large and for $\log_{(k)} T$ denoting the k-th iterated logarithm,

$$\mathbb{E}\Big[\sup_{x\in D}\big|\widehat{\rho}_{\widehat{h}_T,T}(x)-\rho(x)\big|\Big] \in \mathrm{O}\bigg(\bigg(\frac{\log_{(k)} T\log T}{T}\bigg)^{\frac{\beta}{2\beta+d-2}}\bigg),$$

where $k\in\mathbb{N}$ can, in principle, be chosen arbitrarily large—which however decreases the size of the set of candidate bandwidths for the adaptive selection procedure given a finite observation horizon. We emphasize that the logarithmic gap could be avoided if constants appearing in the uniform deviation inequality from Section A.3 were explicitly calculated. This, however, requires exact knowledge of the ergodic and short-time behaviour of the process, contradicting a truly adaptive nature of the approach.

Such sup-norm adaptive multivariate estimation results are completely new and complement adaptive L^2 estimation procedures based on model selection considered in [24] for discrete time mixing chains and in [4] for Lévy driven jump-diffusions. We emphasize that [24] also consider estimation of continuous-time mixing processes in terms of their sampled skeletons. However, our faster adaptive estimation rates in presence of heat kernel bounds demonstrate that such approach can be considerably improved by *not* taking a Markov chain viewpoint under partial observations but by exploiting continuous-time probabilistic structures under full observations.

As a concrete example, we investigate multidimensional SDEs with Lévy-driven jump part, i.e., Markov processes associated to the solution of

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t + \gamma(X_{t-})\,\mathrm{d}Z_t, \quad X_0 = x\in\mathbb{R}^d, \tag{A.1}$$

where $\mathbf{W}$ is d-dimensional Brownian motion and $\mathbf{Z}$ is a pure jump Lévy process independent of $\mathbf{W}$. In Section A.4.2, we investigate Lévy driven Ornstein–Uhlenbeck processes as the basic class of Lévy driven jump diffusions with unbounded drift coefficient. In presence of non-trivial Gaussian part and very mild moment assumptions on the Lévy measure, we infer optimal sup-norm and pointwise L^2 invariant density estimation results in any dimension. In this case, an adaptive estimation procedure is not necessary, since the invariant density is a smooth function. In Section A.4.3, we allow for more flexible dispersion and jump coefficients σ,γ with the price to be paid being boundedness of the drift b. By considering solutions $\boldsymbol{X}$ to (A.1) under appropriate assumptions on the coefficients b,σ,γ and the jump measure associated to $\mathbf{Z}$, we can apply our general statistical results to invariant density estimation for $\boldsymbol{X}$, thus establishing new results on sup-norm adaptive invariant density estimation for such general jump processes.

In the sequel, we concentrate on guiding the reader through the framework and the accompanied mathematical results. All proofs are deferred to the appendices after Section A.4, with more specific references on their exact location at the relevant passages of the main text.

Basic notation. A set $B\in\mathcal{B}(\mathbb{R}^d)$ is called μ-full if $\mu(B)=1$. We say that the Borel right Markov process $\boldsymbol{X}$ is μ-a.s. V-ergodic at speed ξ if, for some μ-full set Λ,

$$\|P_t(x,\cdot)-\mu\|_{\mathrm{TV}} \le CV(x)\xi(t), \quad t\ge 0, x\in\Lambda, \tag{A.2}$$

where $V\colon\mathbb{R}^d\to[0,\infty]$ with $V\mathbb{1}_\Lambda(x)<\infty$ and, for a signed measure ν, $\|\nu\|_{\mathrm{TV}} := \sup_{|f|\le 1}|\nu(f)|$ denotes its total variation norm. If (A.2) holds with $\xi(t)=(1+t)^{-\alpha}$ for some $\alpha>0$, we say that $\boldsymbol{X}$ is μ-a.s. V-polynomially ergodic of degree α. If $\xi(t)=\mathrm{e}^{-\kappa t}$ for some $\kappa>0$, then $\boldsymbol{X}$ is called

μ-a.s. V-exponentially ergodic. When $\Lambda = \mathbb{R}^d$ and $V(x) < \infty$ for any $x \in \mathbb{R}^d$, we just say that $\boldsymbol{X}$ is V-ergodic at speed ξ (resp., V-polynomially ergodic and V-exponentially ergodic).

For any multi-index $\alpha \in \mathbb{N}^d$ and $x \in \mathbb{R}^d$, set $|\alpha| = \sum_{i=1}^d \alpha_i$ and $x^\alpha = \prod_{i=1}^d x_i^{\alpha_i}$. For $\lfloor\!\lfloor \beta \rfloor\!\rfloor$ denoting the largest integer *strictly* smaller than β, introduce the Hölder class on an open domain $D \subset \mathbb{R}^d$

$$\mathcal{H}_D(\beta, \mathsf{L}) = \Big\{ f \in \mathcal{C}^{\lfloor\!\lfloor \beta \rfloor\!\rfloor}(D, \mathbb{R}) : \max_{|\alpha| = \lfloor\!\lfloor \beta \rfloor\!\rfloor} \sup_{x,y \in D, x \neq y} \frac{|f^{(\alpha)}(x) - f^{(\alpha)}(y)|}{|x - y|^{\alpha - \lfloor\!\lfloor \alpha \rfloor\!\rfloor}} \leq \mathsf{L}, \sup_{x \in D} |f(x)| \leq \mathsf{L} \Big\}, \quad \text{(A.3)}$$

where $f^{(\alpha)} := \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$. Recall that a kernel function $K \colon \mathbb{R}^d \to \mathbb{R}$ is said to be of order $\ell \in \mathbb{N}$ if, for any $\alpha \in \mathbb{N}^d$ with $|\alpha| \leq \ell$, $x \mapsto x^\alpha K(x)$ is integrable and, moreover, $\int_{\mathbb{R}^d} K(x)\,\mathrm{d}x = 1$, $\int_{\mathbb{R}^d} K(x) x^\alpha \,\mathrm{d}x = 0$, for $\alpha \in \mathbb{N}^d$, $|\alpha| \in \{1, \dots, \ell\}$.

A.2 Basic framework and variance analysis of integral functionals of general Markov processes

This first section focuses on the analysis of the variance of integral functionals of the form $\int_0^t f(X_s)\,\mathrm{d}s$ for the ergodic process $\boldsymbol{X} = (X_s)_{0 \leq s \leq t}$ under different sets of general assumptions on $\boldsymbol{X}$ that will carry us through the rest of the paper. Such variance bounds are indispensable tools for statistical applications since (as we will see in Section A.3) the variance of integral functionals naturally appears in associated deviation inequalities and related moment bounds and thus requires tight estimates. All proofs for this section can be found in Appendix A.I.2.

A.2.1 *Variance analysis under assumptions on transition and invariant density*

Recall the definition of Assumption (𝒜0) from the introduction. We start by working under the following set of additional assumptions:

(𝒜1) In case $d = 1$, there exists a non-negative, measurable function $\alpha \colon (0, 1] \to \mathbb{R}_+$ such that, for any $t \in (0, 1]$,

$$\sup_{x,y \in \mathbb{R}} p_t(x, y) \leq \alpha(t) \quad \text{and} \quad \int_{0+}^1 \alpha(t)\,\mathrm{d}t = c_1 < \infty,$$

and, in case $d \geq 2$, there exists $c_2 > 0$ such that the following on-diagonal heat kernel estimate holds true:

$$\forall t \in (0, 1] : \sup_{x,y \in \mathbb{R}^d} p_t(x, y) \leq c_2 t^{-d/2}. \quad \text{(A.4)}$$

(𝒜2) There exists a μ-full set Λ such that, for any compact set $\mathcal{S} \subset \mathbb{R}^d$, there exists a non-negative, measurable function $r_{\mathcal{S}} \colon (0, \infty) \to \mathbb{R}_+$ such that

$$\forall t > 1 : \sup_{x \in \mathcal{S} \cap \Lambda, y \in \mathcal{S}} |p_t(x, y) - \rho(y)| \leq r_{\mathcal{S}}(t) \text{ with } \int_1^\infty r_{\mathcal{S}}(t)\,\mathrm{d}t = c_{\mathcal{S}} < \infty. \quad \text{(A.5)}$$

An essential aspect of the statistical analysis of stochastic processes is the influence of the dimension of the underlying process. It is known that certain phenomena (as compared, e.g., to estimation based on i.i.d. observations) occur in the one-dimensional case. However, these phenomena can usually only be detected by means of specific techniques that take advantage of the unique probabilistic characteristics of scalar processes such as local time for one-dimensional diffusion processes. A "standardized" statistical framework which covers all dimensions with similar conditions cannot capture these phenomena. Our assumptions may therefore be understood as an attempt to find general conditions that make no reference to dimension or process specific phenomena, yet yield variance bounds which are tight enough to allow proving optimal convergence rates for nonparametric procedures.

In this regard, they should be compared to the Castellana–Leadbetter condition [18] requiring that

$$\int_{(0,\infty)} \sup_{x,y\in\mathbb{R}^d} |\rho(x)p_t(x,y) - \rho(x)\rho(y)| \,\mathrm{d}t < \infty, \tag{A.6}$$

which allows L^2 estimation of the invariant density via a kernel estimator at parametric (or *superoptimal* [13]) rate $1/T$ in *any* dimension $d \geq 1$. Since (𝒮1) implies that ρ is bounded, (𝒮2) can be understood as a localized, unweighted alternative to (A.6) away from 0, which captures the mixing behaviour of the process as we discuss below. Our assumption (𝒮1) corresponds to the integral part of (A.6) close to 0 and guarantees that the distribution of X_t is not too close to a singular distribution. However, in dimension $d \geq 2$ this assumption is much milder than (A.6) since heat kernel bounds on the transition density are quite common for many multidimensional Markov processes such as strong solutions of (jump) SDEs. On the other hand, (A.6) is too strong for such Markov processes, since, e.g., the minimax optimal L^2 rate for multivariate diffusions processes is known to be worse than $1/T$ and hence the variance bound implied by (A.6) cannot be achieved.

Also note that the transition density bounds formulated in (𝒮1) are weak compared to related literature dealing with statistical estimation of jump processes. E.g., [4] construct their assumptions on the coefficients and the jump measure of a d-dimensional Lévy-driven jump diffusion to guarantee a heat kernel-type estimate of the form

$$p_t(x,y) \lesssim t^{-d/2}\mathrm{e}^{-\lambda\frac{\|y-x\|^2}{t}} + \frac{t}{|\sqrt{t}+\|y-x\||^{d+\alpha}}, \quad x,y\in\mathbb{R}^d, t\in(0,T],$$

for the estimation horizon $T > 0$, where $\alpha \in (0,2)$ is the self-similarity index of a strictly α-stable Lévy process whose Lévy measure is assumed to dominate the Lévy measure governing the jumps of the SDE. Clearly, this condition is stronger than what we require and is fitted to the concrete probabilistic setting. The reason for this specific choice becomes apparent from Corollary A.16 in Section A.4.3, but our approach reveals that (𝒮1) is sufficient to obtain tight variance bounds in a general multivariate setting. Let us now give the variance bounds implied in our framework.

PROPOSITION A.1. *Suppose that (𝒮1) and (𝒮2) are satisfied, and let f be a bounded function with compact support $\mathcal{S}$ fulfilling $\lambda(\mathcal{S}) < 1$. Then, there exists a constant $C > 0$ independent of f, such*

that, for any $T > 0$,

$$\mathrm{Var}\left(\int_0^T f(X_t)\,\mathrm{d}t\right) \le C(1\vee c_{\mathcal{S}})T\|f\|_\infty^2 \lambda(\mathcal{S})\mu(\mathcal{S})\psi_d^2(\lambda(\mathcal{S})), \text{ with } \psi_d(x) := \begin{cases} 1, & d=1,\\ \sqrt{1+\log(1/x)}, & d=2,\\ x^{\frac{1}{d}-\frac{1}{2}}, & d\ge 3, \end{cases} \tag{A.7}$$

where the variance is taken with respect to $\mathbb{P}$.

To get an impression of the usefulness of the above result, let us discuss the relation of the local uniform transition density convergence assumption (𝒜2) to more general and often conveniently verifiable stability conditions on $\boldsymbol{X}$. In [83], conditions on the characteristic function $\varphi^x_{X_t}(\lambda) := \mathbb{E}^x[\exp(\mathrm{i}\langle X_t,\lambda\rangle)]$ of X_t and the Fourier transform $\{\mathcal{F}\mu\}(\lambda) = \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\langle x,\lambda\rangle}\,\mu(\mathrm{d}x)$ were formulated in the scalar setting $d = 1$ that imply finiteness of the integral part away from 0 in the Castellana–Leadbetter condition (A.6). A straightforward adaption to our multivariate localized setting yields the following result, with the proof being omitted.

Lemma A.2. *Suppose that* $\boldsymbol{X}$ *is* V*-polynomially ergodic of degree* $\gamma_1 > q/(q-1)$ *for some locally bounded function* V *and* $q > 1$. *If there exists* $\gamma_2 > qd$ *such that*

(𝒱1) $|\varphi^x_{X_t}(\lambda) - \{\mathcal{F}\mu\}(\lambda)| \le V(x)(1+t)^{-\gamma_1}, \quad t \ge 1, x, \lambda \in \mathbb{R}^d$

(𝒱2) $|\varphi^x_{X_t}(\lambda)| \vee |\{\mathcal{F}\mu\}(\lambda)| \lesssim (1+\|\lambda\|)^{-\gamma_2}, \quad x, \lambda \in \mathbb{R}^d, t \ge 1,$

then (𝒜2) is satisfied with $\Lambda = \mathbb{R}^d$, $r_{\mathcal{S}}(t) \sim \sup_{x\in\mathcal{S}} V(x)(1+t)^{-\gamma_1}$ *for compacts* $\mathcal{S}$.

Note that (𝒱2) implies that the Fourier transforms of $P_t(x,\cdot)$ and μ are integrable and hence the Fourier inversion theorem guarantees that continuous bounded transition and invariant densities exist. Moreover, as remarked in [83], (𝒱1) is fulfilled whenever $\boldsymbol{X}$ is V-polynomially ergodic with rate $\gamma_1 > 1$.

Condition (𝒱2) is quite natural in a statistical estimation context since it essentially encodes a certain amount of smoothness of the transition and stationary density. However, the following simple observation demonstrates that the additional growth conditions on the characteristic function are not needed in presence of sufficiently fast total variation convergence. Concerning the specific set of assumptions (𝒜0)–(𝒜2), it is established with this result in Section A.4.2 that they are satisfied, e.g., for a large class of multivariate Lévy-driven Ornstein–Uhlenbeck processess.

Lemma A.3. *Suppose that* $\|p_1\|_\infty < \infty$ *and that* $\boldsymbol{X}$ *is* μ*-a.s.* V*-ergodic at speed* ξ *such that* $V\mathbb{1}_\Lambda$ *is locally bounded and* $\int_0^\infty \xi(t)\,\mathrm{d}t < \infty$. *Then, (𝒜2) holds with*

$$r_{\mathcal{S}}(t) = 2C\|p_1\|_\infty \sup_{x\in\mathcal{S}\cap\Lambda} V(x)\xi(t-1), \quad t > 1.$$

Recall that the stationary Markov process $\boldsymbol{X}$ is said to be β-mixing if

$$\beta(t) := \int_{\mathbb{R}^d} \|P_t(x,\cdot) - \mu(\cdot)\|_{\mathrm{TV}}\,\mu(\mathrm{d}x) \underset{t\to\infty}{\longrightarrow} 0.$$

Hence, if $\boldsymbol{X}$ is μ-a.s. V-ergodic at speed ξ for a function V such that $\mu(V) < \infty$, then $\beta(t) \lesssim \xi(t)$, i.e., $\boldsymbol{X}$ is β-mixing at speed ξ. If there exist constants $\kappa, c_\kappa > 0$ such that $\beta(t) \le c_\kappa \mathrm{e}^{-\kappa t}$ for any

$t > 0$, then $\boldsymbol{X}$ is said to be exponentially β-mixing. This is always true for μ-a.s. V-exponentially ergodic Markov processes since in this case we can always find a nonnegative function $\widetilde{V} \in L^1(\mu)$ such that $\boldsymbol{X}$ is $\widetilde{V}$-exponentially ergodic as well, which follows from a straightforward extension of [65, Theorem 6.14.(iii)] to the continuous-time case. Conversely, it follows from combining [64, Theorem 1] and [65, Theorem 6.14.(iii)] that if $\boldsymbol{X}$ is exponentially β-mixing, then $\boldsymbol{X}$ is μ-a.s. V-exponentially ergodic for some function V satisfying $\mu(V) < \infty$. See also [20, Lemma 8.9] or [16, Theorem 3.7] for these statements. Exponential β-mixing is formulated as assumption ($\mathscr{A}\beta$) in the next section and will be one of the pillars of our statistical analysis for the sup-norm risk. It is therefore critical for us to understand the exact relationship between exponential β-mixing and ($\mathscr{A}2$). To this end, as a partial converse to Lemma A.3, we explore in Appendix A.I.1 under which additional (quite natural) conditions, ($\mathscr{A}2$) implies the exponential β-mixing property of $\boldsymbol{X}$. Our main findings, taking account of Lemma A.3, Appendix A.I.1 and the developments in Section A.2.2, are summarized in Figure A.1.

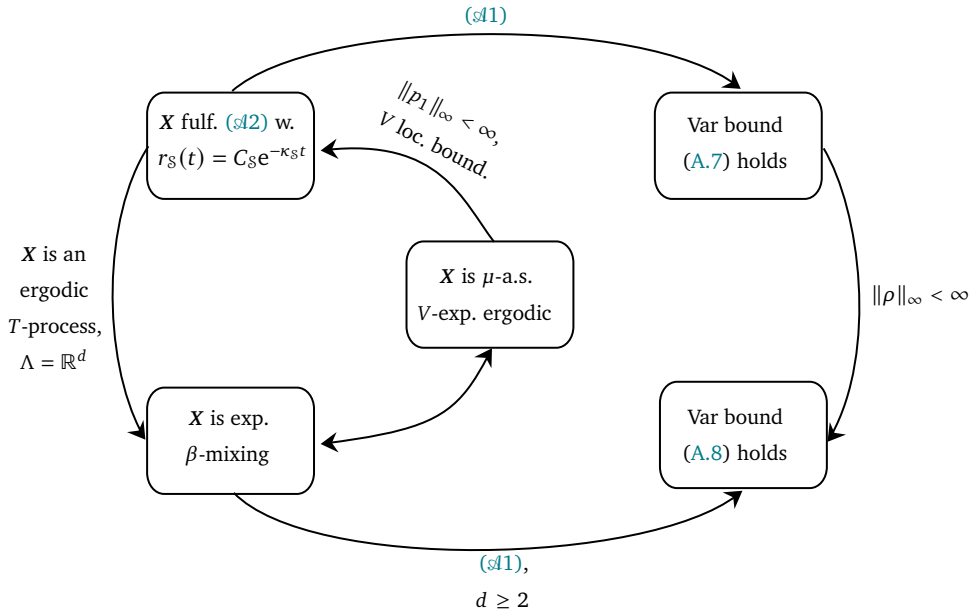

Figure A.1: Overview of interplay between variance bound results, assumptions and stability concepts

A clear picture is drawn, demonstrating that local uniform transition density convergence at exponential speed is intimately connected with exponential β-mixing of the process—both concepts having μ-a.s. exponential ergodicity as the driving force behind them in most concrete applications. Both conditions ($\mathscr{A}2$) and ($\mathscr{A}\beta$) gain substantial additional statistical power via the smoothing assumption ($\mathscr{A}1$), which allows obtaining tight variance bounds that yield superior estimation properties under continuous observations compared to incomplete information via sampling procedures. This will be demonstrated in Section A.4. Moreover, the slightly more specific localized Castellana–Leadbetter condition provides the advantage of optimal estimation also in the scalar case $d = 1$ and wrt the L^2 risk under less restrictive assumptions on the speed of convergence of the process (polynomial is sufficient) in any dimension, which justifies us studying this concept separately from exponential β-mixing.

A.2.2 *Variance analysis under exponential β-mixing*

In this subsection, we specify our study to multidimensional stochastic processes by restricting the analysis to dimension $d \geq 2$. While we further assume that the on-diagonal heat kernel bound on the transition density (A.4) from (𝒜1) still holds, we drop the transition density rate assumption (𝒜2) and instead impose exponential β-mixing of $\boldsymbol{X}$. Note that this is implied by (𝒜2) under suitable technical conditions on $\boldsymbol{X}$ (see Figure A.1 and Propositions A.19 and A.20 in Appendix A.I.1).

(𝒜β) The process $\boldsymbol{X}$ started in the invariant measure μ is exponentially β-mixing, i.e., there exist constants $c_\kappa, \kappa > 0$ such that

$$\int \|P_t(x,\cdot) - \mu\|_{\mathrm{TV}}\, \mu(\mathrm{d}x) \;\leq\; c_\kappa \mathrm{e}^{-\kappa t}, \quad t \geq 0.$$

Let us emphasize that in presence of the heat kernel bound (𝒜1), Lemma A.4 below shows that Assumption (𝒜0) is strengthened to the existence of a *bounded* invariant density since the transition density of any skeleton chain is uniformly bounded for fixed $t > 0$. That is, the following assumption is in place.

(𝒜0+) Assumption (𝒜0) holds and the invariant density has a bounded version ρ, i.e., $\|\rho\|_\infty < \infty$.

Lemma A.4. *Assume that $\boldsymbol{X}$ has an invariant distribution μ and that there is some $\Delta > 0$ such that the transition density p_Δ exists and $\sup_{x,y\in\mathbb{R}^d} p_\Delta(x,y) \leq c$ for some constant $c > 0$. Then, μ admits a bounded density.*

The next result gives a tight variance bound on the integral $\int_0^T f(X_t)\,\mathrm{d}t$ under β-mixing. Its effectiveness for sup-norm estimation of general Markov processes will be demonstrated in Section A.4. Note in particular that, using boundedness of ρ under (𝒜0) and (𝒜1), the same rate can be obtained under (𝒜2) from Proposition A.1. Recall the definition of $\psi_d\colon (0,\mathrm{e}) \to \mathbb{R}_+$ in (A.7).

Proposition A.5. *Grant assumptions (𝒜1) and (𝒜β), and let f be a bounded function with compact support $\mathcal{S}$ fulfilling $\lambda(\mathcal{S}) < 1$. Then, for any $d \geq 2$, there exists a constant $C > 0$ not depending on f such that, for any $T > 0$,*

$$\mathrm{Var}\left(\int_0^T f(X_t)\,\mathrm{d}t\right) \leq CT\|f\|_\infty^2 \|\rho\|_\infty \lambda^2(\mathcal{S})\psi_d^2(\lambda(\mathcal{S})). \tag{A.8}$$

Notation. Throughout the sequel, we denote by $\boldsymbol{\Sigma}$ the class of non-explosive, exponentially β-mixing Borel right Markov processes $\boldsymbol{X}$ such that assumptions (𝒜0) and (𝒜1) hold (and hence (𝒜0+) is in place, i.e., the invariant density ρ is bounded). Moreover, in dimension $d = 1$ we assume that (𝒜2) is in place with a rate function $r_{\mathcal{S}}$ which is monotone wrt the compact sets $\mathcal{S}$ in the sense that

$$\mathcal{S}_1 \subset \mathcal{S}_2 \implies c_{\mathcal{S}_1} = \int_1^\infty r_{\mathcal{S}_1}(t)\,\mathrm{d}t \leq \int_1^\infty r_{\mathcal{S}_2}(t)\,\mathrm{d}t = c_{\mathcal{S}_2} < \infty. \tag{A.9}$$

Alternatively, if we do not want to restrict to exponentially β-mixing processes, consider the class of processes $\boldsymbol{\Theta}$ consisting of d-dimensional non-explosive Borel right processes such that

(𝒜0)–(𝒜2) hold, where again the constants $c_{\mathcal{S}}$ appearing in (𝒜2) satisfy (A.9). Note that if $\widetilde{\boldsymbol{\Theta}}$ is the restriction of $\boldsymbol{\Theta}$ containing the class of processes $\boldsymbol{X}$ satisfying the assumptions of Proposition A.19 or Proposition A.20, then $\widetilde{\boldsymbol{\Theta}} \subset \boldsymbol{\Sigma}$.

A.3 UNIFORM MOMENT BOUNDS FOR PATH INTEGRALS

We now turn to deriving uniform moment bounds for integral functionals of the ergodic process $\boldsymbol{X}$. These are intimately connected with Bernstein-type tail inequalities, which due to their crucial importance for many probabilistic and statistical applications—such as the derivation of limit theorems or upper bound statements for nonparametric estimation procedures—have been excessively studied in the literature (see Section 1.1 of [40] for an overview). Both a Lyapunov function method and a functional inequalities approach can be used for deriving results on the concentration behaviour of additive functionals of $\boldsymbol{X}$. [19] establish non-asymptotic deviation bounds for

$$\mathbb{P}\left(\left|\frac{1}{t}\int_0^t f(X_s)\,\mathrm{d}s - \int f\,\mathrm{d}\mu\right| \geq r\right), \qquad f \in L^1(\mu),$$

using different moment assumptions for f and regularity conditions for μ, "regularity" referring to the condition that μ may satisfy various functional inequalities (F-Sobolev, generalized Poincaré, etc.). In a symmetric Markovian setting and assuming a spectral gap, Lezaud [53] uses Kato's perturbation theory for proving Bernstein-type concentration inequalities for empirical means of the form $\int_0^t f(X_s)\,\mathrm{d}s$, the upper bound depending on the asymptotic variance of f. Amongst other methods, [40] exploit both a Lyapunov function method and a functional inequalities approach for extending Lezaud's result to inequalities for possibly unbounded f. Going beyond the symmetric case, Lyapunov-type conditions can also be used for verifying exponential mixing properties, paving the way to generalizing concentration results based on independent observations to the dependent case. For corresponding results for discrete random (Markov) sequences under different mixing or ergodicity assumptions, we refer to [1, 2, 10, 23, 31, 51, 58, 66, 71].

A.3.1 *General framework*

Our main focus in this subsection is on deriving corresponding uniform moment inequalities of empirical processes, using merely the previously introduced assumptions (in particular, the β-mixing property), and without imposing any additional conditions on the process. We emphasize that for this section no assumption on the existence of transition or invariant densities is needed, but that we only work within an ergodic β-mixing framework. Moreover, the results are established for β-mixing Markov processes with arbitrary topological state space $\mathcal{X}$, not necessarily equal to $\mathbb{R}^d$, and general mixing rate. That is, we suppose in this section that

$$\beta(t) = \int_{\mathcal{X}} \|P_t(x,\cdot) - \mu\|_{\mathrm{TV}}\,\mu(\mathrm{d}x) \leq \Xi(t),$$

for some rate function $\Xi(t)$ decreasing to 0 as $t \to \infty$. We aim to prove moment bounds for suprema of the form

$$\sup_{g\in\mathcal{G}} |\mathbb{G}_t(g)| =: \|\mathbb{G}_t\|_{\mathcal{G}}, \qquad \text{for } \mathbb{G}_t(g) := \frac{1}{\sqrt{t}}\int_0^t g(X_s)\,\mathrm{d}s,$$

where the supremum is taken over entire (possibly infinite-dimensional) function classes $\mathcal{G} \subset \mathcal{B}_b(\mathcal{X})$ of μ-centered measurable bounded functions on $\mathcal{X}$. Similarly to [9] and [33], we apply the generic chaining device for the derivation of our result. The basic strategy of the proof is splitting the integral into blocks of length m_t, constructing an independent Berbee coupling based on the β-mixing property as described in Viennet [84], and then using the classical Bernstein inequality for i.i.d. random variables for the coupled integral blocks to drive the chaining procedure from [33]. The use of Berbee's coupling lemma is a well-established method for studying empirical processes of discrete β-mixing sequences, see [70, Chapter 8], and has recently been employed in [4] for establishing L^2 oracle bounds for an adaptive estimator of the invariant density of a class of exponentially β-mixing Lévy-driven jump diffusions.

We now formulate a crucial tool for deriving upper bounds on the sup-*norm risk* of estimators of the invariant density of processes $\boldsymbol{X} \in \boldsymbol{\Sigma}$. Our final moment bound on the supremum of the process $\mathbb{G}_t$ is stated in terms of entropy integrals of the indexing function class $\mathcal{G}$. In many applications, the corresponding assumption is straightforward to verify. For any given $\varepsilon > 0$, denote by $\mathcal{N}(\varepsilon, \mathcal{G}, d)$ the covering number of $\mathcal{G}$, i.e., the smallest number of balls of d-radius ε needed to cover $\mathcal{G}$. Furthermore, given $f, g \in \mathcal{G}$, let $d_\infty(f,g) := \|f-g\|_\infty$ and

$$d^2_{\mathbb{G},t}(f,g) := \sigma^2_t(f-g), \text{ where } \sigma^2_t(f) := \operatorname{Var}\left(\frac{1}{\sqrt{t}}\int_0^t f(X_s)\,\mathrm{d}s\right).$$

Theorem A.6. *Suppose that $\boldsymbol{X}$ is β-mixing with rate function $\Xi(t)$. Let $\mathcal{G}$ be a countable class of bounded real-valued functions with $\mu(g) = 0$ and let $m_t \in (0, t/4]$. Then, there exist $\tau \in [m_t, 2m_t]$ and constants $\widetilde{C}_1, \widetilde{C}_2 > 0$ such that, for any $1 \le p < \infty$,*

$$\begin{aligned}\left(\mathbb{E}\left[\|\mathbb{G}_t\|^p_{\mathcal{G}}\right]\right)^{1/p} \le{}& \widetilde{C}_1 \int_0^\infty \log \mathcal{N}\big(u, \mathcal{G}, \tfrac{2m_t}{\sqrt{t}} d_\infty\big)\,\mathrm{d}u + \widetilde{C}_2 \int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{G}, d_{\mathbb{G},\tau})}\,\mathrm{d}u \\ &+ 4 \sup_{g\in\mathcal{G}} \left(\frac{2m_t}{\sqrt{t}}\|g\|_\infty \widetilde{c}_1 p + \|g\|_{\mathbb{G},\tau}\widetilde{c}_2\sqrt{p} + \frac{1}{2}\|g\|_\infty \sqrt{t}\,\Xi(m_t)^{1/p}\right),\end{aligned} \tag{A.10}$$

for positive constants $\widetilde{c}_1, \widetilde{c}_2$ defined in (A.33).

Remark A.7. Consider $p = 1$ and the specific choice of $m_t = \kappa^{-1}\log t$ in case of exponential β-mixing rate $\Xi(t) = c_\kappa \exp(-\kappa t)$. Then, the above result implies that

$$\mathbb{E}[\|\mathbb{G}_t\|_{\mathcal{G}}] \lesssim \int_0^\infty \log \mathcal{N}\big(u, \mathcal{G}, \tfrac{\log t}{\sqrt{t}} d_\infty\big)\,\mathrm{d}u + \int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{G}, d_{\mathbb{G},\tau})}\,\mathrm{d}u + \sup_{g\in\mathcal{G}}\left(\frac{\log t}{\sqrt{t}}\|g\|_\infty + \|g\|_{\mathbb{G},\tau}\right).$$

If we considered the related discrete time problem of finding uniform moment bounds for additive functionals $\frac{1}{\sqrt{n}}\sum_{k=0}^n g(X_k)$ of a Markov chain $(X_n)_{n\in\mathbb{N}_0}$ and assumed exponential ergodicity of the chain, using the state of the art Bernstein inequality given in [1, Theorem 6] (see also [51]) for the generic chaining procedure would yield an analogous result with an asymptotic version of the variance norm. In particular, the log-scaling of the sup-norm is also present in the discrete time case as a consequence of exponential ergodicity, whereas in the i.i.d. case this factor would disappear. Our direct coupling approach therefore yields tight uniform moment bounds and makes the contribution of the mixing term transparent, which paves the way for studying nonparametric implications of sub-exponential mixing rates for sup-norm estimation problems in continuous time.

To get a first taste of the consequences of Theorem A.6, consider the trivial situation where $\mathcal{G}$ is a singleton set. This allows the study of rates for the L^p-version of von Neumann's ergodic theorem[1] for continuous-time ergodic Markov processes which states that, for $g \in L^p(\mu)$,

$$\frac{1}{T}\int_0^T g(X_s)\,\mathrm{d}s \underset{t\to\infty}{\longrightarrow} \mu(g), \quad \text{in } L^p(\mathbb{P}).$$

Indeed, β-mixing implies strong mixing such that the σ-algebra of shift invariant sets is $\mathbb{P}$-trivial and hence the ergodic theorem is satisfied.

COROLLARY A.8. *Suppose that $\boldsymbol{X}$ is exponentially β-mixing. Then, there exists a constant $C > 0$ such that, for any $T > 0$, $1 \le p < \infty$ and any bounded, measurable function g,*

$$\left\|\frac{1}{T}\int_0^T g(X_t)\,\mathrm{d}t - \mu(g)\right\|_{L^p(\mathbb{P})} \le Cp\|g\|_\infty \frac{1}{\sqrt{T}}.$$

If $\boldsymbol{X}$ is polynomially mixing of degree $\alpha > 1$, i.e., $\Xi(t) \lesssim t^{-\alpha}$, then for any $p \ge 1$ and $T \ge 4^{(\alpha+p)/\alpha}$ we have

$$\left\|\frac{1}{T}\int_0^T g(X_t)\,\mathrm{d}t - \mu(g)\right\|_{L^p(\mathbb{P})} \lesssim \|g\|_\infty T^{-\left(\frac{1}{2}\wedge\frac{\alpha}{\alpha+p}\right)}.$$

A.3.2 *Deviation inequalities for suprema of empirical Markov processes*

Theorem A.6 provides a foundation for the derivation of deviation inequalities, as they are needed, for example, for bounding the sup-norm risk of estimators and for the convergence analysis of adaptive estimation procedures. We will focus on the question of invariant density estimation for Borel right Markov processes, introduced and discussed in Section A.2. Recall the definition of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Theta}$ at the end of that section. Given the observation $(X_s)_{0\le s\le T}$, a natural kernel estimator for the invariant density ρ on a domain D of a Markov process $\boldsymbol{X} \in \boldsymbol{\Sigma}\cup\boldsymbol{\Theta}$ is given by

$$\widehat{\rho}_{h,T}(x) = \frac{1}{T}\int_0^T K_h(x - X_s)\,\mathrm{d}s, \quad x \in \mathbb{R}^d, \quad \text{where } K_h(\cdot) := h^{-d}K(\cdot/h), \quad h > 0, \tag{A.11}$$

for some smooth, Lipschitz continuous kernel function $K\colon \mathbb{R}^d \to \mathbb{R}$ with compact support $[-1/2, 1/2]^d$. The knowledge of the invariant density is not only a question of its own interest, but is also needed, among other things, for the implementation of drift estimation procedures or data-driven methods of stochastic control. Furthermore, this specific estimation problem can be regarded as an acid test for the quality of the statistical analysis: It is known that the invariant density of (possibly multidimensional) diffusion processes can be estimated with a faster convergence rate than is feasible in the classical discrete i.i.d. or weak dependency context. However, these superior convergence rates can only be verified with sufficiently tight estimates in the proof of the upper bound, more precisely, for the stochastic error part appearing in the decomposition

$$\widehat{\rho}_{h,T}(x) - \rho(x) = \mathbb{H}_{h,T}(x) + (\rho * K_h - \rho)(x), \quad \text{for } \mathbb{H}_{h,T}(x) := \widehat{\rho}_{h,T}(x) - \mathbb{E}[\widehat{\rho}_{h,T}(x)]. \tag{A.12}$$

While the bias part is bounded using standard arguments, tight upper bounds on (the supremum of) the stochastic error require specific probabilistic tools.

[1] Not referring to the L^p-statement as Birkhoff's ergodic theorem is not without reason, see [89].

The following uniform deviation inequality is central for our statistical analysis. Its proof requires bounding $\mathbb{E}[\sup_{x\in D}|\mathbb{H}_{h,T}(x)|^p]$ which is done by applying Theorem A.6 to the function class

$$\mathcal{G} := \left\{\overline{K}((x-\cdot)/h) : x \in D\cap\mathbb{Q}^d\right\}, \quad \text{where } \overline{K}((x-\cdot)/h) = K((x-\cdot)/h) - \mu(K((x-\cdot)/h)), \tag{A.13}$$

for some kernel function K as in (A.11) with Lipschitz constant L wrt to the sup-norm $\|\cdot\|_\infty$, and the bandwidth h chosen in $(0,1)$.

Recall that any $\boldsymbol{X} \in \boldsymbol{\Sigma}$, by definition, is exponentially β-mixing, i.e., β-mixing with rate function $\Xi(t) = c_\kappa \mathrm{e}^{-\kappa t}$ for some constants $c_\kappa, \kappa > 0$.

Lemma A.9. *Suppose that $\boldsymbol{X} \in \boldsymbol{\Theta}\cup\boldsymbol{\Sigma}$ and additionally assume in case $\boldsymbol{X} \in \boldsymbol{\Theta}$ that $\boldsymbol{X}$ is β-mixing with strictly decreasing rate function $\Xi(t)$. Then, for any $u_T \geq 1$ such that $\Xi^{-1}(T^{-u_T}) \in \mathrm{o}(T)$ and $T^{-2} \leq h = h_T \in \mathrm{o}(1)$, there exists a constant $c^* > 0$ such that for large enough T*

$$\mathbb{P}\left(\left\|\widehat{\rho}_{h,T} - \mathbb{E}\widehat{\rho}_{h,T}\right\|_{L^\infty(D)} \geq c^*\left(\frac{u_T + \log T}{Th^d}\Xi^{-1}(T^{-u_T}) + T^{-\frac{1}{2}}\psi_d(h^d)\sqrt{u_T \vee \log(h^{-1})}\right)\right) \leq \mathrm{e}^{-u_T}.$$

In particular, when $\boldsymbol{X} \in \boldsymbol{\Sigma}$, for any $\gamma > 0$ and $u_T \in [1, \gamma\log T]$ there exists a constant $c_\gamma > 0$ such that for large enough T

$$\mathbb{P}\left(\left\|\widehat{\rho}_{h,T} - \mathbb{E}\widehat{\rho}_{h,T}\right\|_{L^\infty(D)} \geq c_\gamma \Upsilon_{h,T}(u_T)\right) \leq \mathrm{e}^{-u_T},$$

where

$$\Upsilon_{h,T}(u) := \frac{u(\log T)^2}{Th^d} + T^{-\frac{1}{2}}\psi_d(h^d)\sqrt{u \vee \log(h^{-1})}, \quad u \geq 1. \tag{A.14}$$

A.4 sup-norm adaptive estimation of the stationary density

In this section, we demonstrate the effectiveness of our previous results and probabilistic tools in a concrete statistical application. We already introduced the general form of the kernel invariant density estimator in (A.11). In order to quantify the speed of convergence, we will now analyse its convergence behaviour under standard Hölder smoothness assumptions, i.e., we focus on the problem of estimating the invariant density ρ on a domain D of a Markov process $\boldsymbol{X} \in \boldsymbol{\Sigma}\cup\boldsymbol{\Theta}$ with $\rho|_D \in \mathcal{H}_D(\beta, \mathsf{L})$ (as introduced in (A.3)). For stating our statistical results, we define

$$\Phi_{d,\beta}(T) := \begin{cases} 1/\sqrt{T}, & d = 1, \\ \sqrt{\frac{\log T}{T}}, & d = 2, \\ T^{-\frac{\beta}{2\beta+d-2}}, & d \geq 3, \end{cases} \quad \text{and} \quad \Psi_{d,\beta}(T) := \begin{cases} \sqrt{\frac{\log T}{T}}, & d = 1, \\ \frac{\log T}{\sqrt{T}}, & d = 2, \\ \left(\frac{\log T}{T}\right)^{\frac{\beta}{2\beta+d-2}}, & d \geq 3. \end{cases} \tag{A.15}$$

Note that these convergence rates have already been identified in the literature as being optimal in the minimax sense, where $\Phi_{d,\beta}(\cdot)$ is associated with the pointwise or L^2 risk, while $\Psi_{d,\beta}(\cdot)$ contains an additional logarithmic factor as it inevitably arises when passing to the examination of the sup-norm risk.

For the case $d = 1$, we refer to [49, Section 4.2]. The multivariate case is less classical. For a class of multivariate reversible diffusion processes satisfying spectral gap and Nash-type

inequalities, $\Psi_{d,\beta}(\cdot)$, $d \geq 3$, has been identified in [76] as the minimax optimal convergence rate (cf. Theorem 3.4 and Theorem 3.6). For $d = 2$, a uniform upper bound of order $\Psi_{2,\beta}(\cdot)$ is stated in (1.7) in [76], which can be complemented with a lower bound obtained using similar arguments to the proof of Theorem 5 from the current reference [5]. Adapting the strategy from [76] to the simpler case of pointwise L^2 risk one can verify minimax optimality of the rates $\Phi_{d,\beta}$ for reversible diffusions, which has also been carried out explicitly in [5] for a class of exponentially ergodic diffusions under explicit boundedness and smoothness constraints on the coefficients.

Given the benchmark results mentioned above, the rates introduced in (A.15) will be referred to as "optimal" in what follows. However, we actually do not target the verification of optimality in the minimax sense, since this would require in particular to verify upper bound statements holding uniformly over entire classes of processes. Controlling the constants involved in mixing inequalities is known to be extremely challenging and can only be achieved by adding further assumptions to the processes under consideration.

All proofs of this section are given in Appendix A.III. Throughout, K denotes a $\|\cdot\|_\infty$-Lipschitz kernel of order ℓ and with Lipschitz constant L that is supported on $[-1/2, 1/2]^d$.

A.4.1 *General framework*

Depending on the concrete application, one might be interested in quantifying the accuracy of estimators in terms of different risk measures. Our findings from Section A.2 immediately imply an upper bound on the classical mean squared error at some fixed point $x \in \mathbb{R}^d$.

COROLLARY A.10. *Suppose that $\boldsymbol{X} \in \boldsymbol{\Sigma} \cup \boldsymbol{\Theta}$. For $x \in \mathbb{R}^d$ such that there exists an open neighbourhood $D \subset \mathbb{R}^d$ of x such that $\rho|_D \in \mathcal{H}_D(\beta, \mathsf{L})$, $\beta \in (0, \ell + 1]$, it holds for the kernel estimator*

$$\mathbb{E}\Big[\big(\widehat{\rho}_{h,T}(x) - \rho(x)\big)^2\Big] \in \mathsf{O}\big(\Phi^2_{d,\beta}(T)\big), \quad \text{if } h = h(T) \sim \begin{cases} T^{-1/\gamma}, & d \leq 2, \gamma \in (0, \beta], \\ T^{-1/(2\beta+d-2)}, & d \geq 3. \end{cases}$$

We now turn our focus to the technically significantly more involved problem of sup-norm adaptive invariant density estimation for processes from the class $\boldsymbol{\Sigma}$ having Hölder continuous invariant densities. We demonstrate that optimal estimation rates in any dimension are achieved by kernel estimators $\widehat{\rho}_{T,h}$ as introduced in (A.11) for a suitable choice of the bandwidth h. While in dimension $d = 1, 2$ the optimal bandwidth has the remarkable property of being independent of the (typically unknown) order β of Hölder smoothness, this is not the case in higher dimensions $d \geq 3$. In order to remove β from the bandwidth choice, we need to find a data-driven substitute for the upper bound on the bias in the balancing process. Heuristically, this is the idea behind the Lepski-type selection procedure suggested now:

1. Specify the discrete set of candidate bandwidths

$$\mathscr{H}_T \equiv \mathscr{H}_T^{(k)} := \left\{ h_l = \eta^{-l} : \ l \in \mathbb{N}_0, \ \eta^{-l} > \left(\frac{\log_{(k)} T (\log T)^5}{T} \right)^{\frac{1}{d+2}} \right\}, \quad \eta > 1 \text{ arbitrary},$$

for arbitrarily chosen $k \in \mathbb{N}$, and denote by $h_{\min}$ the smallest element in the grid $\mathscr{H}_T$. Here, $\log_{(k)} T$ denotes the k-th iterated logarithm, iteratively specified by $\log_{(k)} T := \log \log_{(k-1)} T$ and $\log_{(0)} T = T$, which is well-defined for T large enough.

2. Define $\widehat{h}_T \equiv \widehat{h}_T^{(k)}$ by letting

$$\widehat{h}_T := \max\left\{h \in \mathscr{H}_T : \left\|\widehat{\rho}_{h,T} - \widehat{\rho}_{g,T}\right\|_{L^\infty(D)} \leq \sqrt{\|\widehat{\rho}_{h_{\min},T}\|_{L^\infty(D)}}\sigma(g,T)\ \forall g \leq h,\ g \in \mathscr{H}_T\right\}, \tag{A.16}$$

where, for $\psi_d(\cdot)$ introduced in (A.7),

$$\sigma(h,T) := \frac{\log_{(k)} T(\log T)^2}{Th^d}\log(h^{-1}) + \psi_d(h^d)\sqrt{\frac{\log_{(k)} T\log(h^{-1})}{T}}, \quad h \in \mathscr{H}_T. \tag{A.17}$$

Letting $\|\cdot\|_{L^\infty(D)}$ denote the restriction of the sup-norm to a domain $D \subset \mathbb{R}^d$, we obtain the following result.

THEOREM A.11. *Suppose that $\boldsymbol{X} \in \boldsymbol{\Sigma}$. Let $D \subset \mathbb{R}^d$ be open and bounded. Suppose that $\rho|_D \in \mathcal{H}_D(\beta, \mathsf{L})$ with $\beta \in (1, \ell+1]$ for $d = 1$ and $\beta \in (2, \ell+1]$ for $d \geq 2$. Then, for any $p \geq 1$,*

$$\left(\mathbb{E}\left[\left\|\widehat{\rho}_{h,T} - \rho\right\|^p_{L^\infty(D)}\right]\right)^{1/p} \in \mathsf{O}\big(\Psi_{d,\beta}(T)\big), \quad \text{if } h = h(T) \sim \begin{cases} \log^2 T/\sqrt{T}, & d = 1, \\ \log T/T^{1/4}, & d = 2, \\ (\log T/T)^{1/(2\beta+d-2)}, & d \geq 3. \end{cases}$$

For the adaptive bandwidth scheme, let $\widehat{h}_T = \widehat{h}_T^{(k)}$ be selected according to (A.16) for some $k \in \mathbb{N}$. Then, if $\rho|_D \in \mathcal{H}_D(\beta, \mathsf{L})$ with $\beta \in (2, \ell+1]$, we have in any dimension $d \geq 3$,

$$\mathbb{E}\left[\left\|\widehat{\rho}_{\widehat{h}_T,T} - \rho\right\|_{L^\infty(D)}\right] \in \mathsf{O}\left(\left(\frac{\log_{(k)} T\log T}{T}\right)^{\frac{\beta}{2\beta+d-2}}\right). \tag{A.18}$$

While the scheme of the proof of (A.18) is close to the proof of the result on nonparametric density estimation based on i.i.d. observations in [41], the established convergence rate (recall the definition (A.15)) clearly reflects the fact that the invariant density of stochastic processes can be estimated faster than in the classical i.i.d. context. This is well-known for ergodic continuous diffusion processes (see [28, 76]), but, as we will show in the sequel, the result is fulfilled for a much larger class of stochastic processes. The additional log-factor occurring in the definition of $\Psi_{d,\beta}(\cdot)$ represents the common price to be paid when switching from the pointwise error control (described by $\Phi_{d,\beta}(\cdot)$) to bounding the sup-norm risk.

Remark A.12. (a) The conditions on the Hölder index β stated in Theorem A.11 are due to two different reasons: On the one hand, in dimension $d \leq 2$, we chose a bandwidth not depending on β which still achieves the optimal balance between bias and stochastic error. By choosing a bandwidth dependent on β (as in Corollary A.10), restrictions on β could be avoided. However, for the implementation of estimators it is advantageous to be able to choose a bandwidth independent of the typically unknown smoothness β. On the other hand, in dimension $d \geq 3$, the assumption on β is an unavoidable effect. The coupling error leaves us no other choice but to select the interval block length m_T in the decomposition of (A.11) of order $\log T$, which forces $\beta > 2$ to balance out bias and stochastic sensitivity of the estimator. We emphasize that this is not an artifact of our proof strategy since the additional log-factor also appears in the optimal Bernstein inequalities for geometrically ergodic Markov chains in [1, 51]. The restriction on β can therefore be considered as a price that must be paid for the generality of our exponential β-mixing assumption.

(b) The logarithmic gap (of arbitrary iterative order k) between the adaptive rate (see (A.18)) and the optimal rate $\Psi_{d,\beta}$ in dimension $d \geq 3$ (see (A.15)) is *not* a consequence of suboptimality of arguments used in the proof. Rather, it is a deliberate choice motivated by our desire to introduce a truly adaptive selection procedure that does not rely on the specification of obscure constants. To be more precise, a key step in the proof of the upper bound for the adaptive approach requires quantifying the concentration of the estimator $\widehat{\rho}_{h,T}$ around the variance proxy $\sigma(h,T)$ from (A.17), which is handled with the deviation inequality from Lemma A.9 involving the term $\Upsilon_{h,T}(\gamma \log T)$ (see (B.11)). If we remove the factor $\log_{(k)} T$ in the variance proxy $\sigma(h,T)$, we obtain

$$\frac{(\log T)^2}{Th^d}\log(h^{-1}) + \psi_d(h^d)\sqrt{\frac{\log(h^{-1})}{T}} \simeq \Upsilon_{h,T}(\gamma \log T).$$

In this case, an exact quantification of the constant c_γ from Lemma A.9 is mandatory, which would then be included as an additional factor in the specification of $\widehat{h}_T$ in (A.16). Together with an adjustment of the candidate bandwidths $\mathcal{H}_T$, this would allow us to close the logarithmic gap and hence obtain optimal rates for the adaptive procedure.

However, c_γ is of the form $\gamma \times C(D, L, \kappa, c_\kappa, c_2)$—where we recall that c_κ, κ determine the mixing coefficient and c_2 is a constant appearing in the heat kernel bound from Assumption ($\mathscr{A}1$)—and therefore can only be bounded with explicit knowledge/assumptions on the process. We avoid this fundamental problem in our procedure to not shift the problem from unknown exact smoothness to unknown exact ergodic and small time behaviour, with the price to be paid being a logarithmic loss. In this regard, our approach differs from the bandwidth selection procedure for the L^2 risk in [4], which relies on the choice of a "sufficiently large" constant k that cannot be exactly specified or efficiently chosen in a data-driven way.

Our previous results rely on the very general conditions ($\mathscr{A}0$) and ($\mathscr{A}1$) as well as assumptions related to the speed of convergence to the invariant distribution, ($\mathscr{A}2$) and ($\mathscr{A}\beta$). For statistical purposes, however, it is essential to derive results under conditions on the coefficients of the underlying process as easily verifiable as possible. For this reason, the next two subsections are devoted to investigating specific classes of jump diffusion processes and explicit conditions on their underlying characteristics such that the above assumptions are satisfied and hence statistical conclusions can be drawn from our general theory.

A.4.2 *Example: Lévy-driven Ornstein–Uhlenbeck processes*

As a first example, we discuss estimation rates of d-dimensional Lévy-driven Ornstein–Uhlenbeck processes as representatives of Lévy-driven jump diffusions with unbounded drift coefficient by establishing assumptions on the characteristics of the Lévy process that guarantee $\mathbf{X} \in \boldsymbol{\Sigma} \cup \boldsymbol{\Theta}$.

Let $\mathbf{Z}$ be a d-dimensional Lévy process with generating triplet (a, Q, ν), where $a \in \mathbb{R}^d$, $Q \in \mathbb{R}^{d\times d}$ is a symmetric positive semidefinite matrix and ν is a measure on $\mathbb{R}^d$ satisfying $\nu(\{0\}) = 0$ and $\int_{\mathbb{R}^d}(1 \wedge \|x\|^2)\,\nu(\mathrm{d}x) < \infty$ such that $\mathbb{E}^0[\exp(\mathrm{i}\langle Z_1, \theta\rangle)] = \exp(\psi(\theta))$ with

$$\psi(\theta) = \mathrm{i}\langle a, \theta\rangle - \frac{1}{2}\langle Q\theta, \theta\rangle + \int_{\mathbb{R}^d\setminus\{0\}}\left(\mathrm{e}^{\mathrm{i}\langle x,\theta\rangle} - 1 - \mathrm{i}\langle x, \theta\rangle \mathbb{1}_{B(0,1)}(x)\right)\nu(\mathrm{d}x), \quad \theta \in \mathbb{R}^d,$$

where $B(0,1) = \{x \in \mathbb{R}^d : \|x\| < 1\}$. Then, given some matrix $B \in \mathbb{R}^{d\times d}$, a Lévy-driven Ornstein–Uhlenbeck process $\mathbf{X}$ is a solution to the SDE

$$\mathrm{d}X_t = -BX_t\,\mathrm{d}t + \mathrm{d}Z_t,$$

given by

$$X_t = \mathrm{e}^{-tB}X_0 + \int_0^t \mathrm{e}^{-(t-s)B}\,\mathrm{d}Z_s, \quad t \geq 0.$$

We suppose that the real parts of all eigenvalues of B are positive, implying that $\mathrm{e}^{-tB} \to \mathbf{0}_{d\times d}$ as $t \to \infty$, and assume the following moment condition

$$\int_{\|z\|>2} \log\|z\|\,\nu(\mathrm{d}z) < \infty. \tag{A.19}$$

Then, $\boldsymbol{X}$ is a Markov process on $\mathbb{R}^d$ with invariant distribution μ such that

$$\{\mathcal{F}\mu\}(u) = \exp\Big(\int_0^\infty \psi\big(\mathrm{e}^{-sB^\top}u\big)\,\mathrm{d}s\Big), \quad u \in \mathbb{R}^d,$$
$$\text{and } \varphi_{X_t}^x(u) = \exp\Big(\mathrm{i}\langle x, \mathrm{e}^{-tB^\top}u\rangle + \int_0^t \psi\big(\mathrm{e}^{-sB^\top}u\big)\,\mathrm{d}s\Big), \quad u, x \in \mathbb{R}^d, t > 0,$$

see [72, Theorem 3.1, Theorem 4.1]. Let us now introduce the following conditions.

(𝒪1) $\operatorname{rank}(Q) = d$;

(𝒪2) $\int_{\{\|x\|>1\}} \|x\|^p\,\nu(\mathrm{d}x) < \infty$ for some $p > 0$;

(𝒪3) $\int_{\{\|x\|>1\}} (\log\|x\|)^\alpha\,\nu(\mathrm{d}x) < \infty$ for some $\alpha > 2$.

These assumptions are borrowed from [57], [55] and [48], where (sub-)exponential ergodicity and exponential β-mixing of OU-processes are investigated. (𝒪1) guarantees the strong Feller property of $\boldsymbol{X}$ and the existence of a $\mathcal{C}_b^\infty$-density for $P_t(x,\cdot)$, $x \in \mathbb{R}^d$ ([57, Theorem 3.1]). Similar arguments to the ones in [57, Theorem 3.2] also show that under (𝒪1), μ admits a $\mathcal{C}_b^\infty$-density ρ. (𝒪2) and (𝒪3) are moment assumptions on $\mathbf{Z}$, where (𝒪3) in absence of (𝒪2) corresponds to an extremely heavy-tailed distribution and represents a minor strengthening of the necessary and sufficient criterion (A.19) for stationarity of $\boldsymbol{X}$.

Based on the results from [48, 55, 57] together with our investigations in Sections A.2 and A.4.1, we can obtain the following result which is proved in Appendix A.III.

Theorem A.13. *Suppose that (𝒪1) holds. Then, in any dimension $d \in \mathbb{N}$, (𝒮1) holds with*

$$\sup_{x,y\in\mathbb{R}^d} p_t(x,y) \lesssim t^{-d/2}, \quad t \in (0,1]. \tag{A.20}$$

If, additionally,

(i) (𝒪2) holds for some $p > 0$, then, for any $d \geq 1$, $\boldsymbol{X} \in \boldsymbol{\Sigma} \cap \boldsymbol{\Theta}$;

(ii) (𝒪3) holds, then, for $d = 1$, $\boldsymbol{X} \in \boldsymbol{\Theta}$.

Let $d \geq 1$ in scenario (i) and $d = 1$ in scenario (ii). Then, for arbitrary $\beta \in (0, \ell + 1]$, we obtain for any $x \in \mathbb{R}^d$ that

$$\mathbb{E}\Big[\big(\widehat{\rho}_{h,T}(x) - \rho(x)\big)^2\Big] \in \mathrm{O}\big(\Phi^2_{d,\beta}(T)\big), \quad \text{if } h = h(T) \sim \begin{cases} T^{-1}, & d \leq 2, \\ T^{-1/(2\beta+d-2)}, & d \geq 3, \end{cases}$$

and for any bounded, open domain $D \subset \mathbb{R}^d$ and $p \geq 1$ that in scenario (i)

$$\mathbb{E}\Big[\big\|\widehat{\rho}_{h,T} - \rho\big\|^p_{L^\infty(D)}\Big]^{1/p} \in \mathrm{O}\big(\Psi_{d,\beta}(T)\big), \quad \text{if } h = h(T) \sim \begin{cases} \log^2 T/\sqrt{T}, & d = 1, \\ \log T/T^{1/4}, & d = 2, \\ (\log T/T)^{1/(2\beta+d-2)}, & d \geq 3. \end{cases}$$

Remark A.14. (a) Since we can choose $\beta > 0$ arbitrarily large, we make the remarkable observation that, in the scenarios described above, for any $\varepsilon > 0$ we can obtain the almost superoptimal rates $T^{-(1+\varepsilon)}$ and $(\log T/T)^{1/(2(1+\varepsilon))}$ in any dimension $d \geq 3$ for the pointwise L^2 and sup-norm risk, respectively. Moreover, in any dimension, an adaptive choice of the bandwidth is not necessary.

(b) The result demonstrates that even under much less stringent assumptions (logarithmic moments and unbounded drift) compared to the class of processes studied in the next section, there are examples of jump diffusions with Lévy-driven jump part for which optimal estimation results are feasible. It is therefore an interesting question for future research to determine more general coefficient assumptions based on a linear growth condition on the drift that yield optimal estimation properties.

A.4.3 *Example: Non-reversible Lévy-driven jump diffusion processes*

The goal of this section is to show that solutions of the d-dimensional SDE, $d \in \mathbb{N}$,

$$X_t = X_0 + \int_0^t b(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}W_s + \int_0^t \int_{\mathbb{R}^d} \gamma(X_{s-}) z\, \widetilde{N}(\mathrm{d}s, \mathrm{d}z) \tag{A.21}$$

satisfy assumptions (𝒜0), (𝒜1) and (𝒜β) which then allows using Theorem A.11 to bound the sup-norm risk of the kernel invariant density estimator. Here, $\sigma\colon \mathbb{R}^d \to \mathbb{R}^{d\times d}$, $\gamma\colon \mathbb{R}^d \to \mathbb{R}^{d\times d}$, $b\colon \mathbb{R}^d \to \mathbb{R}^d$, $\mathbf{W}$ denotes an $\mathbb{R}^d$-valued Brownian motion, N is a Poisson random measure on $[0,\infty) \times \mathbb{R}^d\backslash\{0\}$ with intensity measure $\mu(\mathrm{d}s, \mathrm{d}z) = \mathrm{d}s \otimes \nu(\mathrm{d}z)$, and $\widetilde{N}$ denotes the compensated Poisson random measure. Moreover, ν is a Lévy measure and we assume that N, W and X_0 are independent. Note that, if $z \mapsto \gamma(x)z$ is in $L^1(\mathbb{R}^d\backslash B_1, \nu)$ for all $x \in \mathbb{R}^d$, (A.21) is equivalent to

$$\begin{aligned} X_t = X_0 &+ \int_0^t b^*(X_s)\,\mathrm{d}s + \int_0^t \sigma(X_s)\,\mathrm{d}W_s \\ &+ \int_0^t \int_{\|z\|\leq 1} \gamma(X_{s-}) z\, \widetilde{N}(\mathrm{d}s, \mathrm{d}z) + \int_0^t \int_{\|z\|>1} \gamma(X_{s-}) z\, N(\mathrm{d}s, \mathrm{d}z), \end{aligned} \tag{A.22}$$

with $b^*(x) := b(x) - \int_{\|z\|>1} \gamma(x) z\, \nu(dz)$ and $B_1 := \{z \in \mathbb{R}^d : \|z\| \leq 1\}$. We assume the following.

(Ƒ1) The functions b, γ, σ are globally Lipschitz continuous, b and γ are bounded, and, for $\mathbb{I}_{d\times d}$ denoting the $d \times d$-identity matrix, there exists a constant $c \geq 1$ such that

$$c^{-1}\mathbb{I}_{d\times d} \leq \sigma\sigma^\top \leq c\mathbb{I}_{d\times d},$$

where the ordering is in the sense of Loewner for positive semi-definite matrices.

(Ƒ2) ν is absolutely continuous wrt the Lebesgue measure and, for an $\alpha \in (0, 2)$,

$$(x, z) \mapsto \|\gamma(x)z\|^{d+\alpha}\nu(z)$$

is bounded and measurable, where, by abuse of notation, we denoted the density of ν also by ν. Furthermore, if $\alpha = 1$,

$$\int_{r<\|\gamma(x)z\|\leq R} \gamma(x)z\,\nu(\mathrm{d}z) = 0, \quad \text{for any } 0 < r < R < \infty,\ x \in \mathbb{R}^d.$$

(Ƒ3) There exist $c_1, c_2 > 0$ and $\eta_0 > 0$ such that

$$\langle x, b(x)\rangle \leq -c_1\|x\|, \quad \forall x : \|x\| \geq c_2, \quad \text{and} \quad \int_{\mathbb{R}^d} \|z\|^2 \mathrm{e}^{\eta_0\|z\|}\nu(\mathrm{d}z) < \infty.$$

In [4], the authors also investigate L^2 invariant density estimation for jump diffusions and use a similar approach for formulating requirements on the diffusion coefficients which imply their respective heat kernel bound and mixture assumptions. The conditions however are more restrictive and, in particular, the case of continuous diffusions cannot be handled within their framework since it requires $\mathrm{supp}(\nu) = \mathbb{R}^d$ and $\det(\gamma(x)) > c$ for some constant $c > 0$ and all $x \in \mathbb{R}^d$. In [6], the authors improve the L^2 rate for dimension $d = 1$ from [4] to the parametric rate $1/T$ by imposing an additional smoothness restriction on the jump measure. Our main contribution in this section is to show that under the less stringent assumptions above, optimal convergence rates can be achieved not only wrt the L^2 risk but even wrt sup-norm risk in any dimension. In particular, reversible diffusion processes satisfying the drift and dispersion matrix assumptions fall into the above process class, such that the minimax lower bounds from the literature (see section A.4.1) suggest sharpness of our estimation rates.

Note that (Ƒ1) and (Ƒ3) directly imply $\gamma(x)z \in L^1(\mathbb{R}^d\backslash B_1, \nu)$, so (A.21) and (A.22) are equivalent. The subsequent lemma shows that, under the given assumptions, there exists a pathwise unique strong solution for (A.21) and that the conditions of Corollary 1.5 of [21] hold, implying the heat kernel bound (A.23). All proofs can be found in Appendix A.III.

Lemma A.15. *Let (Ƒ1)–(Ƒ3) hold. Then,* (A.21) *admits a càdlàg, non-explosive, pathwise unique, strong solution possessing the strong Markov property, and the assumptions* ($\mathbf{H}^\alpha$) *and* ($\mathbf{H}^\kappa$) *of [21] hold.*

Let $\boldsymbol{X}$ be the unique solution of (A.21) described in Lemma A.15.

Corollary A.16. *Let (Ƒ1)–(Ƒ3) hold. Then, transition densities* $(p_t)_{t>0}$ *exist and there are constants* $C, \lambda > 1$ *such that the solution* $\boldsymbol{X}$ *of* (A.21) *satisfies the following heat kernel estimate for all* $x, y \in \mathbb{R}^d, 0 < t \leq 1$,

$$\begin{aligned} &C^{-1}(t^{-d/2}\exp(-\lambda\|x-y\|^2/t) + (\inf_{x\in\mathbb{R}^d}\operatorname{ess\,inf}_{z\in\mathbb{R}^d}\kappa_\alpha(x,z))t(\|x-y\|+t^{1/2})^{-d-\alpha}) \\ &\quad\leq p_t(x,y) \leq C(t^{-d/2}\exp(-\|x-y\|^2/(\lambda t)) + \|\kappa_\alpha\|_\infty t(\|x-y\|+t^{1/2})^{-d-\alpha}), \end{aligned} \tag{A.23}$$

where $\kappa_\alpha(x,z) = \|\gamma(x)z\|^{d+\alpha}\nu(z)$. *In particular, assumption (𝒜1) is satisfied.*

Now our goal is to show that the solution $\boldsymbol{X}$ of (A.21) fulfills the fundamental assumption (𝒜0+) and exponential ergodicity along with the mixing property (𝒜β). First, observe that (ℱ1) implies that $b \in \mathcal{C}_b(\mathbb{R}^d;\mathbb{R}^d)$ and $\sigma,\gamma \in \mathcal{C}_b(\mathbb{R}^d;\mathbb{R}^{d\times d})$ and hence Theorem 6.7.4 in [7] guarantees that the unique càdlàg Markov process $\boldsymbol{X}$ solving (A.21) is Feller and therefore Borel right. Further, Corollary A.16 in particular implies the existence of bounded transition densities and thus, by Lemma A.4, it suffices to show the existence of an invariant distribution. This will be done as a byproduct while proving exponential ergodicity and the exponential mixing property (𝒜β). For this, we will employ results of Masuda [55] which are again based on the theory of stability of continuous-time Markov processes of Meyn and Tweedie [62]. These lead us to the following proposition.

Proposition A.17. *Grant assumptions (ℱ1)–(ℱ3). Then, an invariant distribution exists,* $\boldsymbol{X}$ *is* V*-exponentially ergodic with locally bounded* V *and the process* $\boldsymbol{X}$ *started in the invariant distribution* μ *is exponentially* β*-mixing.*

Gathering the results of Corollary A.16 and Proposition A.17 and employing Lemma A.3 now yields that (𝒜0)–(𝒜2) and (𝒜β) are fulfilled for the solution $\mathbf{X}$ of (A.21), i.e., $\mathbf{X} \in \boldsymbol{\Sigma} \cap \boldsymbol{\Theta}$. In particular, the results from Section A.4.1 can be applied.

Theorem A.18. *Let* $D \subset \mathbb{R}^d$ *be open and bounded and assume (ℱ1)–(ℱ3). If* $\rho|_D \in \mathcal{H}_D(\beta,\mathsf{L})$ *with* $\beta \in (1,\ell+1]$ *for* $d=1$ *and* $\beta \in (2,\ell+1]$ *for* $d \geq 2$*, then, the* sup*-norm risk of the kernel estimator defined in* (A.11) *is of order*

$$\mathbb{E}\Big[\|\widehat{\rho}_{h,T}-\rho\|^p_{L^\infty(D)}\Big]^{1/p} \in \begin{cases} \mathsf{O}(\sqrt{\log T/T}), & d=1,\\ \mathsf{O}(\log T/\sqrt{T}), & d=2,\\ \mathsf{O}\big((\log T/T)^{\beta/(2\beta+d-2)}\big), & d\geq 3,\end{cases} \quad \text{if } h \sim \begin{cases} \log^2 T/\sqrt{T}, & d=1,\\ \log T/T^{1/4}, & d=2,\\ (\log T/T)^{-1/(2\beta+d-2)}, & d\geq 3.\end{cases}$$

for any $p \geq 1$. *If* $\widehat{h}_T \equiv \widehat{h}^{(k)}_T$ *is chosen adaptively according to* (A.16) *for some* $k \in \mathbb{N}$*, then for any* $d \geq 3$*,*

$$\mathbb{E}\Big[\big\|\widehat{\rho}_{\widehat{h}_T}-\rho\big\|_{L^\infty(D)}\Big] \in \mathsf{O}\left(\left(\frac{\log_{(k)} T \log T}{T}\right)^{\beta/(2\beta+d-2)}\right).$$

Moreover, for any $x \in \mathbb{R}^d$ *such that* $\rho|_D \in \mathcal{H}_D(\beta,\mathsf{L})$ *for some* $\beta \in (0,\ell+1]$ *and a neighborhood* D *of* x*, we have the pointwise* L^2 *risk estimate*

$$\mathbb{E}\Big[\big(\widehat{\rho}_{h,T}(x)-\rho(x)\big)^2\Big] \in \begin{cases} \mathsf{O}(1/T), & d=1,\\ \mathsf{O}(\log T/T), & d=2,\\ \mathsf{O}\big(T^{-2\beta/(2\beta+d-2)}\big), & d\geq 3,\end{cases} \quad \text{if } h \sim \begin{cases} T^{-1/\gamma}, & d\leq 2,\gamma\leq\beta,\\ T^{-1/(2\beta+d-2)}, & d\geq 3.\end{cases}$$

Appendices

A.I Supplements of Section A.2

A.I.1 *Assumption (A2) and the exponential β-mixing property*

As in the rest of the paper, we will assume in this section that $\boldsymbol{X}$ is a Borel right Markov process with unique invariant distribution μ possessing a Lebesgue density ρ. Let us start by collecting some important definitions in the realm of stability theory of Markov processes. We say that $\mathbf{X}$ is *ψ-irreducible* for some σ-finite measure ψ on its state space if $\psi(B) > 0$ for some Borel set B implies

$$U(x, B) := \int_0^\infty P_t(x, B)\,\mathrm{d}t = \mathbb{E}^x[\eta_B] > 0$$

for any $x \in \mathbb{R}^d$, i.e., the expected sojourn time η_B of $\mathbf{X}$ in B (or, equivalently, the potential of B), where $\eta_B = \int_0^\infty \mathbb{1}_{\{X_t \in B\}}\,\mathrm{d}t$, when $\mathbf{X}$ is started in an arbitrary state is strictly positive. If for $B \in \mathcal{B}(\mathbb{R}^d)$, $\psi(B) > 0$ even implies $\mathbb{P}^x(\eta_B = \infty) = 1$ for any $x \in \mathbb{R}^d$, we say that $\mathbf{X}$ is *Harris recurrent* and that ψ is a Harris measure. Harris recurrent Markov processes having an invariant distribution (which is unique in this case) are called positive Harris recurrent. A Borel set C is called *small* if there exists $T > 0$ and a non-trivial measure ν on the state space such that

$$P_T(x, \cdot) \geq \nu(\cdot), \quad x \in C.$$

Petite sets generalize the notion of small sets. We call a Borel set C petite if there exists a sampling distribution a on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ and a non-trivial measure ν_a on the state space s.t.

$$K_a(x, \cdot) := \int_0^\infty P_t(x, \cdot)\,a(\mathrm{d}t) \geq \nu_a(\cdot), \quad x \in C,$$

i.e., small sets are petite sets with sampling distribution $a = \delta_T$ for some $T > 0$. All three concepts have obvious counterparts for discrete-time chains. If moreover the ψ-irreducible process $\mathbf{X}$ possesses a small set C such that $\psi(C) > 0$ and there is $T > 0$ such that $P_t(x, C) > 0$, $\forall x \in C$, $t \geq T$, we say that $\mathbf{X}$ is *aperiodic*.

These notions are of central importance in the theory of stability of Markovian processes on general state spaces in both discrete as well as continuous time. In discrete time, the existence of small sets allows the construction of a related Markov chain via the technique of Nummelin splitting, which shares the same stability properties with the original chain but possesses an atom. This in turn allows to transfer well-known reasoning in Markov chain theory on countable state spaces to the general state space situation with renewal arguments. With the Meyn and Tweedie approach to stability of continuous-time Markov processes, which heavily involves the aforementioned concept of aperiodicity, we can then infer stability properties through sampled chains, generalizing discrete-time results to continuous time. For a complete picture in discrete time, we refer to the monograph [60]. Continuous-time theory was developed in the 1990s in a series of papers [35, 59, 61, 62] and many other subsequent contributions.

We see that these concepts are quite natural when we aim to infer stability of general Markov processes, and we need no more than irreducibility as well as the property that compact sets are small together with exponential decay in (A.5) to infer exponential β-mixing of the process.

Proposition A.19. *Suppose that* $\mathbf{X}$ *is* ψ*-irreducible and that every compact set* $\mathcal{S} \subset \mathbb{R}^d$ *is small. Moreover, let (A.12) be satisfied with* $\Lambda = \mathbb{R}^d$ *and*

$$r_{\mathcal{S}}(t) := C_{\mathcal{S}} \mathrm{e}^{-\kappa_{\mathcal{S}} t}, \quad t > 0, \tag{A.24}$$

with constants $C_{\mathcal{S}}, \kappa_{\mathcal{S}} > 0$*. Then,* $\mathbf{X}$ *is exponentially* β*-mixing.*

Proof. Let $\mathcal{S} \subset \mathbb{R}^d$ be compact such that $\boldsymbol{\lambda}(\mathcal{S}) > 0$. Since $\mathbb{R}^d$ can be covered by countably many compact sets and the irreducibility measure ψ is σ-finite, we can also assume that $\psi(\mathcal{S}) > 0$ and $\mu(\mathcal{S}) > 0$. Letting $(P_t)_{t \geq 0}$ denote the semigroup associated to $\mathbf{X}$, we obtain from (A.5) and (A.24) that, for any $x \in \mathcal{S}$ and $t > 0$,

$$|P_t(x, \mathcal{S}) - \mu(\mathcal{S})| \leq \int_{\mathcal{S}} |p_t(x, y) - \rho(y)| \, \mathrm{d}y \leq C_{\mathcal{S}} \mathrm{e}^{-\kappa_{\mathcal{S}} t} \boldsymbol{\lambda}(\mathcal{S}) = \widetilde{C}_{\mathcal{S}} \mathrm{e}^{-\kappa_{\mathcal{S}} t},$$

with $\widetilde{C}_{\mathcal{S}} = C_{\mathcal{S}} \boldsymbol{\lambda}(\mathcal{S})$. Since $\mu(\mathcal{S}) > 0$, this implies in particular that there exists $T(\mathcal{S}) > 0$ such that $P_t(x, \mathcal{S}) > 0$ for all $t \geq T(\mathcal{S})$ and $x \in \mathcal{S}$. Since $\mathcal{S}$ is small by assumption, it follows that $\mathbf{X}$ is aperiodic. Hence, by Theorem 5.3 in [35] and the remarks thereafter, there exists (a) an extended real-valued measurable function $V \geq 1$ such that, for some $T > 0$, we have

$$P_T V(x) \leq \lambda V(x) + b \mathbf{1}_{\Theta} \tag{A.25}$$

for some $0 < \lambda < 1$, $b \geq 0$ and a small set $\Theta \in \mathcal{B}(\mathbb{R}^d)$ and (b) a set $S_V \subset \{V < \infty\}$, which is full and absorbing—that is, $\mu(S_V) = 1$ and $P_T(x, S_V) = 1$ for any $x \in S_V$—such that $\mathbf{X}$ restricted to S_V is exponentially ergodic in the sense

$$\|P_t(x, \cdot) - \mu\|_{\mathrm{TV}} \leq C V(x) \mathrm{e}^{-\kappa t}, \quad x \in S_V, \tag{A.26}$$

for some constants $C, \kappa > 0$. Noting that (A.25) implies

$$\Delta \widetilde{V} \leq -V + \frac{b}{1 - \lambda} \mathbf{1}_{\Theta}$$

with $\widetilde{V} = V/(1 - \lambda) \geq 0$ and $\Delta := P_T - \mathbb{I}$, it follows from Theorem 14.0.1 in [60] that $\mu(V) < \infty$. The claim on exponential β-mixing of the process now follows from (A.26) since

$$\int_{\mathbb{R}^d} \|P_t(x, \cdot) - \mu\|_{\mathrm{TV}} \, \mu(\mathrm{d}x) = \int_{S_V} \|P_t(x, \cdot) - \mu\|_{\mathrm{TV}} \, \mu(\mathrm{d}x) \leq C \mathrm{e}^{-\kappa t} \int_{S_V} V(x) \, \mu(\mathrm{d}x) = \widetilde{C} \mathrm{e}^{-\kappa t},$$

for any $t \geq 0$, where finiteness of $\widetilde{C} = C\mu(V)$ was discussed above and for the first equality we used that S_V is full. ∎

Compactness of small sets can be inferred for a quite general class of Markov processes. We say that $\mathbf{X}$ is a T-process if there exists a non-trivial continuous component for some sampled chain, i.e., there exists a sampling distribution a on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ and a non-trivial, lower semi-continuous kernel T on the state space s.t.

$$K_a(x, \cdot) \geq T(x, \cdot), \quad x \in \mathbb{R}^d.$$

Many processes in applied probability can be shown to be T-processes such as price processes driven by Lévy risk and return processes [67], certain piecewise deterministic Markov processes

used for MCMC [11] or queuing networks [34]. Moreover, any open set irreducible weak $\mathcal{C}_b$-Feller process is a T-process (cf. [81, Theorem 7.1]). Markov processes having the strong Feller property—that is, the semigroup satisfies $P_t\mathcal{B}_b(\mathbb{R}^d) \subset \mathcal{C}_b(\mathbb{R}^d)$ for all $t \geq 0$—are trivially T-processes, since any operator P_t is a continuous component for itself. Here, we denoted by $\mathcal{C}_b(\mathbb{R}^d)$ the family of bounded, continuous functions on $\mathbb{R}^d$ and by $\mathcal{B}_b(\mathbb{R}^d)$ the family of bounded Borel functions on $\mathbb{R}^d$. The strength of Markov processes with the strong Feller property—and T-processes as a generalization of such processes—comes from making possible to connect distributional properties of the Markov process induced by the semigroup and topological properties of the state space, thus allowing to use knowledge of the topology to infer strong stability results of the Markov process. Classical examples of Markov processes with the strong Feller property are Lévy processes with absolutely continuous semigroup with respect to the Lebesgue measure [45, Theorem 2.2], diffusion processes with hypoelliptic Fisk–Stratonovich-type generator [46, Lemma 5.1], diffusion processes on Hilbert spaces under appropriate assumptions on the coefficients [68, Theorem 1.2], or solutions of different classes of parabolic SPDEs [25, 26, 38, 54]. More recently, the strong Feller property was discussed for switching (jump-)diffusions [86, 88], for jump-diffusions with non-Lipschitz coefficients [87], or Markov semigroups generated by singular SPDEs such as the KPZ equation in Hairer and Mattingly [44]. For an account discussing conditions for which (weak) $\mathcal{C}_b$-Feller processes are even strong Feller, we refer to Schilling and Wang [73].

Let us now infer the exponential β-mixing property for T-processes given exponential decay in (A.5) and, as a natural mixing requirement, ergodicity in the sense of total variation convergence to the invariant distribution, i.e., $\|P_t(x,\cdot) - \mu\|_{\mathrm{TV}} \underset{t\to\infty}{\longrightarrow} 0$, $\forall x \in \mathbb{R}^d$. Note that indeed, dominated convergence shows that any stationary, ergodic Markov process is β-mixing.

PROPOSITION A.20. *Let* $\mathbf{X}$ *be an ergodic* T*-process such that (A2) is satisfied for* r_S *given as in* (A.24) *and* $\Lambda = \mathbb{R}^d$*. Then,* $\mathbf{X}$ *is positive Harris recurrent, every compact set is small and* $\mathbf{X}$ *is exponentially* β*-mixing.*

Proof. For the exponential β-mixing property, it suffices to check that every compact set is small by Proposition A.19, since ergodicity clearly implies μ-irreducibility of $\boldsymbol{X}$. We prove this property together with positive Harris recurrence at once. To this end, for a given $\varepsilon > 0$, choose a compact set $C \subset \mathbb{R}^d$ such that $\mu(C) \geq 1 - \varepsilon$. Then, for fixed $x \in \mathbb{R}^d$, ergodicity guarantees that $\lim_{t\to\infty} \mathbb{P}^x(X_t \in C) \geq 1 - \varepsilon$, and hence $\boldsymbol{X}$ is bounded in probability on average as defined on p. 495 of [61]. Since $\boldsymbol{X}$ is an irreducible T-process, Theorem 3.2 and Theorem 4.1 of the same reference yield Harris recurrence and petiteness of compact sets. It remains to show that small and petite sets coincide for the given process. The reverse implication of Theorem 6.1 in [61] guarantees that there exists an irreducible skeleton $\boldsymbol{X}^\Delta = (X_{n\Delta})_{n\in\mathbb{N}_0}$ for some $\Delta > 0$ thanks to ergodicity and positive Harris recurrence of $\mathbf{X}$. Proposition 6.1 in [61] therefore implies equivalence of small and petite sets, which finishes the proof. ∎

A.I.2 *Proofs for Section A.2*

Proof of Proposition A.1. Without loss of generality, let $T \geq 1$ be fixed. Then, using the Markov property and the invariance of μ, for any $\delta \in [0,1]$,

$$\begin{aligned}
\mathrm{Var}\left(\int_0^T f(X_s)\,\mathrm{d}s\right) &= \mathbb{E}\left[\left(\int_0^T (f(X_s) - \mathbb{E}f(X_0))\,\mathrm{d}s\right)^2\right] \\
&= 2\mathbb{E}\left[\int_0^T \int_0^u (f(X_0) - \mathbb{E}f(X_0))(f(X_{u-s}) - \mathbb{E}f(X_0))\,\mathrm{d}s\,\mathrm{d}u\right] \\
&= 2\int_0^T \int_0^u \Big(\mathbb{E}[f(X_0)f(X_{u-s})] - (\mathbb{E}f(X_0))^2\Big)\,\mathrm{d}s\,\mathrm{d}u \\
&= 2\int_0^T \int_0^u \left[\iint_{\mathbb{R}^{d\times d}} f(x)f(y)p_{u-s}(x,y)\,\mathrm{d}y\mu(\mathrm{d}x) - \int f(x)\,\mu(\mathrm{d}x)\int f(y)\rho(y)\,\mathrm{d}y\right]\mathrm{d}s\,\mathrm{d}u \\
&= 2\int_0^T \int_0^u \int_\Lambda \int_{\mathbb{R}^d} f(x)f(y)(p_{u-s}(x,y) - \rho(y))\,\mathrm{d}y\,\mu(\mathrm{d}x)\,\mathrm{d}s\,\mathrm{d}u \\
&= 2(\mathcal{I}(0,\delta) + \mathcal{I}(\delta,1) + \mathcal{I}(1,T)),
\end{aligned}$$

with (substituting $v = u - s$)

$$\mathcal{I}(a,b) := \int_a^b (T-v)\int_{\mathbb{R}^d}\int_\Lambda f(x)f(y)(p_v(x,y) - \rho(y))\,\mu(\mathrm{d}x)\,\mathrm{d}y\,\mathrm{d}v, \quad 0 \leq a < b \leq T.$$

It follows from the assumption on the convergence of the transition density in (A.5) that

$$\begin{aligned}
\mathcal{I}(1,T) &\leq \int_1^T (T-v)\sup_{x\in\mathcal{S}\cap\Lambda, y\in\mathcal{S}} |p_v(x,y) - \rho(y)|\,\mathrm{d}v \iint_{\mathbb{R}^d\times\mathbb{R}^d} f(x)f(y)\,\mu(\mathrm{d}x)\,\mathrm{d}y \\
&\leq T\|f\|_\infty^2 \lambda(\mathcal{S})\mu(\mathcal{S})\int_1^T r_{\mathcal{S}}(v)\,\mathrm{d}v \leq c_{\mathcal{S}} T\|f\|_\infty^2\lambda(\mathcal{S})\mu(\mathcal{S}).
\end{aligned}$$

It remains to consider the first parts of the integral. We now restrict to dimension $d \geq 3$; the remaining cases are handled with analogous arguments. Note first that

$$\mathcal{I}(0,\delta) \leq T\|f\|_\infty^2 \int_0^\delta \iint_{\mathcal{S}\times\mathbb{R}^d} p_v(x,y)\,\mu(\mathrm{d}x)\,\mathrm{d}y\,\mathrm{d}v = T\|f\|_\infty^2\mu(\mathcal{S})\delta. \tag{A.27}$$

On the other hand, the heat kernel bound (A.4) gives for any $x, y \in \mathbb{R}^d$,

$$\int_\delta^1 p_v(x,y)\,\mathrm{d}v \leq c_2\int_\delta^1 v^{-d/2}\,\mathrm{d}v = c_2'\delta^{1-d/2},$$

where $c_2' = 2/(d-2)c_2$. Letting $\delta = (\lambda(S))^{2/d}$ and exploiting that $\lambda(\mathcal{S}) < 1$, it follows

$$\mathcal{I}(\delta,1) \leq T\|f\|_\infty^2\int_\delta^1 \iint_{\mathcal{S}^2} p_v(x,y)\,\mu(\mathrm{d}x)\,\mathrm{d}y\,\mathrm{d}v \leq c_2' T\|f\|_\infty^2\mu(\mathcal{S})(\lambda(\mathcal{S}))^{\frac{2}{d}}.$$

■

Proof of Lemma A.3. By the semigroup property of $(P_t)_{t\geq 0}$ and invariance of μ we have for any $t > 1$ and $y \in \mathbb{R}^d$ and μ-a.e. $x \in \mathbb{R}^d$,

$$\begin{aligned}|p_t(x,y) - \rho(y)| &\leq \int_{\mathbb{R}^d} p_1(z,y)|p_{t-1}(x,z) - \rho(z)|\,\mathrm{d}z \\ &\leq \|p_1\|_\infty \int_{\mathbb{R}^d} |p_{t-1}(x,z) - \rho(z)|\,\mathrm{d}z \\ &= 2\|p_1\|_\infty \|P_{t-1}(x,\cdot) - \mu\|_{\mathrm{TV}} \leq 2\|p_1\|_\infty CV(x)\xi(t-1),\end{aligned}$$

where the equality follows from Scheffé's theorem, see [80, Lemma 2.1]. Thus, for any compact set $\mathcal{S}$ and $r_{\mathcal{S}}(t) = 2C\|p_1\|_\infty \sup_{x\in\mathcal{S}\cap\Lambda} V(x)\xi(t-1)$, it follows that

$$\int_1^\infty \sup_{x\in\mathcal{S}\cap\Lambda, y\in\mathcal{S}} |p_t(x,y) - \rho(y)|\,\mathrm{d}t \leq \int_1^\infty r_{\mathcal{S}}(t)\,\mathrm{d}t \lesssim \sup_{x\in\mathcal{S}\cap\Lambda} V(x) \int_0^\infty \xi(t) < \infty,$$

by local boundedness of $V\mathbb{1}_\Lambda$ and the convergence assumption on ξ, which yields (𝒮2). ∎

Proof of Lemma A.4. Let $B \in \mathcal{B}(\mathbb{R}^d)$ such that $\boldsymbol{\lambda}(B) = 0$. Then, it holds that

$$\mu(B) = \int_{\mathbb{R}^d} \int_B p_\Delta(x,y)\,\mathrm{d}y\,\mu(\mathrm{d}x) = 0,$$

which yields the existence of a Lebesgue density ρ of μ by the Radon–Nikodym theorem. Now, let $B \in \mathcal{B}(\mathbb{R}^d)$ such that $\boldsymbol{\lambda}(B) > 0$. Arguing as above and using boundedness of p_Δ, we get

$$\frac{\int_B \rho(x)\,\boldsymbol{\lambda}(\mathrm{d}x)}{\boldsymbol{\lambda}(B)} \leq c.$$

Now the Lebesgue differentiation theorem yields $\operatorname{ess\,sup} \rho \leq c$, and defining

$$\rho_b(x) = \rho(x)\mathbf{1}_{[0,c]}(\rho(x)), \quad x \in \mathbb{R}^d,$$

we have $\rho = \rho_b$ almost everywhere and $\rho_b \leq c$ which completes the proof. ∎

Proof of Proposition A.5. Let $0 < \delta < 1 \leq D$. Analogously to the proof of Proposition A.1, one can compute that

$$\begin{aligned}\mathrm{Var}\left(\int_0^T f(X_t)\,\mathrm{d}t\right) &= 2\int_0^T (T-v) \iint_{\mathbb{R}^{d\times d}} f(x)f(y)(p_v(x,y) - \rho(y))\,\mu(\mathrm{d}x)\,\mathrm{d}y\,\mathrm{d}v \\ &\leq 2T\|f\|_\infty^2 \Big(\int_0^D \iint_{\mathcal{S}^2} p_v(x,y)\,\mu(\mathrm{d}x)\,\mathrm{d}y\,\mathrm{d}v + \int_D^T \iint_{\mathcal{S}^2} (p_v(x,y) - \rho(y))\,\mu(\mathrm{d}x)\,\mathrm{d}y\,\mathrm{d}v \Big) \\ &= 2T\|f\|_\infty^2 (\mathcal{I}_\delta + \mathcal{I}_D + \mathcal{I}_T),\end{aligned}$$

where $\mathcal{I}_\delta := \int_0^\delta \iint_{\mathcal{S}^2} p_v(x,y)\mu(\mathrm{d}x)\,\mathrm{d}y\,\mathrm{d}v$, $\mathcal{I}_D := \int_\delta^D \iint_{\mathcal{S}^2} p_v(x,y)\mu(\mathrm{d}x)\,\mathrm{d}y\,\mathrm{d}v$ and

$$\mathcal{I}_T := \int_D^T \int_{\mathcal{S}} (P_v(x,\mathcal{S}) - \mu(\mathcal{S}))\,\mu(\mathrm{d}x)\,\mathrm{d}v.$$

As before (see (A.27)) and under our additional assumption that ρ is bounded, it holds

$$\mathcal{I}_\delta \leq \mu(\mathcal{S})\delta \leq \|\rho\|_\infty \boldsymbol{\lambda}(\mathcal{S})\delta. \tag{A.28}$$

Furthermore, exploiting the mixing property of $\boldsymbol{X}$,

$$\mathcal{I}_T \le \int_D^T \int \|P_v(x,\cdot) - \mu(\cdot)\|_{\mathrm{TV}} \mu(\mathrm{d}x)\,\mathrm{d}v \le c_\kappa \int_D^T \mathrm{e}^{-\kappa v}\,\mathrm{d}v \le \frac{c_\kappa}{\kappa}\mathrm{e}^{-\kappa D}\mathbf{1}_{(D,\infty)}(T). \tag{A.29}$$

By assumption (𝔄1), $p_v(x,y) \le c_2 v^{-d/2}$, for $0 < t \le 1$. Hence, we have $p_{1/2}(x,y) \le c_2 2^{d/2} =: c_p$ which implies

$$p_t(x,y) = \int p_{t-1/2}(x,z)p_{1/2}(z,y)\,\mathrm{d}z \le c_p,$$

for all $t > 1/2$. Since $\delta < 1 \le D$, it follows

$$\int_\delta^D p_v(x,y)\,\mathrm{d}v \le c_2 \int_\delta^1 v^{-d/2}\,\mathrm{d}v + c_p D\mathbf{1}_{(1,\infty)}(D) \le c_{\delta,D}\Big(\int_\delta^1 v^{-d/2}\,\mathrm{d}v + D\mathbf{1}_{(1,\infty)}(D)\Big)$$

for $c_{\delta,D} := c_2 + c_p$. For $d \ge 3$, this implies

$$\begin{aligned}\int_\delta^D p_v(x,y)\,\mathrm{d}v &\le c_{\delta,D}\Big(\int_\delta^1 v^{-d/2}\,\mathrm{d}v + D\mathbf{1}_{(1,\infty)}(D)\Big)\\ &\le c_{\delta,D}\Big((d/2-1)^{-1}\delta^{1-d/2} + D\mathbf{1}_{(1,\infty)}(D)\Big)\\ &\le c'_{\delta,D}\Big(\delta^{1-d/2} + D\mathbf{1}_{(1,\infty)}(D)\Big),\end{aligned} \tag{A.30}$$

where $c'_{\delta,D} := 2c_{\delta,D}$. Letting $\delta = \lambda(\mathbb{S})^{2/d}$, $D = (1 \vee -\frac{2}{\kappa}\log(\lambda(\mathbb{S}))) \wedge T$, (A.30) and $\lambda(\mathbb{S}) < 1$ imply

$$\begin{aligned}\int_\delta^D p_v(x,y)\,\mathrm{d}v &\le c'_{\delta,D}\Big(\lambda(\mathbb{S})^{2/d-1} + \tfrac{2}{\kappa}\log(\lambda(\mathbb{S})^{-1})\Big) \le c'_{\delta,D}\Big(\lambda(\mathbb{S})^{2/d-1} + \tfrac{2}{\kappa(1-2/d)}\lambda(\mathbb{S})^{2/d-1}\Big)\\ &\le c''_{\delta,D}\lambda(\mathbb{S})^{2/d-1}, \quad \text{for } c''_{\delta,D} := c'_{\delta,D}(1 + \tfrac{2}{\kappa(1-2/d)}),\end{aligned}$$

where we have used the well-known inequality $\log x \le nx^{1/n}$, $x, n > 0$. Using Fubini's theorem, this directly implies

$$\mathcal{I}_D = \int_\delta^D \iint_{\mathbb{S}^2} p_v(x,y)\mu(\mathrm{d}x)\,\mathrm{d}y\,\mathrm{d}v \le c''_{\delta,D}\mu(\mathbb{S})\lambda(\mathbb{S})^{2/d} \le c''_{\delta,D}\|\rho\|_\infty \lambda(\mathbb{S})^{2/d+1} \tag{A.31}$$

for $d \ge 3$. Noting that our choice of δ and D implies by (A.28) and (A.29) that

$$\mathcal{I}_\delta \le \|\rho\|_\infty \lambda(\mathbb{S})^{2/d+1} \quad \text{and} \quad \mathcal{I}_T \le \frac{c_\kappa}{\kappa}\lambda(\mathbb{S})^2 \le \frac{c_\kappa}{\kappa}\lambda(\mathbb{S})^{2/d+1},$$

(A.8) follows for any $d \ge 3$ by combining these estimates with (A.31). The case $d = 2$ is treated by similar arguments. ∎

A.II Proofs for Section A.3

Proof of Theorem A.6. We start by splitting the process $(X_s)_{0\le s\le t}$ with Borel state space $\mathcal{X}$ into $2n_t$ parts of length m_t, where $t = 2n_t m_t$, $n_t \in \mathbb{N}$, $m_t \in \mathbb{R}_+$. More precisely, for $j \in \{1,\dots,n_t\}$, define the processes

$$X^{j,1} := (X_s)_{s\in[2(j-1)m_t,(2j-1)m_t]}, \quad X^{j,2} := (X_s)_{s\in[(2j-1)m_t,2jm_t]}.$$

Since $\boldsymbol{X}$ is a stationary Markov process, the β-mixing assumption is equivalent to

$$\Xi(s) \geq \int_{\mathbb{R}^d} \|P_s(x,\cdot) - \mu\|_{\mathrm{TV}}\, \mu(\mathrm{d}x) = \mathbb{E}\Big[\big\|\mathbb{P}(\cdot|\mathcal{F}_0) - \mathbb{P}\big\|_{\mathrm{TV}|\overline{\mathcal{F}}_s}\Big] = \mathbb{E}\Big[\big\|\mathbb{P}(\cdot|\mathcal{F}_t) - \mathbb{P}\big\|_{\mathrm{TV}|\overline{\mathcal{F}}_{t+s}}\Big],$$

for any $s,t > 0$, see Proposition 1 in [29]. Here, $(\mathcal{F}_t = \sigma(X_s, s \leq t))_{t\geq 0}$ denotes the natural filtration of $\boldsymbol{X}$, $(\overline{\mathcal{F}}_t = \sigma(X_s, s \geq t))_{t\geq 0}$ the filtration of the future of $\boldsymbol{X}$ and, for a signed measure μ and a sub-σ-algebra $\mathcal{A}$ on a measure space $(\Omega, \mathcal{F})$, $\|\mu\|_{\mathrm{TV}|\mathcal{A}}$ denotes the total variation norm of μ restricted to $\mathcal{A}$. As demonstrated in [85, Lemma 1.4],

$$\mathbb{E}\Big[\big\|\mathbb{P}(\cdot|\mathcal{F}_t) - \mathbb{P}\big\|_{\mathrm{TV}|\overline{\mathcal{F}}_{t+s}}\Big] = \beta(\mathcal{F}_t, \overline{\mathcal{F}}_{t+s}),$$

where for two sub-σ-algebras $\mathcal{A}, \mathcal{B} \subset \mathcal{G}$ and a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{G})$, the classical β-mixing coefficient $\beta(\mathcal{A}, \mathcal{B})$ is given by

$$\beta(\mathcal{A}, \mathcal{B}) = \sup_{C \in \mathcal{A}\otimes\mathcal{B}} \big|\mathbb{P}|_{\mathcal{A}\otimes\mathcal{B}}(C) - \mathbb{P}|_{\mathcal{A}} \otimes \mathbb{P}|_{\mathcal{B}}(C)\big|.$$

Here, $\mathbb{P}|_{\mathcal{A}\otimes\mathcal{B}}$ is the restriction to $(\Omega\times\Omega, \mathcal{A}\otimes\mathcal{B})$ of the image measure of $\mathbb{P}$ under the canonical injection $\iota(\omega) = (\omega,\omega)$. Clearly, if $\mathcal{A}_1 \subset \mathcal{A}_2$, we have $\beta(\mathcal{A}_1, \mathcal{B}) \leq \beta(\mathcal{A}_2, \mathcal{B})$. Observe that $X^{j,1}$, as a mapping from Ω to $\mathcal{X}^{[2(j-1)m_t, (2j-1)m_t]}$, is both $\mathcal{F}_{(2j-1)m_t}$-measurable and $\overline{\mathcal{F}}_{2(j-1)m_t}$-measurable. It now follows from the above discussion for $j,k \in \{1,\dots,n_t\}$, $j<k$, that

$$\begin{aligned}\beta(X^{j,1}, X^{k,1}) &:= \beta(\sigma(X^{j,1}), \sigma(X^{k,1})) \leq \beta(\mathcal{F}_{(2j-1)m_t}, \overline{\mathcal{F}}_{2(k-1)m_t}) \\ &\leq \Xi((2(k-j)-1)m_t) \leq \Xi((k-j)m_t).\end{aligned}$$

In the same way, we obtain $\beta(X^{j,2}, X^{k,2}) \leq \Xi((k-j)m_t)$. Arguing as in the proof of Proposition 5.1 of [84], we can then construct a process $(\widehat{X}_s)_{0\leq s\leq t}$ by Berbee's coupling method, such that for $k = 1, 2$,

1. $X^{j,k} \overset{(\mathrm{d})}{=} \widehat{X}^{j,k}$, for all $j \in \{1,\dots,n_t\}$,
2. $\mathbb{P}(X^{j,k} \neq \widehat{X}^{j,k}) \leq \Xi(m_t)$ for all $j \in \{1,\dots,n_t\}$,
3. $\widehat{X}^{1,k}, \dots, \widehat{X}^{n_t,k}$ are independent,

where $\widehat{X}^{j,k}$ is defined analogously to $X^{j,k}$ for $j \in \{1,\dots,n_t\}$ and $k = 1, 2$. In order to ease the notation, define for $j \in \{1,\dots,n_t\}$

$$I_g(X^{j,1}) := \int_{2(j-1)m_t}^{(2j-1)m_t} g(X_s)\,\mathrm{d}s, \quad I_g(X^{j,2}) := \int_{(2j-1)m_t}^{2jm_t} g(X_s)\,\mathrm{d}s,$$

and, analogously, define $I_g(\widehat{X}^{j,k})$ for $k = 1, 2$, $j \in \{1,\dots,n_t\}$. Fix $p \geq 1$. Then,

$$\begin{aligned}\Big(\mathbb{E}\Big[\|\mathbb{G}_t\|_{\mathcal{G}}^p\Big]\Big)^{1/p} &\leq \left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{\sqrt{t}}\int_0^t g(\widehat{X}_s)\,\mathrm{d}s\right|^p\right]\right)^{1/p} + \left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{\sqrt{t}}\int_0^t (g(X_s) - g(\widehat{X}_s))\,\mathrm{d}s\right|^p\right]\right)^{1/p} \\ &= \left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{\sqrt{t}}\sum_{k=1}^{2}\sum_{j=1}^{n_t} I_g(\widehat{X}^{j,k})\right|^p\right]\right)^{1/p} \\ &\qquad + \left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{\sqrt{t}}\sum_{k=1}^{2}\sum_{j=1}^{n_t} (I_g(X^{j,k}) - I_g(\widehat{X}^{j,k}))\right|^p\right]\right)^{1/p}.\end{aligned} \tag{A.32}$$

The classical Bernstein inequality (cf. Theorem 2.10 in [15]) implies that for $u > 0$

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{t}}\sum_{j=1}^{n_t} I_g(\widehat{X}^{j,k})\right| > \sqrt{\frac{\mathrm{Var}\left(\int_0^{m_t} g(X_s)\,\mathrm{d}s\right)u}{m_t}} + \frac{m_t\|g\|_\infty u}{\sqrt{t}}\right) \le 2\mathrm{e}^{-u},$$

where we used that $2n_t/t = 1/m_t$. Consequently, denoting

$$\widetilde{c}_1 := 2\mathrm{e}^{1/(2\mathrm{e})}\sqrt{2\pi}\mathrm{e}^{-11/12}, \quad \widetilde{c}_2 := 2(2\mathrm{e})^{-1/2}\mathrm{e}^{1/(2\mathrm{e})}\sqrt{\pi}\mathrm{e}^{1/6}, \tag{A.33}$$

Lemma A.2 in [33] gives, for $k \in \{1,2\}$,

$$\left(\mathbb{E}\left[\left|\frac{1}{\sqrt{t}}\sum_{j=1}^{n_t} I_g(\widehat{X}^{j,k})\right|^p\right]\right)^{1/p} \le \|g\|_\infty \frac{m_t}{\sqrt{t}}\widetilde{c}_1 p + \sqrt{\mathrm{Var}\left(\frac{1}{\sqrt{m_t}}\int_0^{m_t} g(X_s)\,\mathrm{d}s\right)}\widetilde{c}_2\sqrt{p}. \tag{A.34}$$

In addition, Theorem 3.5 in [33] implies that there exist positive constants $\widetilde{C}_1, \widetilde{C}_2$ such that

$$\begin{aligned}\left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{\sqrt{t}}\sum_{j=1}^{n_t} I_g(\widehat{X}^{j,k})\right|^p\right]\right)^{1/p} \le{}& \frac{\widetilde{C}_1}{2}\int_0^\infty \log \mathcal{N}\big(u, \mathcal{G}, \tfrac{m_t}{\sqrt{t}}d_\infty\big)\,\mathrm{d}u + \frac{\widetilde{C}_2}{2}\int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{G}, d_{\mathbb{G},m_t})}\,\mathrm{d}u \\ &+ 2\sup_{g\in\mathcal{G}}\left(\mathbb{E}\left[\left|\frac{1}{\sqrt{t}}\sum_{j=1}^{n_t} I_g(\widehat{X}^{j,k})\right|^p\right]\right)^{1/p}.\end{aligned} \tag{A.35}$$

Here, we bounded the γ_α-functionals appearing in the original statement of the theorem by the corresponding entropy integrals. Note further that the last term on the rhs of (A.32) is upper bounded by

$$\begin{aligned}\left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{\sqrt{t}}\sum_{k=1}^{2}\sum_{j=1}^{n_t}\left(I_g(X^{j,k}) - I_g(\widehat{X}^{j,k})\right)\cdot \mathbf{1}_{X^{j,k}\neq\widehat{X}^{j,k}}\right|^p\right]\right)^{1/p} &\le \frac{4n_t m_t}{\sqrt{t}}\sup_{g\in\mathcal{G}}\|g\|_\infty\left(\mathbb{P}\left(X^{j,k}\neq\widehat{X}^{j,k}\right)\right)^{1/p} \\ &\le 2\sup_{g\in\mathcal{G}}\|g\|_\infty\sqrt{t}\Xi(m_t)^{1/p}.\end{aligned} \tag{A.36}$$

Plugging the upper bounds (A.34), (A.35) and (A.36) into (A.32) yields

$$\begin{aligned}\left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}|\mathbb{G}_t(g)|^p\right]\right)^{1/p} \le{}& \widetilde{C}_1\int_0^\infty \log \mathcal{N}\big(u, \mathcal{G}, \tfrac{m_t}{\sqrt{t}}d_\infty\big)\,\mathrm{d}u + \widetilde{C}_2\int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{G}, d_{\mathbb{G},m_t})}\,\mathrm{d}u \\ &+ 4\sup_{g\in\mathcal{G}}\left(\frac{m_t}{\sqrt{t}}\|g\|_\infty\widetilde{c}_1 p + \|g\|_{\mathbb{G},m_t}\widetilde{c}_2\sqrt{p} + \frac{1}{2}\|g\|_\infty\sqrt{t}\Xi(m_t)^{1/p}\right).\end{aligned} \tag{A.37}$$

For general $m_t \in (0, \frac{t}{4}]$, let $\widetilde{n}_t = \lfloor \frac{t}{2m_t}\rfloor$, where $\lfloor x\rfloor$ denotes the largest integer smaller or equal to $x \ge 1$. Then, for $\widetilde{m}_t := \frac{t}{2\widetilde{n}_t}$, we have $m_t \le \widetilde{m}_t$, and from $\widetilde{n}_t \ge \frac{t}{2m_t} - 1 = \frac{t-2m_t}{2m_t}$ and $m_t \le \frac{t}{4}$, we get

$$\widetilde{m}_t = \frac{t}{2\widetilde{n}_t} \le \frac{tm_t}{t-2m_t} \le 2m_t.$$

Since $\widetilde{n}_t \in \mathbb{N}$, (A.37) holds with $\tau = \widetilde{m}_t \in [m_t, 2m_t]$ and m_t being replaced by $\widetilde{m}_t$, and combining this with the computations above yields

$$\begin{aligned}\left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}|\mathbb{G}_t(g)|^p\right]\right)^{1/p} &\le \widetilde{C}_1\int_0^\infty \log \mathcal{N}\big(u,\mathcal{G},\tfrac{\tau}{\sqrt{t}}d_\infty\big)\,\mathrm{d}u + \widetilde{C}_2\int_0^\infty \sqrt{\log \mathcal{N}(u,\mathcal{G},d_{\mathbb{G},\tau})}\,\mathrm{d}u \\ &\quad + 4\sup_{g\in\mathcal{G}}\left(\frac{\tau}{\sqrt{t}}\|g\|_\infty\widetilde{c}_1 p + \|g\|_{\mathbb{G},\tau}\widetilde{c}_2\sqrt{p} + \frac{1}{2}\|g\|_\infty\sqrt{t}\Xi(\tau)^{1/p}\right) \\ &\le \widetilde{C}_1\int_0^\infty \log \mathcal{N}\big(u,\mathcal{G},\tfrac{2m_t}{\sqrt{t}}d_\infty\big)\,\mathrm{d}u + \widetilde{C}_2\int_0^\infty \sqrt{\log \mathcal{N}(u,\mathcal{G},d_{\mathbb{G},\tau})}\,\mathrm{d}u \\ &\quad + 4\sup_{g\in\mathcal{G}}\left(\frac{2m_t}{\sqrt{t}}\|g\|_\infty\widetilde{c}_1 p + \|g\|_{\mathbb{G},\tau}\widetilde{c}_2\sqrt{p} + \frac{1}{2}\|g\|_\infty\sqrt{t}\Xi(m_t)^{1/p}\right),\end{aligned}$$

which completes the proof. ∎

Proof of Corollary A.8. In case of exponential β-mixing we obtain, similarly to the proof of Proposition A.5, for any $t > 0$,

$$\|g\|_{\mathbb{G},t}^2 = \frac{1}{t}\mathrm{Var}\left(\int_0^t g(X_s)\,\mathrm{d}s\right) \le 2\|g\|_\infty^2\int_0^t\int \|P_s(x,\cdot)-\mu\|_{\mathrm{TV}}\,\mu(\mathrm{d}x)\,\mathrm{d}s \le 2\|g\|_\infty^2\frac{c_\kappa}{\kappa}.$$

Choosing $m_T = \sqrt{T}$ and plugging this into (A.10) therefore yields the assertion for the exponential mixing case. For the α-polynomial case we obtain the assertion similarly by the minimizing choice $m_T = T^{p/(\alpha+p)}$, where $T \ge 4^{(\alpha+p)/\alpha}$ guarantees that $m_T \le T/4$ and the assumption $\alpha > 1$ is needed to guarantee uniform boundedness of $\|g\|_{\mathbb{G},t}^2$ in t. ∎

For the proof of bounds on the stochastic error $\mathbb{H}_{h,T}$ defined in (A.12), we start with the following preparatory lemma that provides bounds of the covering numbers of the function class $\mathcal{G}$ introduced in (A.13) with respect to the norms appearing in Theorem A.6. By a slight abuse of notation, we do not distinguish notationally between the sup-norm on $\mathbb{R}^d$ and the function space $\mathcal{B}_b(\mathbb{R}^d)$.

Lemma A.21. *Let $D \subset \mathbb{R}^d$ be a bounded set and, given some Lipschitz continuous kernel K with Lipschitz constant L and compact support $[-1/2,1/2]^d$, define the function class $\mathcal{G}$ according to (A.13). Then, for any $\varepsilon > 0$,*

$$\mathcal{N}(\varepsilon,\mathcal{G},\|\cdot\|_{d_\infty}) \le \left(\frac{4L\mathrm{diam}(D)}{\varepsilon h}\right)^d,$$

and if moreover $\boldsymbol{X} \in \boldsymbol{\Sigma}\cup\boldsymbol{\Theta}$, then there exists a constant $\mathbb{A} > 0$ such that, for any $\varepsilon > 0$ and $t > 0$,

$$\mathcal{N}(\varepsilon,\mathcal{G},\|\cdot\|_{\mathbb{G},t}) \le \left(\frac{2L\mathrm{diam}(D)\sqrt{\mathbb{A}\|\rho\|_\infty}h^{d-1}\psi_d(h^d)}{\varepsilon}\right)^d.$$

Proof. For $x \in \mathbb{R}^d$, we obtain by Lipschitz continuity of K that

$$\begin{aligned}B_{d_\infty}\big(\overline{K}\big((x-\cdot)/h\big),\varepsilon\big) &= \big\{\overline{K}\big((y-\cdot)/h\big) : y\in\mathbb{R}^d, \big\|\overline{K}\big((x-\cdot)/h\big)-\overline{K}\big((y-\cdot)/h\big)\big\|_\infty \le \varepsilon\big\} \\ &\supset \big\{\overline{K}\big((y-\cdot)/h\big) : y\in\mathbb{R}^d, \|x-y\|_\infty \le \varepsilon h/(2L)\big\}.\end{aligned} \tag{A.38}$$

Let $Q \supset D$ be a cube of side length $\mathrm{diam}(D) < \infty$ and choose for

$$\overline{n} := \left(\left\lfloor \frac{2L\mathrm{diam}(D)}{\varepsilon h} \right\rfloor \right)^d$$

points $x_1, \dots, x_{\overline{n}} \in Q$ such that $\{B_{d_\infty}(x_i, \varepsilon h/(2L)) : i = 1, \dots \overline{n}\}$ covers Q and therefore D. From (A.38), it follows that $\{B_{d_\infty}(\overline{K}((x_i - \cdot)/h), \varepsilon) : i = 1, \dots \overline{n}\}$ is an external covering of $\mathcal{G}$. The external covering number $\mathcal{N}_{\mathrm{ext}}(\varepsilon, \mathcal{G}, d_\infty)$ is thus bounded by $(2L\mathrm{diam}(D)/(\varepsilon h))^d$. Hence,

$$\mathcal{N}(\varepsilon, \mathcal{G}, d_\infty) \le \mathcal{N}_{\mathrm{ext}}(\varepsilon/2, \mathcal{G}, d_\infty) \le \left(\frac{4L\mathrm{diam}(D)}{\varepsilon h} \right)^d.$$

Similarly, for

$$\widetilde{\mathcal{G}} = \{K(x - \cdot)/h) : x \in D \cap \mathbb{Q}^d\}, \tag{A.39}$$

we obtain

$$\mathcal{N}(\varepsilon, \widetilde{\mathcal{G}}, d_\infty) \le \left(\frac{2L\mathrm{diam}(D)}{\varepsilon h} \right)^d.$$

The variance term is bounded by means of Propositions A.1 and A.5, respectively. In case $d = 1$ for $\boldsymbol{X} \in \boldsymbol{\Sigma}$ or any dimension for $\boldsymbol{X} \in \boldsymbol{\Theta}$, boundedness of ρ, Proposition A.1 and (A.9) yield that, for $h \in (0, 1)$ and some constant C independent of $\lambda(\mathrm{supp}(K((x - \cdot)/h))) = h^d$,

$$\mathrm{Var}\left(\int_0^T K\left(\frac{x - X_t}{h} \right) \mathrm{d}t \right) \le C(1 \vee c_{\widetilde{D}})T\|K\|_\infty^2 \|\rho\|_\infty h^{2d} \psi_d^2(h^d),$$

where $\widetilde{D}$ is a compact set containing $D + [-1/2, 1/2]^d$. Hence, for any dimension d and $\boldsymbol{X} \in \boldsymbol{\Sigma} \cup \boldsymbol{\Theta}$, we obtain together with Proposition A.5 that there exists some global constant $\mathbb{A}$ independent of h such that for any $h \in (0, 1)$, $t > 0$ and $g \in \widetilde{\mathcal{G}}$,

$$\mathrm{Var}\left(\frac{1}{\sqrt{t}} \int_0^t g(X_s) \, \mathrm{d}s \right) \le \mathbb{A}\|g\|_\infty^2 \|\rho\|_\infty h^{2d} \psi_d^2(h^d), \tag{A.40}$$

and hence

$$\|g\|_{\mathbb{G},t} \le \sqrt{\mathbb{A}\|\rho\|_\infty} h^d \psi_d(h^d) \|g\|_\infty. \tag{A.41}$$

Consequently, with the first part of the proof we obtain

$$\begin{aligned} \mathcal{N}(\varepsilon, \mathcal{G}, \|\cdot\|_{\mathbb{G},t}) &= \mathcal{N}(\varepsilon, \widetilde{\mathcal{G}}, \|\cdot\|_{\mathbb{G},t}) \\ &\le \mathcal{N}(\varepsilon(\sqrt{\mathbb{A}\|\rho\|_\infty} h^d \psi_d(h^d))^{-1}, \widetilde{\mathcal{G}}, \|\cdot\|_\infty) \\ &\le \left(\frac{2L\mathrm{diam}(D)\sqrt{\mathbb{A}\|\rho\|_\infty} h^{d-1} \psi_d(h^d)}{\varepsilon} \right)^d. \end{aligned}$$

■

Proof of Lemma A.9. Let $\boldsymbol{X} \in \boldsymbol{\Theta} \cup \boldsymbol{\Sigma}$. We start with bounding $\mathbb{E}[\sup_{x \in D} |\mathbb{H}_{h,T}(x)|^p]$. Let $m_T \in (0, T/4]$ and $\tau \in [m_T, 2m_T]$ as in Theorem A.6. Using (A.41) and $\sup_{f,g \in \widetilde{\mathcal{G}}} \|f - g\|_\infty \le 2\|K\|_\infty$ for $\widetilde{\mathcal{G}}$ defined in (A.39), we obtain

$$\sup_{f,g \in \widetilde{\mathcal{G}}} \|f - g\|_{\mathbb{G},\tau} \le \sqrt{\mathbb{A}\|\rho\|_\infty} \sup_{f,g \in \widetilde{\mathcal{G}}} \|f - g\|_\infty h^d \psi_d(h^d) \le 2\sqrt{\mathbb{A}\|\rho\|_\infty} \|K\|_\infty h^d \psi_d(h^d) =: \mathbb{V}(h), \tag{A.42}$$

such that $\mathcal{N}(u, \widetilde{\mathcal{G}}, \|\cdot\|_{\mathbb{G},\tau}) = 1$ for $u \geq \mathbb{V}(h)$. Consequently, using $\int_0^C \sqrt{\log(M/u)}\,\mathrm{d}u \leq 4C\sqrt{\log(M/C)}$ provided $\log(M/C) \geq 2$, see e.g. p. 592 of Giné and Nickl [41], and the covering number bound from Lemma A.21, it follows for $h \leq \mathrm{e}^{-2}L\mathrm{diam}(D)/\|K\|_\infty$ that

$$\begin{aligned}\int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{G}, d_{\mathbb{G},\tau})}\,\mathrm{d}u &= \int_0^\infty \sqrt{\log \mathcal{N}(u, \widetilde{\mathcal{G}}, d_{\mathbb{G},\tau})}\,\mathrm{d}u \leq \int_0^{\mathbb{V}(h)} \sqrt{d\,\log\left(\frac{L\mathrm{diam}(D)\mathbb{V}(h)}{uh\|K\|_\infty}\right)}\,\mathrm{d}u \\ &\leq 2\mathbb{V}(h)\sqrt{d\log\left(\frac{L\mathrm{diam}(D)}{\|K\|_\infty h}\right)}.\end{aligned}$$

Moreover, since $\sup_{f,g\in\mathcal{G}}\|f-g\|_\infty \leq 4\|K\|_\infty$, it follows that $\mathcal{N}(u, \mathcal{G}, d_\infty) = 1$ for all $u \geq 4\|K\|_\infty$ and hence we obtain by the covering number bound with respect to the sup-norm from Lemma A.21 and elementary calculations

$$\int_0^\infty \log \mathcal{N}\big(u, \mathcal{G}, \tfrac{2m_T}{\sqrt{T}}d_\infty\big)\,\mathrm{d}u = 2\tfrac{m_T}{\sqrt{T}}\int_0^{4\|K\|_\infty} \log \mathcal{N}(u, \mathcal{G}, d_\infty)\,\mathrm{d}u \leq 8\tfrac{m_T}{\sqrt{T}}d\|K\|_\infty\Big(1+\log\left(\frac{L\mathrm{diam}(D)}{\|K\|_\infty h}\right)\Big).$$

Denseness of $\mathbb{Q}^d$ in $\mathbb{R}^d$, continuity of $x \mapsto \mathbb{H}_{h,T}(x)$ and Theorem A.6 thus imply for $h \leq \mathrm{e}^{-2}L\mathrm{diam}(D)/\|K\|_\infty$

$$\begin{aligned}\Big(\mathbb{E}\Big[\sup_{x\in D}|\mathbb{H}_{h,T}(x)|^p\Big]\Big)^{1/p} &= \Big(\mathbb{E}\Big[\sup_{x\in D\cap\mathbb{Q}^d}|\mathbb{H}_{h,T}(x)|^p\Big]\Big)^{1/p} \\ &\leq \frac{1}{\sqrt{T}h^d}\Bigg(8\widetilde{C}_1\frac{m_T}{\sqrt{T}}d\|K\|_\infty\Big(1+\log\Big(\frac{L\mathrm{diam}(D)}{\|K\|_\infty h}\Big)\Big) + 2\widetilde{C}_2\mathbb{V}(h)\sqrt{d\log\Big(\frac{L\mathrm{diam}(D)}{\|K\|_\infty h}\Big)} \\ &\quad + 16\frac{m_T}{\sqrt{T}}\|K\|_\infty\widetilde{c}_1 p + 2\mathbb{V}(h)\widetilde{c}_2\sqrt{p} + 4\|K\|_\infty\sqrt{T}\Xi(m_T)^{1/p}\Bigg),\end{aligned} \tag{A.43}$$

for $\mathbb{V}(h)$ introduced in (A.42). Now, let $p = u_T \geq 1$ be such that $\Xi^{-1}(T^{-u_T}) \in \mathrm{o}(T)$. Then, for T large enough, (A.43), $h \geq T^{-2}$ and $h \in \mathrm{o}(1)$ imply for the choice $m_T = \Xi^{-1}(T^{-u_T})$ that

$$\begin{aligned}&\mathbb{E}\big[\big\|\widehat{\rho}_{h,T} - \mathbb{E}\widehat{\rho}_{h,T}\big\|^{u_T}_{L^\infty(D)}\big] \\ &\leq c^{u_T}\Bigg(\frac{\log T}{Th^d}\Xi^{-1}(T^{-u_T}) + T^{-\frac{1}{2}}\psi_d(h^d)\sqrt{\log(h^{-1})} + \frac{u_T}{Th^d}\Xi^{-1}(T^{-u_T}) + T^{-\frac{1}{2}}\psi_d(h^d)\sqrt{u_T} + h^{-d}T^{-1}\Bigg)^{u_T} \\ &\leq c^{u_T}\Bigg(\frac{\log T + u_T}{Th^d}\Xi^{-1}(T^{-u_T}) + T^{-\frac{1}{2}}\psi_d(h^d)\big(\sqrt{\log(h^{-1})} + \sqrt{u_T}\big)\Bigg)^{u_T},\end{aligned}$$

where the value of the constant c changes from line to line. Hence Markov's inequality implies that there exists some constant $c^* > 0$ such that

$$\mathbb{P}\Bigg(\big\|\widehat{\rho}_{h,T} - \mathbb{E}\widehat{\rho}_{h,T}\big\|_{L^\infty(D)} \geq c^*\Bigg(\frac{u_T + \log T}{Th^d}\Xi^{-1}(T^{-u_T}) + T^{-\frac{1}{2}}\psi_d(h^d)\sqrt{u_T \vee \log(h^{-1})}\Bigg)\Bigg) \leq \mathrm{e}^{-u_T}. \tag{A.44}$$

Suppose now that $\boldsymbol{X} \in \boldsymbol{\Sigma}$. Then, $\boldsymbol{X}$ is exponentially β-mixing, i.e., $\Xi(t) = c_\kappa \mathrm{e}^{-\kappa t}$, where without loss of generality we may assume that $c_\kappa \geq 1$. Then, for any $\gamma > 0$ and $1 \leq u_T \leq \gamma \log T$, it follows from $\Xi^{-1}(T^{-u_T}) \leq u_T \log T/\kappa$ and (A.44) that there exists some constant $c_\gamma > 0$ such that

$$\mathbb{P}\Bigg(\big\|\widehat{\rho}_{h,T} - \mathbb{E}\widehat{\rho}_{h,T}\big\|_{L^\infty(D)} \geq c_\gamma\Bigg(\frac{u_T(\log T)^2}{Th^d} + T^{-\frac{1}{2}}\psi_d(h^d)\sqrt{u_T \vee \log(h^{-1})}\Bigg)\Bigg) \leq \mathrm{e}^{-u_T}.$$

■

A.III PROOFS FOR SECTION A.4

Proof of Corollary A.10. Fix x such that there exists an open neighbourhood D of x such that $\rho|_D \in \mathcal{H}_D(\beta, \mathsf{L})$. The usual bias-variance decomposition gives

$$\mathbb{E}\Big[\big(\widehat{\rho}_{h,T}(x) - \rho(x)\big)^2\Big] = \big(\rho * K_h(x) - \rho(x)\big)^2 + \mathrm{Var}\big(\widehat{\rho}_{h,T}(x)\big). \tag{A.45}$$

For the bias term, since $\lfloor \beta \rfloor \leq \ell$, there exists a universal constant $M > 0$ such that

$$|(\rho * K_h - \rho)(x)| = \left| h^{-d} \int K\Big(\frac{x-y}{h}\Big)(\rho(y) - \rho(x))\,\mathrm{d}y \right| \leq M h^{\beta}, \tag{A.46}$$

see Proposition 1.2 in [80] for the case $d = 1$ and the analogous estimator for discrete observations, which can be extended to the general multivariate case under continuous observations without much effort. Moreover, for any dimension d and $\boldsymbol{X} \in \boldsymbol{\Sigma} \cup \boldsymbol{\Theta}$, it follows from (A.40) that for any $h \in (0,1)$

$$\mathrm{Var}\left(\frac{1}{T}\int_0^T K_h(x - X_t)\,\mathrm{d}t\right) \lesssim T^{-1}\|K\|_\infty^2 \|\rho\|_\infty \psi_d^2(h^d).$$

The claim follows by plugging the specific choice of h into (A.46) and (A.40) and using (A.45). ■

Proof of Theorem A.11. Fix $p \geq 1$, and recall the decomposition (A.12). By the assumption on the order of the kernel K, the bias term $\rho * K_h - \rho$ is bounded by $B(h) := Mh^\beta$ for some universal constant $M > 0$ as in the pointwise case (see (A.46)), while the upper bound on the stochastic error $\mathbb{H}_{h,T}$ relies on a suitable specification on the upper bound in (A.43). For $d \geq 3$, set $h = h(T) = (\log T/T)^{1/(2\beta+d-2)}$ and $m_T = p \log T/\kappa$ such that

$$\frac{1}{\sqrt{T}}\psi_d(h^d) \in \mathsf{O}\big(T^{-\beta/(2\beta+d-2)}\big) \quad \text{and} \quad \frac{m_T}{Th^d} = \left(\frac{\log T}{T}\right)^{\frac{2(\beta-1)}{2(\beta-1)+d}}.$$

Upon noting that $\beta > 2$ implies $2(\beta-1) > \beta$, it follows from (A.43) that

$$\left(\mathbb{E}\Big[\sup_{x\in D}|\mathbb{H}_{h,T}(x)|^p\Big]\right)^{1/p} \in \mathsf{O}\left(\left(\frac{\log T}{T}\right)^{\beta/(2\beta+d-2)}\right). \tag{A.47}$$

Since $h^\beta = (\log T/T)^{\beta/(2\beta+d-2)}$, (A.12), (A.46) and (A.47) finally give $\mathbb{E}[\|\widehat{\rho}_{h,T} - \rho\|^p_{L^\infty(D)}]^{1/p} \in \mathsf{O}\big(\Psi_{d,\beta}(T)\big)$ for $d \geq 3$. For $d = 1$ and $d = 2$, the assertion follows by analogous arguments.

We now proceed with the proof of the convergence rate of the adaptive scheme for $d \geq 3$. For the variance, we obtain from (A.43) that, for $m_T := 2\log_{(k)} T(\log T)^2/\kappa$ and whenever $h \leq \mathrm{e}^{-2}L\mathrm{diam}(D)/\|K\|_\infty$, there exists some constant $\mathfrak{C} > 0$ such that

$$\mathbb{E}\Big[\big\|\widehat{\rho}_{h,T} - \mathbb{E}\widehat{\rho}_{h,T}\big\|^2_{L^\infty(D)}\Big] = \mathbb{E}\Big[\sup_{x\in D}|\mathbb{H}_{h,T}(x)|^2\Big] \leq \mathfrak{C}^2\sigma^2(h,T),$$

where $\sigma^2(\cdot,\cdot)$ is defined according to (A.17). Define h_ρ by the balance equation

$$h_\rho := \max\left\{h \in \mathscr{H}_T : B(h) \leq \frac{1}{4}\sqrt{0.8\mathcal{M}}\sigma(h,T)\right\}, \quad \text{where } \mathcal{M} := \|\rho\|_{L^\infty(D)}.$$

This definition implies that $B(h_\rho) \simeq \sqrt{0.8\mathcal{M}}\sigma(h_\rho, T)/4$ and, since $\mathcal{H}_T \ni h_\rho > \left(\frac{\log_{(k)} T(\log T)^5}{T}\right)^{\frac{1}{d+2}}$,

$$h_\rho^{2\beta+d-2} \simeq \frac{\log_{(k)} T \log T}{T} \quad \text{and} \quad \sigma(h_\rho, T) \simeq \left(\frac{\log_{(k)} T \log T}{T}\right)^{\frac{\beta}{2\beta+d-2}}.$$

To justify this, define $h_0 := (\log_{(k)} T \log T/T)^{1/(2\beta+d-2)}$. For large enough T, we have $\log(\log_{(k)} T \log T) \leq (\log T)/2$ and hence

$$\begin{aligned}\sigma(h_0, T) &= \frac{\log_{(k)} T(\log T)^2}{Th_0^d}\log(h_0^{-1}) + \psi_d(h_0^d)\sqrt{\frac{\log_{(k)} T \log(h_0^{-1})}{T}} \\ &\geq \sqrt{\frac{\log_{(k)} T \log T}{2(2\beta+d-2)T}}\psi_d(h_0^d) = \sqrt{\frac{1}{2(2\beta+d-2)}}\left(\frac{\log_{(k)} T \log T}{T}\right)^{\frac{\beta}{2\beta+d-2}} = \mathcal{L}^{-1}B(h_0),\end{aligned}$$

for $\mathcal{L} = \sqrt{2(2\beta+d-2)M^2}$. Additionally, we get, since $\beta > 2$,

$$\sigma(h_0, T) = \frac{\log_{(k)} T(\log T)^2}{Th_0^d}\log(h_0^{-1}) + \psi_d(h_0^d)\sqrt{\frac{\log_{(k)} T \log(h_0^{-1})}{T}} \simeq \left(\frac{\log_{(k)} T \log T}{T}\right)^{\frac{\beta}{2\beta+d-2}}.$$

In particular, it holds that $h_0 \lesssim h_\rho$, which is clear if $\mathcal{L} \leq \frac{1}{4}\sqrt{0.8\mathcal{M}}$, and else follows by the fact that, for any $0 < \lambda < 1$,

$$B(\lambda h_0) = \lambda^\beta B(h_0) \leq \lambda^\beta \mathcal{L}\sigma(h_0, T) \leq \lambda^\beta \mathcal{L}\sigma(\lambda h_0, T).$$

Lastly, we show $h_\rho \lesssim h_0$ by proving $h_\rho^{2\beta+d-2}h_0^{-(2\beta+d-2)} \in \mathsf{O}(1)$. Indeed, by the definition of h_ρ,

$$\begin{aligned}h_\rho^{2\beta+d-2} &\lesssim h_\rho^{d-2}\sigma^2(h_\rho, T) \\ &\lesssim h_\rho^{d-2}\left(\frac{\log_{(k)} T(\log T)^3}{T}h_\rho^{-d} + \psi_d(h_\rho^d)\sqrt{\frac{\log_{(k)} T \log T}{T}}\right)^2 \\ &\lesssim \frac{(\log_{(k)} T)^2(\log T)^6}{T^2}h_\rho^{-(2+d)} + h_\rho^{d-2}\psi_d^2(h_\rho^d)\frac{\log_{(k)} T \log T}{T},\end{aligned}$$

and thus it holds that

$$h_\rho^{2\beta+d-2}h_0^{-(2\beta+d-2)} \lesssim \frac{\log_{(k)} T(\log T)^5}{T}h_\rho^{-(2+d)} + h_\rho^{d-2}\psi_d^2(h_\rho^d) \in \mathsf{O}(1),$$

thanks to $h_\rho > (\log_{(k)} T(\log T)^5/T)^{1/(d+2)}$.

Case 1: We first consider the case where $\widehat{h}_T \geq h_\rho$. To shorten notation, denote $\widetilde{\mathcal{M}} := \|\widehat{\rho}_{h_{\min},T}\|_{L^\infty(D)}$. Then, exploiting the definition of $\widehat{h}_T$ according to (A.16) and the bias and variance bounds,

$$\mathbb{E}\left[\|\widehat{\rho}_{\widehat{h}_T,T} - \rho\|_{L^\infty(D)} \cdot \mathbf{1}_{\{\widehat{h}_T \geq h_\rho\}\cap\{\widetilde{\mathcal{M}} \leq 1.2\mathcal{M}\}}\right]$$

$$\leq \mathbb{E}\Big[\Big(\|\widehat{\rho}_{\widehat{h}_T,T} - \widehat{\rho}_{h_\rho,T}\|_{L^\infty(D)} + \|\widehat{\rho}_{h_\rho,T} - \mathbb{E}\widehat{\rho}_{h_\rho,T}\|_{L^\infty(D)} + B(h_\rho)\Big)\mathbf{1}_{\{\widehat{h}_T \geq h_\rho\}\cap\{\widetilde{\mathcal{M}}\leq 1.2\mathcal{M}\}}\Big]$$
$$\leq \sqrt{1.2\mathcal{M}}\sigma(h_\rho,T) + \mathfrak{C}\sigma(h_\rho,T) + \frac{1}{4}\sqrt{0.8\mathcal{M}}\sigma(h_\rho,T) \in \mathrm{O}(\sigma(h_\rho,T)).$$

Similarly,

$$\mathbb{E}\Big[\|\widehat{\rho}_{\widehat{h}_T,T} - \rho\|_{L^\infty(D)} \cdot \mathbf{1}_{\{\widehat{h}_T \geq h_\rho\}\cap\{\widetilde{\mathcal{M}}>1.2\mathcal{M}\}}\Big]$$
$$\leq \sum_{h\in\mathscr{H}_T : h\geq h_\rho} \mathbb{E}\Big[\big(\|\widehat{\rho}_{h,T} - \mathbb{E}\widehat{\rho}_{h,T}\|_{L^\infty(D)} + B(h)\big)\cdot \mathbf{1}_{\{\widehat{h}_T=h\}\cap\{\widetilde{\mathcal{M}}>1.2\mathcal{M}\}}\Big]$$
$$\lesssim \log T\big(\mathfrak{C}\sigma(h_\rho,T) + B(1)\big)\sqrt{\mathbb{P}(\widetilde{\mathcal{M}} > 1.2\mathcal{M})}.$$

Now, for any T large enough,

$$\begin{aligned}
\mathbb{P}\Big(\big|\widetilde{\mathcal{M}} - \mathcal{M}\big| > 0.2\|\rho\|_{L^\infty(D)}\Big) &= \mathbb{P}\Big(\big|\|\widehat{\rho}_{h_{\min},T}\|_{L^\infty(D)} - \|\rho\|_{L^\infty(D)}\big| > 0.2\mathcal{M}\Big)\\
&\leq \mathbb{P}\Big(\big\|\widehat{\rho}_{h_{\min},T} - \rho\big\|_{L^\infty(D)} > 0.2\|\rho\|_{L^\infty(D)}\Big)\\
&\leq \mathbb{P}\Big(\big\|\widehat{\rho}_{h_{\min},T} - \mathbb{E}\widehat{\rho}_{h_{\min},T}\big\|_{L^\infty(D)} > 0.2\|\rho\|_{L^\infty(D)} - B(h_{\min})\Big)\\
&\leq \mathbb{P}\Big(\big\|\widehat{\rho}_{h_{\min},T} - \mathbb{E}\widehat{\rho}_{h_{\min},T}\big\|_{L^\infty(D)} > 0.1\|\rho\|_{L^\infty(D)}\Big)\\
&\leq \mathbb{P}\Big(\big\|\widehat{\rho}_{h_{\min},T} - \mathbb{E}\widehat{\rho}_{h_{\min},T}\big\|_{L^\infty(D)} > \Upsilon_{h_{\min},T}(\log T)\Big)\\
&\leq T^{-1},
\end{aligned} \tag{A.48}$$

where, for the function $\Upsilon_{h_{\min},T}(\cdot)$ defined according to (B.11), the last inequality follows from Lemma A.9 and the last but one inequality holds since there exists some constant C such that

$$\begin{aligned}
\Upsilon_{h_{\min},T}(\log T) &\leq CT^{-\frac{2}{d+2}}\Big((\log T)^{\frac{6-2d}{d+2}}(\log_{(k)} T)^{-\frac{d}{d+2}} + (\log T)^{\frac{6-3d}{d+2}}(\log_{(k)} T)^{\frac{2-d}{2(d+2)}}\Big)\\
&\leq 0.2\|\rho\|_{L^\infty(D)},
\end{aligned}$$

for T sufficiently large. Thus, we can conclude that $\mathbb{E}\Big[\|\widehat{\rho}_{\widehat{h}_T,T} - \rho\|_{L^\infty(D)} \cdot \mathbf{1}_{\{\widehat{h}_T \geq h_\rho\}}\Big] \in \mathrm{O}(\sigma(h_\rho,T))$.

Case 2: For the case $\widehat{h}_T < h_\rho$, note first that the previous bias and variance bounds together with (A.48) imply that

$$\mathbb{E}\Big[\|\widehat{\rho}_{\widehat{h}_T,T} - \rho\|_{L^\infty(D)} \cdot \mathbf{1}_{\{\widehat{h}_T < h_\rho\}\cap\{\widetilde{\mathcal{M}}<0.8\mathcal{M}\}}\Big]$$
$$\leq \sum_{h\in\mathscr{H}_T : h< h_\rho} \mathbb{E}\Big[\big(\|\widehat{\rho}_{h,T} - \mathbb{E}\widehat{\rho}_{h,T}\|_{L^\infty(D)} + B(h)\big)\cdot \mathbf{1}_{\{\widehat{h}_T=h\}\cap\{\widetilde{\mathcal{M}}<0.8\mathcal{M}\}}\Big]$$
$$\lesssim \log T\big(\mathfrak{C}\sigma(h_{\min},T) + B(h_\rho)\big)\sqrt{\mathbb{P}(\widetilde{\mathcal{M}} < 0.8\mathcal{M})} = O(\sigma(h_\rho,T)).$$

On the other hand,

$$\mathbb{E}\Big[\|\widehat{\rho}_{\widehat{h}_T,T} - \rho\|_{L^\infty(D)} \cdot \mathbf{1}_{\{\widehat{h}_T < h_\rho\}\cap\{0.8\mathcal{M}\leq\widetilde{\mathcal{M}}\}}\Big]$$

$$\leq \sum_{h\in\mathscr{H}_T:h<h_\rho} \mathbb{E}\Big[\big(\|\widehat{\rho}_{h,T}-\mathbb{E}\widehat{\rho}_{h,T}\|_{L^\infty(D)}+B(h)\big)\cdot \mathbf{1}_{\{\widehat{h}_T=h\}\cap\{0.8\mathcal{M}\leq\widetilde{\mathcal{M}}\}}\Big]$$
$$\leq \sum_{h\in\mathscr{H}_T:h<h_\rho} \sqrt{\mathbb{E}\Big[\|\widehat{\rho}_{h,T}-\mathbb{E}\widehat{\rho}_{h,T}\|^2_{L^\infty(D)}\Big]}\sqrt{\mathbb{E}\Big[\mathbf{1}_{\{\widehat{h}_T\geq h_\rho\}\cap\{0.8\mathcal{M}\leq\widetilde{\mathcal{M}}\}}\Big]}+B(h_\rho)$$
$$\leq \sum_{h\in\mathscr{H}_T:h<h_\rho} \mathfrak{C}\sigma(h,T)\sqrt{\mathbb{P}\Big(\{\widehat{h}_T\geq h_\rho\}\cap\{0.8\mathcal{M}\leq\widetilde{\mathcal{M}}\}\Big)}+\mathrm{O}(\sigma(h_\rho,T)).$$

Pick any $h\in\mathscr{H}_T$ such that $h<h_\rho$ and denote $h^+:=\min\{g\in\mathscr{H}_T: g>h\}=\eta h$. It is then shown as in the proof of Theorem 2 in [41] that the verification of the fact that the first sum on the rhs of the last display is of order $\mathrm{O}(\sigma(h_\rho,T))$ boils down to proving that

$$\sum_{h\in\mathscr{H}_T:h<h_\rho}\sigma(h,T)\Bigg(\sum_{g\in\mathscr{H}_T:g\leq h}\mathbb{P}\Big(\big\|\widehat{\rho}_{h^+,T}-\widehat{\rho}_{g,T}\big\|_{L^\infty(D)}>\sqrt{0.8\mathcal{M}}\sigma(g,T)\Big)\Bigg)^{1/2}\in\mathrm{O}(\sigma(h_\rho,T)).$$

Following again the lines of [41], we obtain

$$\mathbb{P}\Big(\big\|\widehat{\rho}_{h^+,T}-\widehat{\rho}_{g,T}\big\|_{L^\infty(D)}>\sqrt{0.8\mathcal{M}}\sigma(g,T)\Big)\leq \mathbb{P}\Big(\big\|\widehat{\rho}_{h^+,T}-\mathbb{E}\widehat{\rho}_{h^+,T}\big\|_{L^\infty(D)}>\frac{1}{4}\sqrt{0.8\mathcal{M}}\sigma(h^+,T)\Big)$$
$$+\mathbb{P}\Big(\big\|\widehat{\rho}_{g,T}-\mathbb{E}\widehat{\rho}_{g,T}\big\|_{L^\infty(D)}>\frac{1}{4}\sqrt{0.8\mathcal{M}}\sigma(g,T)\Big).$$

Let $\gamma\geq 1$. Clearly, by definition of $\sigma(g,T)$, there exists $T(\gamma)>0$ such that, for any $T\geq T(\gamma)$ and any $g\leq h_\rho$, $g\in\mathscr{H}_T$, we have

$$\frac{1}{4}\sqrt{0.8\mathcal{M}}\sigma(g,T)\geq c_\gamma\Upsilon_{g,T}(\gamma\log(g^{-1}))=c_\gamma\frac{\gamma\log(g^{-1})(\log T)^2}{Tg^d}+\psi_d(g^d)\sqrt{\frac{\gamma\log(g^{-1})}{T}},$$

where c_γ is the constant appearing in Lemma A.9. Thus, using Lemma A.9, we obtain for $T\geq T(\gamma)$ that

$$\mathbb{P}\Big(\big\|\widehat{\rho}_{g,T}-\mathbb{E}[\widehat{\rho}_{g,T}]\big\|_{L^\infty(D)}>\frac{1}{4}\sqrt{0.8\mathcal{M}}\sigma(g,T)\Big)\leq \mathrm{e}^{-\gamma\log(g^{-1})}=g^\gamma=:\iota_\gamma(g)$$

and hence

$$\sum_{g\in\mathscr{H}_T:g\leq h}\mathbb{P}\Big(\big\|\widehat{\rho}_{h^+,T}-\widehat{\rho}_{g,T}\big\|_{L^\infty(D)}>\sqrt{0.8\mathcal{M}}\sigma(g,T)\Big)\leq\sum_{g\in\mathscr{H}_T:g\leq h}(\iota_\gamma(g)+\iota_\gamma(h^+))\leq 2\iota_\gamma(h)\log T.$$

Thus, choosing γ large enough demonstrates that

$$\sum_{h\in\mathscr{H}_T:h<h_\rho}\sigma(h,T)\Bigg(\sum_{g\in\mathscr{H}_T:g\leq h}\mathbb{P}\Big(\big\|\widehat{\rho}_{h^+,T}-\widehat{\rho}_{g,T}\big\|_{L^\infty(D)}>\sqrt{0.8\mathcal{M}}\sigma(g,T)\Big)\Bigg)^{1/2}$$
$$\leq\sum_{h\in\mathscr{H}_T:h<h_\rho}\sigma(h,T)\sqrt{2\iota_\gamma(h)\log T}\leq\sqrt{2h_\rho^\gamma(\log T)^3}\sigma(h_{\min},T)\in\mathrm{O}(\sigma(h_\rho,T)),$$

as desired. ∎

Proof of Theorem A.13. Let us first verify that under (𝔒1) the heat kernel bound (𝔄1) holds. Arguing as in the proof of Theorem 3.2 of Masuda [57], we see that $\mathfrak{F}\mu$ and $\varphi^x_{X_t}$ are integrable for any $x \in \mathbb{R}$ and $t > 0$ and hence we can obtain the invariant density ρ and the transition density p_t of $\boldsymbol{X}$ via inverse Fourier transformation through

$$\rho(y) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \mathrm{e}^{-\mathrm{i}\langle y,\lambda\rangle} \{\mathfrak{F}\mu\}(\lambda)\, \mathrm{d}\lambda, \quad y \in \mathbb{R}^d,$$

and

$$p_t(x,y) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}\langle y,\lambda\rangle} \varphi^x_{X_t}(\lambda)\, \mathrm{d}\lambda, \quad x, y \in \mathbb{R}^d, t > 0.$$

Again, as in the proof of Theorem 3.2 in [57], it follows that under (𝔒1),

$$\big|\varphi^x_{X_t}(\lambda)\big| \le \exp\Big(-\frac{1}{2}\lambda^\top \Big(\int_0^t \mathrm{e}^{-sB} Q \mathrm{e}^{-sB^\top}\, \mathrm{d}s \Big)\lambda \Big), \quad x, \lambda \in \mathbb{R}^d, t > 0.$$

Thus, using the characterization of the multivariate normal distribution, we obtain

$$\begin{aligned} p_t(x,y) &\le \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \exp\Big(-\frac{1}{2}\lambda^\top \Big(\int_0^t \mathrm{e}^{-sB} Q \mathrm{e}^{-sB^\top}\, \mathrm{d}s \Big)\lambda \Big)\, \mathrm{d}\lambda \\ &= \frac{1}{(2\pi)^{d/2}} \Big(\det\Big(\int_0^t \mathrm{e}^{-sB} Q \mathrm{e}^{-sB^\top}\, \mathrm{d}s \Big) \Big)^{-1/2}. \end{aligned}$$

Observing that

$$\begin{aligned} \lim_{t\downarrow 0} t^{d/2} \Big(\det\Big(\int_0^t \mathrm{e}^{-sB} Q \mathrm{e}^{-sB^\top}\, \mathrm{d}s \Big) \Big)^{-1/2} &= \Big(\det\Big(\lim_{t\downarrow 0} \frac{1}{t} \int_0^t \mathrm{e}^{-sB} Q \mathrm{e}^{-sB^\top}\, \mathrm{d}s \Big) \Big)^{-1/2} \\ &= \det(Q)^{-1/2} < \infty, \end{aligned}$$

where finiteness is a consequence of invertibility of Q by (𝔒1), it follows that for any $d \ge 1$, there exists a constant $c > 0$ such that

$$\sup_{x,y\in\mathbb{R}^d} p_t(x,y) \le c t^{-d/2}, \quad t \in (0,1].$$

Thus indeed, for any dimension $d \in \mathbb{N}$, (𝔄1) holds. Next, in scenario (i), [57, Theorem 4.3] gives the exponential β-mixing property and the proof of Theorem 2.6 in [55] along with [55, Proposition 3.8] yields V-exponential ergodicity with $V(x) \sim (1 + \|x\|^p)$. This together with (A.20) entails that in scenario (i), we have $\mathbf{X} \in \boldsymbol{\Sigma} \cap \boldsymbol{\Theta}$. Finally, $\boldsymbol{X} \in \boldsymbol{\Theta}$ in scenario (ii) follows from the considerations above and Lemma A.3 due to the fact that the combination of (𝔒1) and the logarithmic moment condition imply that every compact set is small and hence petite since $\boldsymbol{X}$ is strong Feller and by [47, Theorem 3.1] ergodic (see Proposition A.20) and hence (𝔒3) implies V-polynomial ergodicity of degree $\alpha - 1 > 1$ with $V(x) = C(\log|x|)^\alpha$ in dimension $d = 1$ by [48, Corollary 1]. The statements on the estimation rates are now an immediate consequence of Corollary A.10 and Theorem A.11 and the fact that $\rho \in \mathcal{C}^\infty_b$ has arbitrary Hölder smoothness. ∎

Proof of Lemma A.15. We will employ Theorem 6.2.9 and Exercise 6.4.7 of [7] to show the first assertion. So first we must verify that condition **(C1)** on page 365 of [7] holds. Since (𝔍1) holds, we only have to show that there exists a constant $K_1 > 0$ such that, for all $x, y \in \mathbb{R}^d$,

$$\sum_{i,j=1}^d (\sigma_{i,j}(x) - \sigma_{i,j}(y))^2 + \int_{\mathbb{R}^d} \|\gamma(x)z - \gamma(y)z\|^2\, \nu(\mathrm{d}z) \le K_1 \|x - y\|^2,$$

where $\sigma_{i,j}(x)$ denotes the components of $\sigma(x) \in \mathbb{R}^{d\times d}$ for any $x \in \mathbb{R}^d$. (Ƒ1) implies that there exists a finite constant $L_{i,j} > 0$ for any $i, j \in \{1, \dots, d\}$, such that $\sigma_{i,j} \colon \mathbb{R}^d \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant $L_{i,j} > 0$ and hence we have for $x, y \in \mathbb{R}^d$

$$\sum_{i,j=1}^{d} (\sigma_{i,j}(x) - \sigma_{i,j}(y))^2 \le 2d \max_{i,j\in\{1,\dots,d\}} L_{i,j}^2 \|x - y\|^2.$$

Furthermore, we have for $x, y \in \mathbb{R}^d$ by the Lipschitz continuity of γ

$$\int_{\mathbb{R}^d} \|\gamma(x)z - \gamma(y)z\|^2 \, \nu(\mathrm{d}z) \le L_\gamma^2 \|x - y\|^2 \int_{\mathbb{R}^d} \|z\|^2 \, \nu(\mathrm{d}z),$$

where we denote the Lipschitz constant of γ by L_γ. By (Ƒ3), $\int_{\mathbb{R}^d} \|z\|^2 \, \nu(\mathrm{d}z)$ is finite and hence **(C1)** holds. To verify the growth condition **(C2)** on page 366 of [7], we have to show that there exists a constant K_2 such that, for all $x \in \mathbb{R}^d$,

$$\int_{\mathbb{R}^d} \|\gamma(x)z\|^2 \, \nu(\mathrm{d}z) \le K_2(1 + \|x\|^2).$$

Since γ is Lipschitz continuous by (Ƒ1), there exists a constant $K > 0$ such that the linear growth condition $\|\gamma(x)\| \le K(1 + \|x\|)$ holds for all $x \in \mathbb{R}^d$, and thus we have, for $x \in \mathbb{R}^d$,

$$\int_{\mathbb{R}^d} \|\gamma(x)z\|^2 \, \nu(\mathrm{d}z) \le 2K^2(1 + \|x\|^2) \int_{\mathbb{R}^d} \|z\|^2 \, \nu(\mathrm{d}z).$$

Again by (Ƒ3), $\int_{\mathbb{R}^d} \|z\|^2 \, \nu(\mathrm{d}z)$ is finite and hence **(C2)** holds for $K_2 = 2K^2 \int_{\mathbb{R}^d} \|z\|^2 \, \nu(\mathrm{d}z)$. Since Assumption 6.2.8 in [7] is trivially fulfilled, the first assertion follows by Theorem 6.2.9 and Exercise 6.4.7 of [7].

We proceed by showing the second assertion. Equation (1.21) of [21] is in the setting of (A.21) equivalent to $\kappa_\alpha(x, z) = \|\gamma(x)z\|^{d+\alpha}\nu(z) \ge 0$ for all $x \in \mathbb{R}^d$ and almost every $z \in \mathbb{R}^d$. Since ν is a density, this assumption is fulfilled.

For assumption $(\mathbf{H}^a)$ of [21] to hold, we only need to show that there exists a $\beta \in (0, 1)$ such that the function $a(x) \coloneqq \sigma(x)\sigma^\top(x)$ is β-Hölder continuous. However this follows directly from the Lipschitz continuity and the boundedness of σ imposed in (Ƒ1), as can be seen in the proof of Lemma 1 of [4]. Now we note that assumption $(\mathbf{H}^\kappa)$ of [21] follows by (Ƒ2). ∎

Proof of Corollary A.16. Since (Ƒ1) and (Ƒ3) imply that b^* is bounded, arguing as in the proof of Lemma 1 of [4] and using Lemma A.15 entails that b^* belongs to the Kato class $\mathbb{K}_2$ for $d \ge 2$. For the definition of $\mathbb{K}_2$, see (2.28) in [21]. Existence of transition densities and the heat kernel estimate now follow directly from Corollary 1.5 of [21] and Lemma A.15 for $d \ge 2$ and as described in Lemma 1 of [4], the same conclusions may be drawn for dimension $d = 1$ by adapting the arguments in [21]. Now note that (A.23), $t \le 1$ and $\alpha \in (0, 2)$ imply

$$\begin{aligned} p_t(x, y) &\le C(t^{-d/2} \exp(-\|x - y\|^2/(\lambda t)) + \|\kappa_\alpha\|_\infty t(\|x - y\| + t^{1/2})^{-d-\alpha}) \\ &\le C(t^{-d/2} + t^{1-(d+\alpha)/2}) \le Ct^{-d/2}, \end{aligned}$$

where the value of C changes from line to line. This completes the proof. ∎

Proof of Proposition A.17. To verify the assertion, we show that the solution of (A.21) $\boldsymbol{X}$ satisfies the assumptions of Theorem 2.2 (ii) of [55] which are Assumption 1, 2(a)' and 3* of [55] and [56], respectively. Assumption 1 follows directly from (𝒥1). Now, define $b_u^*(x) := b^*(x) - \int_{u<\|z\|\leq 1} \gamma(x)z\,\nu(\mathrm{d}z) = b(x) - \int_{\|z\|>u} \gamma(x)z\,\nu(\mathrm{d}z)$, and let the diffusion process $Y^u = (Y_t^u)_{t\geq 0}$ be given by

$$Y_t^u = x + \int_0^t b_u^*(x)(Y_s^u)\,\mathrm{d}s + \int_0^t \sigma(Y_s^u)\,\mathrm{d}W_s.$$

For Assumption 2(a)' to be fulfilled, we first have to show that, for any $u \in (0,1)$, there exists $\Delta > 0$ such that $\mathbb{P}_x(Y_\Delta^u \in B) > 0$ for any $x \in \mathbb{R}^d$ and any nonempty open set $B \subset \mathbb{R}^d$. Since Y^u is a continuous diffusion process with bounded and Lipschitz coeffcients b_u^*, σ and $a = \sigma\sigma^T$ is uniformly elliptic, it follows from classical results, see e.g. [75, Theorem A], that for any $x \in \mathbb{R}^d$ and $t > 0$, the transition function $P_t^u(x,\cdot)$ of Y^u has a transition density with full support and hence any Δ-skeleton of Y^u is open set irreducible, showing that Assumtpion 2(a)' is in place. It remains to show that Assumption 3* of [55] is satisfied which is, that there exists a function $V \in Q^*$, where

$$Q^* := \Big\{ f\colon \mathbb{R}^d \to \mathbb{R}_+ : f \in \mathcal{C}^2, f(x) \to \infty \text{ as } \|x\| \to \infty, \text{ and there exists a locally bounded, measurable function } \bar{f}, \text{ such that } \int_{\|z\|>1} f(x+\gamma(x)z)\,\nu(\mathrm{d}z) \leq \bar{f}(x),\ \forall x \in \mathbb{R}^d \Big\},$$

such that there are constants $c_1, c_2 > 0$, for which the Lyapunov drift criterion

$$\mathsf{A}V \leq -c_1 V + c_2 \tag{A.49}$$

holds, where A denotes the extended generator of $\boldsymbol{X}$ acting on Q^* by

$$\mathsf{A}f(x) = \langle \nabla f(x), b^*(x)\rangle + \tfrac{1}{2}\operatorname{tr}(\nabla^2 f(x)\sigma(x)\sigma^T(x)) + \int_{\mathbb{R}^d} f(x+\gamma(x)z) - f(x) - \mathbf{1}_{\|z\|\leq 1}\langle \nabla f(x), \gamma(x)z\rangle \nu(\mathrm{d}z), \quad x \in \mathbb{R}^d, f \in Q^*.$$

Now, for $\eta \in (0, \eta_0 c_\gamma^{-1} \wedge 1)$, where $c_\gamma := \|\gamma\|_\infty$, let V^η be a positive and increasing function in $\mathcal{C}^2(\mathbb{R}^d, \mathbb{R})$ such that $V^\eta = \mathrm{e}^{\eta\|x\|}$ for all $\|x\| > c_V$, where $c_V > 0$. Then, it holds for $i, j \in \{1, \dots, d\}$ and $\|x\| > c_V$,

$$\begin{aligned} \partial_i V^\eta(x) &= \eta \mathrm{e}^{\eta\|x\|}\frac{x_i}{\|x\|}, \\ \partial_{ij}^2 V^\eta(x) &= \eta^2 \mathrm{e}^{\eta\|x\|}\frac{x_i x_j}{\|x\|^2} - \eta \mathrm{e}^{\eta\|x\|}\frac{x_i x_j}{\|x\|^3} + \eta \mathrm{e}^{\eta\|x\|}\|x\|^{-1}\delta_{ij}, \end{aligned} \tag{A.50}$$

where δ_{ij} denotes the Kronecker delta. Furthermore, since $V^\eta \in \mathcal{C}^2(\mathbb{R}^d; \mathbb{R})$, for $i, j \in \{1, \dots, d\}$ the functions $V^\eta, \partial_i V^\eta, \partial_{ij}^2 V^\eta$ are bounded by a constant $c_D > 0$ for $\|x\| \leq c_V$ and hence

$$\begin{aligned} \int_{\|z\|>1} V^\eta(x+\gamma(x)z)\,\nu(\mathrm{d}z) &\leq \int_{\|z\|>1} \left(\mathrm{e}^{\eta\|x+\gamma(x)z\|} + c_D\right)\nu(\mathrm{d}z) \\ &\leq \mathrm{e}^{\eta\|x\|}\int_{\|z\|>1} \mathrm{e}^{c_\gamma \eta\|z\|}\nu(\mathrm{d}z) + c_D\nu(\mathbb{R}^d \setminus B_1), \end{aligned}$$

implying that $V_S^\eta \in Q^*$ for all $\eta \leq \frac{\eta_0}{c_\gamma}$. This last condition is satisfied by our choice of η. To conclude the proof, the only thing left to show is that there exists $0 < \eta \leq \frac{\eta_0}{c_\gamma}$ such that (A.49) holds for V^η. Note that, by the mean value theorem, the definition of b^* and the Cauchy–Schwarz inequality, we have for any $f \in Q^*$

$$\begin{aligned}
\mathsf{A}f(x) &= \langle \nabla f(x), b(x)\rangle + \tfrac{1}{2}\operatorname{tr}(\nabla^2 f(x)\sigma(x)\sigma^\top(x)) \\
&\qquad + \int_{\mathbb{R}^d} f(x+\gamma(x)z) - f(x) - \langle \nabla f(x), \gamma(x)z\rangle \nu(\mathrm{d}z) \\
&\leq \langle \nabla f(x), b(x)\rangle + \tfrac{1}{2}\operatorname{tr}(\nabla^2 f(x)\sigma(x)\sigma^\top(x)) \\
&\qquad + \int_{\mathbb{R}^d} \sup_{t\in[0,1]} \|\nabla f(x+t\gamma(x)z) - \nabla f(x)\| \|\gamma(x)z\| \nu(\mathrm{d}z) \\
&\leq \mathsf{A}_c f(x) + \mathsf{A}_d f(x),
\end{aligned}$$

where, for $\mathsf{H}^2 f(x)$ denoting the Hessian of f evaluated at x,

$$\begin{aligned}
\mathsf{A}_c f(x) &:= \langle \nabla f(x), b(x)\rangle + \tfrac{1}{2}\operatorname{tr}(\nabla^2 f(x)\sigma(x)\sigma^T(x)), \\
\mathsf{A}_d f(x) &:= c_\gamma^2 \int_{\mathbb{R}^d} \sup_{t\in[0,1]} \|\mathsf{H}^2 f(x+t\gamma(x)z)\| \|z\|^2 \nu(\mathrm{d}z).
\end{aligned}$$

We start by investigating the jump part. By (A.50) and the fact that the operator norm can be bounded by the Frobenius norm $\|\cdot\|_F$, we get for $\|x\| > c_V$

$$\begin{aligned}
\|\mathsf{H}^2 V^\eta(x)\| &\leq \|\mathsf{H}^2 \mathrm{e}^{\eta\|x\|}\|_F = \left(\sum_{i,j=1}^d \left(\eta^2 \mathrm{e}^{\eta\|x\|}\frac{x_i x_j}{\|x\|^2} - \eta \mathrm{e}^{\eta\|x\|}\frac{x_i x_j}{\|x\|^3} + \eta \mathrm{e}^{\eta\|x\|}\|x\|^{-1}\delta_{ij}\right)^2\right)^{\frac{1}{2}} \\
&\leq 2\eta \mathrm{e}^{\eta\|x\|}\left(\sum_{i,j=1}^d \left(\eta^2 \frac{x_i^2 x_j^2}{\|x\|^4} + \frac{x_i^2 x_j^2}{\|x\|^6} + \|x\|^{-2}\delta_{ij}\right)\right)^{\frac{1}{2}} \leq 2^{3/2}\sqrt{d}\eta \mathrm{e}^{\eta\|x\|}\left(\eta^2 + 2\|x\|^{-2}\right)^{\frac{1}{2}}.
\end{aligned}$$

Since we can choose c_V to be large, we can without loss of generality assume $c_V \geq \sqrt{2}\eta^{-1}$ and, additionally, $V^\eta \in \mathcal{C}^2$ implies that there exists a real-valued function $c_{\mathsf{H}}(\eta) > 0$ on $(0,\infty)$ such that $\|\mathsf{H}^2 V^\eta(x)\| < c_{\mathsf{H}}(\eta)$ for all $\|x\| \leq c_V$. Thus, we have $\|\mathsf{H}^2 V^\eta(x)\| \leq 4\sqrt{d}\eta^2 \mathrm{e}^{\eta\|x\|} + c_{\mathsf{H}}(\eta)$, $x \in \mathbb{R}^d$, and we can conclude

$$\begin{aligned}
\mathsf{A}_d V^\eta(x) &\leq 4c_\gamma^2\sqrt{d}\eta^2 \int_{\mathbb{R}^d} \sup_{t\in[0,1]} \mathrm{e}^{\eta\|x+t\gamma(x)z\|}\|z\|^2\nu(\mathrm{d}z) + c_\gamma^2 c_{\mathsf{H}}(\eta)\int_{\mathbb{R}^d}\|z\|^2\nu(\mathrm{d}z) \\
&\leq \eta^2 \mathrm{e}^{\eta\|x\|} 4c_\gamma^2\sqrt{d}\int_{\mathbb{R}^d} \mathrm{e}^{\eta_0\|z\|}\|z\|^2\nu(\mathrm{d}z) + c_\gamma^2 c_{\mathsf{H}}(\eta)\int_{\mathbb{R}^d}\|z\|^2\nu(\mathrm{d}z) =: c_{d,1}\eta^2\mathrm{e}^{\eta\|x\|} + c_{d,2}(\eta),
\end{aligned} \tag{A.51}$$

where $c_{d,1}, c_{d,2}(\eta)$ are positive and finite because of (𝔍3) and $\eta < \eta_0 c_\gamma^{-1}$. Now we turn our attention to the continuous part. From now on, without loss of generality, we assume that $c_V \geq c_1$ in (𝔍3). Then, for $\|x\| > c_V \geq \eta^{-1}$, we have by (𝔍1), (𝔍3) and (A.50)

$$\mathsf{A}_c V^\eta(x) \leq -c_1\eta\mathrm{e}^{\eta\|x\|} + \frac{c_2}{2}\sum_{k=1}^d \left|\eta^2\mathrm{e}^{\eta\|x\|}\frac{x_i^2}{\|x\|^2} + \eta\mathrm{e}^{\eta\|x\|}\|x\|^{-1} - \eta\mathrm{e}^{\eta\|x\|}\frac{x_i^2}{\|x\|^3}\right|$$

$$\leq \eta \mathrm{e}^{\eta\|x\|}\left(-c_1 + \frac{3c_2 d}{2}\eta\right),$$

and since $V^\eta \in \mathcal{C}^2(\mathbb{R}^d;\mathbb{R})$, there exists a real-valued function $c_c(\eta)$ on $(0,\infty)$ such that $\mathsf{A}_c V^\eta(x) \leq c_c(\eta)$ for all $\|x\| \leq c_V$. Hence, we have

$$\mathsf{A}_c V^\eta(x) \leq \eta \mathrm{e}^{\eta\|x\|}\left(-c_1 + \frac{3c_2 d}{2}\eta\right) + c_C(\eta) + c_1 \mathrm{e}^{c_V} =: \eta \mathrm{e}^{\eta\|x\|}\left(-c_1 + \frac{3c_2 d}{2}\eta\right) + c_{c,1}(\eta), \tag{A.52}$$

where we used that $\eta < 1$, by assumption. Combining (A.51) and (A.52) yields

$$\mathsf{A} V^\eta(x) \leq \eta \mathrm{e}^{\eta\|x\|}\left(-c_1 + \eta\left(\frac{3c_2 d}{2} + c_{d,1}\right)\right) + c_{d,2}(\eta) + c_{c,1}(\eta).$$

Choosing $\eta^* = 1 \wedge \eta_0 c_\gamma^{-1} \wedge \frac{c_1}{3c_2 d + 2c_{d,1}}$ implies

$$\mathsf{A} V^{\eta^*}(x) \leq -\frac{c_1 \eta^*}{2}\mathrm{e}^{\eta^*\|x\|} + c_{d,2}(\eta^*) + c_{c,1}(\eta^*),$$

and thus (A.49) holds for $V^{\eta^*} \in Q^*$. Now, Theorem 2.2 (ii) and Proposition 3.8 of [55] show the required assertion. ∎

References

[1] R. Adamczak. "A tail inequality for suprema of unbounded empirical processes with applications to Markov chains". In: *Electron. J. Probab.* 13 (2008), no. 34, 1000–1034.

[2] R. Adamczak and W. Bednorz. "Exponential concentration inequalities for additive functionals of Markov chains". In: *ESAIM Probab. Stat.* 19 (2015), pp. 440–481.

[3] L. H. R. Alvarez. "Stochastic forest stand value and optimal timber harvesting". In: *SIAM J. Control Optim.* 42.6 (2004), pp. 1972–1993.

[4] C. Amorino and A. Gloter. "Invariant density adaptive estimation for ergodic jump-diffusion processes over anisotropic classes". In: *J. Statist. Plann. Inference* 213 (2021), pp. 106–129.

[5] C. Amorino and A. Gloter. *Minimax rate of estimation for invariant densities associated to continuous stochastic differential equations over anisotropic Holder classes*. arXiv:2110.02774. 2021. arXiv: 2110.02774 [math.ST].

[6] C. Amorino and E. Nualart. "Optimal convergence rates for the invariant density estimation of jump-diffusion processes". arXiv:2101.08548. 2021. arXiv: 2101.08548.

[7] D. Applebaum. *Lévy Processes and Stochastic Calculus*. 2nd ed. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2009.

[8] D. Bakry, P. Cattiaux, and A. Guillin. "Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré". In: *J. Funct. Anal.* 254.3 (2008), pp. 727–759.

[9] Y. Baraud. "A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression". In: *Bernoulli* 16.4 (2010), pp. 1064–1085.

[10] P. Bertail and G. Ciołek. "New Bernstein and Hoeffding type inequalities for regenerative Markov chains". In: *ALEA Lat. Am. J. Probab. Math. Stat.* 16.1 (2019), pp. 259–277.

[11] J. Bierkens, G. O. Roberts, and P.-A. Zitt. "Ergodicity of the zigzag process". In: *Ann. Appl. Probab.* 29.4 (2019), pp. 2266–2301.

[12] R. M. Blumenthal and R. K. Getoor. *Markov processes and potential theory*. Pure and Applied Mathematics, Vol. 29. Academic Press, New York-London, 1968, pp. x+313.

[13] D. Bosq. *Nonparametric statistics for stochastic processes*. 2nd ed. Vol. 110. Lecture Notes in Statistics. Estimation and prediction. Springer-Verlag, New York, 1998, pp. xvi+210.

[14] D. Bosq. "Parametric rates of nonparametric estimators and predictors for continuous time processes". In: *Ann. Statist.* 25.3 (1997), pp. 982–1000.

[15] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, Oxford, 2013, pp. x+481.

[16] R. C. Bradley. "Basic properties of strong mixing conditions. A survey and some open questions". In: *Probab. Surv.* 2 (2005). Update of, and a supplement to, the 1986 original, pp. 107–144.

[17] N. Brosse, A. Durmus, and E. Moulines. "The promises and pitfalls of Stochastic Gradient Langevin Dynamics". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2018, pp. 8268–8278.

[18] J. V. Castellana and M. R. Leadbetter. "On smoothed probability density estimation for stationary processes". In: *Stochastic Process. Appl.* 21.2 (1986), pp. 179–193.

[19] P. Cattiaux and A. Guillin. "Deviation bounds for additive functionals of Markov processes". In: *ESAIM Probab. Stat.* 12 (2008), pp. 12–29.

[20] M.-F. Chen. *Eigenvalues, inequalities, and ergodic theory*. Probability and its Applications (New York). Springer-Verlag London, Ltd., London, 2005, pp. xiv+228.

[21] Z.-Q. Chen, E. Hu, L. Xie, and X. Zhang. "Heat kernels for non-symmetric diffusion operators with jumps". In: *Journal of Differential Equations* 263.10 (2017), pp. 6576 –6634.

[22] S. Christensen, C. Strauch, and L. Trottner. "Learning to reflect: A unifying approach for data-driven stochastic control strategies". arXiv:2104.11496. 2021. arXiv: 2104.11496.

[23] S. J. M. Clémencon. "Moment and probability inequalities for sums of bounded additive functionals of regular Markov chains via the Nummelin splitting technique". In: *Statist. Probab. Lett.* 55.3 (2001), pp. 227–238.

[24] F. Comte and F. Merlevède. "Adaptive estimation of the stationary density of discrete and continuous time mixing processes". In: *ESAIM Probab. Statist.* 6 (2002). New directions in time series analysis (Luminy, 2001), pp. 211–238.

[25] G. Da Prato, K. D. Elworthy, and J. Zabczyk. "Strong Feller property for stochastic semilinear equations". In: *Stochastic Analysis and Applications* 13.1 (1995), pp. 35–45.

[26] G. Da Prato and J. Zabczyk. "Smoothing properties of transition semigroups in Hilbert spaces". In: *Stochastics and Stochastic Reports* 35.2 (1991), pp. 63–77.

[27] A. Dalalyan. "Sharp adaptive estimation of the drift function for ergodic diffusions". In: *Ann. Statist.* 33.6 (2005), pp. 2507–2528.

[28] A. Dalalyan and M. Reiß. "Asymptotic statistical equivalence for ergodic diffusions: the multidimensional case". In: *Probab. Theory Relat. Fields* 137.1 (2007), pp. 25–47.

[29] J. A. Davydov. "Mixing conditions for Markov chains". In: *Teor. Verojatnost. i Primenen.* 18 (1973), pp. 321–338.

[30] J. Dedecker, P. Doukhan, G. Lang, J. R. León R., S. Louhichi, and C. Prieur. *Weak dependence: with examples and applications*. Vol. 190. Lecture Notes in Statistics. Springer, New York, 2007, pp. xiv+318.

[31] J. Dedecker and S. Gouëzel. "Subgaussian concentration inequalities for geometrically ergodic Markov chains". In: *Electron. Commun. Probab.* 20 (2015), no. 64, 12.

[32] N. Dexheimer and C. Strauch. *Estimating the characteristics of stochastic damping Hamiltonian systems from continuous observations*. arXiv:2109.13190. 2021. arXiv: 2109.13190 [math.ST].

[33] S. Dirksen. "Tail bounds via generic chaining". In: *Electron. J. Probab.* 20 (2015), no. 53, 29.

[34] D. Down and S. P. Meyn. "Piecewise linear test functions for stability and instability of queueing networks". In: *Queueing Systems Theory Appl.* 27.3-4 (1997), 205–226 (1998).

[35] D. Down, S. P. Meyn, and R. L. Tweedie. “Exponential and uniform ergodicity of Markov processes”. In: *Ann. Probab.* 23.4 (1995), pp. 1671–1691.

[36] N. H. Du, N. T. Dieu, and N. N. Nhu. “Conditions for permanence and ergodicity of certain SIR epidemic models”. In: *Acta Appl. Math.* 160 (2019), pp. 81–99.

[37] A. Durmus and E. Moulines. “High-dimensional Bayesian inference via the unadjusted Langevin algorithm”. In: *Bernoulli* 25.4A (2019), pp. 2854–2882.

[38] J.-P. Eckmann and M. Hairer. “Uniqueness of the invariant measure for a stochastic PDE driven by degenerate noise”. In: *Comm. Math. Phys.* 219.3 (2001), pp. 523–565.

[39] C. Fuchs. *Inference for diffusion processes – With applications in life sciences*. Springer, Heidelberg, 2013, pp. xx+430.

[40] F. Gao, A. Guillin, and L. Wu. “Bernstein type’s concentration inequalities for symmetric Markov processes”. In: *Teor. Veroyatnost. i Primenen* 58.3 (2013), pp. 521–549.

[41] E. Giné and R. Nickl. “An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation”. In: *Probab. Theory Relat. Fields* 143.3-4 (2009), pp. 569–596.

[42] M. Giordano and K. Ray. *Nonparametric Bayesian inference for reversible multi-dimensional diffusions*. 2020. arXiv: 2012.12083.

[43] U. Grenander and M. I. Miller. “Representations of knowledge in complex systems”. In: *J. Roy. Statist. Soc. Ser. B* 56.4 (1994). With discussion and a reply by the authors, pp. 549–603.

[44] M. Hairer and J. Mattingly. “The strong Feller property for singular stochastic PDEs”. In: *Ann. Inst. Henri Poincaré Probab. Stat.* 54.3 (2018), pp. 1314–1340.

[45] J. Hawkes. “Potential theory of Lévy processes”. In: *Proc. London Math. Soc. (3)* 38.2 (1979), pp. 335–352.

[46] K. Ichihara and H. Kunita. “A classification of the second order degenerate elliptic operators and its probabilistic characterization”. In: *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* 30 (1974), pp. 235–254.

[47] G. Jongbloed, F. H. van der Meulen, and A. W. van der Vaart. “Nonparametric inference for Lévy-driven Ornstein-Uhlenbeck processes”. In: *Bernoulli* 11.5 (2005), pp. 759–791.

[48] P. Kevei. “Ergodic properties of generalized Ornstein-Uhlenbeck processes”. In: *Stochastic Process. Appl.* 128.1 (2018), pp. 156–181.

[49] Y. A. Kutoyants. *Statistical Inference for Ergodic Diffusion Processes*. Springer Series in Statistics. New York: Springer, 2004.

[50] F. Leblanc. “Density estimation for a class of continuous time processes”. In: *Math. Methods Statist.* 6.2 (1997), pp. 171–199.

[51] M. Lemanczyk. “General Bernstein-Like Inequality for Additive Functionals of Markov Chains”. In: *Journal of Theoretical Probability* (2020).

[52] O. Lepski. “Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure”. In: *Ann. Statist.* 41.2 (2013), pp. 1005–1034.

[53] P. Lezaud. "Chernoff and Berry–Esséen inequalities for Markov processes". In: *ESAIM: Probability and Statistics* 5 (2001), pp. 183–201.

[54] B. Maslowski. "Strong Feller property for semilinear stochastic evolution equations and applications". In: *Stochastic systems and optimization (Warsaw, 1988)*. Vol. 136. Lect. Notes Control Inf. Sci. Springer, Berlin, 1989, pp. 210–224.

[55] H. Masuda. "Ergodicity and exponential β-mixing bounds for multidimensional diffusions with jumps". In: *Stochastic Process. Appl.* 117.1 (2007), pp. 35 –56.

[56] H. Masuda. "Erratum to: "Ergodicity and exponential β -mixing bound for multidimensional diffusions with jumps" [Stochastic Process. Appl. 117 (2007) 35–56]". In: *Stochastic Processes and their Applications* 119.2 (2009), pp. 676 –678.

[57] H. Masuda. "On multidimensional Ornstein-Uhlenbeck processes driven by a general Lévy process". In: *Bernoulli* 10.1 (2004), pp. 97–120.

[58] F. Merlevède, M. Peligrad, and E. Rio. "Bernstein inequality and moderate deviations under strong mixing conditions". In: *High dimensional probability V: the Luminy volume*. Vol. 5. Inst. Math. Stat. (IMS) Collect. Inst. Math. Statist., Beachwood, OH, 2009, pp. 273–292.

[59] S. P. Meyn and R. L. Tweedie. "Generalized resolvents and Harris recurrence of Markov processes". In: *Doeblin and modern probability (Blaubeuren, 1991)*. Vol. 149. Contemp. Math. Amer. Math. Soc., Providence, RI, 1993, pp. 227–250.

[60] S. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Second. With a prologue by Peter W. Glynn. Cambridge University Press, Cambridge, 2009, pp. xxviii+594.

[61] S. P. Meyn and R. L. Tweedie. "Stability of Markovian processes. II. Continuous-time processes and sampled chains". In: *Adv. in Appl. Probab.* 25.3 (1993), pp. 487–517.

[62] S. P. Meyn and R. L. Tweedie. "Stability of Markovian processes. III. Foster-Lyapunov criteria for continuous-time processes". In: *Adv. in Appl. Probab.* 25.3 (1993), pp. 518–548.

[63] R. Nickl and K. Ray. "Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions". In: *Ann. Statist.* 48.3 (2020), pp. 1383–1408.

[64] E. Nummelin and R. L. Tweedie. "Geometric ergodicity and R-positivity for general Markov chains". In: *Ann. Probability* 6.3 (1978), pp. 404–420.

[65] E. Nummelin. *General irreducible Markov chains and nonnegative operators*. Vol. 83. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1984, pp. xi+156.

[66] D. Paulin. "Concentration inequalities for Markov chains by Marton couplings and spectral methods". In: *Electron. J. Probab.* 20 (2015), no. 79, 32.

[67] J. Paulsen. "Sharp conditions for certain ruin in a risk process with stochastic return on investments". In: *Stochastic Process. Appl.* 75.1 (1998), pp. 135–148.

[68] S. Peszat and J. Zabczyk. "Strong Feller property and irreducibility for diffusions on Hilbert spaces". In: *Ann. Probab.* 23.1 (1995), pp. 157–172.

[69] E. Rio. *Asymptotic theory of weakly dependent random processes*. Vol. 80. Probability Theory and Stochastic Modelling. Translated from the 2000 French edition [MR2117923]. Springer, Berlin, 2017, pp. xviii+204.

[70] E. Rio. *Théorie asymptotique des processus aléatoires faiblement dépendants*. Mathématiques & Applications 31. Berlin: Springer, 2000.

[71] P.-M. Samson. "Concentration of measure inequalities for Markov chains and Φ-mixing processes". In: *Ann. Probab.* 28.1 (2000), pp. 416–461.

[72] K.-i. Sato and M. Yamazato. "Operator-self-decomposable distributions as limit distributions of processes of Ornstein-Uhlenbeck type". In: *Stochastic Process. Appl.* 17.1 (1984), pp. 73–100.

[73] R. L. Schilling and J. Wang. "Strong Feller continuity of Feller processes and semigroups". In: *Infin. Dimens. Anal. Quantum Probab. Relat. Top.* 15.2 (2012), pp. 1250010, 28.

[74] M. Sharpe. *General theory of Markov processes*. Vol. 133. Pure and Applied Mathematics. Academic Press Inc., Boston, MA, 1988, pp. xii+419.

[75] S. J. Sheu. "Some estimates of the transition density of a nondegenerate diffusion Markov process". In: *Ann. Probab.* 19.2 (1991), pp. 538–561.

[76] C. Strauch. "Adaptive invariant density estimation for ergodic diffusions over anisotropic classes". In: *Ann. Statist.* 46.6B (2018), pp. 3451 –3480.

[77] C. Strauch. "Exact adaptive pointwise drift estimation for multidimensional ergodic diffusions". In: *Probab. Theory Relat. Fields* 164.1-2 (2016), pp. 361–400.

[78] C. Strauch. "Sharp adaptive drift estimation for ergodic diffusions: The multivariate case". In: *Stochastic Process. Appl.* 125.7 (2015), pp. 2562 –2602.

[79] Y. Tamura and S. Yamada. "Reliability Analysis Based on a Jump Diffusion Model with Two Wiener Processes for Cloud Computing with Big Data". In: *Entropy* 17.7 (2015), pp. 4533–4546.

[80] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009, pp. xii+214.

[81] R. L. Tweedie. "Topological conditions enabling use of Harris methods in discrete and continuous time". In: *Acta Appl. Math.* 34.1-2 (1994), pp. 175–188.

[82] B. Tzen and M. Raginsky. "Neural Stochastic Differential Equations: Deep Latent Gaussian Models in the Diffusion Limit". arXiv: 1905.09883. 2019. arXiv: 1905.09883.

[83] A. Y. Veretennikov. "On Castellana-Leadbetter's condition for diffusion density estimation". In: *Stat. Inference Stoch. Process.* 2.1 (1999), 1–9 (2000).

[84] G. Viennet. "Inequalities for absolutely regular sequences: application to density estimation". In: *Probab. Theory Relat. Fields* 107.4 (1997), pp. 467–492.

[85] V. A. Volkonskiĭ and J. A. Rozanov. "Some limit theorems for random functions. II". In: *Teor. Verojatnost. i Primenen.* 6 (1961), pp. 202–215.

[86] F. Xi and G. Yin. "The strong Feller property of switching jump-diffusion processes". In: *Statist. Probab. Lett.* 83.3 (2013), pp. 761–767.

[87] F. Xi and C. Zhu. "Jump type stochastic differential equations with non-Lipschitz coefficients: non-confluence, Feller and strong Feller properties, and exponential ergodicity". In: *J. Differential Equations* 266.8 (2019), pp. 4668–4711.

[88] C. Zhu and G. Yin. "On strong Feller, recurrence, and weak stabilization of regime-switching diffusions". In: *SIAM J. Control Optim.* 48.3 (2009), pp. 2003–2031.

[89] J. D. Zund. "George David Birkhoff and John von Neumann: a question of priority and the ergodic theorems, 1931–1932". In: *Historia Math.* 29.2 (2002), pp. 138–156.

Estimating the characteristics of stochastic damping Hamiltonian systems from continuous observations

B

Niklas Dexheimer and Claudia Strauch

Abstract

We consider nonparametric invariant density and drift estimation for a class of multidimensional degenerate resp. hypoelliptic diffusion processes, so-called stochastic damping Hamiltonian systems or kinetic diffusions, under anisotropic smoothness assumptions on the unknown functions. The analysis is based on continuous observations of the process, and the estimators' performance is measured in terms of the sup-norm loss. Regarding invariant density estimation, we obtain highly nonclassical results for the rate of convergence, which reflect the inhomogeneous variance structure of the process. Concerning estimation of the drift vector, we suggest both non-adaptive and fully data-driven procedures. All of the aforementioned results strongly rely on tight uniform moment bounds for empirical processes associated to deterministic and stochastic integrals of the investigated process, which are also proven in this paper.

B.1 Introduction

Diffusion processes have been in the center of attention of the statistical analysis of stochastic processes for a long time due to their various fields of application, e.g. meteorology, genetics, financial mathematics and neuroscience. A standard assumption often imposed for the in-depth investigation is the strict ellipticity of the diffusion operator. In particular, this regularity condition (together with other assumptions) allows to verify helpful analytical tools such as the existence of a spectral gap. However, the nondegeneracy assumption excludes many processes which are of great importance for applications, for an overview see e.g. [10]. A prominent example are so-called *stochastic damping Hamiltonian systems*, which are often interpreted as a model for the coupled velocity $\mathbf{Y}$ and position $\boldsymbol{X}$ of some object. More specifically, one considers a multivariate diffusion process $\mathbf{Z} = (Z_t)_{t\geq 0} = (X_t, Y_t)_{t\geq 0} = (\boldsymbol{X}, \mathbf{Y})$, which is governed by the stochastic differential equation (SDE)

$$\begin{aligned} \mathrm{d}X_t &= Y_t \,\mathrm{d}t, \\ \mathrm{d}Y_t &= -(c(X_t, Y_t)Y_t + \nabla V(X_t))\,\mathrm{d}t + \sigma(X_t, Y_t)\,\mathrm{d}W_t. \end{aligned} \tag{B.1}$$

Here, $c\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}, V\colon \mathbb{R}^d \to \mathbb{R}, \sigma\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$, and $(W_t)_{t\geq 0}$ is a d-dimensional Brownian motion independent of the random variables X_0, Y_0. In particular, the velocity process $\mathbf{Y}$ is given as a nondegenerate diffusion process, whereas the definition of the position process $\boldsymbol{X}$ implies $\mathbf{Z}$ to be degenerate, respectively hypoelliptic.

In the following, we focus on the nonparametric estimation of the invariant density and the drift of a multidimensional diffusion $\mathbf{Z}$ described by (B.1), assuming a continuous record of observations is available. Invariant density estimation for stochastic damping Hamiltonian systems has been investigated before by [3] and [5], focusing however on estimation based on discrete (and partial) observations. The framework of continuous observations considered here

is different in some respects. On the one hand, it has been shown in the current reference [7] in the scalar framework that the assumption of continuous observations allows for a refined variance analysis and, as a consequence, for the derivation of nonclassical convergence rates for nonparametric estimation of the invariant density. On the other hand, the question of estimation under partial observations, i.e., based only on observations of the position $\boldsymbol{X}$, does not need to be considered in our setting: The assumption of continuous observations trivialises this problem, as the path of $\mathbf{Y}$ is easily computable by derivation of $\boldsymbol{X}$ in this case. Similar to [7]'s approach, we employ a kernel estimator for estimating the invariant density, considering now however a general d-dimensional framework and the sup-norm as a risk criterion. To find a good compromise between appropriate generality and technical complexity, we work with an anisotropic mixture of isotropic smoothness assumptions, i.e., we assume that the components $\boldsymbol{X}$ and $\mathbf{Y}$ are associated with Hölder-smoothness coefficient β_1 and β_2, respectively, where $\beta_1 = \beta_2$ does not necessarily hold (see Definition B.6 for the precise description). The upper bounds on the sup-norm rates of convergence which we derive are nonclassical in multiple ways. Firstly, they depend largely on the set, respectively point, where the invariant density is estimated. Secondly, the ratio of the smoothness parameters of the invariant density's assumed anisotropic Hölder regularity also plays a vital role in the rate of convergence. The function Υ introduced in (B.11) reflects these specifities. Lastly, the proven rates of convergence are faster than in the classical case of kernel density estimation based on a discrete set of i.i.d. observations. All of the aforementioned nonclassicalities stem from the variance bounds for the kernel density estimator in Section B.3.1, which are a result of the degeneracy of $\mathbf{Z}$. For a more thorough explanation of this, see Section B.3.2. The last observation of the rate of convergence being faster than in the classical case of nonparametric density estimation has also been made for nondegenerate diffusion processes, see, e.g., [6] or [17], or more generally for suitable Markov processes [8].

Our (auxiliary) results for estimating the invariant density are not only of theoretical interest, but are also directly applied to the second question addressed in this paper, namely the estimation of the drift function

$$b\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d, \quad b(x,y) \mapsto -(c(x,y)y + \nabla V(x)).$$

To the best of our knowledge, nonparametric drift estimation for stochastic damping Hamiltonian systems has so far only been considered in [2] who prove asymptotic normality of their estimator. By way of contrast, we derive an upper bound on the convergence rate with respect to the the sup-norm risk and, in addition, suggest a sup-norm adaptive estimation scheme. The analysis in particular relies on an extension of the uniform moment bounds for additive functionals of exponentially β-mixing Markov processes obtained in [8] to uniform moment bounds of stochastic integrals with respect to $\mathbf{Y}$, notably exploiting the nondegeneracy of $\mathbf{Y}$. Our tools turn out to be sufficiently tight to permit the formulation of a sup-norm adaptive drift estimation procedure as well. Remarkably, although the analysis is essentially based on the fact that the process $\mathbf{Z}$ is exponentially β-mixing, the associated mixing constants are not relevant for the adaptive procedure. Note that this is in contrast to adaptive estimation procedures for the invariant density of corresponding processes.

The structure of the paper is as follows. In Section B.2, the necessary assumptions on $\mathbf{Z}$ are denoted, which are of a quite technical nature, followed by a collection of explicit conditions on the coefficients of (B.1), which imply the necessary assumptions. Section B.3 consists of the results regarding invariant density estimation, with Section B.3.1 containing the aforementioned

variance bounds for the kernel density estimator and Section B.3.2 introducing the concrete rates of convergence. Lastly, Section B.4 accommodates the results on drift estimation, with Section B.4.1 presenting the extended uniform moment bounds, Section B.4.2 containing the results on the rate of convergence and Section B.4.3 describing our results on adaptive drift estimation. For the sake of readability, all proofs have been deferred to the appendix.

B.2 Preliminaries

Throughout the paper, we assume that the SDE (B.1) admits a unique, non-explosive weak solution whose semigroup is strongly Feller. Furthermore, we impose the following assumptions on the associated solution $\mathbf{Z}$.

($\mathscr{A}1$) The marginal laws of $\mathbf{Z}$ are absolutely continuous, i.e., for any $t > 0$ and $z \in \mathbb{R}^{2d}$, there exists a measurable function $p_t \colon \mathbb{R}^{2d} \times \mathbb{R}^{2d} \to \mathbb{R}_+$ such that

$$P_t(z_1, B) = \int_B p_t(z_1, z_2)\,\mathrm{d}z_2, \quad B \in \mathcal{B}(\mathbb{R}^{2d}),$$

and, moreover, $\mathbf{Z}$ admits a unique absolutely continuous invariant probability measure μ, i.e., there exists a density $\rho \colon \mathbb{R}^{2d} \to \mathbb{R}_+$ such that $\mathrm{d}\mu = \rho\,\mathrm{d}\lambda$ and, for any Borel set B,

$$\mathbb{P}^\mu(Z_t \in B) := \int_{\mathbb{R}^{2d}} P_t(z_1, B)\,\mu(\mathrm{d}z_1) = \int_{\mathbb{R}^{2d}} \int_B p_t(z_1, z_2)\rho(z_1)\,\mathrm{d}z_2\,\mathrm{d}z_1 = \int_B \rho(z_1)\,\mathrm{d}z_1 = \mu(B).$$

($\mathscr{A}2$) For any bounded set $D \subset \mathbb{R}^{2d}$, there exist constants $c_U, c_G > 0$ depending on D such that, for all $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2) \in D$, $t \in (0, 1]$,

$$p_t(z_1, z_2) \le p_t^G(z_1, z_2) + p_t^U(z_1, z_2),$$

where

$$p_t^G(x_1, y_1; x_2, y_2) = c_G t^{-2d} \exp\left(-c_G^{-1}\left(\frac{\|y_1 - y_2\|^2}{4t} + \frac{3\|x_2 - x_1 - \frac{t(y_1+y_2)}{2}\|^2}{t^3}\right)\right),$$

and p_t^U is a non-negative, measurable function such that, for any $z_1 \in D$, $t \in (0, 1]$,

$$\int_{\mathbb{R}^{2d}} p_t^U(z_1, z_2)\,\mathrm{d}z_2 \le c_U \exp\left(-\frac{1}{c_U t}\right).$$

($\mathscr{A}3$) The process $\mathbf{Z}$ started in the invariant measure μ is exponentially β-mixing, i.e., there exist constants $c_\kappa, \kappa > 0$ such that

$$\int \|P_t(z, \cdot) - \mu(\cdot)\|_{\mathrm{TV}}\,\mu(\mathrm{d}z) \;\le\; c_\kappa \mathrm{e}^{-\kappa t}, \quad t \ge 0,$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation norm.

We will refer to this framework as $\mathscr{A}$. Contrary to the rather general and standard assumptions ($\mathscr{A}1$) and ($\mathscr{A}3$), the heat kernel bound ($\mathscr{A}2$) is specifically tailored for the study of kinetic diffusions. This will be more evident from the new set of assumptions $\widetilde{\mathscr{A}}$ introduced below, which contains *explicit* conditions on the coefficients of the SDE (B.1). As will be shown $\widetilde{\mathscr{A}}$ implies $\mathscr{A}$, and thus our results also hold true under these more practical and verifiable assumptions.

($\widetilde{\mathscr{A}}_V$) V is twice continuously differentiable and lower bounded.

($\widetilde{\mathscr{A}}_c$) c is continuously differentiable and uniformly bounded. Furthermore, there exist $c_1, l > 0$ such that $c^s(x,y) \geq c_1 \mathbb{I}_{d\times d}$ for all $|x| > l, y \in \mathbb{R}^d$.

($\widetilde{\mathscr{A}}_\sigma$) σ is uniformly elliptic, symmetric and infinitely differentiable. Additionally, there exists $c_2 > 0$ such that $\sigma(x,y) \leq c_2 \mathbb{I}_{d\times d}$.

($\widetilde{\mathscr{A}}_{\text{Erg}}$) $|x|^{-1}\langle \nabla V(x), x\rangle \to \infty$ as $|x| \to \infty$.

Here, $c^s(x,y)$ denotes the symmetrization of the matrix $c(x,y)$, i.e., $c^s(x,y) := \frac{1}{2}(c_{ij}(x,y) + c_{ji}(x,y))_{1\leq i,j\leq d}$, $\mathbb{I}_{d\times d}$ is the identity matrix in $\mathbb{R}^{d\times d}$, and the order relation on symmetric matrices is the usual one defined by definite non-negativeness. It is easy to see that $\widetilde{\mathscr{A}}$ is a multidimensional generalization of the assumptions **HReg** and **HErg** in [7]. Hence, our results (such as upper bounds on the sup-norm risk for invariant density and drift estimators) also hold for the class of processes investigated in their recent paper.

The next auxiliary result confirms that the explicit assumptions $\widetilde{\mathscr{A}}$ indeed provide the technical framework required for our analysis.

LEMMA B.1. *$\widetilde{\mathscr{A}}$ implies $\mathscr{A}$.*

Proof. The assertion follows by results of [22], [14] and [3]. To be more precise, the fact that $\widetilde{\mathscr{A}}$ implies that (B.1) has a unique, non-explosive weak solution with strong Feller semigroup follows by Lemma 1.1 and Proposition 1.2 in [22], and the fact that a unique invariant probability measure exists follows by Theorem 3.1, once we note that the conditions (3.1) and (3.2) therein are fulfilled when $\widetilde{\mathscr{A}}$ holds (see Remark 3.2 in [22]). The mixing property follows by the proof of Theorem 3.1 in [22], which implies the existence of a Lyapunov function larger than 1. Thus, the process is exponentially ergodic due to Theorem 2.4 in [22], which is based on results of [11], and since Theorem 2.4 also states that any Lyapunov function is integrable with respect to the invariant measure, the mixing property follows. The heat kernel bound in $\mathscr{A}$ was shown in Theorem 2.1 of [14] for differentiable, globally Lipschitz continuous and uniformly bounded coefficients and extended in Corollary 2.12 of [3] to hold under local conditions. Noting that the results of [14] only require differentiability of the coefficients, the proof of Corollary 2.12 in [3] still holds true under $\widetilde{\mathscr{A}}$. Additionally, these results imply that the invariant distribution admits a density with respect to the Lebesgue measure. Hence, $\mathscr{A}$ holds if $\widetilde{\mathscr{A}}$ holds. ∎

Additional assumptions and notation In the following, we will always assume $Z_0 \sim \mu$, i.e., the process $\mathbf{Z}$ is stationary, and we denote $\mathbb{P}^\mu = \mathbb{P}$, $\mathbb{E}^\mu = \mathbb{E}$. Additionally, we define $\mu(g) := \int g \, \mathrm{d}\mu$ for $g \in L^1(\mu)$, and we introduce

$$a\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}, \quad a(x,y) \mapsto \sigma(x,y)\sigma^\top(x,y).$$

Given a class of functions $\mathcal{G}$ and a function b, we set $\mathcal{G}b := \{gb : g \in \mathcal{G}\}$, and for $x \in \mathbb{R}^d, \varepsilon > 0$, we denote the open ball with radius ε around x by $B(x, \varepsilon)$. Throughout all proofs, c will denote a positive constant, whose value may change from line to line, whereas specific constants are denoted by a c with additional subscript. Furthermore, we denote the restriction of the sup-norm to a domain $D \subset \mathbb{R}^{2d}$ by $\|\cdot\|_{L^\infty(D)}$. Lastly, the sup-norm risk of an estimator $\widehat{f}$ of a function f on a domain D is denoted by

$$\mathcal{R}_\infty^{(p)}\big(\widehat{f}, f; D\big) := \mathbb{E}\Big[\|\widehat{f} - f\|_{L^\infty(D)}^p\Big]^{\frac{1}{p}}, \quad p \geq 1.$$

B.3 Invariant density estimation

We start by investigating the issue of estimating the invariant density of $\mathbf{Z}$ on some bounded domain $D \subset \mathbb{R}^{2d}$. As mentioned earlier, our particular interest lies in identifying the effect of a continuous observation scheme on the convergence rate. Reflecting the two-component structure of the process $\mathbf{Z} = (\boldsymbol{X}, \mathbf{Y})$, our estimator has the form

$$\widehat{\rho}_{h_1,h_2,T}(x, y) = \frac{1}{T}\int_0^T K_{h_1,h_2}(x - X_s, y - Y_s)\,\mathrm{d}s, \quad x, y \in \mathbb{R}^d, \tag{B.2}$$

where

$$K_{h_1,h_2} \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \quad K_{h_1,h_2}(x, y) \mapsto (h_1 h_2)^{-d} K_1\Big(\frac{x}{h_1}\Big) K_2\Big(\frac{y}{h_2}\Big),$$

with $K_1, K_2 \colon \mathbb{R}^d \to \mathbb{R}$ bounded functions with $\operatorname{supp}(K_i) \subset [-1/2, 1/2]^d$, $i = 1, 2$, and bandwidths $h_1, h_2 > 0$. When the context is clear, we will often abbreviate this (by a slight abuse of notation) with $K_{\boldsymbol{h}}(z)$, where $\boldsymbol{h} = (h_1, h_2)$, $z = (x, y)$, $x, y \in \mathbb{R}^d$. Usually, h_1, h_2 are functions depending on T, however (by another abuse of notation), we often suppress this dependence.

B.3.1 *Variance bounds*

The proof of tight upper bounds on the speed of convergence of estimators requires sufficiently neat variance bounds. To achieve these, we will extend the results of Propositions 2 and 3 of [7], where remarkable bounds are shown for fixed values of (x, y) and $d = 1$. It is particularly interesting that the bounds are different for $y = 0$ and $y \neq 0$.

In the sequel, we will also consider the multidimensional case $d > 1$, and we will provide a *uniform* generalization for values of (x, y) in some bounded set $D \subset \mathbb{R}^{2d}$. The bandwidths will be assumed to belong to the set

$$\begin{aligned}\mathcal{H} = \mathcal{H}(Q_1, Q_2) := \Big\{h \colon [0, \infty) \to (0, \infty) : \exists Q_1, Q_2 > 0 \text{ such that } \forall T > 0 :\\ h^{-1}(T) \leq Q_1(1 + T)^{Q_1} \text{ and } h(T) \leq Q_2 T^{-Q_2} \wedge 1\Big\},\end{aligned} \tag{B.3}$$

thus confining their speed of convergence to 0 to be approximately polynomial. We start with investigating the more general case.

Proposition B.2. *Assume* 𝒜, $\|\rho\|_\infty < \infty$, *and let* $f \colon \mathbb{R}^d \times \mathbb{R}^d \times [0, \infty) \to \mathbb{R}$ *be a bounded function such that there exist* $n \in \mathbb{N}$, $\mathsf{x}_1, \mathsf{y}_1, \ldots, \mathsf{x}_n, \mathsf{y}_n \in \mathbb{R}^d$ *and functions* $\mathsf{s}_1, \mathsf{s}_2 \colon [0, \infty) \to (0, \infty) \in \mathcal{H}$

fulfilling

$$\operatorname{supp}(f(\cdot,\cdot,T)) \subset \bigcup_{i=1}^{n} B(x_i, s_1(T)) \times B(y_i, s_2(T)).$$

Then, there exists a constant $c > 0$ such that, for large enough T,

$$\operatorname{Var}\left(T^{-1}\int_0^T f(X_s, Y_s, T)\,\mathrm{d}s\right) \le cT^{-1}\|f\|_\infty^2 \times \begin{cases} s_1^2(T)\log(T) & d = 1, \\ s_1^{d+1}(T)s_2^{d-1}(T), & d \ge 2, \end{cases} \tag{B.4}$$

and

$$\operatorname{Var}\left(T^{-1}\int_0^T f(X_s, Y_s, T)\,\mathrm{d}s\right) \le cT^{-1}\|f\|_\infty^2 \times \begin{cases} s_1^{4/3}(T)s_2^2(T), & d = 1, \\ s_1^2(T)s_2^4(T)\log(T), & d = 2, \\ s_1^d(T)s_2^{d+2}(T), & d \ge 3. \end{cases} \tag{B.5}$$

The assumption on the support of f in Proposition B.2 is tailored to functions of the form

$$f(x, y, t) = \sum_{i=1}^{n} K_1((x_i - \cdot)/s_1(t))K_2(y_i - \cdot)/s_2(t)),$$

where $K_1, K_2 \colon \mathbb{R}^d \to \mathbb{R}$ are bounded functions with compact support, and $s_1, s_2 \in \mathcal{H}$. In particular the function K_{h_1,h_2} corresponding to the estimator introduced in (B.2) is of such a form as soon as the two bandwidths are elements of $\mathcal{H}$. The results in [7] now suggest that the variance bounds (B.4) and (B.5) can be improved if $\min_{i=1,\dots,n} \|y_i\| > 0$. However, as in the referred work, the proof of these refined results largely relies on the upper bound $\|x - x'\| \lesssim s_1$, being valid for all x, x' in the support of f (see (B.34) and the arguments thereafter). Now note that if $n > 1$, there exist $(x, y), (x', y') \in \operatorname{supp}(f)$, such that $x \in B(x_1, s_1(T)), x' \in B(x_2, s_1(T))$, which implies $\|x - x'\| \ge \|x_1 - x_2\| - 2s_1(T)$ by the reverse triangle inequality. This contradicts the needed upper bound for small enough values of s_1 and thus the following variance bound only concerns the case $n = 1$.

PROPOSITION B.3. *Let everything be given as in Proposition B.2 with $n = 1$ and assume, additionally, that $\|y_1\| > 0$. Then, there exists a constant $c > 0$ such that, for large enough T,*

$$\operatorname{Var}\left(T^{-1}\int_0^T f(X_s, Y_s, T)\,\mathrm{d}s\right) \le cT^{-1}\|f\|_\infty^2 s_1^{d+1}(T)s_2^d(T), \tag{B.6}$$

and, additionally, for $d = 1$,

$$\operatorname{Var}\left(T^{-1}\int_0^T f(X_s, Y_s, T)\,\mathrm{d}s\right) \le cT^{-1}\|f\|_\infty^2 s_1^{3/2}(T)s_2^2(T). \tag{B.7}$$

An interpretation of the highly nonclassical results stated in Propositions B.2 and B.3 will be given after our main results on the rate of convergence in the next section (see Theorem B.7).

B.3.2 *Rate of convergence*

With the introduced variance bounds, we are able to bound the convergence rate of the estimator under specific assumptions on the invariant density ρ, resulting in new upper bounds. In order to use the different results of Propositions B.2 and B.3, we also introduce the functions $\psi_d(x,y,t) \coloneqq \psi_{1,d}(x,y,t) \wedge \psi_{2,d}(x,y,t)$, where

$$\psi_{1,d}(x,y,t) \coloneqq \begin{cases} y^{-1}\sqrt{\log t}, & d=1, \\ x^{(1-d)/2}y^{-(1+d)/2}, & d \geq 2, \end{cases} \quad \text{and} \quad \psi_{2,d}(x,y,t) \coloneqq \begin{cases} x^{-1/3}, & d=1, \\ x^{-1}\sqrt{\log t}, & d=2, \\ x^{-d/2}y^{1-d/2}, & d \geq 3, \end{cases}$$

and $\psi_d^\circ(x,y,t) \coloneqq \psi_{1,d}^\circ(x,y,t) \wedge \psi_{2,d}^\circ(x,y,t)$, with

$$\psi_{1,d}^\circ(x,y,t) \coloneqq x^{(1-d)/2}y^{-d/2} \quad \text{and} \quad \psi_{2,d}^\circ(x,y,t) \coloneqq \begin{cases} x^{-1/4}, & d=1, \\ \psi_{2,d}(x,y,t), & d \geq 2. \end{cases}$$

Note that ψ_d^2 and $(\psi_d^\circ)^2$ represent the variance bounds (up to the term T^{-1}) for the estimator $\widehat{\rho}_{h_1,h_2,t}$ implied by Propositions B.2 and B.3. One remarkable fact in this context is that

$$\psi_d^2(h_1,h_2) < (h_1h_2)^{-d},$$

i.e., our obtained variance bounds are tighter compared to the classical one obtained for the kernel density estimator $\widehat{\rho}_{h_1,h_2,t}$. This is one of the reasons for the faster rates of convergence we will see later, compared to the classical nonparametric rate of convergence.

Proposition B.4. *Let $1 \leq p \leq \gamma \log T$, for $\gamma > 0$, and $D \subset \mathbb{R}^{2d}$ be a bounded, open set, assume A1, and choose $h_1 = h_1(T)$, $h_2 = h_2(T) \in \mathcal{H}$. Then, there exists a constant $c > 0$ independent of p such that, for large enough T,*

$$\mathcal{R}_\infty^{(p)}\big(\widehat{\rho}_{h_1,h_2,T}, \rho; D\big) \leq c\left(\mathcal{B}_\rho(h_1,h_2) + \frac{p\log T}{T(h_1h_2)^d}(\log T + p) + \frac{\psi_d(h_1,h_2,T)}{\sqrt{T}}\left(\sqrt{\log T} + \sqrt{p}\right)\right), \tag{B.8}$$

where the bias term is given as $\mathcal{B}_\rho(\boldsymbol{h}) = \mathcal{B}_\rho(h_1,h_2) \coloneqq \sup_{z\in\mathbb{R}^{2d}}\big|(\rho * K_{h_1,h_2} - \rho)(z)\big|$.

The refined variance bounds stated in Proposition B.3 imply the following result:

Proposition B.5. *Let $D \subset \mathbb{R}^{2d}$ be a bounded, open set such that $\inf_{(x,y)\in D}\|y\| > 0$, assume A1, and choose $h_1 = h_1(T)$, $h_2 = h_2(T) \in \mathcal{H}$ such that*

$$\frac{(\log T)^{3/2}}{\sqrt{T}(h_1h_2)^d\psi_d^\circ(h_1,h_2,T)} \longrightarrow 0, \quad \text{as } T \to \infty. \tag{B.9}$$

Then, there exists a constant $c > 0$ such that, for large enough T,

$$\mathcal{R}_\infty^{(1)}\big(\widehat{\rho}_{h_1,h_2,T}, \rho; D\big) \leq c\left(\mathcal{B}_\rho(h_1,h_2) + \psi_d^\circ(h_1,h_2,T)\sqrt{\frac{\log T}{T}}\right). \tag{B.10}$$

Note that Assumption (B.9) reflects the upper bound of Proposition B.4, since it implies the following for large enough values of T

$$\frac{(\log T)^2}{T(h_1h_2)^d} \leq \sqrt{\frac{\log T}{T}}\psi_d^\circ(h_1,h_2,T).$$

In fact this condition is only a marginal restriction as will be made clear in Theorem B.7. Furthermore, the results of Propositions B.4 and B.5 reflect the classical bias-variance decomposition, with the term $\mathcal{B}_\rho$ denoting the bias term and $\psi_d(\cdot,T)\sqrt{T^{-1}\log T}$ representing the stochastic error. For translating the above results into concrete upper bounds on the convergence rate, we will work under classical Hölder smoothness assumptions for the invariant density, with a small adjustment due to our concrete problem: In order to reflect the specific form of the process $\mathbf{Z} = (\boldsymbol{X}, \mathbf{Y})$, we will use a mixture of isotropic and anisotropic Hölder conditions as described in the following definition.

DEFINITION B.6. Let $\beta_1, \beta_2, \mathcal{L}_1, \mathcal{L}_2 > 0$ and $D \subset \mathbb{R}^{2d}$ be an open set. A function $g\colon \mathbb{R}^{2d} \to \mathbb{R}$ is said to belong to the anisotropic Hölder class $\mathcal{H}_D(\beta_1,\beta_2,\mathcal{L}_1,\mathcal{L}_2)$ if, for all $i = 1,\dots,d$,

$$\begin{aligned}\|D_i^k g\|_{L^\infty(D)} &\leq \mathcal{L}_1, \quad \forall k = 0,\dots,\lfloor\!\lfloor\beta_1\rfloor\!\rfloor,\\ \|D_i^{\lfloor\!\lfloor\beta_1\rfloor\!\rfloor} g(\cdot + t\mathsf{e}_i) - D_i^{\lfloor\!\lfloor\beta_1\rfloor\!\rfloor} g(\cdot)\|_{L^\infty(D)} &\leq \mathcal{L}_1|t|^{\beta_1 - \lfloor\!\lfloor\beta_1\rfloor\!\rfloor}, \quad \forall t \in \mathbb{R},\end{aligned}$$

and, for all $i = d+1,\dots,2d$,

$$\begin{aligned}\|D_i^k g\|_{L^\infty(D)} &\leq \mathcal{L}_2, \quad \forall k = 0,\dots,\lfloor\!\lfloor\beta_2\rfloor\!\rfloor,\\ \|D_i^{\lfloor\!\lfloor\beta_2\rfloor\!\rfloor} g(\cdot + t\mathsf{e}_i) - D_i^{\lfloor\!\lfloor\beta_2\rfloor\!\rfloor} g(\cdot)\|_{L^\infty(D)} &\leq \mathcal{L}_2|t|^{\beta_2 - \lfloor\!\lfloor\beta_2\rfloor\!\rfloor}, \quad \forall t \in \mathbb{R},\end{aligned}$$

where $D_i^k g$ is the k-th order partial derivative of g with respect to the i-th component, $\lfloor\!\lfloor\beta\rfloor\!\rfloor$ denotes the largest integer *strictly* smaller than β and $\mathsf{e}_1,\dots,\mathsf{e}_{2d}$ is the canonical basis in $\mathbb{R}^{2d}$.

For estimating the invariant density ρ of the process $\mathbf{Z}$ on a domain D, assuming that $\rho \in \mathcal{H}_D(\beta_1,\beta_2,\mathcal{L}_1,\mathcal{L}_2)$, we choose K_1, K_2 to be smooth Lipschitz continuous kernel functions of order $\lfloor\!\lfloor\beta_1\rfloor\!\rfloor, \lfloor\!\lfloor\beta_2\rfloor\!\rfloor$. Recall that a kernel $K\colon \mathbb{R}^d \to \mathbb{R}$ is said to be of order $\ell \in \mathbb{N}$ if, for any $\alpha \in \mathbb{N}^d$ with $|\alpha| \leq \ell$, $x \mapsto x^\alpha K(x)$ is integrable and, moreover,

$$\int_{\mathbb{R}^d} K(x)\,\mathrm{d}x = 1, \quad \int_{\mathbb{R}^d} K(x)x^\alpha\,\mathrm{d}x = 0, \quad \alpha \in \mathbb{N}^d,\ |\alpha| \in \{1,\dots,\ell\},$$

where $|\alpha| = \sum_{i=1}^d \alpha_i$ and $x^\alpha = \prod_{i=1}^d x_i^{\alpha_i}$ for all $x \in \mathbb{R}^d$, $\alpha \in \mathbb{N}^d$. For notational convenience, we denote the harmonic mean of the smoothness parameters β_1 and β_2 by

$$\overline{\beta}_{1,2} := 2(\beta_1^{-1} + \beta_2^{-1})^{-1}.$$

When the context is clear, we will omit the index in this notation. For stating our results on the convergence rate in a compact way, it is also useful to introduce the functions $\Upsilon\colon \mathbb{R}_0^+ \times \mathbb{R}_0^+ \times \mathbb{N} \times \mathbb{R}_0^+ \to \mathbb{R}_0^+$, $\Phi\colon \mathbb{R}_0^+ \times \mathbb{R}_0^+ \times \mathbb{R}_0^+ \times \mathbb{N} \times \mathbb{R}_0^+ \to \mathbb{R}_0^+$ and $\chi_{\mathcal{B}}\colon \mathbb{R}_0^+ \times \mathbb{R}_0^+ \times \mathbb{R}_0^+ \times \mathbb{N} \times \mathbb{R}_0^+ \to \mathbb{R}_0^+$, specified

as

$$\Upsilon(\beta_1,\beta_2,d,\varepsilon) := \begin{cases} \frac{2}{3}\frac{3\beta_1+\beta_2}{\beta_1+\beta_2}, & 3\beta_1 \geq \beta_2, d = 1, \varepsilon = 0, \\ \frac{4\beta_1}{\beta_1+\beta_2}, & 3\beta_1 \geq \beta_2, d \geq 2, \varepsilon = 0, \\ 2\frac{\beta_2-\beta_1}{\beta_1+\beta_2}, & 3\beta_1 < \beta_2, \varepsilon = 0, \\ \frac{2\beta_1+\beta_2}{\beta_1+\beta_2}, & 2\beta_1 \geq \beta_2, d = 1, \varepsilon > 0, \\ \frac{4\beta_1}{\beta_1+\beta_2}, & 2\beta_1 \geq \beta_2, d \geq 2, \varepsilon > 0, \\ \frac{2\beta_2}{\beta_1+\beta_2}, & 2\beta_1 < \beta_2, \varepsilon > 0, \end{cases} \tag{B.11}$$

$$\Psi(T,\beta_1,\beta_2,d,\varepsilon) := \left(\frac{\log T}{T}\right)^{\frac{\overline{\beta}}{2(\overline{\beta}+d)-\Upsilon(\beta_1,\beta_2,d,\varepsilon)}}, \tag{B.12}$$

$$\chi_{\mathcal{B}}(T,\beta_1,\beta_2,d,\varepsilon) := \begin{cases} 1, & (\beta_1,\beta_2,d,\varepsilon) \notin \mathcal{B} \\ \sqrt{\log T}, & (\beta_1,\beta_2,d,\varepsilon) \in \mathcal{B}, \end{cases}$$

where the set $\mathcal{B}$ is given as

$$\mathcal{B} := \left\{ (\beta_1,\beta_2,d,\varepsilon) \in \mathbb{R}^4, \text{such that one of the following holds: } \begin{cases} 3\beta_1 < \beta_2 \wedge d = 1 \wedge \varepsilon = 0 \\ 3\beta_1 > \beta_2 \wedge d = 2 \wedge \varepsilon = 0 \\ 2\beta_1 > \beta_2 \wedge d = 2 \wedge \varepsilon > 0 \end{cases} \right\}. \tag{B.13}$$

We are now ready to state the first of the bounds on convergence rates announced in the introduction. In line with the two different variance bounds in Propositions B.2 and B.3, we will consider both the general case and the case where $\inf_{(x,y)} \|y\| > 0$. The proof of the upper bounds on the classical sup-norm risk $\mathcal{R}_\infty^{(1)}$ (stated in part (b) below) relies on a classical combination of Proposition B.3 with a discretization of the domain and the exploitation of concentration results. For the general case, we even obtain an upper bound for arbitrary p-th moments, $p \geq 1$ (see part (a)): Since Proposition B.2 permits to bound the *difference* of kernels, we are able to bound the entropy integrals in the uniform moment bounds obtained in [8], which then yields (B.15).

THEOREM B.7. *Let $D \subset \mathbb{R}^{2d}$ be a bounded, open set, and assume $\mathcal{A}$ and $\rho \in \mathcal{H}_D(\beta_1,\beta_2,\mathcal{L}_1,\mathcal{L}_2)$ for $\beta_1 > 1$, $\beta_2 > 2$.*

(a) *If the bandwidth is chosen such that*

$$h_i \sim \Psi(T,\beta_1,\beta_2,d,0)^{\frac{1}{\beta_i}}, \quad i = 1,2, \tag{B.14}$$

then the associated invariant density estimator fulfills

$$\mathcal{R}_\infty^{(p)}\left(\widehat{\rho}_{h_1,h_2,T},\rho;D\right) \in \mathrm{O}((\Psi\chi_{\mathcal{B}})(T,\beta_1,\beta_2,d,0)), \quad p \geq 1. \tag{B.15}$$

(b) *Define $\varepsilon_D := \inf_{(x,y)\in D} \|y\|$. Then, specifying*

$$h_i \sim \Psi(T,\beta_1,\beta_2,d,\varepsilon_D)^{\frac{1}{\beta_i}}, \quad i = 1,2, \tag{B.16}$$

yields

$$\mathcal{R}_\infty^{(1)}\left(\widehat{\rho}_{h_1,h_2,T},\rho;D\right) \in \mathrm{O}((\Psi\chi_{\mathcal{B}})(T,\beta_1,\beta_2,d,\varepsilon_D)).$$

Remark B.8. (a) Note that, for the proposed specification of bandwidths, the rate of convergence in certain cases only depends on *one* of the smoothness parameters. More precisely, the convergence rate is specified as $(\Psi\chi_{\mathcal{B}})(T,\beta_1,\beta_2,d,\varepsilon_D) = (\log T/T)^{\alpha}\chi_{\mathcal{B}}(T,\beta_1,\beta_2,d,\varepsilon_D)$ with

$$\alpha = \alpha(\beta_1,\beta_2,d,\varepsilon_D) := \begin{cases} \frac{\beta_1}{2\beta_1+(2/3)}, & 3\beta_1 \geq \beta_2, d = 1, \varepsilon_D = 0, \\ \frac{\beta_1}{2\beta_1+2}, & 3\beta_1 \geq \beta_2, d = 2, \varepsilon_D = 0, \\ \frac{\beta_2}{2\beta_2+2}, & 3\beta_1 < \beta_2, d = 1, \varepsilon_D = 0, \\ \frac{\beta_1}{2\beta_1+(1/2)}, & 2\beta_1 \geq \beta_2, d = 1, \varepsilon_D > 0, \\ \frac{\beta_1}{2\beta_1+2}, & 2\beta_1 \geq \beta_2, d = 2, \varepsilon_D > 0, \\ \frac{\beta_2}{2\beta_2+1}, & 2\beta_1 < \beta_2, d = 1, \varepsilon_D > 0. \end{cases}$$

Similar results were also obtained in [7] for the pointwise risk in the scalar case.

(b) Although the function Υ introduced in (B.11) may seem like a technical artifact of our procedures, it is regular in the sense of being continuous in the smoothness parameters β_1 and β_2 for fixed values of d and ε. The only thing that counteracts this regularity in the derived convergence rate is the appearance of an additional logarithmic term in some cases, described by the set $\mathcal{B}$ and the function $\chi_{\mathcal{B}}$. However, this concerns only some cases in a low-dimensional setting ($d = 1$ or $d = 2$) and was also observed in [7].

(c) In order to translate the above result into a statement on minimax optimality, two steps are necessary: The upper bounds have to be verified uniformly for the class of all diffusions satisfying Assumption $\mathcal{A}$, and the upper bound has to be complemented by a corresponding lower bound. It is very challenging to obtain the mixing control uniformly over a class of diffusions. Instead of directing our efforts in this direction, we focus on constructive aspects: In the upcoming Section B.4, we study the issue of nonparametric drift estimation, for which we even propose an adaptive procedure. In particular, since our primary interest is not in optimality issues, we refrain from proving lower bounds. However, it is to be expected that such statements can be derived by (elaborate) adaptations of the procedures of [7], who considered the case $d = 1$ and the pointwise risk.

(d) Let us finally compare one aspect of Theorem B.7 to the scalar, pointwise risk estimates in Theorems 1 and 2 of [7]. In these theorems, one of the bandwidths (depending on the ratio of the two smoothness parameters) can be chosen rather freely, as long as it fulfills some regularity assumptions. However, in the much more delicate context considered in Theorem B.7 (we investigate the sup-norm risk in a multidimensional situation), we specify both bandwidths explicitly. The reason for this is the bound obtained in Proposition B.4, respectively Assumption (B.9), which are also the reasons for assuming $\beta_1 > 1$ and $\beta_2 > 2$. In fact, these assumptions on β_1 and β_2 can be relaxed slightly in some cases, but for the sake of brevity and since this does not offer much further insight, we decided to omit this result.

We continue by providing interpretations of the functions Ψ, Υ and the set $\mathcal{B}$ introduced in (B.11), (B.12) and (B.13), respectively, and explaining our reasons for this particular form of notation.

Remark B.9. In the classical setting of n d-dimensional, i.i.d. observations, the minimax optimal convergence rate for the sup-norm risk, given the estimated density belongs to an isotropic Hölder class with smoothness β, is given by $(\log n/n)^{\beta/(2\beta+d)}$, where the logarithmic term in the convergence rate stems from investigating the sup-norm risk. Thus, for our specific problem, an analogous rate would be of the form $(\log T/T)^{\beta/(2\beta+2d)}$ (recall that $\mathbf{Z}$ is $2d$-dimensional), where the smoothness index β is replaced by the harmonic mean of the smoothness indices in the anisotropic framework. Note now that the function Ψ with $\Upsilon \equiv 0$ corresponds to this classical nonparametric rate of convergence. However, it has already been observed that this rate can be improved for invariant density estimation of diffusion-type processes when *continuous* observations are available, corresponding to Υ being strictly positive in our notation. Specifically, we refer, e.g., to Corollary 1 in [6] for a result on the convergence rate of the pointwise risk in the continuous diffusion context, Theorem 3.4 in [17], which concerns the rate of convergence of the sup-norm risk for an adaptive estimator of the invariant density of a continuous diffusion under anisotropic Hölder assumptions, or Theorem 4.3 in [8], which bounds the rate with respect to the sup-norm risk for a more general class of exponentially β-mixing Markov processes. In all these cases, the rate of convergence is essentially given by Ψ with $\Upsilon \equiv 2$. In particular, contrary to our result, Υ does not depend on the dimension, the smoothness indices or any other entities, especially not even in the anisotropic framework considered in [17]. For a summary of the mentioned results in this paragraph see the following table, which contains the polynomic rates of convergence:

	(invariant) density	drift vector
nondegenerate diffusion	$\frac{\bar\beta}{2(\bar\beta+d)-2}$	$\frac{\bar\beta}{2(\bar\beta+d)}$
kinetic diffusion	$\frac{\bar\beta}{2(\bar\beta+d)-\Upsilon(\beta_1,\beta_2,d,\varepsilon)}$	$\frac{\bar\beta}{2(\bar\beta+d)}$
i.i.d. case	$\frac{\bar\beta}{2(\bar\beta+d)}$	-

As can be seen in the proof of the variance bounds in Propositions B.2 and B.3, which are the quintessential reason for our results, the particular form of Υ in our case is caused by the heat kernel bound in $\mathcal{A}$. More specifically, the function p_t^G suggests that the variances of the processes $\boldsymbol{X}$ and $\mathbf{Y}$ are of a different order. To illustrate this further, we refer to the following example taken from [3].

Example B.10 (Example 2.9 in [3]). Let $d = 1$ and $c = V = 0$. Then, Z_t is a two-dimensional Gaussian vector with

$$\mathbb{E}[X_t] = x_0 + y_0 t, \quad \mathbb{E}[Y_t] = y_0,$$

and

$$\mathrm{Var}(X_t) = \frac{t^3}{3}, \quad \mathrm{Var}(Y_t) = t, \quad \mathrm{Cov}(X_t, Y_t) = \frac{t^2}{2}.$$

B.4 Drift estimation

We now turn to the question of proposing a nonparametric estimator of the drift function appearing in (B.1), specified as $b(x,y) = -(c(x,y)y + \nabla V(x))$, $x, y \in \mathbb{R}^d$. Throughout this entire section, we will assume b to be locally bounded and σ to be uniformly bounded. Note that these assumptions are satisfied under $\widetilde{\mathcal{A}}$.

Given two bounded kernel functions $K_1, K_2 : \mathbb{R}^d \to \mathbb{R}$ with compact support, $x, y \in \mathbb{R}^d$ and $j \in \{1, \dots, d\}$, set

$$\overline{b}_{j,h_1,h_2,T}(x,y) := \frac{1}{T}\int_0^T K_{h_1,h_2}(x - X_u, y - Y_u)\,\mathrm{d}Y_u^j, \quad \text{where} \quad K_{h_1,h_2} := (h_1h_2)^{-d}K_1\Big(\frac{x}{h_1}\Big)K_2\Big(\frac{y}{h_2}\Big).$$

For some strictly positive $r_T \in o(1)$, an estimator of the j-th component of the drift vector b is then given by a Nadaraya–Watson-type estimator of the form

$$\widehat{b}_{j,\boldsymbol{h},T,r_T} := \frac{\overline{b}_{j,h_1,h_2,T}}{|\widehat{\rho}_{h_1^{(\rho)},h_2^{(\rho)},T}| + r_T}, \quad x, y \in \mathbb{R}^d, \boldsymbol{h} := (h_1, h_2, h_1^{(\rho)}, h_2^{(\rho)}). \tag{B.17}$$

Note that our bounds on the rate of convergence of the estimator $\widehat{\rho}$ stated in the previous section continue to hold for $|\widehat{\rho}|$. Strict positivity of r_T ensures that the drift estimator $\widehat{b}$ introduced in (B.17) is well-defined. However, since $\overline{b}$ is defined via some stochastic integral, a crucial point for deriving upper bounds on the rate of convergence of this estimator will be uniform moment bounds of stochastic integrals with respect to $\mathbf{Y}$ over countable classes of bounded functions. This will be the main focus of the subsequent section. In principle, all the applied techniques would also be suitable for the estimation of a drift function b, which is not in the specified form. However, as the results of [22], which in particular imply the exponential β-mixing property, only consider such drifts, we focus our analysis on this case.

B.4.1 *Uniform moment bounds*

In Section 3 of [8], uniform moment bounds over countable classes of bounded functions $\mathcal{G}$ were derived for suprema of functionals of the form

$$\sup_{g\in\mathcal{G}}|\mathbb{G}_t(g)|, \quad \text{where} \quad \mathbb{G}_t(g) := \frac{1}{\sqrt{t}}\int_0^t g(X_s)\,\mathrm{d}s,\ g \in L_0^2(\mu),$$

under the assumption of $\boldsymbol{X}$ being exponentially β-mixing. For the reader's convenience, we start this section with a reminder of the relevant results. As the bounds are derived via an application of the generic chaining device based on [9], they are stated in terms of covering numbers, so recall that, for any given $\varepsilon > 0$, the covering number $\mathcal{N}(\varepsilon, \mathcal{G}, d)$ of $\mathcal{G}$ denotes the smallest number of balls of d-radius ε needed to cover $\mathcal{G}$. Furthermore, given $f, g \in \mathcal{G}$, we define the following semi-metrics,

$$\begin{aligned} d_\infty(f,g) &:= \|f - g\|_\infty, \quad d^p_{L^p(\mu)}(f,g) := \mu((f-g)^p) \quad p \geq 1, \\ d^2_{\mathbb{G},t}(f,g) &:= \mathrm{Var}\bigg(\frac{1}{\sqrt{t}}\int_0^t (f-g)(X_s)\,\mathrm{d}s\bigg), \end{aligned} \tag{B.18}$$

with $\boldsymbol{X}$ being the Markov process in Theorem B.11.

THEOREM B.11 (Theorem 3.2 in [8]). *Suppose that $\boldsymbol{X}$ is an exponentially β-mixing Markov process. Let $\mathcal{G}$ be a countable class of bounded real-valued functions with $\mu(g) = 0$, and let $m_t \in [0, t/4)$. Then, there exist $\tau \in [m_t, 2m_t]$ and constants $\widetilde{C}_1, \widetilde{C}_2 > 0$ such that, for any $1 \leq p < \infty$,*

$$\bigg(\mathbb{E}\bigg[\sup_{g\in\mathcal{G}}|\mathbb{G}_t(g)|^p\bigg]\bigg)^{1/p} \leq \widetilde{C}_1\int_0^\infty \log\mathcal{N}(u, \mathcal{G}, \frac{2m_t}{\sqrt{t}}d_\infty)\,\mathrm{d}u + \widetilde{C}_2\int_0^\infty \sqrt{\log\mathcal{N}(u, \mathcal{G}, d_{\mathbb{G},\tau})}\,\mathrm{d}u$$

$$+ 4 \sup_{g \in \mathcal{G}} \left(\frac{2m_t}{\sqrt{t}} \|g\|_\infty \widetilde{c}_1 p + \|g\|_{\mathbb{G},\tau} \widetilde{c}_2 \sqrt{p} + \frac{1}{2} \|g\|_\infty c_\kappa \sqrt{t} \mathrm{e}^{-\frac{\kappa m_t}{p}} \right),$$

where $\widetilde{c}_1, \widetilde{c}_2$ are positive constants, defined in equation (B.3) of [8], and $c_\kappa, \kappa > 0$ are specified in Assumption ($\mathcal{A}\beta$) of [8].

One of the main tools in the derivation of this result was the following Bernstein-type concentration inequality.

LEMMA B.12 (Lemma 3.1 in [8]). *Suppose that $\boldsymbol{X}$ is an exponentially β-mixing Markov process, and let g be a bounded, measurable function fulfilling $\mu(g) = 0$. Then, for any $t, u > 0$ and $m_t \in (0, \frac{t}{4}]$, there exists $\tau \in [m_t, 2m_t]$ such that*

$$\mathbb{P}\left(\frac{1}{\sqrt{t}} \int_0^t g(X_s) \, \mathrm{d}s > u \right) \leq 2 \exp\left(- \frac{u^2}{32 \left(\mathrm{Var}\left(\frac{1}{\sqrt{\tau}} \int_0^\tau g(X_s) \, \mathrm{d}s \right) + 2u \|g\|_\infty \frac{m_t}{\sqrt{t}} \right)} \right) + \frac{t}{m_t} c_\kappa \mathrm{e}^{-\kappa m_t} \mathbb{1}_{(0, 4\sqrt{t}\|g\|_\infty)}(u).$$

The term m_t in the previous statements arises from the use of the classical Bernstein inequality for independent random variables in the proofs. It can thus be interpreted as a kind of loss compared to results concerning i.i.d. random variables. However, since $\boldsymbol{X}$ is exponentially β-mixing, the decay in m_t is exponentially fast, resulting in this loss being almost negligible. In fact, m_t can be viewed as a tuning parameter in our concentration results, with the typical choice being $m_t = c \log t$, where $c > 0$ is suitably large so that the error term decays with an adequately fast polynomial rate.

We will extend Theorem B.11 in our specific framework to functionals of the form

$$\sup_{g \in \mathcal{G}} |\mathbb{H}_t^j(g)|, \quad \text{where} \quad \mathbb{H}_t^j(g) := \frac{1}{\sqrt{t}} \int_0^t g(X_s, Y_s) \, \mathrm{d}Y_s^j, g \in L_0^2(\mu), j \in \{1, \ldots, d\}. \tag{B.19}$$

Once again, a Bernstein-type concentration inequality will play a vital role in our proofs, namely the Bernstein inequality for continuous martingales (see, e.g., p. 153 in [16]). Given a continuous local martingale $(M_t)_{t \geq 0}$, it states that

$$\forall t, x, y > 0, \quad \mathbb{P}(M_t \geq x, \langle M \rangle_t \leq y) \leq \exp\left(-\frac{x^2}{2y} \right), \tag{B.20}$$

where $(\langle M \rangle_t)_{t \geq 0}$ denotes the quadratic variation process of $(M_t)_{t \geq 0}$. Combining the concentration inequalities in Lemma B.12 and equation (B.20) will enable us to again employ the generic chaining device for the derivation of the required uniform moment bounds, similar to [9]. They are stated in terms of entropy integrals with respect to the semi-metrics defined in (B.18), with the semi-metric $d_{\mathbb{G}}$ induced by the variance of the integral functional now being specified as

$$d_{\mathbb{G},t}^2(f, g) := \sigma_t^2(f - g), \text{ where } \sigma_t^2(f) := \mathrm{Var}\left(\frac{1}{\sqrt{t}} \int_0^t f(X_s, Y_s) \, \mathrm{d}s \right).$$

The main result of this section is the following theorem.

THEOREM B.13. *Assume 𝒜, let 𝒢 be a countable class of bounded real-valued functions such that gb is bounded for all g ∈ 𝒢, and let* $m_t, \widetilde{m}_t \in (0, t/4]$. *Then, there exist* $\tau \in [m_t, 2m_t], \widetilde{\tau} \in [\widetilde{m}_t, 2\widetilde{m}_t]$ *and a constant c > 0 such that, for large enough t, any* $1 \le p < \infty$ *and* $j \in \{1, \dots, d\}$,

$$\begin{aligned}\left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}|\mathbb{H}_t^j(g) - \sqrt{t}\mu(gb^j)|^p\right]\right)^{1/p} \le c\Bigg(&\int_0^\infty \log \mathcal{N}\big(u, \mathcal{G}b^j, \tfrac{m_t}{\sqrt{t}}d_\infty\big)\,\mathrm{d}u + \int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{G}b^j, d_{\mathbb{G},\tau})}\,\mathrm{d}u \\ &+ \int_0^\infty \log \mathcal{N}\big(u, \mathcal{G}, t^{-1/4}d_\infty + t^{-1/8}d_{L^4(\mu)}\big)\,\mathrm{d}u \\ &+ \int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{G}, d_{L^2(\mu)} + t^{-1/8}d_{L^4(\mu)})}\,\mathrm{d}u \\ &+ \sup_{g\in\mathcal{G}}\Bigg(\frac{m_t}{\sqrt{t}}\|g\|_\infty p + \|g\|_{\mathbb{G},\tau}\widetilde{c}_2\sqrt{p} + \frac{1}{2}\|g\|_\infty\sqrt{t}\mathrm{e}^{-\frac{\kappa m_t}{p}} \\ &+ p\sqrt{\frac{\widetilde{m}_t\|a_{jj}\|_\infty}{t}}\|g\|_\infty + p^{3/4}(\widetilde{\tau}/t)^{1/4}\|g\|_{L^4(\mu)} \\ &+ \sqrt{p\|a_{jj}\|_\infty}\|g\|_\infty\mathrm{e}^{-\frac{\kappa\widetilde{m}_t}{2p}} + \sqrt{p\|a_{jj}\|_\infty}\|g\|_{L^2(\mu)}\Bigg)\Bigg).\end{aligned}$$

The form of the upper bound in Theorem B.13 reflects the result obtained in Theorem 3.5 of [9], where the generic chaining device is applied on stochastic processes with a mixed tail behaviour. As we perform the chaining procedure twice using two different concentration inequalities, it is not surprising that the obtained result contains four different entropy integrals, each concerning a different distance.

B.4.2 *Rate of convergence*

Applying the powerful result stated in Theorem B.13 together with bounds on the involved covering numbers already yields a bound on the rate of convergence of the sup-norm risk of the estimator $\overline{b}_{j,h_1,h_2,T}$ of $b^j\rho$, for adequately chosen bandwidths h_1, h_2.

PROPOSITION B.14. *Let* $D \subset \mathbb{R}^{2d}$ *be an open and bounded set, assume 𝒜, and let* $h_1, h_2 \in \mathcal{H}$ *such that* $(h_1h_2)^d \ge T^{-\frac{1}{2}}\log(h_1^{-1} + h_2^{-1})$. *Then, for any γ > 0, there exists a constant* c_γ *such that, for any* $1 \le p \le \gamma \log T$, *it holds for large enough T*

$$\mathcal{R}_\infty^{(p)}\big(\overline{b}_{j,h_1,h_2,T}, b^j\rho; D\big) \le \mathcal{B}_{b^j\rho}(h_1, h_2) + c_\gamma(h_1h_2)^{-d/2}T^{-1/2}\sqrt{\log(h_1^{-1} + h_2^{-1})}.$$

Combining Theorem B.7 with Proposition B.14 and a specific choice of r_T then yields our next main result, an upper bound on the rate of convergence for a weighted version of the sup-norm risk of the drift estimator $\widehat{b}$ introduced in (B.17).

THEOREM B.15. *Let* $D \subset \mathbb{R}^{2d}$ *be a bounded, open set, fix* $j \in \{1, \dots, d\}$, *assume 𝒜,* $b^j\rho, \rho \in \mathcal{H}_D(\beta_1, \beta_2, \mathcal{L}_1, \mathcal{L}_2)$, *and set*

$$r_T := (\Psi\chi_{\mathcal{B}})(\beta_1, \beta_2, d, 0)\exp\Big(\sqrt{\log T}\Big).$$

For defining the drift estimator, choose $K_1, K_2, h_1^{(\rho)}, h_2^{(\rho)}$ *as in Theorem B.7(a), and specify*

$$h_i \sim \left(\frac{\log T}{T}\right)^{\frac{\bar{\beta}}{2\beta_i(\bar{\beta}+d)}}, \quad i = 1, 2. \tag{B.21}$$

Then, if $\overline{\beta} > d$, $\beta_1 > 1$, $\beta_2 > 2$, *it holds*

$$\mathbb{E}\Big[\|(\widehat{b}_{j,\boldsymbol{h},T,r_T}(z) - b^j(z))\rho(z)\|_{L^\infty(D)}\Big] \in \mathrm{O}\left(\left(\frac{\log T}{T}\right)^{\frac{\bar{\beta}}{2(\bar{\beta}+d)}}\right).$$

The entire proof can again be found in the appendix. One may wonder why we arrive at the classical nonparametric rate of convergence even though highly nonclassical results were obtained in Corollary B.7. The technical reason for this is the occurrence of the covering number with respect to $d_{L^2(\mu)}$ in Theorem B.13. Contrary to the approach taken to find the variance bounds in Propositions B.2 and B.3, we cannot use the exponential β-mixing property of $\mathbf{Z}$ to bound this, and since the heat kernel bound in (A2) only applies to values of t in $(0, 1]$, it cannot be used either. The fact that a faster convergence rate for the invariant density estimate is not equivalent to a faster convergence rate for the drift estimate is well-known. In particular, it has been shown in some cases that the classical nonparametric convergence rate is optimal in the minimax sense (see, e.g., [18, 19]).

B.4.3 *Adaptive estimation scheme*

We now address the question of finding a data-driven approach to drift estimation. Our interest is in bounding the sup-norm risk $\mathcal{R}_\infty^{(p)}(\widehat{b}_{j,\boldsymbol{h},T}, b^j; D)$, $1 \le p < \infty$, of (the components of) the drift estimator $\widehat{b}_{j,\boldsymbol{h},T}$ over an open and bounded set $D \subset \mathbb{R}^{2d}$. The bandwidths specified in Theorem B.15 (see (B.21)) clearly depend on the typically unknown smoothness of $b^j\rho$.

For defining an adaptive drift estimator which relies on bandwidths specified in a data-driven way, consider some symmetric, Lipschitz continuous kernel functions $K_1, K_2 \colon \mathbb{R}^d \to \mathbb{R}$ of order ℓ_1, ℓ_2 fulfilling

$$\int K_i \,\mathrm{d}\lambda = 1, \quad \|K_i\|_\infty < \infty \quad \text{and} \quad \operatorname{supp}(K_i) \subset [-1/2, 1/2]^d, \quad i = 1, 2.$$

For any bandwidths $(h_1, h_2)^\top, (\eta_1, \eta_2)^\top \in (0, 1]^2$ and any points $x, y \in \mathbb{R}^d$, denote

$$\begin{aligned}\big(K_{h_1,h_2} \star K_{\eta_1,\eta_2}\big)(x, y) &:= \big(K_{h_1} * K_{\eta_1}\big)(x) \cdot \big(K_{h_2} * K_{\eta_2}\big)(y) \\ &= \int_{\mathbb{R}^d} K_{h_1}(u - x)K_{\eta_1}(u)\,\mathrm{d}u \int_{\mathbb{R}^d} K_{h_2}(u - y)K_{\eta_2}(u)\,\mathrm{d}u.\end{aligned}$$

For $x, y \in \mathbb{R}^d$, $j \in \{1, \dots, d\}$, define the kernel estimators

$$\begin{aligned}\overline{b}_{j,h_1,h_2,t}(x, y) = \overline{b}_{j,\boldsymbol{h}}(x, y) &:= \frac{1}{t}\int_0^t K_{h_1,h_2}(x - X_s, y - Y_s)\,\mathrm{d}Y_s^j, \\ \overline{b}_{j,h_1,h_2,\eta_1,\eta_2}(x, y) = \overline{b}_{j,\boldsymbol{h},\boldsymbol{\eta}}(x, y) &:= \frac{1}{t}\int_0^t \big(K_{h_1,h_2} \star K_{\eta_1,\eta_2}\big)(X_s - x, Y_s - y)\,\mathrm{d}Y_s^j.\end{aligned}$$

Specify the set $\mathcal{H}_t$ of candidate bandwidths for some arbitrary $\eta > 1$ as

$$\mathcal{H}_t := \Big\{\boldsymbol{h} = (h_1, h_2)^\top \in (0,1]^2 : h_i = \eta^{-k_i} \text{ with } k_i \in \mathbb{N}_0, \eta^{d(k_1+k_2)} \le t^{\frac{1}{2}} \log(\eta^{k_1} + \eta^{k_2})^{-1}\Big\},$$

choose $q \ge 1$, and let

$$\widehat{\Delta}_t^j(\boldsymbol{h}) := \sup_{\boldsymbol{\eta}=(\eta_1,\eta_2)\in\mathcal{H}_t} \Big\{\Big[\|\overline{b}_{j,\boldsymbol{h},\boldsymbol{\eta}} - \overline{b}_{j,\boldsymbol{\eta}}\|_\infty - A_t^{(q)}(\eta_1,\eta_2)\Big]_+\Big\},$$

for

$$A_t^{(q)}(\boldsymbol{\eta}) = A_t^{(q)}(\eta_1,\eta_2) := \mathrm{e}\sqrt{32d\|a_{jj}\|_\infty\|\rho\|_\infty}\Big(\widetilde{C}_1\sqrt{192}\|K\|_\infty + \widetilde{C}_2\sqrt{q}\|K\|_{L^2(\lambda)}\Big)\sqrt{\frac{\log(\eta_1^{-1}+\eta_2^{-1})}{t(\eta_1\eta_2)^d}}, \tag{B.22}$$

where the constants $\widetilde{C}_1, \widetilde{C}_2$ are specified in the proof of Theorem B.13 in Appendix B. Finally, define $\widehat{\boldsymbol{h}}^j = (\widehat{h}_1^j, \widehat{h}_2^j)$ by setting

$$\widehat{\Delta}_t^j(\widehat{\boldsymbol{h}}^j) + A_t^{(q)}(\widehat{\boldsymbol{h}}^j) = \inf_{\boldsymbol{h}=(h_1,h_2)\in\mathcal{H}_t}\Big\{\widehat{\Delta}_t^j(\boldsymbol{h}) + A_t^{(q)}(\boldsymbol{h})\Big\}.$$

Our approach is based on the work of [15], which itself relies on ideas developed in [13]. Intuitively speaking, we make use of the classical decomposition of the error into a bias term and a stochastic error by approximating it through the bias proxy $\widehat{\Delta}_t$ and $A_t^{(q)}$, which mimics the stochastic error (see Proposition B.14).

PROPOSITION B.16. *Let $D \subset \mathbb{R}^{2d}$ be an open and bounded set, assume $\mathcal{A}$, and let $K_1, K_2 : \mathbb{R}^d \to \mathbb{R}$ be symmetric, Lipschitz continuous kernel functions. Then, there exists a constant c such that, for any $t > 0$ sufficiently large,*

$$\mathcal{R}_\infty^{(p)}(\overline{b}_{j,\widehat{\boldsymbol{h}}}, b^j\rho; D) \le c\Big(\mathfrak{R}_t(b^j\rho) + (\log t)^{2/q+1/2}t^{-1/2}\Big), \quad \forall 1 \le p \le q, \tag{B.23}$$

where

$$\mathfrak{R}_t(b^j\rho) := \inf_{\boldsymbol{h}=(h_1,h_2)\in\mathcal{H}_t}\left\{\mathcal{B}_{b^j\rho}(\boldsymbol{h}) + (h_1h_2)^{-d/2}\sqrt{\frac{\log(h_1^{-1}+h_2^{-1})}{t}}\right\}.$$

In the last step of our investigation, we transfer the above result to a finding on the original question of drift estimation. Given a bounded, open set $D \subset \mathbb{R}^{2d}$, denote by $\rho_\star > 0$ an a priori lower bound on the invariant density fulfilling $\inf_{(x,y)\in D}\rho(x,y) \ge \rho_\star$. Similarly to (B.17), define

$$\widehat{b}_{j,h_1,h_2,t} = \widehat{b}_{j,\boldsymbol{h},t} := \frac{\overline{b}_{j,h_1,h_2,t}}{\widehat{\rho}_{h_1,h_2,t}\vee\rho_\star} = \frac{\overline{b}_{j,\boldsymbol{h},t}}{\widehat{\rho}_{\boldsymbol{h},t}\vee\rho_\star}, \quad \boldsymbol{h} = (h_1,h_2)^\top \in \overline{\mathcal{H}}_t := \mathcal{H}_t \cap \mathcal{H}(Q_1,Q_2).$$

THEOREM B.17. *Grant the assumptions of Proposition B.16, and assume, in addition, that $\rho, b^j\rho \in \mathcal{H}_D(\beta_1,\beta_2,\mathcal{L}_1,\mathcal{L}_2)$ with $\overline{\beta} > d, \beta_1 \le \ell_1, \beta_2 \le \ell_2$. Defining the bandwidth $\widehat{\boldsymbol{h}} = \widehat{\boldsymbol{h}}^j$ via*

$$\widehat{\Delta}_t^j(\widehat{\boldsymbol{h}}^j) + A_t^{(q)}(\widehat{\boldsymbol{h}}^j) = \inf_{\boldsymbol{h}=(h_1,h_2)\in\overline{\mathcal{H}}_t}\Big\{\widehat{\Delta}_t^j(\boldsymbol{h}) + A_t^{(q)}(\boldsymbol{h})\Big\}$$

then yields

$$\mathcal{R}_\infty^{(p)}(\widehat{b}_{j,\widehat{\boldsymbol{h}}^j,t}, b^j; D) \in \mathrm{O}\Big((\log t/t)^{\frac{\overline{\beta}}{2(\overline{\beta}+d)}}\Big), \quad \forall 1 \le p \le q. \tag{B.24}$$

Appendices

B.I Proofs for Section B.3

We will require the following auxiliary result for the proof of the variance bounds in Propositions B.2 and B.3.

Lemma B.18. *Let D be a bounded subset of $\mathbb{R}^{2d}$. Then, there exists a constant $c_D > 0$ such that, for all $0 < v < 1, v \le t$, for all $z \in \mathbb{R}^{2d}$ and any bounded measurable function f with support D,*

$$|P_t(f)(z)| \le c_D\left(\frac{\|f\|_{L^1(\mathbb{R}^{2d})}}{v^{2d}} + \|f\|_\infty \exp\left(-\frac{1}{c_D v}\right)\right).$$

Proof. We start by proving the assertion for $v = t$ and $z \in \widetilde{D}$, where $\widetilde{D} := \{z \in \mathbb{R}^{2d} : d(z, D) \le 1\}$, with $d(z, D) := \inf_{x\in D} \|z - x\|$ denoting the distance of the point z to the set D. Since $v < 1$, (𝒜2) implies that there exists a constant c depending on $\widetilde{D}$ such that

$$|P_v(f)(z)| \le \int |f(z')| p_v(z, z')\, \mathrm{d}z' \le cv^{-2d} \int |f(z')|\, \mathrm{d}z' + c\|f\|_\infty \exp\left(-\frac{1}{cv}\right),$$

thus completing the proof in this case. For analysing the case $z \notin \widetilde{D}$, introduce the first hitting time of $\widetilde{D}$, defined as $\tau_{\widetilde{D}} := \inf\{t \ge 0 : Z_t \in \widetilde{D}\}$. Continuity and the strong Markov property of $\mathbf{Z}$ imply

$$P_v(f)(z) = \mathbb{E}_z\left[f(Z_v)\mathbf{1}_{[0,v]}(\tau_{\widetilde{D}})\right] = \mathbb{E}_z\left[P_{v-\tau_{\widetilde{D}}}(f)(Z_{\tau_{\widetilde{D}}})\mathbf{1}_{[0,v]}(\tau_{\widetilde{D}})\right],$$

and $v - \tau_{\widetilde{D}} \in (0, v)$, $d(Z_{\tau_{\widetilde{D}}}, D) = 1$. Thus, it is enough to find a bound for $|P_s(f)(z')|$ such that $s \in (0, v)$, $z' \in \widetilde{D} : d(z', D) = 1$. Assumption (𝒜2) now implies for this case

$$|P_s(f)(z')| \le \int |f(w)| p_s(z', w)\, \mathrm{d}w \le c\left(\mathcal{P}\int |f(w)|\, \mathrm{d}w + \|f\|_\infty \exp\left(-\frac{1}{cD}\right)\right),$$

where

$$\mathcal{P} := \sup\left\{ s^{-2d} \exp\left(-c^{-1}\left(\frac{\|z_2' - w_2\|^2}{4s} + \frac{3\|w_1 - z_1' - \frac{s(z_2'+w_2)}{2}\|^2}{s^3}\right)\right) : \begin{array}{l} s \in (0, v), (w_1, w_2) \in D, \\ (z_1', z_2') \in \widetilde{D} : d(z', D) = 1 \end{array} \right\}$$

and the constant c only depends on $\widetilde{D}$. To show that $\mathcal{P}$ is finite, fix $w = (w_1, w_2) \in D$, $s \in (0, 1)$, $z' = (z_1', z_2') \in \widetilde{D} : d(z', D) = 1$. Then, the reverse triangle inequality and the inequality $(A - B)^2 \ge A^2\frac{s}{s+1} - B^2 s$, valid for any $A, B \in \mathbb{R}$, imply

$$\begin{aligned}\frac{\|z_2' - w_2\|^2}{4s} + \frac{3\|w_1 - z_1' - \frac{s(z_2'+w_2)}{2}\|^2}{s^3} &\ge \frac{1}{4}\left(\frac{\|z_2' - w_2\|^2}{s} + \frac{(\|w_1 - z_1'\| - \|\frac{s(z_2'+w_2)}{2}\|)^2}{s^3}\right) \\ &\ge \frac{1}{4}\left(\frac{\|z_2' - w_2\|^2}{s} + \frac{\|w_1 - z_1'\|^2}{s^2(s+1)} - \frac{\|\frac{s(z_2'+w_2)}{2}\|^2 s}{s^3}\right)\end{aligned}$$

$$\geq \frac{1}{4}\left(\frac{\|z_2' - w_2\|^2 + \|w_1 - z_1'\|^2}{2s} - c_2\right) = \frac{1}{4}\left(\frac{\|w - z'\|^2}{2s} - c_2\right)$$
$$\geq \frac{1}{8s} - \frac{c_2}{4},$$

where c_2 denotes some uniform bound of $\|z_2' + w_2\|$ which is finite because D is bounded. Thus, $\mathcal{P}$ is indeed bounded by a finite constant (depending on D and d) and hence the assertion also follows in this case because $v < 1$. For the case $v < t$, we have

$$|P_t(f)(z)| \leq \int p_{t-v}(z, z')|P_v(f)(z')|\,\mathrm{d}z'.$$

Thus, the assertion follows by the bound derived above. This completes the proof. ∎

Proofs of Propositions B.2 and B.3. Throughout the whole proof, we will suppress the dependence of functions on T for notational convenience.

Proof of (B.4) We start with the following well-known bound of the variance functional

$$\mathrm{Var}\left(\frac{1}{T}\int_0^T f(X_s, Y_s)\,\mathrm{d}s\right) \leq \frac{2}{T}\int_0^T \underbrace{|\mathrm{Cov}(f(X_s, Y_s), f(X_0, Y_0))|}_{=:\mathcal{C}(s)}\,\mathrm{d}s. \tag{B.25}$$

We proceed by splitting the integral, using integral bounds $0 \leq \delta_0 \leq \delta \leq D_1 \leq D_2 \leq T$. Note that stationarity of $\mathbf{Z}$, boundedness of ρ and the Cauchy–Schwarz inequality imply

$$\mathcal{C}(s) \leq \mathrm{Var}(f(X_0, Y_0)) \leq \mathbb{E}\left[f^2(X_0, Y_0)\right] \leq c\|f\|_\infty^2 \lambda(\mathcal{S}),$$

for some suitable constant $c > 0$, where $\mathcal{S} := \mathrm{supp}(f)$. Hence,

$$\int_0^{\delta_0} \mathcal{C}(s)\,\mathrm{d}s \leq c\delta_0\|f\|_\infty^2\lambda(\mathcal{S}). \tag{B.26}$$

Furthermore, the heat kernel bound (A2) and boundedness of ρ imply for $\delta_0 < s < \delta < 1$

$$\begin{aligned}
\mathcal{C}(s) &\leq \mathbb{E}[f(X_s, Y_s)f(X_0, Y_0)] + \mathbb{E}[f(X_0, Y_0)]^2 \\
&\leq c\int_{\mathbb{R}^{4d}} |f(x', y')||f(x'', y'')|\mathbf{1}_{\mathcal{S}}(x', y')\mathbf{1}_{\mathcal{S}}(x'', y'') \\
&\qquad \times s^{-2d}\exp\left(-c^{-1}\left(\frac{\|y' - y''\|^2}{4s} + \frac{3\|x'' - x' - \frac{s(y'+y'')}{2}\|^2}{s^3}\right)\right)\mathrm{d}x''\,\mathrm{d}y''\rho(x', y')\,\mathrm{d}x'\,\mathrm{d}y' \\
&\qquad + c\|f\|_\infty^2\lambda(\mathcal{S})\exp\left(-(cs)^{-1}\right) + c\|f\|_\infty^2\lambda^2(\mathcal{S}) \\
&\leq c\|f\|_\infty^2 \int_{\mathbb{R}^{2d}} \mathbf{1}_{\bigcup_{i=1}^n B(x_i, \mathfrak{s}_1)}(x')\mathbf{1}_{\bigcup_{i=1}^n B(x_i, \mathfrak{s}_1)}(x'')\mathbf{1}_{\bigcup_{i=1}^n B(y_i, \mathfrak{s}_2)}(y')\mathbf{1}_{\bigcup_{i=1}^n B(y_i, \mathfrak{s}_2)}(y'') \\
&\qquad \times \int_{\mathbb{R}^{2d}} s^{-2d}\exp\left(-c^{-1}\left(\frac{\|y' - y''\|^2}{4s} + \frac{3\|x'' - x' - \frac{s(y'+y'')}{2}\|^2}{s^3}\right)\right)\mathrm{d}y'\,\mathrm{d}y''\,\mathrm{d}x'\,\mathrm{d}x'' \\
&\qquad + c\|f\|_\infty^2\lambda(\mathcal{S})\exp\left(-(cs)^{-1}\right) + c\|f\|_\infty^2\lambda^2(\mathcal{S})
\end{aligned}$$

For bounding the inner integral, note that

$$\begin{aligned}
&\int_{\mathbb{R}^{2d}} \exp\left(-c^{-1}\left(\frac{\|y'-y''\|^2}{4s} + \frac{3\|x''-x'-\frac{s(y'+y'')}{2}\|^2}{s^3}\right)\right) \mathrm{d}y'\,\mathrm{d}y'' \\
&\quad = s^d \int_{\mathbb{R}^{2d}} \exp\left(-c^{-1}\left(\|w\|^2 + \|(x''-x')s^{-3/2} - \frac{w'}{2}\|^2\right)\right) \mathrm{d}w\,\mathrm{d}w' \\
&\quad = s^d \int_{\mathbb{R}^{2d}} \exp\left(-c^{-1}\left(\|w\|^2 + \|\frac{w'}{2}\|^2\right)\right) \mathrm{d}w\,\mathrm{d}w',
\end{aligned}$$

where we used the transformations $w = (y'-y'')/\sqrt{s}$, $w' = (y'+y'')/\sqrt{s}$ and the invariance of the Lebesgue measure under translation. Using polar coordinates, it is easy to see that the integral in the last line is bounded by some finite constant independent of x' and x''. Hence, we obtain that the inner integral is bounded by cs^d for some positive finite constant c. Thus,

$$\begin{aligned}
\mathcal{C}(s) &\le cs^{-d}\|f\|_\infty^2 \int_{\mathbb{R}^{2d}} \left(\sum_{i=1}^n \mathbf{1}_{B(x_i,\mathfrak{s}_1)}(x')\right)\left(\sum_{i=1}^n \mathbf{1}_{B(x_i,\mathfrak{s}_1)}(x'')\right) \mathrm{d}x'\,\mathrm{d}x'' \\
&\qquad + c\|f\|_\infty^2 \boldsymbol{\lambda}(\mathcal{S}) \exp\left(-(cs)^{-1}\right) + c\|f\|_\infty^2 \boldsymbol{\lambda}^2(\mathcal{S}) \\
&\le cn^2 \mathfrak{s}_1^{2d} s^{-d}\|f\|_\infty^2 \; + c\|f\|_\infty^2 \boldsymbol{\lambda}(\mathcal{S}) \exp\left(-(cs)^{-1}\right) + c\|f\|_\infty^2 \boldsymbol{\lambda}^2(\mathcal{S}),
\end{aligned} \tag{B.27}$$

which implies

$$\int_{\delta_0}^{\delta} \mathcal{C}(s)\,\mathrm{d}s \le c\|f\|_\infty^2 \left(\mathfrak{s}_1^{2d}(\log(\delta/\delta_0)\mathbb{1}_{d=1} + \delta_0^{1-d}\mathbb{1}_{d\ge 2}) + \boldsymbol{\lambda}(\mathcal{S}) \exp\left(-(c\delta)^{-1}\right) + \delta\boldsymbol{\lambda}^2(\mathcal{S})\right). \tag{B.28}$$

Subsequently, for $\delta \le s \le D_1 < 1$, (𝒜2) implies as in (B.27)

$$\begin{aligned}
\mathcal{C}(s) &\le c \int_{\mathbb{R}^{2d}\times\mathbb{R}^{2d}} |f(x',y')||f(x'';y'')| \\
&\qquad \times s^{-2d} \exp\left(-c^{-1}\left(\frac{\|y'-y''\|^2}{4s} + \frac{3\|x''-x'-\frac{s(y'+y'')}{2}\|^2}{s^3}\right)\right) \mathrm{d}x''\,\mathrm{d}y''\rho(x',y')\,\mathrm{d}x'\,\mathrm{d}y' \\
&\qquad + c\|f\|_\infty^2 \boldsymbol{\lambda}(\mathcal{S}) \exp\left(-(cs)^{-1}\right) + c\|f\|_\infty^2 \boldsymbol{\lambda}^2(\mathcal{S}) \\
&\le c\left(s^{-2d} \int_{\mathbb{R}^{4d}} |f(x',y')||f(x'';y'')|\,\mathrm{d}x''\,\mathrm{d}y''\,\mathrm{d}x'\,\mathrm{d}y' + \|f\|_\infty^2 \boldsymbol{\lambda}(\mathcal{S}) \exp\left(-(cs)^{-1}\right) + \|f\|_\infty^2 \boldsymbol{\lambda}^2(\mathcal{S})\right) \\
&\le c\|f\|_\infty^2 \left(s^{-2d}\boldsymbol{\lambda}^2(\mathcal{S}) + \boldsymbol{\lambda}(\mathcal{S}) \exp\left(-(cs)^{-1}\right) + \boldsymbol{\lambda}^2(\mathcal{S})\right),
\end{aligned}$$

and thus

$$\begin{aligned}
\int_{\delta}^{D_1} \mathcal{C}(s)\,\mathrm{d}s &\le c\|f\|_\infty^2 \int_{\delta}^{D_1} s^{-2d}\left(\boldsymbol{\lambda}^2(\mathcal{S}) + \boldsymbol{\lambda}(\mathcal{S}) \exp\left(-(cs)^{-1}\right) + \boldsymbol{\lambda}^2(\mathcal{S})\right) \mathrm{d}s \\
&\le c\|f\|_\infty^2 \left(\delta^{1-2d}\boldsymbol{\lambda}^2(\mathcal{S}) + \boldsymbol{\lambda}(\mathcal{S}) \exp\left(-(cD_1)^{-1}\right) + D_1\boldsymbol{\lambda}^2(\mathcal{S})\right) \\
&\le c\|f\|_\infty^2 \left(\delta^{1-2d}\boldsymbol{\lambda}^2(\mathcal{S}) + \boldsymbol{\lambda}(\mathcal{S}) \exp\left(-(cD_1)^{-1}\right)\right),
\end{aligned} \tag{B.29}$$

where we used $D_1 < 1$. We continue by investigating the integral from D_1 to D_2. Arguing as in the derivation of (B.27), we get

$$\begin{aligned}\mathcal{C}(s) &\le c\int_{\mathbb{R}^{2d}\times\mathbb{R}^{2d}} |f(x',y')||f(x'';y'')|p_s(x',y',x'',y'')\,\mathrm{d}x''\,\mathrm{d}y''\rho(x',y')\,\mathrm{d}x'\,\mathrm{d}y' + c\|f\|_\infty^2\lambda^2(\mathcal{S})\\ &= c\int_{\mathbb{R}^{2d}} |f(x',y')|P_s(|f|)(x',y')\rho(x',y')\,\mathrm{d}x'\,\mathrm{d}y' + c\|f\|_\infty^2\lambda^2(\mathcal{S}).\end{aligned}$$

Thus, Lemma B.18 implies

$$\begin{aligned}\int_{D_1}^{D_2}\mathcal{C}(s)\,\mathrm{d}s &\le c\int_{D_1}^{D_2}\int_{\mathbb{R}^{2d}} |f(x',y')|P_s(|f|)(x',y')\rho(x',y')\,\mathrm{d}x'\,\mathrm{d}y'\,\mathrm{d}s + cD_2\|f\|_\infty^2\lambda^2(\mathcal{S})\\ &\le c\int_{D_1}^{D_2} D_1^{-2d}\|f\|_\infty\lambda(\mathcal{S})\int_{\mathbb{R}^{2d}}|f(x',y')|\rho(x',y')\,\mathrm{d}x'\,\mathrm{d}y'\,\mathrm{d}s\\ &\qquad + c\int_{D_1}^{D_2}\|f\|_\infty\exp\left(-\frac{1}{cD_1}\right)\int_{\mathbb{R}^{2d}}|f(x',y')|\rho(x',y')\,\mathrm{d}x'\,\mathrm{d}y'\,\mathrm{d}s + c\|f\|_\infty^2\lambda^2(\mathcal{S})D_2\\ &\le c\|f\|_\infty^2\lambda(\mathcal{S})\left(D_2D_1^{-2d}\lambda(\mathcal{S}) + D_2\exp\left(-\frac{1}{cD_1}\right) + D_2\lambda(\mathcal{S})\right),\end{aligned} \tag{B.30}$$

where the value of c only depends on $\mathcal{S}$ and ρ. For the remaining part of the integral, we make use of the mixing property (A3), which implies that there exists $\kappa > 0$ such that

$$\int_{D_2}^{T}\mathcal{C}(s)\,\mathrm{d}s \le c\int_{D_2}^{T}\|f\|_\infty^2\mathrm{e}^{-\kappa s}\,\mathrm{d}s \le c\|f\|_\infty^2\mathrm{e}^{-\kappa D_2}. \tag{B.31}$$

The fact that exponential β-mixing implies a covariance bound of the above form follows from the proof on page 479 in [21] and is also described in equation (5) of [5]. Combining (B.25), (B.26), (B.30), (B.31), (B.29) and (B.28) then yields that there exist $c,\kappa > 0$ such that, for $0 \le \delta_0 \le \delta \le D_1 < 1 \le D_2 \le T$,

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{T}\int_0^T f(X_s,Y_s)\,\mathrm{d}s\right) \le cT^{-1}\|f\|_\infty^2\Big(&\delta_0\lambda(\mathcal{S}) + \mathsf{s}_1^{2d}(\log(\delta/\delta_0)\mathbb{1}_{d=1} + \delta_0^{1-d}\mathbb{1}_{d\ge2})\\ &+ \lambda(\mathcal{S})\exp\left(-(c\delta)^{-1}\right) + \lambda^2(\mathcal{S})\delta + \delta^{1-2d}\lambda^2(\mathcal{S})\\ &+ \lambda(\mathcal{S})\exp\left(-(cD_1)^{-1}\right) + D_2D_1^{-2d}\lambda^2(\mathcal{S})\\ &+ D_2\lambda(\mathcal{S})\exp\left(-(cD_1)^{-1}\right) + D_2\lambda^2(\mathcal{S}) + \mathrm{e}^{-\kappa D_2}\Big),\end{aligned}$$

and choosing $D_1 = (-c\log(n(\mathsf{s}_1\mathsf{s}_2)^d))^{-1}$, $D_2 = -2\kappa^{-1}\log(n(\mathsf{s}_1\mathsf{s}_2)^d)$ we get

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{T}\int_0^T f(X_s,Y_s)\,\mathrm{d}s\right) \le cT^{-1}\|f\|_\infty^2\Big(&\delta_0 n(\mathsf{s}_1\mathsf{s}_2)^d + \mathsf{s}_1^{2d}(\log(\delta/\delta_0)\mathbb{1}_{d=1} + \delta_0^{1-d}\mathbb{1}_{d\ge2})\\ &+ n(\mathsf{s}_1\mathsf{s}_2)^d\exp\left(-(c\delta)^{-1}\right) + \delta^{1-2d}n^2(\mathsf{s}_1\mathsf{s}_2)^{2d}\\ &+ \log(n^{-1}(\mathsf{s}_1\mathsf{s}_2)^{-d})^{2d+1}n^2(\mathsf{s}_1\mathsf{s}_2)^{2d}\Big),\end{aligned} \tag{B.32}$$

where we used that $\delta < 1$ and $s_1, s_2 \in \mathcal{H}$. Choosing $\delta_0 = s_1, \delta = s_2$ if $s_1 < s_2$ now entails for large enough T in the case $d = 1$

$$\operatorname{Var}\left(\frac{1}{T}\int_0^T f(X_s, Y_s)\,\mathrm{d}s\right) \le cT^{-1}\|f\|_\infty^2 s_1^2 \log T.$$

We will explain at the end of the proof why the assumption $s_1 < s_2$ is indeed without loss of generality. For $d \ge 2$, the choice $\delta_0 = s_1 s_2^{-1}$, $\delta = D_1 = (-c\log(n(s_1 s_2)^d))^{-1}$ gives

$$\operatorname{Var}\left(\frac{1}{T}\int_0^T f(X_s, Y_s)\,\mathrm{d}s\right) \le cT^{-1}\|f\|_\infty^2 s_1^{d+1} s_2^{d-1},$$

where we used $s_1, s_2 \in \mathcal{H}$.

Proof of (B.5) Again we split up the covariance integral from (B.25) into five parts. The only new bound concerns the integral from δ_0 to δ. Arguing as in (B.27), we obtain for $0 < s < \delta < 1$

$$\begin{aligned}\mathcal{C}(s) \le c\int_{\mathbb{R}^{4d}} |f(x', y')||f(x''; y'')| p_s^1(x', y', x'', y')\,\mathrm{d}x''\,\mathrm{d}y''\rho(x', y')\,\mathrm{d}x'\,\mathrm{d}y' \\ + c\|f\|_\infty^2 \lambda(\mathcal{S}) \exp\left(-(cs)^{-1}\right) + c\|f\|_\infty^2 \lambda^2(\mathcal{S}),\end{aligned}$$

where

$$p_s^1(x', y', x'', y') := s^{-2d}\exp\left(-c^{-1}\left(\frac{\|y' - y''\|^2}{4s} + \frac{3\|x'' - x' - \frac{s(y'+y'')}{2}\|^2}{s^3}\right)\right).$$

We proceed by finding a bound for $p_s^1(x', y', x'', y'') = s^{-d/2} q_s(x''|x', y', y'')$, where

$$q_s(x''|x', y', y'') := s^{-(3/2)d}\exp\left(-c^{-1}\left(\frac{3\|x'' - x' - \frac{s(y'+y'')}{2}\|^2}{s^3}\right)\right).$$

Since q_s resembles the density of a multidimensional normal distribution, we get

$$\begin{aligned}&\sup_{s\in(0,1)}\sup_{x', y', y''\in\mathbb{R}^d}\int q_s(x''|x', y', y')\,\mathrm{d}x'' \\ &\qquad \le \sup_{s\in(0,1)}\sup_{x', y', y''\in\mathbb{R}^d} s^{-(3/2)d}\int \exp\left(-c^{-1}\left(\frac{\|x'' - x' - \frac{s(y'+y'')}{2}\|^2}{s^3}\right)\right)\mathrm{d}x'' \le c.\end{aligned}$$

Hence, we can infer

$$\begin{aligned}\mathcal{C}(s) &\le cs^{-d/2}\int_{\mathbb{R}^{4d}} |f(x''; y'')| q_s(x''|x', y', y'')\,\mathrm{d}x''\,\mathrm{d}y''|f(x', y')|\rho(x', y')\,\mathrm{d}x'\,\mathrm{d}y' \\ &\quad + c\|f\|_\infty^2 \lambda(\mathcal{S})\exp\left(-(cs)^{-1}\right) + c\|f\|_\infty^2 \lambda^2(\mathcal{S}) \\ &\le c\left(\|f\|_\infty s_2^d s^{-d/2}\int_{\mathbb{R}^{2d}} |f(x', y')|\rho(x', y')\,\mathrm{d}x'\,\mathrm{d}y' + \|f\|_\infty^2 \lambda(\mathcal{S})\exp\left(-(cs)^{-1}\right) + \|f\|_\infty^2 \lambda^2(\mathcal{S})\right)\end{aligned}$$

$$\leq c\|f\|_\infty^2(\mathsf{s}_2^d s^{-d/2}\lambda(\mathcal{S}) + \lambda(\mathcal{S})\exp\Big(-(cs)^{-1}\Big) + \lambda^2(\mathcal{S})).$$

Consequently, for $\delta_0 \leq \delta$,

$$\begin{aligned}\int_{\delta_0}^{\delta} |\mathcal{C}(s)|\,\mathrm{d}s &\leq c\|f\|_\infty^2 \int_{\delta_0}^{\delta} \Big(\mathsf{s}_2^d s^{-d/2}\lambda(\mathcal{S}) + \lambda(\mathcal{S})\exp\Big(-(cs)^{-1}\Big) + \lambda^2(\mathcal{S})\Big)\,\mathrm{d}s \\ &\leq c\|f\|_\infty^2 \int_{\delta_0}^{\delta} \Big(\mathsf{s}_2^d s^{-d/2}\lambda(\mathcal{S}) + \lambda(\mathcal{S})s^{-2}\exp\Big(-(cs)^{-1}\Big) + \lambda^2(\mathcal{S})\Big)\,\mathrm{d}s \\ &\leq c\|f\|_\infty^2 \Big(\mathsf{s}_2^d\lambda(\mathcal{S})(\sqrt{\delta}\mathbf{1}_{d=1} + \log(\delta/\delta_0)\mathbf{1}_{d=2} + \delta_0^{1-d/2}\mathbf{1}_{d\geq 3}) \\ &\qquad + \lambda(\mathcal{S})\exp\Big(-(c\delta)^{-1}\Big) + \lambda^2(\mathcal{S})\delta\Big). \end{aligned} \tag{B.33}$$

Combining (B.25), (B.26), (B.29), (B.30), (B.31) and (B.33), we get by choosing $D_1 = (-c\log(n(\mathsf{s}_1\mathsf{s}_2)^d))^{-1}$, $D_2 = -2\kappa^{-1}\log(n(\mathsf{s}_1\mathsf{s}_2)^d)$ as in (B.32)

$$\begin{aligned}\mathrm{Var}\Big(\frac{1}{T}\int_0^T f(X_s,Y_s)\,\mathrm{d}s\Big) \leq cT^{-1}\|f\|_\infty^2\Big(&\delta_0 n(\mathsf{s}_1\mathsf{s}_2)^d + n\mathsf{s}_1^d\mathsf{s}_2^{2d}(\sqrt{\delta}\mathbf{1}_{d=1} + \log(\delta/\delta_0)\mathbf{1}_{d=2} + \delta_0^{1-d/2}\mathbf{1}_{d\geq 3}) \\ &+ n(\mathsf{s}_1\mathsf{s}_2)^d\exp\Big(-(c\delta)^{-1}\Big) + \delta^{1-2d}n^2(\mathsf{s}_1\mathsf{s}_2)^{2d} \\ &+ \log(n^{-1}(\mathsf{s}_1\mathsf{s}_2)^{-d})^{2d+1}n^2(\mathsf{s}_1\mathsf{s}_2)^{2d}\Big).\end{aligned}$$

Choosing $\delta_0 = 0, \delta = \mathsf{s}_1^{2/3}$ for $d = 1$ then yields

$$\mathrm{Var}\Big(\frac{1}{T}\int_0^T f(X_s,Y_s)\,\mathrm{d}s\Big) \leq cT^{-1}\|f\|_\infty^2\mathsf{s}_1^{4/3}\mathsf{s}_2^2.$$

For $d = 2$, we choose $\delta_0 = \mathsf{s}_2^2$, $\delta = \mathsf{s}_1^{2/3}$ if $\mathsf{s}_2 \leq \mathsf{s}_1^{1/3}$ which entails

$$\mathrm{Var}\Big(\frac{1}{T}\int_0^T f(X_s,Y_s)\,\mathrm{d}s\Big) \leq cT^{-1}\|f\|_\infty^2\mathsf{s}_1^2\mathsf{s}_2^4\log(T),$$

and, for $d \geq 3$, we set $\delta_0 = \mathsf{s}_2^2, \delta = D_1$, yielding

$$\mathrm{Var}\Big(\frac{1}{T}\int_0^T f(X_s,Y_s)\,\mathrm{d}s\Big) \leq cT^{-1}\|f\|_\infty^2\mathsf{s}_1^d\mathsf{s}_2^{d+2}.$$

As with the assumption $\mathsf{s}_1 < \mathsf{s}_2$ in the verification of (B.4), we will explain at the end of the proof why the assumption $\mathsf{s}_2 \leq \mathsf{s}_1^{1/3}$ is in fact only temporary.

Proof of (B.6) For proving (B.6), we set $\delta_0 = \delta$. Then, we only need to find a new bound for the covariance integral from δ to D_1. Arguing as in the derviation of (B.27), we get for $0 < s < 1$

$$\mathcal{C}(s) \leq c\|f\|_\infty^2 \int_{\mathbb{R}^{2d}} \mathbf{1}_{\cup_{i=1}^n B(x_i,\mathsf{s}_1)}(x')\mathbf{1}_{\cup_{i=1}^n B(x_i,\mathsf{s}_1)}(x'')\mathbf{1}_{\cup_{i=1}^n B(y_i,\mathsf{s}_2)}(y')\mathbf{1}_{\cup_{i=1}^n B(y_i,\mathsf{s}_2)}(y'')$$

$$\times \int_{\mathbb{R}^{2d}} s^{-2d} \exp\left(-c^{-1}\left(\frac{\|y'-y''\|^2}{4s} + \frac{3\|x''-x'-\frac{s(y'+y'')}{2}\|^2}{s^3}\right)\right) \mathrm{d}y'\,\mathrm{d}y''\,\mathrm{d}x'\,\mathrm{d}x''$$
$$+ c\|f\|_\infty^2 \lambda(\mathcal{S}) \exp\left(-(cs)^{-1}\right) + c\|f\|_\infty^2 \lambda^2(\mathcal{S}).$$

We continue by bounding the exponent in the inner integral. Under the given assumptions on f, the relation $f(x',y')f(x'',y'') \neq 0$ implies

$$\exists \mathsf{x}, \mathsf{y} \in \mathbb{R}^{2d}: \quad \|x'-\mathsf{x}\| < \mathsf{s}_1, \quad \|x''-\mathsf{x}\| < \mathsf{s}_1, \quad \|y'-\mathsf{y}\| < \mathsf{s}_2, \quad \|y''-\mathsf{y}\| < \mathsf{s}_2.$$

The reverse triangle inequality and the well-known inequality $(a+b)^2 \leq 2(a^2+b^2)$ additionally yield if $\mathsf{s}_2 \leq \|\mathsf{y}\|/2$, i.e., T is large enough,

$$\begin{aligned}
\frac{\|y'-y''\|^2}{4s} + \frac{3\|x''-x'-\frac{s(y'+y'')}{2}\|^2}{s^3} &\geq \frac{\|y'-y''\|^2}{4s} + \frac{3(\|x''-x'\| - \|\frac{s(y'+y'')}{2}\|)^2}{s^3} \\
&\geq \frac{(\|y'-y''\| + \|y'+y''\|)^2}{8s} \\
&\quad + \frac{3\|x''-x'\|}{s^3}(\|x''-x'\| - s\|y'+y''\|) \\
&\geq \frac{\|y'\|^2}{2s} + \frac{3\|x''-x'\|}{s^3}(\|x''-x'\| - s\|y'+y''\|) \\
&\geq \frac{\|\mathsf{y}\|^2}{8s} + \frac{3\|x''-x'\|}{s^3}(\|x''-x'\| - s\|y'+y''\|). \qquad \text{(B.34)}
\end{aligned}$$

Hence, we have for $s \geq 288\mathsf{s}_1/\|\mathsf{y}\|$

$$\begin{aligned}
\frac{\|y'-y''\|^2}{4s} + \frac{3\|x''-x'-\frac{s(y'+y'')}{2}\|^2}{s^3} &\geq \frac{\|\mathsf{y}\|^2}{8s} - \frac{3\|x''-x'\|\|y'+y''\|}{s^2} \\
&\geq \frac{\|\mathsf{y}\|^2}{8s} - \frac{18\mathsf{s}_1\|\mathsf{y}\|}{s^2} \\
&\geq \frac{\|\mathsf{y}\|^2}{16s}.
\end{aligned}$$

Thus, for $288\mathsf{s}_1/\|\mathsf{y}\| \leq s < 1$, it holds

$$\begin{aligned}
\mathcal{C}(s) &\leq c\|f\|_\infty^2 \int_{\mathbb{R}^{4d}} \mathbf{1}_{\bigcup_{i=1}^n B(x_i,\mathsf{s}_1)}(x') \mathbf{1}_{\bigcup_{i=1}^n B(x_i,\mathsf{s}_1)}(x'') \mathbf{1}_{\bigcup_{i=1}^n B(y_i,\mathsf{s}_2)}(y') \mathbf{1}_{\bigcup_{i=1}^n B(y_i,\mathsf{s}_2)}(y'') \\
&\qquad \times s^{-2d} \exp\left(-\frac{1}{cs}\right) \mathrm{d}y'\,\mathrm{d}y''\,\mathrm{d}x'\,\mathrm{d}x'' \\
&\quad + c\|f\|_\infty^2 \lambda(\mathcal{S}) \exp\left(-(cs)^{-1}\right) + c\|f\|_\infty^2 \lambda^2(\mathcal{S}) \\
&\leq c\|f\|_\infty^2 \left(s^{-2d}(\mathsf{s}_1\mathsf{s}_2)^{2d} \exp\left(-(cs)^{-1}\right) + (\mathsf{s}_1\mathsf{s}_2)^d \exp\left(-(cs)^{-1}\right) + (\mathsf{s}_1\mathsf{s}_2)^{2d}\right),
\end{aligned}$$

which implies for $288\mathsf{s}_1/\|\mathsf{y}\| \leq \delta < D_1 < 1$

$$\int_\delta^{D_1} \mathcal{C}(s)\,\mathrm{d}s \leq c\|f\|_\infty^2 (\mathsf{s}_1\mathsf{s}_2)^d \int_\delta^{D_1} \left((\mathsf{s}_1\mathsf{s}_2)^d s^{-2d} \exp\left(-(cs)^{-1}\right) + \exp\left(-(cs)^{-1}\right) + (\mathsf{s}_1\mathsf{s}_2)^d\right) \mathrm{d}s$$

$$
\begin{aligned}
&\leq c\|f\|_\infty^2(s_1s_2)^d\left(\int_{(cD_1)^{-1}}^{(c\delta)^{-1}}\left((s_1s_2)^d s^{2(d-1)}e^{-s}+s^2e^{-s}\right)\mathrm{d}s + D_1(s_1s_2)^d\right)\\
&\leq c\|f\|_\infty^2(s_1s_2)^d\left((s_1s_2)^d\Gamma(2d-1,(cD_1)^{-1})+\Gamma(3,(cD_1)^{-1})+D_1(s_1s_2)^d\right)\\
&\leq c\|f\|_\infty^2(s_1s_2)^d\left((s_1s_2)^d\exp\left(-\frac{1}{cD_1}\right)\sum_{k=0}^{2(d-1)}(cD_1)^{-k}\right.\\
&\qquad\qquad\left.+\exp\left(-\frac{1}{cD_1}\right)\sum_{k=0}^{2}(cD_1)^{-k}+D_1(s_1s_2)^d\right)\\
&\leq c\|f\|_\infty^2(s_1s_2)^d\left((s_1s_2)^d\exp\left(-\frac{1}{cD_1}\right)D_1^{-2(d-1)}+\exp\left(-\frac{1}{cD_1}\right)D_1^{-2}+D_1(s_1s_2)^d\right).
\end{aligned}
\tag{B.35}
$$

Here, $\Gamma(\cdot,\cdot)$ denotes the upper incomplete gamma function, whose explicit values are well-known if the first argument is an integer. Combining now (B.25), (B.26), (B.30), (B.31) and (B.35), we get by choosing
$D_1 = (-c\log((s_1s_2)^d))^{-1}$, $D_2 = -2\kappa^{-1}\log((s_1s_2)^d)$ as in (B.32)

$$
\mathrm{Var}\left(\frac{1}{T}\int_0^T f(X_s,Y_s)\,\mathrm{d}s\right)\leq cT^{-1}\|f\|_\infty^2\left(\delta(s_1s_2)^d+\log((s_1s_2)^{-d})^{2d+1}(s_1s_2)^{2d}\right),
$$

and thus choosing $\delta = 288s_1/\|y\|$ entails for large enough T

$$
\mathrm{Var}\left(\frac{1}{T}\int_0^T f(X_s,Y_s)\,\mathrm{d}s\right)\leq cT^{-1}\|f\|_\infty^2 s_1(s_1s_2)^d.
$$

Proof of (B.7) To prove the assertion, it suffices to combine equations (B.25), (B.26), (B.30), (B.31), (B.33) and (B.35) with the choices $\delta_0 = 0$, $\delta = 288s_1/\|y\|$, $D_1 = (-c\log((s_1s_2)^d))^{-1}$, $D_2 = -2\kappa^{-1}\log((s_1s_2)^d)$.
To conclude the proof it only remains to consider why the assumptions $s_1 < s_2$ in the proof of (B.4) and $s_2 \leq s_1^{1/3}$ for $d = 2$ in the proof of (B.5) are negligible. For this note that if one of these assumptions fails to hold, the corresponding other assumption is fulfilled and hence the other variance bound holds, which then yields a tighter bound. ∎

Define the function class

$$
\mathcal{G} := \left\{K((x-\cdot)/h_1,(y-\cdot)/h_2) : (x,y)\in D\cap\mathbb{Q}^{2d}\right\},\quad h_1,h_2\in(0,1),
$$

where $K(x,y) = K_1(x)K_2(y)$ with K_1, K_2 being Lipschitz continuous, bounded functions of compact support with Lipschitz constants L_1, L_2 wrt to the sup-norm $\|\cdot\|_\infty$.

LEMMA B.19. *Let $D\subset\mathbb{R}^{2d}$ be a bounded set, assume $\mathcal{A}$, and let f be a locally bounded function. Then, for large enough t, it holds for any $\varepsilon > 0$*

$$
\mathcal{N}(\varepsilon,\mathcal{G}f,d_\infty)\leq\left(\frac{2L_K\sup_{z\in\mathbb{K}}|f(z)|(h_1^{-1}+h_2^{-1})\,\mathrm{diam}(D)}{\varepsilon}\right)^{2d},
$$

$$\mathcal{N}(\varepsilon, \mathcal{G}f, d_{L^2(\mu)}) \le \left(\frac{2L_K \operatorname{diam}(D) \sqrt{\|\rho\|_\infty} \sup_{z\in\mathcal{K}} |f(z)| (h_1 h_2)^{d/2} (h_1^{-1} + h_2^{-1})}{\varepsilon} \right)^{2d},$$

$$\mathcal{N}(\varepsilon, \mathcal{G}f, d_{\mathbb{G},t}) \le \left(\frac{2L_K c_D \sup_{z\in\mathcal{K}} |f(z)| (h_1 h_2)^{d} \psi_d(h_1, h_2, t)(h_1^{-1} + h_2^{-1}) \operatorname{diam}(D)}{\varepsilon} \right)^{2d},$$

where

$$\mathcal{K} := \bigcup_{(x,y)\in D\cap\mathbb{Q}^{2d}} \operatorname{supp}(K((x-\cdot)/h_1, (y-\cdot)/h_2)), \quad L_K := L_1\|K_2\|_\infty + L_2\|K_1\|_\infty.$$

Proof. Fix $\varepsilon > 0$. For all $(x, y) \in D \cap \mathbb{Q}^{2d}$, $K((x-\cdot)/h_1, (y-\cdot)/h_2)$ is Lipschitz continuous with Lipschitz constant $L_K(h_1^{-1} + h_2^{-1})$ wrt to the sup-norm. Hence, for $(x, y) \in \mathbb{R}^{2d}$

$$\begin{aligned} &B_{d_\infty}(K((x-\cdot)/h_1, (y-\cdot)/h_2), \varepsilon) \\ &\quad \supset \{K((a-\cdot)/h_1, (b-\cdot)/h_2) : \|K((x-\cdot)/h_1, (y-\cdot)/h_2) - K((a-\cdot)/h_1, (b-\cdot)/h_2)\|_\infty < \varepsilon\} \\ &\quad \supset \left\{K((a-\cdot)/h_1, (b-\cdot)/h_2) : \|(x, y) - (a, b)\|_\infty < \varepsilon (L_K(h_1^{-1} + h_2^{-1}))^{-1}\right\}. \end{aligned} \tag{B.36}$$

Let $Q \supset D$ be a cube of side length $\operatorname{diam}(D) < \infty$. Then, for

$$\overline{n} := \left(\left\lfloor \frac{L_K(h_1^{-1} + h_2^{-1}) \operatorname{diam}(D)}{\varepsilon} \right\rfloor \right)^{2d},$$

there exist points $(x_1, y_1) \ldots, (x_{\overline{n}}, y_{\overline{n}}) \in Q$ such that $D \subset Q \subset \bigcup_{i=1}^{\overline{n}} B_{d_\infty}((x_i, y_i), \varepsilon(L_K(h_1^{-1} + h_2^{-1}))^{-1})$. It now follows from (B.36) that $\{B_{d_\infty}(K((x_i-\cdot)/h_1, (y_i-\cdot)/h_2), \varepsilon) : i = 1, \ldots, \overline{n}\}$ is an external covering of $\mathcal{G}$. Hence, we obtain

$$\mathcal{N}(\varepsilon, \mathcal{G}, d_\infty) \le \mathcal{N}_{\text{ext}}(\varepsilon/2, \mathcal{G}, d_\infty) \le \left(\frac{2L_K(h_1^{-1} + h_2^{-1}) \operatorname{diam}(D)}{\varepsilon} \right)^{2d}.$$

Now, let $\mathcal{F}$ be an $\frac{\varepsilon}{\sup_{z\in\mathcal{K}} |f(z)|}$-cover of $\mathcal{G}$ with respect to d_∞, where we can assume without losing generality that $\sup_{z\in\mathcal{K}} |f(z)| > 0$. Then, for any $(x, y) \in D \cap \mathbb{Q}^{2d}$, there exists $g \in \mathcal{F}$ such that

$$d_\infty(fg, fK((x-\cdot)/h_1, (y-\cdot)/h_2)) \le \sup_{z\in\mathcal{K}} |f(z)| d_\infty(g, K((x-\cdot)/h_1, (y-\cdot)/h_2)) \le \varepsilon.$$

Now note that for $(x_1, y_1), (x_2, y_2) \in D \cap \mathbb{Q}^{2d}$

$$\begin{aligned} &d_{L^2(\mu)}(fK((x_1-\cdot)/h_1, (y_1-\cdot)/h_2), fK((x_2-\cdot)/h_1, (y_2-\cdot)/h_2)) \\ &\quad \le \sqrt{\|\rho\|_\infty} \sup_{z\in\mathcal{K}} |f(z)| (h_1 h_2)^{d/2} d_\infty(K(x_1/h_1 - \cdot, y_1/h_2 - \cdot), K(x_2/h_1 - \cdot, y_2/h_2 - \cdot)), \end{aligned}$$

and hence

$$\mathcal{N}(\varepsilon, \mathcal{G}f, d_{L^2(\mu)}) \le \mathcal{N}\left(\varepsilon \left(\sqrt{\|\rho\|_\infty} \sup_{z\in\mathcal{K}} |f(z)| (h_1 h_2)^{d/2} \right)^{-1}, \mathcal{G}, d_\infty \right)$$

$$\leq \left(\frac{2L_K \operatorname{diam}(D) \sqrt{\|\rho\|_\infty} \sup_{z\in\mathcal{K}} |f(z)| (h_1 h_2)^{d/2} (h_1^{-1} + h_2^{-1})}{\varepsilon} \right)^{2d}.$$

Similarly, we obtain from Proposition B.2 for $(x_1, y_1), (x_2, y_2) \in D \cap \mathbb{Q}^{2d}$ that there exists a constant $c_D > 0$, depending on D and K, such that for large enough t

$$\begin{aligned} &d_{\mathbb{G},t}((fK((x_1 - \cdot)/h_1, (y_1 - \cdot)/h_2), fK((x_2 - \cdot)/h_1, (y_2 - \cdot)/h_2)) \\ &\leq c_D \sup_{z\in\mathcal{K}} |f(z)| (h_1 h_2)^d \psi_d(h_1, h_2, t) d_\infty((K((x_1 - \cdot)/h_1, (y_1 - \cdot)/h_2) K((x_2 - \cdot)/h_1, (y_2 - \cdot)/h_2)). \end{aligned} \tag{B.37}$$

Thus, we have for large enough t,

$$\begin{aligned} \mathcal{N}(\varepsilon, \mathcal{G}f, d_{\mathbb{G},t}) &\leq \mathcal{N}\left(\varepsilon \left(c_D \sup_{z\in\mathcal{K}} |f(z)| (h_1 h_2)^d \psi_d(h_1, h_2, t) \right)^{-1}, \mathcal{G}, d_\infty \right) \\ &\leq \left(\frac{2L_K c_D \sup_{z\in\mathcal{K}} |f(z)| (h_1 h_2)^d \psi_d(h_1, h_2, t)(h_1^{-1} + h_2^{-1}) \operatorname{diam}(D)}{\varepsilon} \right)^{2d}, \end{aligned}$$

which completes the proof. ∎

Remark B.20. In equation (B.37), we implicitly used that there exists a *uniform* constant, such that the results of Propositions B.2 and B.3 hold for all $g \in \mathcal{G}$. A look at the proof of these assertions shows that this is indeed true, since we can find a bounded set $\widetilde{D} \subset \mathbb{R}^{2d}$ fulfilling

$$\bigcup_{(x,y)\in D\cap\mathbb{Q}^{2d}} \operatorname{supp}(K((x - \cdot)/h_1, (y - \cdot)/h_2)) \subset \widetilde{D}.$$

Furthermore, for the case $\inf_{(x,y)\in D} \|y\| > \varepsilon > 0$ and $n = 1$, the constants depending on $\|y_1\|$ in Proposition B.3 can all be replaced by analogous constants with respect to ε. This observation will also be used in the proof of Proposition B.5.

Proof of Proposition B.4. First note that the decomposition into bias and stochastic error yields

$$\mathcal{R}_\infty^{(p)}(\widehat{\rho}_{h_1,h_2,T}, \rho; D) \leq \mathbb{E}\left[\sup_{z\in D} |\widehat{\rho}_{h_1,h_2,T}(z) - \mu(\widehat{\rho}_{h_1,h_2,T})|^p \right]^{\frac{1}{p}} + \mathcal{B}_\rho(h_1, h_2).$$

We continue by bounding the first term. Denseness of $\mathbb{Q}$ in $\mathbb{R}$ gives

$$\mathbb{E}\left[\sup_{z\in D} |\widehat{\rho}_{h_1,h_2,T}(z) - \mu(\widehat{\rho}_{h_1,h_2,T})|^p \right]^{\frac{1}{p}} = T^{-1/2} (h_1 h_2)^{-d} \mathbb{E}\left[\sup_{g\in\overline{\mathcal{G}}} \|\mathbb{G}_T(g)\|^p \right]^{1/p},$$

where $\overline{\mathcal{G}} = \{K((x - \cdot)/h_1, (y - \cdot)/h_2) - \mu(K((x - \cdot)/h_1, (y - \cdot)/h_2)) : (x, y) \in D \cap \mathbb{Q}^{2d}\}$. Now, since $\mathbf{Z}$ is exponentially β-mixing, Theorem 3.2 in [8] implies that for $m_T \in (0, T/4]$ there exist $\tau \in [m_T, 2m_T]$ and a constant $c > 0$ such that

$$\mathbb{E}\left[\sup_{g\in\overline{\mathcal{G}}} \|\mathbb{G}_T(g)\|^p \right]^{1/p} \leq c \left(\int_0^\infty \log \mathcal{N}(u, \overline{\mathcal{G}}, \tfrac{2m_t}{\sqrt{T}} d_\infty)\, du + \int_0^\infty \sqrt{\log \mathcal{N}(u, \overline{\mathcal{G}}, d_{\mathbb{G},\tau})}\, du \right.$$

$$+ 4 \sup_{g \in \overline{\mathcal{G}}} \left(\tfrac{2m_T}{\sqrt{T}} \|g\|_\infty p + \|g\|_{\mathbb{G},\tau} \sqrt{p} + \tfrac{1}{2} \|g\|_\infty \sqrt{T} \mathrm{e}^{-\frac{\kappa m_T}{p}} \right) \Bigg). \tag{B.38}$$

Obviously, the results of Lemma B.19 continue to hold with different constants for $\overline{\mathcal{G}}$. Thus, for large enough T,

$$\begin{aligned} \int_0^\infty \log \mathcal{N}(u, \overline{\mathcal{G}}, \tfrac{2m_T}{\sqrt{T}} d_\infty) \, \mathrm{d}u &= c \frac{m_T}{\sqrt{T}} \int_0^c \log \mathcal{N}(u, \overline{\mathcal{G}}, d_\infty) \, \mathrm{d}u \\ &\leq c \frac{m_T}{\sqrt{T}} \int_0^c \log \left(\frac{c(h_1^{-1} + h_2^{-1})}{u} \right) \mathrm{d}u \leq c \frac{m_T}{\sqrt{T}} \log T. \end{aligned}$$

From Proposition B.2 and the inequality

$$\int_0^C \sqrt{\log(M/u)} \, \mathrm{d}u \leq 4C \sqrt{\log(M/C)} \quad \text{if} \quad \log(M/C) \geq 2 \tag{B.39}$$

(see, e.g., p. 592 of [12]), we get for large enough T

$$\begin{aligned} \int_0^\infty \sqrt{\log \mathcal{N}(u, \overline{\mathcal{G}}, d_{\mathbb{G},\tau})} \, \mathrm{d}u &\leq c \int_0^{c(h_1 h_2)^d \psi_d(h_1,h_2,T)} \sqrt{\log \frac{c(h_1 h_2)^d \psi_d(h_1,h_2,T)(h_1^{-1} + h_2^{-1})}{u}} \, \mathrm{d}u \\ &\leq c(h_1 h_2)^d \psi_d(h_1, h_2, T) \sqrt{\log T}, \end{aligned}$$

where we used (B.39) and $h_1, h_2 \in \mathcal{H}$. Letting $m_T = (p/\kappa) \log T$ yields together with Proposition B.2 and (B.38) for large enough T and $1 \leq p \leq \gamma \log T$, with $\gamma > 0$,

$$\mathbb{E}\left[\sup_{g \in \overline{\mathcal{G}}} \|\mathbb{G}_T(g)\|^p \right]^{1/p} \leq c \left(\frac{p(\log T)^2}{\sqrt{T}} + (h_1 h_2)^d \psi_d(h_1, h_2, t) \sqrt{\log T} + \frac{\log T}{\sqrt{T}} p^2 + (h_1 h_2)^d \psi_d(h_1, h_2, T) \sqrt{p} \right),$$

which completes the proof. ∎

Proof of Proposition B.5. Denoting $G_{h_1,h_2,T}(z) := \widehat{\rho}_{h_1,h_2,T}(z) - \mathbb{E}[\widehat{\rho}_{h_1,h_2,T}(z)]$, we obtain

$$\widehat{\rho}_{h_1,h_2,T}(z) - \rho(z) = G_{h_1,h_2,T}(z) + (\rho * K_{h_1,h_2} - \rho)(z), \quad \forall z \in D. \tag{B.40}$$

For bounding $\mathbb{E}[\|G_{h_1,h_2,T}(z)\|_{L^\infty(D)}]$, we discretize D by means of a finite set $D_T \subset D$ such that any point $z \in D$ fulfills $\inf_{\widetilde{z} \in D_T} |z - \widetilde{z}| \leq \delta_T$, which can be done with $\mathrm{card}(D_T) \leq c\delta_T^{-2d}$. Exploiting Lipschitz continuity of K_1, K_2 yields

$$\sup_{z \in D} |G_{h_1,h_2,T}(z)| - \sup_{z \in D_T} |G_{h_1,h_2,T}(z)| \leq c(h_1^{-1} + h_2^{-1})(h_1 h_2)^{-d} \delta_T.$$

Now, Proposition B.3 and Lemma B.12 imply that, for h_1, h_2 small enough and $m_t \in (0, \frac{t}{4}]$, there exists $\tau \in [m_t, 2m_t]$ such that

$$\mathbb{P}\left(\sup_{z \in D_T} |G_{h_1,h_2,T}(z)| > \left(r \sqrt{\frac{\log T}{T}} \psi_d^\circ(h_1, h_2) \right) \right)$$

$$
\begin{aligned}
&\leq 2\sum_{z\in D_T}\left(2\exp\left(-\frac{r^2\log T\psi_d^\circ(h_1,h_2)^2}{32(\tau\mathrm{Var}(\widehat{\rho}_{h_1,h_2,\tau}(z))+4\|K_h\|_\infty r\sqrt{\frac{\log T}{T}}\psi_d^\circ(h_1,h_2)m_T)}\right)\right.\\
&\qquad\left.+\frac{T}{m_T}c_\kappa \mathrm{e}^{-\kappa m_T}\mathbb{1}_{(0,8\|K_{h_1,h_2}\|_\infty)}\left(r\sqrt{\frac{\log T}{T}}\psi_d^\circ(h_1,h_2)\right)\right)\\
&\leq c\delta_T^{-2d}\left(\exp\left(-\frac{r^2\log T\psi_d^\circ(h_1,h_2)^2}{c(\psi_d^\circ(h_1,h_2)^2+(h_1h_2)^{-d}r\sqrt{\frac{\log T}{T}}\psi_d^\circ(h_1,h_2)m_T)}\right)\right.\\
&\qquad\left.+\frac{T}{m_T}\mathrm{e}^{-\kappa m_T}\mathbb{1}_{(0,c(h_1h_2)^{-d})}\left(r\sqrt{\frac{\log T}{T}}\psi_d^\circ(h_1,h_2)\right)\right)\\
&= c\delta_T^{-2d}\left(\exp\left(-\frac{r^2\log T}{c(1+(h_1h_2)^{-d}r\sqrt{\frac{\log T}{T}}\psi_d^\circ(h_1,h_2)^{-1}m_T)}\right)\right.\\
&\qquad\left.+\frac{T}{m_T}\mathrm{e}^{-\kappa m_T}\mathbb{1}_{(0,c(h_1h_2)^{-d}T^{1/2}\log T^{-1/2}\psi_d^\circ(h_1,h_2)^{-1})}(r)\right),
\end{aligned}
$$

where we assume that $m_T = c_m \log T$ for some $c_m > 0$. Then, the well-known inequality

$$
\mathbb{E}[Y] \leq a + \int_a^\infty \mathbb{P}(Y > r)\,\mathrm{d}r, \quad a \geq 0,
$$

implies for $Y := \sup_{z\in D_T}|G_{h_1,h_2,T}(z)|\sqrt{\frac{T}{\log T}}\psi_d^\circ(h_1,h_2)^{-1}$, with a suitable constant $c_a > 0$ depending on a,

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{z\in D_T}|G_{h_1,h_2,T}(z)|\right]\\
&\qquad\leq\left(\sqrt{\frac{\log T}{T}}\psi_d^\circ(h_1,h_2)\right)\left(a+c\delta_T^{-2d}\int_a^\infty\left(\exp\left(-\frac{r^2\log T}{c(1+(h_1h_2)^{-d}r\sqrt{\frac{\log T}{T}}\psi_d^\circ(h_1,h_2)^{-1}m_T)}\right)\right.\right.\\
&\qquad\qquad\left.\left.+\frac{T}{m_T}\mathrm{e}^{-\kappa m_T}\mathbb{1}_{(0,c(h_1h_2)^{-d}T^{1/2}\log T^{-1/2}\psi_d^\circ(h_1,h_2)^{-1})}(r)\right)\mathrm{d}r\right)\\
&\qquad\leq\left(\sqrt{\frac{\log T}{T}}\psi_d^\circ(h_1,h_2)\right)\left(a+c\delta_T^{-2d}\left(c_aT^{-\frac{a^2}{c}}+(h_1h_2)^{-d}T^{3/2-\kappa c_m}\log T^{-3/2}\psi_d^\circ(h_1,h_2)^{-1}\right)\right),
\end{aligned}
$$

where we used Assumption (B.9). Lipschitz continuity of K_{h_1,h_2} then yields

$$
\begin{aligned}
\mathbb{E}\left[\|G_{h_1,h_2,T}(z)\|_{L^\infty(D)}\right] &\leq \mathbb{E}\left[\left|\sup_{z\in D}|G_{h_1,h_2,T}(z)|-\sup_{z\in D_T}|G_{h_1,h_2,T}(z)|\right|\right]+\mathbb{E}\left[\sup_{z\in D_T}|G_{h_1,h_2,T}(z)|\right]\\
&\leq c(h_1h_2)^{-d}(h_1^{-1}+h_2^{-1})\delta_T\\
&\qquad+\sqrt{\frac{\log T}{T}}\psi_d(h_1,h_2)
\end{aligned}
$$

$$\times \left(a + c\delta_T^{-2d} \left(c_a T^{-\frac{a^2}{c}} + (h_1 h_2)^{-d} T^{3/2-\kappa c_m} \log T^{-3/2} \psi_d^\circ(h_1, h_2)^{-1} \right) \right).$$

Then, choosing $\delta_T = (h_1 h_2)^{d+1} \sqrt{T^{-1} \log T} \psi_d^\circ(h_1, h_2)$ immediately yields

$$(h_1 h_2)^{-d} (h_1^{-1} + h_2^{-1}) \delta_T \in \mathrm{O}\left(\sqrt{\frac{\log T}{T}} \psi_d^\circ(h_1, h_2) \right).$$

Now note that $h_1, h_2 \in \mathcal{H}$ implies the existence of $c, Q > 0$ such that $\delta_T^{-2d} \leq cT^Q$, for large enough T. Hence, choosing $a^2 = cQ$, $c_m = \kappa^{-1}(\frac{3}{2} + Q)$ yields

$$\mathbb{E}[\|G_{h_1,h_2,T}(z)\|_{L^\infty(D)}] \in \mathrm{O}\left(\sqrt{\frac{\log T}{T}} \psi_d^\circ(h_1, h_2) \right).$$

The assertion now follows by combining this with decomposition (B.40). ∎

Proof of Theorem B.7. It is well-known that there exists a constant $c > 0$, such that, for h_1, h_2 small enough and for all $z \in D$,

$$\mathcal{B}_\rho(h_1, h_2) = |(\rho * K_{h_1,h_2} - \rho)(z)| \leq c(h_1^{\beta_1} + h_2^{\beta_2}) \tag{B.41}$$

(see, e.g., Proposition 1 in [4]). Plugging this bound and h_1, h_2 as specified in (B.12) and (B.14) into (B.8) and (B.10), the assertion follows, since $\beta_1 > 1, \beta_2 > 2$ implies that (B.9) is satisfied. ∎

B.II Proofs for Section B.4

The proof of Theorem B.13 will require the following Lemma.

Lemma B.21. *Suppose that* $\mathbf{Z}$ *is exponentially* β*-mixing, and let* $\mathcal{G}$ *be a countable class of bounded real-valued functions. Then, for* $m_t \in (0, t/4)$*, there exists* $\tau \in [m_t, 2m_t]$ *such that, for any* $p \geq 1$*,*

$$\sup_{g \in \mathcal{G}} \left(\mathbb{E}\left[\left| \int_0^t g(Z_s)\, \mathrm{d}s \right|^p \right] \right)^{1/p} \leq \sup_{g \in \mathcal{G}} \left(c_1 m_t \|g\|_\infty p + c_2 \sqrt{tp} \|g\|_{\mathbb{G},\tau} + 2c_\kappa t \|g\|_\infty \mathrm{e}^{-\frac{\kappa m_t}{p}} + t|\mu(g)| \right),$$

where $c_1 = \frac{8}{3}\mathrm{e}^{1/2\mathrm{e}}\sqrt{2}\mathrm{e}^{1/(12)-1}$, $c_2 = 2(2\mathrm{e})^{-1/2}\mathrm{e}^{1/(2\mathrm{e})}\sqrt{\pi}\mathrm{e}^{1/6}$.

Proof. We start by splitting the process $(Z_s)_{0 \leq s \leq t}$ into $2n_t$ parts of length m_t, where $t = 2n_t m_t$, $n_t \in \mathbb{N}$, i.e., for $j \in \{1, \ldots, n_t\}$, we define the processes

$$Z^{j,1} := (Z_s)_{s \in [2(j-1)m_t, (2j-1)m_t]}, \quad Z^{j,2} := (Z_s)_{s \in [(2j-1)m_t, 2jm_t]}.$$

Analogously to the proof of Lemma 3.1 and Theorem 3.2 of [8], we use arguments of the proof of Proposition 5.2 of [21], yielding the existence of a process $(\widehat{Z}_s)_{0 \leq s \leq t}$ such that, for $k = 1, 2$,

(1) $Z^{j,k} \overset{(\mathrm{d})}{=} \widehat{Z}^{j,k}$ for all $j \in \{1, \ldots, n_t\}$,

(2) $\exists c_\kappa, \kappa > 0 : \mathbb{P}(Z^{j,k} \neq \widehat{Z}^{j,k}) \leq c_\kappa \mathrm{e}^{-\kappa m_t}$ for all $j \in \{1, \ldots, n_t\}$,

(3) $\widehat{Z}^{1,k}, \dots, \widehat{Z}^{n_t,k}$ are independent,

where $\widehat{Z}^{j,k}$ is defined analogously to $Z^{j,k}$ for $j \in \{1, \dots, n_t\}$, $k = 1, 2$. Furthermore, define

$$I_g(Z^{j,1}) := \int_{2(j-1)m_t}^{(2j-1)m_t} g(Z_s)\,\mathrm{d}s, \quad I_g(Z^{j,2}) := \int_{(2j-1)m_t}^{2jm_t} g(Z_s)\,\mathrm{d}s, \quad j = 1, \dots, n_t,$$

and, analogously, define $I_g(\widehat{Z}^{j,k})$ for $k = 1, 2$, $j \in \{1, \dots, n_t\}$. Then, for fixed $p \geq 1$, $g \in \mathcal{G}$, it holds

$$\begin{aligned}\left(\mathbb{E}\left[\left|\int_0^t g(Z_s)\,\mathrm{d}s\right|^p\right]\right)^{1/p} &\leq \left(\mathbb{E}\left[\left|\sum_{k=1}^{2}\sum_{j=1}^{n_t}(I_g(Z^{j,k}) - I_g(\widehat{Z}^{j,k}))\right|^p\right]\right)^{1/p} + \left(\mathbb{E}\left[\left|\sum_{k=1}^{2}\sum_{j=1}^{n_t} I_g(\widehat{Z}^{j,k})\right|^p\right]\right)^{1/p} \\ &\leq 2m_t\|g\|_\infty \sum_{k=1}^{2}\sum_{j=1}^{n_t}\mathbb{P}(Z^{j,k} \neq \widehat{Z}^{j,k})^{1/p} + \left(\mathbb{E}\left[\left|\sum_{k=1}^{2}\sum_{j=1}^{n_t} I_g(\widehat{Z}^{j,k})\right|^p\right]\right)^{1/p} \\ &\leq 2c_\kappa t\|g\|_\infty \mathrm{e}^{-\frac{\kappa m_t}{p}} + \sum_{k=1}^{2}\left(\mathbb{E}\left[\left|\sum_{j=1}^{n_t}(I_g(\widehat{Z}^{j,k}) - m_t\mu(g))\right|^p\right]\right)^{1/p} + t|\mu(g)|.\end{aligned}$$

Since $\widehat{Z}^{1,k}, \dots, \widehat{Z}^{n_t,k}$ are independent, the classical Bernstein inequality gives for $u > 0$

$$\mathbb{P}\left(\left|\sum_{j=1}^{n_t}(I_g(\widehat{Z}^{j,k}) - m_t\mu(g))\right| > \sqrt{2n_t\mathrm{Var}\left(\int_0^{m_t} g(Z_s)\,\mathrm{d}s\right)u} + \frac{4}{3}m_t\|g\|_\infty u\right) \leq \mathrm{e}^{-u},$$

and thus Lemma A.2 in [9] implies

$$\sum_{k=1}^{2}\left(\mathbb{E}\left[\left|\sum_{j=1}^{n_t}(I_g(\widehat{Z}^{j,k}) - m_t\mu(g))\right|^p\right]\right)^{1/p} \leq c_1' m_t\|g\|_\infty p + c_2'\sqrt{t\mathrm{Var}\left(\frac{1}{\sqrt{m_t}}\int_0^{m_t} g(Z_s)\,\mathrm{d}s\right)}\sqrt{p},$$

where $c_1' = \frac{16}{3}\mathrm{e}^{1/2\mathrm{e}}(\sqrt{2}\mathrm{e}^{1/(12p)})^{1/p}\mathrm{e}^{-1}$, $c_2' = 2(2\mathrm{e})^{-1/2}\mathrm{e}^{1/(2\mathrm{e})}(\sqrt{\pi}\mathrm{e}^{1/(6p)})^{1/p}$. The generalization to $m_t \in (0, t/4)$ is now analogous to the proof of Lemma 3.1 and Theorem 3.2 of [8]. ∎

Proof of Theorem B.13. We start by noting that, letting

$$\mathbb{I}_{t,\sigma}^j(g) := \frac{1}{\sqrt{t}}\int_0^t g(Z_s)\sum_{k=1}^{d}\sigma_{jk}(Z_s)\,\mathrm{d}W_s^k, \quad g \in \mathcal{G}, j \in \{1, \dots, d\},$$

we obtain for any $p \geq 1$

$$\left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}|\mathbb{H}_t^j(g) - \sqrt{t}\mu(gb^j)|^p\right]\right)^{1/p} \leq \left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}|\mathbb{G}_t(gb^j - \mu(gb^j))|^p\right]\right)^{1/p} + \left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}|\mathbb{I}_{t,\sigma}^j(g)|^p\right]\right)^{1/p}. \tag{B.42}$$

Theorem 3.2 of [8] then implies that there is a constant $c > 0$ such that, for any $m_t \in (0, t/4]$, there exists $\tau \in [m_t, 2m_t]$ such that, for any $p \geq 1$,

$$\left(\mathbb{E}\left[\sup_{g\in\mathcal{G}} |\mathbb{G}_t(gb^j - \mu(gb^j))|^p\right]\right)^{1/p} \leq c\Bigg(\int_0^\infty \log \mathcal{N}\big(u, \mathcal{G}b^j, \tfrac{m_t}{\sqrt{t}} d_\infty\big)\, \mathrm{d}u + \int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{G}b^j, d_{\mathbb{G},\tau})}\, \mathrm{d}u + \sup_{g\in\mathcal{G}} \left(\frac{m_t}{\sqrt{t}} \|gb^j\|_\infty p + \|gb^j\|_{\mathbb{G},\tau}\sqrt{p} + \|gb^j\|_\infty c_\kappa \sqrt{t} \mathrm{e}^{-\frac{\kappa m_t}{p}}\right)\Bigg). \tag{B.43}$$

It thus remains to bound $\left(\mathbb{E}\left[\sup_{g\in\mathcal{G}} |\mathbb{I}^j_{t,\sigma}(g)|^p\right]\right)^{1/p}$. Since, for any $g \in \mathcal{G}$, $\int_0^t g(Z_s) \sum_{k=1}^d \sigma_{jk}(Z_s)\, \mathrm{d}W^k_s$ is a continuous martingale, (B.20) yields

$$\mathbb{P}\Big(|\mathbb{I}^j_{t,\sigma}(g)| > u\Big) \leq 2\mathrm{e}^{-\frac{tu^2}{2y}} + \mathbb{P}\left(\int_0^t g^2(Z_s) a_{jj}(Z_s)\, \mathrm{d}s > y\right), \quad u, y > 0,$$

where $a = \sigma\sigma^\top$. Additionally, Lemma B.12 yields for $y > 0$ that, for any $m_{t,2} \in (0, t/4)$, there exists $\tau_2 \in [m_{t,2}, 2m_{t,2}]$ such that

$$\mathbb{P}\left(\int_0^t \Big(g^2 a_{jj}(Z_s)\Big)\, \mathrm{d}s > y + t\mu(g^2 a_{jj})\right) \leq 2\exp\left(-\frac{y^2}{32t\big(\mathrm{Var}\big(\frac{1}{\sqrt{\tau_2}}\int_0^{\tau_2} (g^2 a_{jj})(Z_s)\, \mathrm{d}s\big) + 2y\|g^2 a_{jj}\|_\infty \frac{m_{t,2}}{t}\big)}\right) + \frac{t}{m_{t,2}} c_\kappa \mathrm{e}^{-\kappa m_{t,2}} \mathbb{1}_{(0,4t\|g^2 a_{jj}\|_\infty)}(y),$$

and, letting

$$y_{u,t} := \sqrt{2\mathrm{Var}\left(\frac{1}{\sqrt{\tau_2}}\int_0^{\tau_2} (g^2 a_{jj})(Z_s)\, \mathrm{d}s\right)} + \sqrt{256u}\|g^2 a_{jj}\|_\infty \frac{m_{t,2}}{\sqrt{t}},$$

we get

$$\mathbb{P}\left(\int_0^t g^2(Z_s) a_{jj}(Z_s)\, \mathrm{d}s > \sqrt{32ut} y_{u,t} + t\mu(g^2 a_{jj})\right) \leq 2\mathrm{e}^{-u} + \frac{t}{m_{t,2}} c_\kappa \mathrm{e}^{-\kappa m_{t,2}} \mathbb{1}_{(u,\infty)}\big(\tfrac{t}{16 m_{t,2}}\big).$$

The choice $m_{t,2} = \frac{\sqrt{t}}{2\sqrt{\kappa}}$ then yields, for large enough t,

$$\mathbb{P}\left(\int_0^t g^2(Z_s) a_{jj}(Z_s)\, \mathrm{d}s > \sqrt{32ut} y_{u,t} + t\mu(g^2 a_{jj})\right) \leq 2\mathrm{e}^{-u} + 2c_\kappa\sqrt{\kappa t}\mathrm{e}^{-\frac{\sqrt{\kappa t}}{2}} \mathbb{1}_{(u,\infty)}\big(\tfrac{\sqrt{\kappa t}}{8}\big) \leq 2\mathrm{e}^{-u} + 2\mathrm{e}^{-\frac{\sqrt{\kappa t}}{8}} \mathbb{1}_{(u,\infty)}\big(\tfrac{\sqrt{\kappa t}}{8}\big) \leq 4\mathrm{e}^{-u}.$$

Hence, we have for $r > 0$

$$\mathbb{P}(|\mathbb{I}^j_{t,\sigma}(g)| > r) \leq 2\exp\left(-\frac{r^2}{\frac{16\sqrt{u}}{\sqrt{t}}\|g^2 a_{jj}\|_{\mathbb{G},\tau_2} + \sqrt{8192\kappa^{-1}u}\|g^2 a_{jj}\|_\infty \frac{1}{\sqrt{t}} + 2\mu(g^2 a_{jj})}\right) + 4\mathrm{e}^{-u},$$

and, thus, it holds for large enough t

$$\begin{aligned}
6\mathrm{e}^{-u} &\geq \mathbb{P}\left(|\mathbb{I}^j_{t,\sigma}(g)| > \frac{4(\sqrt{u}+u)}{t^{1/4}}\sqrt{\|g^2 a_{jj}\|_{\mathbb{G},\tau_2}} + u\left(\frac{8192\|a_{jj}\|^2_\infty}{\kappa t}\right)^{1/4}\|g\|_\infty + \sqrt{u2\|a_{jj}\|_\infty \mu(g^2)}\right) \\
&\geq \mathbb{P}\left(|\mathbb{I}^j_{t,\sigma}(g)| > \frac{4\sqrt{\|a_{jj}\|_\infty}(\sqrt{u}+u)}{(\kappa t)^{1/8}}\mu(g^4)^{1/4} + u\left(\frac{8192\|a_{jj}\|^2_\infty}{\kappa t}\right)^{1/4}\|g\|_\infty + \sqrt{u2\|a_{jj}\|_\infty \mu(g^2)}\right) \\
&\geq \mathbb{P}\left(|\mathbb{I}^j_{t,\sigma}(g)| > u\left(\frac{256\|a_{jj}\|^2_\infty}{\sqrt{\kappa t}}\right)^{1/4}\left(\mu(g^4)^{1/4} + \left(\frac{32}{\sqrt{\kappa t}}\right)^{1/4}\|g\|_\infty\right)\right. \\
&\qquad \left. + \sqrt{u\|a_{jj}\|_\infty}\left(\sqrt{2\mu(g^2)} + 4\left(\frac{\mu(g^4)}{\sqrt{\kappa t}}\right)^{1/4}\right)\right),
\end{aligned} \tag{B.44}$$

where we used Jensen's inequality and Fubini's theorem for showing

$$\begin{aligned}
\sqrt{\|g^2 a_{jj}\|_{\mathbb{G},\tau_2}} &\leq \left(\frac{1}{\tau_2}\mathbb{E}\left[\left(\int_0^{\tau_2}\left(g^2 a_{jj}\right)(Z_s)\,\mathrm{d}s\right)^2\right]\right)^{1/4} = \left(\tau_2\mathbb{E}\left[\left(\frac{1}{\tau_2}\int_0^{\tau_2}\left(g^2 a_{jj}\right)(Z_s)\,\mathrm{d}s\right)^2\right]\right)^{1/4} \\
&\leq \left(\mathbb{E}\left[\int_0^{\tau_2}\left(g^4 a^2_{jj}\right)(Z_s)\,\mathrm{d}s\right]\right)^{1/4} \leq \tau_2^{1/4}\|a_{jj}\|^{1/2}_\infty \mu(g^4)^{1/4} \leq \left(\frac{t}{\kappa}\right)^{1/8}\|a_{jj}\|^{1/2}_\infty \mu(g^4)^{1/4}.
\end{aligned}$$

We now want to use Theorem 3.5 in [9] which requires a bound of the form $2\exp(-u)$. However, inspection of the proof of this theorem and, in particular, the proof of Lemma A.4 of [9] used therein shows that the bound in (B.44) suffices. Thus, we have that there exist constants $c_1, c_2 > 0$ and $t_0 > 0$ such that, for any $p \geq 1$ and $t \geq t_0$,

$$\begin{aligned}
\left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}|\mathbb{I}^j_{t,\sigma}(g)|^p\right]\right)^{1/p} &\leq c_1\int_0^\infty \log \mathcal{N}\big(u,\mathcal{G},t^{-1/4}d_\infty + t^{-1/8}d_{L^4(\mu)}\big)\,\mathrm{d}u \\
&\quad + c_2\int_0^\infty \sqrt{\log \mathcal{N}(u,\mathcal{G},d_{L^2(\mu)} + t^{-1/8}d_{L^4(\mu)})}\,\mathrm{d}u \\
&\quad + 2\sup_{g\in\mathcal{G}}\left(\mathbb{E}\left[|\mathbb{I}^j_{t,\sigma}(g)|^p\right]\right)^{1/p} \\
&\leq c_1\int_0^\infty \log \mathcal{N}\big(u,\mathcal{G},t^{-1/4}d_\infty + t^{-1/8}d_{L^4(\mu)}\big)\,\mathrm{d}u \\
&\quad + c_2\int_0^\infty \sqrt{\log \mathcal{N}(u,\mathcal{G},d_{L^2(\mu)} + t^{-1/8}d_{L^4(\mu)})}\,\mathrm{d}u \\
&\quad + \frac{2}{\sqrt{t}}C_p\sup_{g\in\mathcal{G}}\left(\mathbb{E}\left[\left(\int_0^t g^2(Z_s)a_{jj}(Z_s)\,\mathrm{d}s\right)^{p/2}\right]\right)^{1/p} \\
&\leq c_1\int_0^\infty \log \mathcal{N}\big(u,\mathcal{G},t^{-1/4}d_\infty + t^{-1/8}d_{L^4(\mu)}\big)\,\mathrm{d}u \\
&\quad + c_2\int_0^\infty \sqrt{\log \mathcal{N}(u,\mathcal{G},d_{L^2(\mu)} + t^{-1/8}d_{L^4(\mu)})}\,\mathrm{d}u \\
&\quad + \frac{2}{\sqrt{t}}C_p\sup_{g\in\mathcal{G}}\left(\mathbb{E}\left[\left(\int_0^t g^2(Z_s)a_{jj}(Z_s)\,\mathrm{d}s\right)^{p}\right]\right)^{1/(2p)},
\end{aligned} \tag{B.45}$$

where we used the Burkholder–Davis–Gundy inequality (with $C_p > 0$ denoting the corresponding constant) and Hölder's inequality. Additionally, we bounded the γ_α functionals appearing in Theorem 3.5 of [9] by the corresponding entropy integrals (see Section 1.2 in [20]). Thus, we can see that c_1, c_2 can be set to

$$c_1 = 4\widetilde{C}_0\left(\frac{\|a_{jj}\|_\infty^2}{\sqrt{\kappa}}\right)^{1/4}\left(1+\left(\frac{32}{\sqrt{\kappa}}\right)^{1/4}\right), \quad c_2 = \widetilde{C}_1\sqrt{\|a_{jj}\|_\infty}\left(\sqrt{2}+4\kappa^{-1/8}\right),$$

where $\widetilde{C}_0, \widetilde{C}_1$ represent the universal constants from Theorem 3.5 in [9], adjusted to the bound in (B.44) and multiplied by the respective constants involved in bounding the γ_α functionals. Furthermore, combining Proposition 4.2 in [1] with the Hölder inequality shows that there exists a universal constant $\widetilde{C}_2 > 0$ such that $C_p \le \widetilde{C}_2\sqrt{p}$ and, thus, Lemma B.21 implies that for any $\widetilde{m}_t \in (0, t/4)$ there exists $\widetilde{\tau} \in [\widetilde{m}_t, 2\widetilde{m}_t]$ such that

$$\begin{aligned}
\frac{2}{\sqrt{t}}C_p \sup_{g\in\mathcal{G}}\left(\mathbb{E}\left[\left(\int_0^t g^2(Z_s)a_{jj}(Z_s)\,\mathrm{d}s\right)^p\right]\right)^{1/(2p)} &\le \frac{2\widetilde{C}_2\sqrt{p}}{\sqrt{t}}\sup_{g\in\mathcal{G}}\left(\mathbb{E}\left[\left(\int_0^t g^2(Z_s)a_{jj}(Z_s)\,\mathrm{d}s\right)^p\right]\right)^{1/(2p)} \\
&\le \frac{2\widetilde{C}_2\sqrt{p}}{\sqrt{t}}\sup_{g\in\mathcal{G}}\left(c\left(\widetilde{m}_t\|g^2a_{jj}\|_\infty p + \sqrt{tp}\|g^2a_{jj}\|_{\mathbb{G},\widetilde{\tau}} + t\|g^2a_{jj}\|_\infty \mathrm{e}^{-\frac{\kappa\widetilde{m}_t}{p}}\right) + t|\mu(g^2a_{jj})|\right)^{1/2} \\
&\le \sup_{g\in\mathcal{G}}\left(c\left(p\sqrt{\frac{\widetilde{m}_t\|a_{jj}\|_\infty}{t}}\|g\|_\infty + p^{3/4}(\widetilde{\tau}/t)^{1/4}\|g\|_{L^4(\mu)} + \sqrt{p\|a_{jj}\|_\infty}\|g\|_\infty \mathrm{e}^{-\frac{\kappa\widetilde{m}_t}{2p}}\right)\right. \qquad \text{(B.46)} \\
&\qquad \left. + 2\widetilde{C}_2\sqrt{p\|a_{jj}\|_\infty}\|g\|_{L^2(\mu)}\right).
\end{aligned}$$

Combining (B.42), (B.43), (B.45) and (B.46) now yields the required assertion. ∎

Proof of Proposition B.14. We start with the usual decomposition

$$\begin{aligned}
\mathcal{R}_\infty^{(p)}\big(\overline{b}_{j,h_3,h_4,T}, b^j\rho; D\big) \le \Bigg(\mathbb{E}\left[\left\|\overline{b}_{j,h_1,h_2,T} - \mu(K_{h_1,h_2}(z-\cdot)b^j)\right\|_{L^\infty(D)}^p\right]\Bigg)^{\frac{1}{p}} \\
+ \underbrace{\left\|\mu(K_{h_1,h_2}(z-\cdot)b^j) - b^j\rho\right\|_{L^\infty(D)}}_{=\mathcal{B}_{b^j\rho}(h_1,h_2)}.
\end{aligned}$$

Now denseness of $\mathbb{Q}$, the dominated convergence theorem for stochastic integrals and Theorem B.13 yield

$$\begin{aligned}
&\left(\mathbb{E}\left[\left\|\overline{b}_{j,h_1,h_2,T} - \mu(K_{h_1,h_2}(z-\cdot)b^j)\right\|_{L^\infty(D)}^p\right]\right)^{1/p} \\
&\quad = (h_1h_2)^{-d}T^{-1/2}\left(\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\mathscr{H}_t^j(K((z-\cdot)/(h_1h_2))) - \sqrt{T}\mu(K((z-\cdot)/(h_1h_2))b^j)\right|^p\right]\right)^{1/p} \\
&\quad \le c(h_1h_2)^{-d}T^{-1/2}\left(\int_0^\infty \log\mathcal{N}\big(u, \mathcal{G}b^j, \tfrac{m_T}{\sqrt{T}}d_\infty\big)\,\mathrm{d}u + \int_0^\infty \sqrt{\log\mathcal{N}(u, \mathcal{G}b^j, d_{\mathbb{G},\tau})}\,\mathrm{d}u\right. \\
&\qquad \left. + \int_0^\infty \log\mathcal{N}\big(u, \mathcal{G}, T^{-1/4}d_\infty + T^{-1/8}d_{L^4(\mu)}\big)\,\mathrm{d}u\right)
\end{aligned}$$

$$+\int_0^\infty \sqrt{\log \mathcal{N}(u,\mathcal{G},d_{L^2(\mu)}+T^{-1/8}d_{L^4(\mu)})}\,\mathrm{d}u$$
$$+\sup_{g\in\mathcal{G}}\Big(\frac{m_T}{\sqrt{T}}\|g\|_\infty p+\|g\|_{\mathbb{G},\tau}\sqrt{p}+\frac{1}{2}\|g\|_\infty\sqrt{T}\mathrm{e}^{-\frac{\kappa m_T}{p}}+p\sqrt{\frac{\widetilde{m}_T\|a_{jj}\|_\infty}{T}}\|g\|_\infty$$
$$+p^{3/4}(\widetilde{\tau}/T)^{1/4}\|g\|_{L^4(\mu)}+\sqrt{p\|a_{jj}\|_\infty}\|g\|_\infty\mathrm{e}^{-\frac{\kappa\widetilde{m}_T}{2p}}+\sqrt{p\|a_{jj}\|_\infty}\|g\|_{L^2(\mu)}\Big)\Big).$$

We continue by bounding the entropy integrals. Elementary calculations and Lemma B.19 yield

$$\begin{aligned}\int_0^\infty \log\mathcal{N}\big(u,\mathcal{G}b^j,\tfrac{m_T}{\sqrt{T}}d_\infty\big)\,\mathrm{d}u&=\frac{m_T}{\sqrt{T}}\int_0^{2\sup_{x\in\mathcal{K}}|b^j(x)|\|K\|_\infty}\log\mathcal{N}\big(u,\mathcal{G}b^j,d_\infty\big)\,\mathrm{d}u\\&\le c\frac{m_T}{\sqrt{T}}\int_0^{2\sup_{x\in\mathcal{K}}|b^j(x)|\|K\|_\infty}\log\Big(\frac{c(h_1^{-1}+h_2^{-1})}{u}\Big)\,\mathrm{d}u\\&\le c\frac{m_T}{\sqrt{T}}\Big(1+\log\Big(h_1^{-1}+h_2^{-1}\Big)\Big).\end{aligned}$$

Analogously, we get

$$\begin{aligned}&\int_0^\infty \log\mathcal{N}\big(u,\mathcal{G},T^{-1/4}d_\infty+T^{-1/8}d_{L^4(\mu)}\big)\,\mathrm{d}u\\&\quad\le\int_0^\infty\log\mathcal{N}\big(u,\mathcal{G},T^{-1/8}d_{L^4(\mu)}\big)\,\mathrm{d}u+\int_0^\infty\log\mathcal{N}\big(u,\mathcal{G},T^{-1/4}d_\infty\big)\,\mathrm{d}u\\&\quad\le\int_0^\infty\log\mathcal{N}\big(T^{1/8}(ch_1h_2)^{-d/4}u,\mathcal{G},d_\infty\big)\,\mathrm{d}u+cT^{-1/4}\Big(1+\log\Big(h_1^{-1}+h_2^{-1}\Big)\Big)\\&\quad\le c\Big(1+\log\Big(h_1^{-1}+h_2^{-1}\Big)\Big)\Big(T^{-1/4}+T^{-1/8}(h_1h_2)^{d/4}\Big).\end{aligned}$$

Furthermore, Proposition B.2 yields for $f,g\in\mathcal{G}$ and large enough T

$$d_{\mathbb{G},t}(fb^j,gb^j)\le c(h_1h_2)^d\psi_d(h_1,h_2)=:\mathbb{V},$$

and, hence, using (B.39) we get by Lemma B.19 for large enough T

$$\begin{aligned}\int_0^\infty\sqrt{\log\mathcal{N}(u,\mathcal{G}b^j,d_{\mathbb{G},\tau})}\,\mathrm{d}u&\le\int_0^{\mathbb{V}}\sqrt{\log\mathcal{N}(u\psi_d(h_1,h_2)^{-1},\mathcal{G}b^j,d_\infty)}\,\mathrm{d}u\\&\le\int_0^{\mathbb{V}}\sqrt{\log\Big(\frac{c(h_1^{-1}+h_2^{-1})(h_3h_4)^d\psi_d(h_1,h_2)}{u}\Big)}\,\mathrm{d}u\\&\le c(h_1h_2)^d\psi_d(h_1,h_2)\sqrt{\log\big(h_1^{-1}+h_2^{-1}\big)}.\end{aligned}$$

For the remaining integral, we argue similarly and get for large enough T

$$\begin{aligned}&\int_0^\infty\sqrt{\log\mathcal{N}(u,\mathcal{G},d_{L^2(\mu)}+T^{-1/8}d_{L^4(\mu)})}\,\mathrm{d}u\\&\le\int_0^{c(h_1h_2)^{d/2}\|K\|_\infty}\sqrt{\log\mathcal{N}(c(h_1h_2)^{-d/2}u,\mathcal{G},d_\infty)}\,\mathrm{d}u\end{aligned}$$

$$
\begin{aligned}
&\quad + \int_0^{c(h_1h_2)^{d/4}T^{-1/8}\|K\|_\infty} \sqrt{\log \mathcal{N}(c(h_1h_2)^{-d/4}T^{1/8}u, \mathcal{G}, d_\infty)}\,\mathrm{d}u \\
&\leq \int_0^{c(h_1h_2)^{d/2}\|K\|_\infty} \sqrt{\log\left(\frac{c(h_1^{-1}+h_2^{-1})(h_1h_2)^{d/2}}{u}\right)}\,\mathrm{d}u \\
&\quad + \int_0^{c(h_1h_2)^{d/4}T^{-1/8}\|K\|_\infty} \sqrt{\log\left(\frac{c(h_1^{-1}+h_2^{-1})(h_1h_2)^{d/4}}{T^{1/8}u}\right)}\,\mathrm{d}u \\
&\leq c\sqrt{\log\big(h_1^{-1}+h_2^{-1}\big)}\Big((h_1h_2)^{d/2} + (h_1h_2)^{d/4}T^{-1/8}\Big).
\end{aligned}
$$

Combining everything above and choosing $m_T = \widetilde{m}_T = \frac{p}{\kappa}\log T$, we obtain for T large enough and $p \leq \gamma \log T$, with $\gamma > 0$,

$$
\begin{aligned}
&\left(\mathbb{E}\left[\left\|\overline{b}_{j,h_1,h_2,T} - \mu(K_{h_1,h_2}(z-\cdot)b^j)\right\|^p_{L^\infty(D)}\right]\right)^{1/p} \\
&\quad \leq c(h_1h_2)^{-d}T^{-1/2}\Big(\frac{p\log\big(h_1^{-1}+h_2^{-1}\big)\log T}{\sqrt{T}} + (h_1h_2)^d\psi_d(h_1,h_2)\sqrt{\log\big(h_1^{-1}+h_2^{-1}\big)} \\
&\qquad + T^{-1/8}\Big(T^{-1/8} + (h_1h_2)^{d/4}\Big)\log\Big(h_1^{-1}+h_2^{-1}\Big) + (h_1h_2)^{d/2}\sqrt{\log\big(h_1^{-1}+h_2^{-1}\big)} \\
&\qquad + \frac{p^2\log T}{\sqrt{T}} + \sqrt{p}(h_1h_2)^d\psi_d(h_1,h_2) + T^{-1/2} + \sqrt{\frac{p^3\log T}{T}} \\
&\qquad + p\left(\frac{\log T}{T}\right)^{1/4}(h_1h_2)^{d/4} + \sqrt{p}T^{-1/2} + \sqrt{p}(h_1h_2)^{d/2}\Big) \\
&\quad \leq c(h_1h_2)^{-d}T^{-1/2}\Big(\frac{\log T^3}{\sqrt{T}} + (h_1h_2)^{d/2}\sqrt{\log\big(h_1^{-1}+h_2^{-1}\big)} \\
&\qquad + T^{-1/8}\Big(T^{-1/8} + (h_1h_2)^{d/4}\Big)\log\Big(h_1^{-1}+h_2^{-1}\Big) + (h_1h_2)^{d/2}\sqrt{\log\big(h_1^{-1}+h_2^{-1}\big)} \\
&\qquad + \frac{\log T^3}{\sqrt{T}} + \sqrt{\log T}(h_1h_2)^d\psi_d(h_1,h_2) + T^{-1/2} + \sqrt{\frac{\log T^4}{T}} \\
&\qquad + \log T\left(\frac{\log T}{T}\right)^{1/4}(h_1h_2)^{d/4} + \sqrt{\frac{\log T}{T}} + \sqrt{\log T}(h_1h_2)^{d/2}\Big) \\
&\quad \leq c_\gamma(h_1h_2)^{-d/2}T^{-1/2}\sqrt{\log\big(h_1^{-1}+h_2^{-1}\big)},
\end{aligned}
\tag{B.47}
$$

where we used that $h_1, h_2 \in \mathcal{H}$ and $(h_1h_2)^d \geq T^{-1/2}\log(h_1^{-1}+h_2^{-1})$, and where the constant c_γ depends on γ. ∎

Proof of Theorem B.15. Introduce the set $B_T := \{\|\widehat{\rho}_{h_1,h_2,T}(z) - \rho(z)\|_{L^\infty(D)} \leq r_T\}$. Markov's inequality and Theorem B.7 then imply, for large enough T and some constant c which is independent of $1 \leq p \leq c_p\sqrt{\log T}$,

$$
\mathbb{P}(B_T^{\mathrm{c}}) \leq c(\Psi\chi_{\mathcal{B}})^p(\beta_1,\beta_2,d,0)r_T^{-p} = c\exp\Big(-p\sqrt{\log T}\Big).
$$

Note furthermore that $\bar{\beta} > d$ implies $(h_1 h_2)^d \geq T^{-1/2} \log(h_1^{-1} + h_2^{-1})$, for large enough T. Thus, for large enough T, it holds on the event B_T^{c}

$$\begin{aligned}
&\mathbb{E}\left[\sup_{z\in D} |(\widehat{b}_{j,\boldsymbol{h},T,r_T}(z) - b(z))\rho(z)| \mathbf{1}_{B_T^{\mathrm{c}}}\right] \\
&\quad \leq \mathbb{E}\left[\sup_{z\in D} |\widehat{b}_{j,\boldsymbol{h},T,r_T}(z)\rho(z)| \mathbf{1}_{B_T^{\mathrm{c}}}\right] + \mathbb{E}\left[\sup_{z\in D} |b(z)\rho(z)| \mathbf{1}_{B_T^{\mathrm{c}}}\right] \\
&\quad \leq \mathbb{E}\left[\sup_{z\in D} |\widehat{b}_{j,\boldsymbol{h},T,r_T}(z)\rho(z)|^2\right]^{\frac{1}{2}} \mathbb{P}(B_T^{\mathrm{c}})^{\frac{1}{2}} + c\mathbb{P}(B_T^{\mathrm{c}}) \\
&\quad \leq c\left(r_T^{-1}\left(\left(\frac{\log T}{T}\right)^{\frac{\bar{\beta}}{2(\bar{\beta}+d)}} + 1\right)\exp\left(-(p/2)\sqrt{\log T}\right) + \exp\left(-p\sqrt{\log T}\right)\right) \\
&\quad \leq c\left(\sqrt{T}\exp\left(-(p/2+1)\sqrt{\log T}\right) + \exp\left(-p\sqrt{\log T}\right)\right),
\end{aligned}$$

where we used equation (B.47) and the Minkowski inequality in the second to last line. Choosing $p = 5\sqrt{\log T}$ now gives

$$\mathbb{E}\left[\sup_{z\in D} |(\widehat{b}_{j,\boldsymbol{h},T,r_T}(z) - b(z))\rho(z)| \mathbf{1}_{B_T^{\mathrm{c}}}\right] \in \mathrm{O}(T^{-2}) \subset \mathrm{O}\left(\left(\frac{\log T}{T}\right)^{\frac{\bar{\beta}}{2(\bar{\beta}+d)}}\right).$$

On the other hand, on the event B_T it holds $\rho/(\widehat{\rho}_{h_1^{(\rho)},h_2^{(\rho)},T} + r_T) \leq 1$. Thus, by Theorem B.7 and Proposition B.14,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{z\in D} |(\widehat{b}_{j,\boldsymbol{h},T,r_T}(z) - b(z))\rho(z)| \mathbf{1}_{B_T}\right] \\
&\quad \leq \mathbb{E}\left[\sup_{z\in D} |\widehat{b}_{j,\boldsymbol{h},T,r_T}(z) - \tfrac{b(z)\rho(z)}{\widehat{\rho}_{h_1^{(\rho)},h_2^{(\rho)},T}(z)+r_T}))\rho(z)| \mathbf{1}_{B_T}\right] + \mathbb{E}\left[\sup_{z\in D} |(\tfrac{b(z)\rho(z)}{\widehat{\rho}_{h_1^{(\rho)},h_2^{(\rho)},T}+r_T} - b(z))\rho(z)| \mathbf{1}_{B_T}\right] \\
&\quad \leq c\left(\left(\frac{\log T}{T}\right)^{\frac{\bar{\beta}}{2(\bar{\beta}+d)}} + \mathbb{E}[\sup_{z\in D} |\rho(z) - \widehat{\rho}_{h_1^{(\rho)},h_2^{(\rho)},T} - r_T| \mathbf{1}_{B_T}]\right) \\
&\quad \leq c\left(\left(\frac{\log T}{T}\right)^{\frac{\bar{\beta}}{2(\bar{\beta}+d)}} + r_T + (\Psi\chi_{\mathcal{B}})(T,\beta_1,\beta_2,d,0)\right),
\end{aligned}$$

where we used the bias bound (B.41). The assertion now follows since $\Upsilon > 0$ (recall (B.11), (B.12)) implies

$$(\Psi\chi_{\mathcal{B}})(T,\beta_1,\beta_2,d,0) + r_t \in \mathrm{O}\left(\left(\frac{\log T}{T}\right)^{\frac{\bar{\beta}}{2(\bar{\beta}+d)}}\right).$$

■

Proof of Proposition B.16. Fix $j \in \{1,\dots,d\}$. In what follows, the dependencies on j and q will be regularly suppressed in the notation. We start with stating an important auxiliary result.

LEMMA B.22. *Let* $\mathcal{G}_{\boldsymbol{h}} := \{K_1((x-\cdot)/h_1)K_2((y-\cdot)/h_2) : (x,y) \in D \cap \mathbb{Q}^{2d}\}$, $\boldsymbol{h} = (h_1,h_2) \in \mathcal{H}_t$, *and recall the definition of* $\mathbb{H}_t^j$ *(see* (B.19)*). Then, for any* $\gamma > 0$ *and large enough* t*, it holds*

$$\forall u_t \in [1, \gamma \log(t)], \quad \mathbb{P}\left(\sup_{g\in\mathcal{G}_{\boldsymbol{h}}} |\mathcal{H}_t^j(g) - \sqrt{t}\mu(gb^j)| > \Delta_{\boldsymbol{h},t}(u_t)\right) \le \mathrm{e}^{-u_t},$$

where

$$\Delta_{\boldsymbol{h},t}(u) := 4\mathrm{e}\sqrt{\|\rho\|_\infty \|a_{jj}\|_\infty (h_1h_2)^d}\left(\widetilde{C}_1\sqrt{384d\log(h_1^{-1}+h_2^{-1})}\|K\|_\infty + \widetilde{C}_2\|K\|_{L^2(\lambda)}u^{1/2}\right). \quad \text{(B.48)}$$

Proof. To prove the assertion, we want to combine Markov's inequality with the uniform moment bounds derived in Theorem B.13. Choosing $p = p(t) = u_t \le \gamma \log t$ for fixed $\gamma > 0$, we get as in the derivation of equation (B.47) that there exist $c_1, c_2 > 0$ such that, for large enough t,

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}_{\boldsymbol{h}}} |\mathcal{H}_t^j(g) - \sqrt{t}\mu(gb^j)|^p\right]^{1/p} \le 2(h_1h_2)^{d/2}\left(c_1\sqrt{\log(h_1^{-1}+h_2^{-1})} + c_2\sqrt{u_t}\right),$$

where the constants c_1, c_2 need to satisfy

$$\widetilde{C}_1\int_0^\infty \sqrt{6\|a_{jj}\|_\infty \log \mathcal{N}(u, \mathcal{G}_{\boldsymbol{h}}, d_{L^2(\mu)})}\,\mathrm{d}u \le c_1(h_1h_2)^{d/2}\sqrt{\log(h_1^{-1}+h_2^{-1})},$$
$$2\widetilde{C}_2\sqrt{\|a_{jj}\|_\infty}\|g\|_{L^2(\mu)} \le c_2(h_1h_2)^{d/2},$$

for large enough t. Here, $\widetilde{C}_1, \widetilde{C}_2 > 0$ correspond to the constants obtained in the proof of Theorem B.13. Now

$$\|g\|_{L^2(\mu)} \le (h_1h_2)^{d/2}\|\rho\|_\infty^{1/2}\|K\|_{L^2(\lambda)}$$

shows that $c_2 = 2\widetilde{C}_2\sqrt{\|a_{jj}\|_\infty\|\rho\|_\infty}\|K\|_{L^2(\lambda)}$ is an adequate choice. Additionally, straightforward computations using (B.39) and Lemma B.19 show that $c_1 = \widetilde{C}_1\sqrt{1536d\|a_{jj}\|_\infty\|\rho\|_\infty}\|K\|_\infty$ also satisfies the given requirement. Hence, defining $\Delta_{\boldsymbol{h},t}$ as in (B.48) implies the assertion through Markov's inequality. ∎

For any $\boldsymbol{h} = (h_1,h_2)^\top, \boldsymbol{\eta} = (\eta_1,\eta_2)^\top \in (0,1]^2$, set

$$s_{\boldsymbol{h}}(\cdot,\cdot) = s_{h_1,h_2}(\cdot,\cdot) := \int_{\mathbb{R}^{2d}} K_{h_1,h_2}(u-\cdot, v-\cdot)(b^j\rho)(u,v)\,\mathrm{d}u\,\mathrm{d}v,$$
$$s^\star_{\boldsymbol{h},\boldsymbol{\eta}}(\cdot,\cdot) = s^\star_{h_1,h_2,\eta_1,\eta_2}(\cdot,\cdot) := \int_{\mathbb{R}^{2d}} \left(K_{h_1,h_2} \star K_{\eta_1,\eta_2}\right)(u-\cdot, v-\cdot)(b^j\rho)(u,v)\,\mathrm{d}u\,\mathrm{d}v.$$

For any kernel estimator

$$\overline{b}_{\boldsymbol{h}}(x,y) = \overline{b}_{j,\boldsymbol{h}}(x,y) \equiv \overline{b}_{j,h_1,h_2,t}(x,y) = \frac{1}{t}\int_0^t K_{h_1,h_2}(x-X_u, y-Y_u)\,\mathrm{d}Y_u^j$$

of $b^j\rho$, denote its stochastic error by $\xi_{\boldsymbol{h}}(\cdot,\cdot) := \overline{b}_{\boldsymbol{h}}(\cdot,\cdot) - s_{\boldsymbol{h}}(\cdot,\cdot)$, and set

$$\zeta_t := \sup_{(\eta_1,\eta_2)\in\mathcal{H}_t}\left\{\left[\|\xi_{\eta_1,\eta_2}\|_\infty - A_t(\eta_1,\eta_2)\right]_+\right\},$$

where $A_t(\cdot,\cdot)$ is defined as in (B.22). The triangle inequality implies that, for any $\boldsymbol{h} \in \mathcal{H}_t$,

$$\|\overline{b}_{\widehat{\boldsymbol{h}}} - b^j\rho\|_\infty \le \|\overline{b}_{\widehat{\boldsymbol{h}}} - \overline{b}_{\boldsymbol{h},\widehat{\boldsymbol{h}}}\|_\infty + \|\overline{b}_{\boldsymbol{h},\widehat{\boldsymbol{h}}} - \overline{b}_{\boldsymbol{h}}\|_\infty + \|\overline{b}_{\boldsymbol{h}} - b^j\rho\|_\infty.$$

Since $\widehat{\boldsymbol{h}} \in \mathcal{H}_t$, we have

$$\|\overline{b}_{\widehat{\boldsymbol{h}}} - \overline{b}_{\boldsymbol{h},\widehat{\boldsymbol{h}}}\|_\infty \le \sup_{\boldsymbol{\eta}\in\mathcal{H}_t}\left\{\left[\|\overline{b}_{\boldsymbol{\eta}} - \overline{b}_{\boldsymbol{h},\boldsymbol{\eta}}\|_\infty - A_t(\boldsymbol{\eta})\right]_+\right\} + A_t(\widehat{\boldsymbol{h}}) = \widehat{\Delta}_t(\boldsymbol{h}) + A_t(\widehat{\boldsymbol{h}}),$$

and, since $\overline{b}_{\boldsymbol{h},\widehat{\boldsymbol{h}}} = \overline{b}_{\widehat{\boldsymbol{h}},\boldsymbol{h}}$,

$$\begin{aligned}\|\overline{b}_{\widehat{\boldsymbol{h}}} - b^j\rho\|_\infty &\le \widehat{\Delta}_t(\boldsymbol{h}) + A_t(\widehat{\boldsymbol{h}}) + \widehat{\Delta}_t(\widehat{\boldsymbol{h}}) + A_t(\boldsymbol{h}) + \|\overline{b}_{\boldsymbol{h}} - b^j\rho\|_\infty \\ &\le 2\Big(\widehat{\Delta}_t(\boldsymbol{h}) + A_t(\boldsymbol{h})\Big) + \|\overline{b}_{\boldsymbol{h}} - b^j\rho\|_\infty.\end{aligned}$$

In view of

$$\|\overline{b}_{\boldsymbol{h}} - b^j\rho\|_\infty \le \|\xi_{\boldsymbol{h}}\|_\infty + \mathcal{B}_{b^j\rho}(\boldsymbol{h}) \le \zeta_t + \mathcal{B}_{b^j\rho}(\boldsymbol{h}) + A_t(\boldsymbol{h}),$$

it remains to bound $\widehat{\Delta}_t(\boldsymbol{h}) + A_t(\boldsymbol{h})$. For doing so, note first that, for any $\boldsymbol{h}, \boldsymbol{\eta} \in (0,1]^2$,

$$\begin{aligned}\|\overline{b}_{\boldsymbol{h},\boldsymbol{\eta}} - s^*_{\boldsymbol{h},\boldsymbol{\eta}}\|_\infty &= \sup_{(x,y)\in\mathbb{R}^{2d}}\left|\int_{\mathbb{R}^{2d}} K_{\eta_1,\eta_2}(x-u, y-v)\xi_{h_1,h_2}(u,v)\,\mathrm{d}u\,\mathrm{d}v\right| \le \mathrm{k}_1\,\|\xi_{\boldsymbol{h}}\|_\infty, \\ \|s^\star_{\boldsymbol{h},\boldsymbol{\eta}} - s_{\boldsymbol{\eta}}\|_\infty &\le \mathrm{k}_1\,\mathcal{B}_{b^j\rho}(\boldsymbol{h}).\end{aligned}$$

Thus,

$$\begin{aligned}\|\overline{b}_{\boldsymbol{h},\boldsymbol{\eta}} - \overline{b}_{\boldsymbol{\eta}}\|_\infty &\le \|\overline{b}_{\boldsymbol{h},\boldsymbol{\eta}} - s^\star_{\boldsymbol{h},\boldsymbol{\eta}}\|_\infty + \|s^\star_{\boldsymbol{h},\boldsymbol{\eta}} - s_{\boldsymbol{\eta}}\|_\infty + \|s_{\boldsymbol{\eta}} - \overline{b}_{\boldsymbol{\eta}}\|_\infty \\ &\le \mathrm{k}_1\Big(\|\xi_{\boldsymbol{h}}\|_\infty + \mathcal{B}_{b^j\rho}(\boldsymbol{h})\Big) + \|\xi_{\boldsymbol{\eta}}\|_\infty \;\le\; \mathrm{k}_1\Big(\zeta_t + A_t(\boldsymbol{h}) + \mathcal{B}_{b^j\rho}(\boldsymbol{h})\Big) + \zeta_t + A_t(\boldsymbol{\eta}),\end{aligned}$$

and

$$\begin{aligned}\widehat{\Delta}_t(\boldsymbol{h}) = \sup_{\boldsymbol{\eta}\in\mathcal{H}_t}\left\{\left[\|\overline{b}_{\boldsymbol{h},\boldsymbol{\eta}} - \overline{b}_{\boldsymbol{\eta}}\|_\infty - A_t(\boldsymbol{\eta})\right]_+\right\} &\le \sup_{\boldsymbol{\eta}\in\mathcal{H}_t}\left\{\mathrm{k}_1\Big(\zeta_t + A_t(\boldsymbol{h}) + \mathcal{B}_{b^j\rho}(\boldsymbol{h})\Big) + \zeta_t\right\} \\ &\le\; (1\vee\mathrm{k}_1)\Big(2\zeta_t + A_t(\boldsymbol{h}) + \mathcal{B}_{b^j\rho}(\boldsymbol{h})\Big),\end{aligned}$$

giving

$$\widehat{\Delta}_t(\boldsymbol{h}) + A_t(\boldsymbol{h}) \le (1\vee\mathrm{k}_1)\Big(2\zeta_t + 2A_t(\boldsymbol{h}) + \mathcal{B}_{b^j\rho}(\boldsymbol{h})\Big).$$

Consequently,

$$\begin{aligned}\|\overline{b}_{\widehat{\boldsymbol{h}}} - b^j\rho\|_\infty &\le 2\Big(\widehat{\Delta}_t(\boldsymbol{h}) + A_t(\boldsymbol{h})\Big) + \|\overline{b}_{\boldsymbol{h}} - b^j\rho\|_\infty \\ &\le 2(1\vee\mathrm{k}_1)\Big(2\zeta_t + 2A_t(\boldsymbol{h}) + \mathcal{B}_{b^j\rho}(\boldsymbol{h})\Big) + \zeta_t + \mathcal{B}_{b^j\rho}(\boldsymbol{h}) + A_t(\boldsymbol{h}) \\ &\le (1\vee\mathrm{k}_1)\Big(5\zeta_t + 5A_t(\boldsymbol{h}) + 3\mathcal{B}_{b^j\rho}(\boldsymbol{h})\Big),\end{aligned}$$

and, for any $\boldsymbol{h} \in \mathcal{H}_t$,

$$\mathbb{E}\Big[\|\overline{b}_{\widehat{\boldsymbol{h}}} - b^j\rho\|^p_\infty\Big]^{1/p} \le (1\vee\mathrm{k}_1)\Big(5A_t(\boldsymbol{h}) + 3\mathcal{B}_{b^j\rho}(\boldsymbol{h})\Big) + (1\vee\mathrm{k}_1)5\big(\mathbb{E}[\zeta_t^p]\big)^{1/p}.$$

It remains to bound $\mathbb{E}[\zeta_t^p]$. We start by writing

$$\mathbb{E}[\zeta_t^q] = \mathbb{E}\left[\sup_{(\eta_1,\eta_2)\in\mathcal{H}_t}\left\{\left[\|\xi_{\eta_1,\eta_2}\|_\infty - A_t(\eta_1,\eta_2)\right]_+^q\right\}\right] \le \sum_{\boldsymbol{\eta}\in\mathcal{H}_t}\mathbb{E}\left[\left[\|\xi_{\boldsymbol{\eta}}\|_\infty - A_t(\boldsymbol{\eta})\right]_+^q\right].$$

Now Lemma B.22 implies for large enough t

$$\mathbb{P}(\|\xi_{\boldsymbol{\eta}}\|_\infty > A_t(\boldsymbol{\eta})) = \mathbb{P}\left(\sup_{g\in\mathcal{G}_{\boldsymbol{\eta}}}|\mathscr{H}_t^j(g) - \sqrt{t}\mu(gb^j)| \ge \Delta_{\boldsymbol{\eta},t}(2dq\log(\eta_1^{-1}+\eta_2^{-1}))\right) \le \left(\frac{1}{\eta_1^{-1}+\eta_2^{-1}}\right)^{2dq}.$$

Then, Hölder's inequality and equation (B.47) entail for large enough t

$$\begin{aligned}\mathbb{E}\left[\left[\|\xi_{\boldsymbol{\eta}}\|_\infty - A_t(\boldsymbol{\eta})\right]_+^q\right] &\le \mathbb{E}\left[(\|\xi_{\boldsymbol{\eta}}\|_\infty - A_t(\boldsymbol{\eta}))^{2q}\right]^{1/2}\mathbb{P}(\|\xi_{\boldsymbol{\eta}}\|_\infty \ge A_t(\boldsymbol{\eta}))^{1/2}\\ &\le c(\eta_1\eta_2)^{-dq/2}t^{-q/2}\log((\eta_1^{-1}+\eta_2^{-1}))^{q/2}\left(\frac{1}{\eta_1^{-1}+\eta_2^{-1}}\right)^{dq}\\ &= c\left(\sqrt{\frac{\eta_1}{\eta_2}}+\sqrt{\frac{\eta_2}{\eta_1}}\right)^{-dq}t^{-q/2}\log((\eta_1^{-1}+\eta_2^{-1}))^{q/2} \in \mathrm{O}(((\log t)t^{-1})^{q/2}).\end{aligned}$$

Finally, $\mathbb{E}[\zeta_t^p]^{1/p} \le \mathbb{E}[\zeta_t^q]^{1/q} \lesssim \mathrm{card}(\mathcal{H}_t)^{1/q}\left(\frac{\log t}{t}\right)^{1/2} \lesssim (\log t)^{2/q+1/2}t^{-1/2}$. ∎

Proof of Theorem B.17. Fix $j \in \{1,\dots,d\}$. For the proof of (B.24), note first that, for any $\boldsymbol{h} \in \overline{\mathcal{H}_t}$,

$$\begin{aligned}|\widehat{b}_{j,\boldsymbol{h},t} - b^j| &= \left|\frac{\left(\overline{b}_{j,\boldsymbol{h},t} - b^j\rho\right) + b^j\left(\rho - \widehat{\rho}_{\boldsymbol{h},t}\vee\rho_\star\right)}{\widehat{\rho}_{\boldsymbol{h},t}\vee\rho_\star}\right|\\ &\le \frac{1\vee|b^j|}{\rho_\star}\left(|\overline{b}_{j,\boldsymbol{h},t} - b^j\rho| + \left|\rho - \frac{1}{2}\left(\widehat{\rho}_{\boldsymbol{h},t} + \rho_\star + |\widehat{\rho}_{\boldsymbol{h},t} - \rho_\star|\right)\right|\right)\\ &\le \frac{1\vee|b^j|}{\rho_\star}\left(|\overline{b}_{j,\boldsymbol{h},t} - b^j\rho| + \frac{1}{2}\left(|\rho - \widehat{\rho}_{\boldsymbol{h},t}| + |\rho - \rho_\star - |\widehat{\rho}_{\boldsymbol{h},t} - \rho_\star|\|\right)\right)\\ &= \frac{1\vee|b^j|}{\rho_\star}\left(|\overline{b}_{j,\boldsymbol{h},t} - b^j\rho| + \frac{1}{2}\left(|\rho - \widehat{\rho}_{\boldsymbol{h},t}| + \|\rho - \rho_\star| - |\widehat{\rho}_{\boldsymbol{h},t} - \rho_\star|\|\right)\right)\\ &\le \frac{1\vee|b^j|}{\rho_\star}\left(|\overline{b}_{j,\boldsymbol{h},t} - b^j\rho| + |\rho - \widehat{\rho}_{\boldsymbol{h},t}|\right).\end{aligned}$$

Thus,

$$\begin{aligned}&\mathbb{E}\left[\|\widehat{b}_{j,\boldsymbol{h},t} - b^j\|_{L^\infty(D)}^p\right]^{1/p}\\ &\le \frac{1\vee\sup_{(x,y)\in D}|b^j(x,y)|}{\rho_\star}\left(\mathbb{E}\left[\|\overline{b}_{j,\boldsymbol{h},t} - b^j\rho\|_{L^\infty(D)}^p\right]^{1/p} + \mathbb{E}\left[\|\widehat{\rho}_{\boldsymbol{h},t} - \rho\|_{L^\infty(D)}^p\right]^{1/p}\right).\end{aligned}$$

Letting $\Phi_{d,\beta}(t) := (\log t/t)^{\frac{\overline{\beta}}{2(\overline{\beta}+d)}}$, it remains to verify $\mathcal{R}_\infty^{(p)}(\overline{b}_{j,\widehat{\boldsymbol{h}},t}, b^j\rho; D) \vee \mathcal{R}_\infty^{(p)}(\widehat{\rho}_{\widehat{\boldsymbol{h}},t}, \rho; D) \in \mathrm{O}(\Phi_{d,\beta}(t))$. The first term is bounded by means of Proposition B.16. The smoothness assumption

on $b^j\rho$ implies that there exists some positive constant $\mathfrak{c}$, depending only on $\mathfrak{b}$, K and d such that $\mathcal{B}_{b^j\rho}(\boldsymbol{h}) \le \mathfrak{c}(\mathcal{L}_1 h_1^{\beta_1} + \mathcal{L}_2 h_2^{\beta_2})$. The bandwidth $\widehat{\boldsymbol{h}} = (h_1, h_2)^\top$ is then chosen by solving

$$\mathcal{L}_j \widehat{h}_j^{\beta_j} = \left(\widehat{h}_1\widehat{h}_2\right)^{-\frac{d}{2}} \sqrt{\frac{\log\left(\widehat{h}_1^{-1} + \widehat{h}_2^{-1}\right)}{t}} \quad \text{such that} \quad \widehat{h}_j \sim \left(\frac{\log t}{t}\right)^{\frac{\overline{\beta}}{2\beta_j(\overline{\beta}+d)}}, \quad j = 1, 2.$$

The obtained solution belongs to $\overline{\mathcal{H}}_t$, and plugging the specified bandwidths into the rhs of (B.23), we obtain $\mathcal{R}_\infty^{(p)}(\overline{b}_{j,\widehat{\boldsymbol{h}},t}, b^j\rho; D) \in \mathsf{O}(\Phi_{d,\beta}(t))$. Furthermore, since $\overline{\mathcal{H}}_t \subset \mathcal{H}(Q_1, Q_2)$, it follows from Proposition B.4 that

$$\begin{aligned}\mathcal{R}_\infty^{(p)}(\widehat{\rho}_{\widehat{\boldsymbol{h}},t}, \rho; D) &\lesssim \widehat{h}_1^{\beta_1} + \widehat{h}_2^{\beta_2} + \frac{\log(t)^2}{T(\widehat{h}_1\widehat{h}_2)^d} + \psi_d(\widehat{h}_1, \widehat{h}_2, t)\sqrt{\frac{\log t}{t}} \\ &\lesssim \left(\frac{\log t}{t}\right)^{\frac{\overline{\beta}}{2(\overline{\beta}+d)}} + \frac{\log(t)^2}{\sqrt{T}},\end{aligned}$$

where we used $\widehat{h}_1\widehat{h}_2 \ge t^{-1/(2d)}$ and $\psi_d(\widehat{h}_1, \widehat{h}_2) \le \psi_{2,d}(\widehat{h}_1, \widehat{h}_2) \le (\widehat{h}_1\widehat{h}_2)^{-d/2}$. ∎

References

[1] M. Barlow and M Yor. “Semi-martingale inequalities via the Garsia–Rodemich–Rumsey lemma, and applications to local times”. In: *J. Funct. Anal.* 49.2 (1982), pp. 198 –229.

[2] P. Cattiaux, J. León, and C. Prieur. “Estimation for Stochastic Damping Hamiltonian Systems under Partial Observation. II. Drift term”. In: *ALEA Lat. Am. J. Probab. Math. Stat.* 11 (July 2014).

[3] P. Cattiaux, J. R. León, and C. Prieur. “Estimation for stochastic damping Hamiltonian systems under partial observation. I. Invariant density”. In: *Stochastic Process. Appl.* 124.3 (2014), pp. 1236–1260.

[4] F. Comte and C. Lacour. “Anisotropic adaptive kernel deconvolution”. en. In: *Ann. Inst. H. Poincaré Probab. Statist.* 49.2 (2013), pp. 569–609.

[5] F. Comte, C. Prieur, and A. Samson. “Adaptive estimation for stochastic damping Hamiltonian systems under partial observation”. In: *Stochastic Process. Appl.* 127.11 (2017), pp. 3689 –3718.

[6] A. Dalalyan and M. Reiß. “Asymptotic statistical equivalence for ergodic diffusions: the multidimensional case”. In: *Probab. Theory Relat. Fields* 137.1 (2007), pp. 25–47.

[7] S. Delattre, A. Gloter, and N. Yoshida. *Rate of Estimation for the Stationary Distribution of Stochastic Damping Hamiltonian Systems with Continuous Observations*. Preprint. Jan. 28, 2020. arXiv: 2001.10423 [math.ST].

[8] N. Dexheimer, C. Strauch, and L. Trottner. *Mixing it up: A general framework for Markovian statistics (v1)*. Preprint. Oct. 31, 2020. arXiv: 2011.00308 [math.ST].

[9] S. Dirksen. “Tail bounds via generic chaining”. In: *Electron. J. Probab.* 20 (2015), no. 53, 29.

[10] S. Ditlevsen and A. Samson. “Hypoelliptic diffusions: filtering and inference from complete and partial observations”. In: *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 81.2 (2019), pp. 361–384.

[11] D. Down, S. P. Meyn, and R. L. Tweedie. “Exponential and uniform ergodicity of Markov processes”. In: *Ann. Probab.* 23.4 (1995), pp. 1671–1691.

[12] E. Giné and R. Nickl. “An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation”. In: *Probab. Theory Relat. Fields* 143.3-4 (2009), pp. 569–596.

[13] A. Goldenshluger and O. Lepski. “Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality”. In: *Ann. Statist.* 39.3 (2011), pp. 1608–1632.

[14] V. Konakov, S. Menozzi, and S. Molchanov. “Explicit parametrix and local limit theorems for some degenerate diffusion processes”. In: *Ann. Inst. H. Poincaré Probab. Statist.* 46.4 (Nov. 2010), pp. 908–923.

[15] O. Lepski. “Multivariate density estimation under sup-norm loss: oracle approach, adaptation and independence structure”. In: *Ann. Statist.* 41.2 (2013), pp. 1005–1034.

[16] D. Revuz and M. Yor. *Continuous martingales and Brownian motion*. Third. Vol. 293. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1999, pp. xiv+602.

[17] C. Strauch. "Adaptive invariant density estimation for ergodic diffusions over anisotropic classes". In: *Ann. Statist.* 46.6B (2018), pp. 3451 –3480.

[18] C. Strauch. "Exact adaptive pointwise drift estimation for multidimensional ergodic diffusions". In: *Probab. Theory Relat. Fields* 164.1-2 (2016), pp. 361–400.

[19] C. Strauch. "Sharp adaptive drift estimation for ergodic diffusions: The multivariate case". In: *Stochastic Process. Appl.* 125.7 (2015), pp. 2562 –2602.

[20] M. Talagrand. *The Generic Chaining*. Springer Monographs in Mathematics. Berlin: Springer, 2005.

[21] G. Viennet. "Inequalities for absolutely regular sequences: application to density estimation". In: *Probab. Theory Relat. Fields* 107.4 (1997), pp. 467–492.

[22] L. Wu. "Large and moderate deviations and exponential convergence for stochastic damping Hamiltonian systems". In: *Stochastic Process. Appl.* 91.2 (2001), pp. 205 –238.

On Lasso and Slope drift estimators for Lévy-driven Ornstein–Uhlenbeck processes

Niklas Dexheimer and Claudia Strauch

C

Abstract

We investigate the problem of estimating the drift parameter of a high-dimensional Lévy-driven Ornstein–Uhlenbeck process under sparsity constraints. It is shown that both Lasso and Slope estimators achieve the minimax optimal rate of convergence (up to numerical constants), for tuning parameters chosen independently of the confidence level, which improves the previously obtained results for standard Ornstein–Uhlenbeck processes. The results are nonasymptotic and hold both in probability and conditional expectation with respect to an event resembling the restricted eigenvalue condition.

C.1 Introduction

Due to increasing computational power, there has been an immense recent interest in high-dimensional statistical models, with many research efforts being made to understand statistical problems in a framework where the number of model parameters can be much larger than the number of observations. For classical models such as linear regression, issues such as how to construct procedures which are both computationally efficient and show optimal statistical performance (as quantified in terms of convergence rates) are now well understood. In contrast, only few deep statistical results are available as regards the high-dimensional modelling of continuous-time processes, even though these types of models can often be very well motivated from an application point of view. A classic example of a continuous-time model of great practical relevance is the Ornstein–Uhlenbeck (OU) process, which is given as the solution of the stochastic differential equation (SDE)

$$\mathrm{d}X_t = -\mathbf{A}X_t\,\mathrm{d}t + \mathbf{\Sigma}\,\mathrm{d}W_t, \quad t \geq 0, \tag{C.1}$$

where $\mathbf{A}, \mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ and $(W_t)_{t\geq 0}$ is a d-dimensional Wiener process. In the scalar case, this process is referred to as the Vasicek model when applied to model interest rates. In its multivariate version, it is frequently used, among many other applications, to model interbank lending (see [10], [7]). Since the matrix $\mathbf{A}$ then describes the interactions between (a possibly very large number of) different banks, the question of estimating it from observations of $\boldsymbol{X} = (X_t)_{t\geq 0}$ naturally arises. Given that banks often have only a limited number of lending partners, it is also natural to assume sparsity of $\mathbf{A}$, which is a classical assumption in the field of high-dimensional statistics as it allows to overcome the curse of dimensionality to some extent. Given the availability of a continuous record of observations of (C.1) with $\mathbf{\Sigma} = \mathbb{I}_{d\times d}$ on some time interval $[0, T]$ and assuming sparsity of the interaction matrix, the issue of estimating $\mathbf{A}$ is investigated in [11] and [9]. The proposed estimators are of Lasso- (in the classical and its adaptive version) and Dantzig-type, since these estimators are known to induce sparse results.

At first glance, it may come as a surprise that theoretical studies on high-dimensional versions of the basic model (C.1) are relatively recent. In fact, however, the investigation brings with it

specific probabilistic challenges. From the classical context of linear regression, it is well-known that convex penalisation methods (such as Lasso or Dantzig selectors) are efficient to compute, but show good statistical performance only under restrictive assumptions on the underlying design. An exemplary requirement for the Lasso estimator is the restricted eigenvalue property (see, e.g., (3.1), (4.2) and the beginning of Section 6 in [4] or Section 3 in [5]). Verifying a corresponding analogue in the context of continuous-time high-dimensional models amounts to the demanding task of establishing concentration of measure phenomena for unbounded functionals of the underlying process. In the Gaussian OU model, [11] succeeded in showing by means of the log-Sobolev inequality that the restricted eigenvalue property follows directly from the model assumptions as soon as $\mathbf{A}$ is symmetric, while [9] were even able to demonstrate (using Malliavin methods) that ergodicity of $\boldsymbol{X}$ already ensures the requested property. Remarkably, unlike sparse linear regression, one thus does not have to impose the restricted eigenvalue property, as it can be derived directly from the model assumptions of the standard OU model. Based on this, high probability estimates for the Lasso estimator $\widehat{\mathbf{A}}_{\text{lasso}}$ are proven in both [11] and [9]. In particular, denoting by $\|\cdot\|_2$ the Frobenius norm and by s the sparsity of $\mathbf{A}$, Corollary 4.3 in [9] gives the tightest available bound by stating that there exists some constant $c > 0$ such that

$$\|\widehat{\mathbf{A}}_{\text{lasso}} - \mathbf{A}\|_2^2 \leq \frac{cs}{T} \log\left(\frac{d^2}{\varepsilon_0}\right) \tag{C.2}$$

holds true with probability larger than $1 - \varepsilon_0$, for observation time T larger than some T_0 and adequately chosen tuning parameter, both depending on the confidence level $\varepsilon_0 > 0$. The upper bound in (C.2) almost matches the well-known minimax optimal rate of estimation in sparse linear regression (see the introduction of [4] and references therein), which in the given setting corresponds to

$$\frac{s}{T} \log\left(\frac{d^2}{s}\right). \tag{C.3}$$

The aims of this paper are now threefold. Firstly, we want to deduce the analysis of penalised estimators of the drift parameter for the more general class of Lévy-driven OU processes, i.e., we replace the driving Wiener process in (C.1) by a general Lévy process $(Z_t)_{t\geq 0}$, resulting in

$$\mathrm{d}X_t = -\mathbf{A}X_t \,\mathrm{d}t + \mathrm{d}Z_t, \quad t \geq 0. \tag{C.4}$$

Secondly, regarding the rates of convergence, we aim at closing the gap between (C.2) and (C.3), while also choosing the tuning parameter of the penalised estimators independently of the confidence level ε_0, which corresponds to our third objective. To achieve the latter goals, a suitable candidate is the Slope estimator, introduced in [6] as a weighted refinement of the Lasso estimator, which was shown to be minimax optimal for sparse linear regression in [4]. Another result of this reference is that the Lasso estimator also attains the optimal convergence rate, but with the downside that the sparsity of the unknown parameter needs to be known for choosing suitable values for the tuning parameter, which is not the case for the Slope estimator. Furthermore, it is demonstrated that the tuning parameters for both Lasso and Slope estimators can be chosen independently of the confidence level ε_0. At the heart of the proof of these results is a refined deviation inequality for the stochastic error term (Theorem 4.1 in [4]), which in turn relies heavily on the (sub-) Gaussianity of the noise in the considered model. Since the stochastic error in the setting of (C.4) studied here corresponds to an Itō integral with non-deterministic

integrand, reaching a result similar to the one obtained in the linear regression framework is not straightforward. For overcoming this challenge, we apply Talagrand's generic chaining device together with the restricted eigenvalue property, which then allows us to find a sufficiently tight result by bounding the Gaussian width of a given set. In fact, using our methods, we succeed in defining estimators of the drift parameter $\mathbf{A}$ for the Lévy-driven OU process (C.4) that have the desired properties and, in particular, achieve the optimal convergence rate.

The structure of this paper is as follows. In Section C.1.1, we introduce the mathematical setting and notation of this paper, and in C.1.2 we continue by introducing the two bespoke estimators. Section C.2 contains our main results on the performance of both Lasso and Slope estimators in the form of oracle inequalities resp. bounds in various norms. We also discuss the optimality of the derived upper bounds on the rates of convergence. The subsequent section consists of the required deviation inequalities for the results in Section C.2, namely a property of restricted eigenvalue-type (Section C.3.1) and the aforementioned deviation inequality for the stochastic error term (Section C.3.2). As explained in Section C.2, our results rely on the concentration assumption $(\mathcal{H})$, which is discussed in more detail in Section C.4.1. The paper concludes by a brief simulation study in Section C.5, where we compare the error of the maximum likelihood estimator to both Lasso and Slope estimators in various dimensions. The appendix contains basic probabilistic results for the processes considered, as well as some longer proofs.

C.1.1 *Preliminaries and notation*

In the following, $\mathbf{Z} = (Z_t)_{t\geq 0}$ will denote a d-dimensional Lévy process on a given filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t), \mathbb{P})$, adapted to the filtration $(\mathcal{F}_t)_{t\geq 0}$. For $\mathbf{A} \in \mathbb{R}^{d\times d}$, we call a strong solution $\boldsymbol{X} = (X_t)_{t\geq 0}$ of the SDE

$$\mathrm{d}X_t = -\mathbf{A}X_t\,\mathrm{d}t + \mathrm{d}Z_t, \quad t > 0,\ X_0 \sim \pi, \tag{C.5}$$

an Ornstein–Uhlenbeck (OU) process with background driving Lévy process (BDLP) $\mathbf{Z}$, initial distribution π and parameter $\mathbf{A}$. The initial condition X_0 is assumed to be independent of $\mathbf{Z}$. It follows from Itō's formula that an explicit solution of (C.5) is given by (see e.g. equations (1.1) and (1.2) in [14])

$$X_t = \mathrm{e}^{-t\mathbf{A}}X_0 + \int_0^t \mathrm{e}^{-(t-s)\mathbf{A}}\,\mathrm{d}Z_s, \quad t > 0. \tag{C.6}$$

Denote by $(b, \mathbf{C} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top, \nu)$ the generating triplet of $\mathbf{Z}$, i.e., $b = (b_k)_{k=1}^d \in \mathbb{R}^d$, $\mathbf{C} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top = (\mathbf{C}_{kl})_{k,l=1}^d$ is a $d\times d$ symmetric non-negative definite matrix and ν is a Lévy measure, i.e., a σ-finite measure on $\mathbb{R}^d$ satisfying

$$\nu(\{\mathbf{0}\}) = 0 \quad \text{and} \quad \int_{\mathbb{R}^d} \min\{1, \|z\|^2\}\nu(\mathrm{d}z) < \infty.$$

Recall that, by the Lévy–Itō decomposition (see e.g. Theorem 2.4.16 in [1]), it then holds

$$Z_t = bt + \boldsymbol{\Sigma}W_t + \int_0^t \int_{\|z\|\geq 1} zN(\mathrm{d}s, \mathrm{d}z) + \int_0^t \int_{\|z\|<1} z\widetilde{N}(\mathrm{d}s, \mathrm{d}z), \quad t \geq 0,$$

where $(W_s)_{s\geq 0}$ denotes a d-dimensional Wiener process, N is a Poisson random measure on $[0,\infty)\times\mathbb{R}^d$ with intensity measure given by $\lambda\otimes\nu$, and $\widetilde{N}$ denotes its compensated counterpart.

Furthermore, denote by $\mathbb{P}_t^{\mathbf{A}}$ the restriction of the measure $\mathbb{P}^{\mathbf{A}}$ induced by (C.5) on the path space to $\mathcal{F}_t$. For $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$, we define

$$\|\mathbf{A}\|_0 := \sum_{1 \le i \le d_1, 1 \le j \le d_2} \mathbb{1}\{A_{ij} \neq 0\}, \qquad \|\mathbf{A}\|_p := \left(\sum_{1 \le i \le d_1, 1 \le j \le d_2} |A_{ij}|^p \right)^{1/p}, \quad p \ge 1,$$

and set $\|\mathbf{A}\|_{\mathrm{Sp}}$ to be the spectral norm. To the Frobenius norm $\|\cdot\|_2$, we associate the scalar product

$$\langle \mathbf{A}_1, \mathbf{A}_2 \rangle_2 := \mathrm{tr}\big(\mathbf{A}_1^\top \mathbf{A}_2\big), \quad \mathbf{A}_1, \mathbf{A}_2 \in \mathbb{R}^{d_1 \times d_2},$$

for $\mathrm{tr}(\cdot)$ denoting the trace. Additionally, for $p \in [1, \infty) \cup \{0\}$ and $r > 0$, set

$$\mathbb{B}_p(r) := \big\{ \mathbf{B} \in \mathbb{R}^{d \times d} : \|\mathbf{B}\|_p \le r \big\}.$$

For a symmetric matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, we write $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ for the largest and the smallest eigenvalue of $\mathbf{A}$, respectively. Denote by $M_+(\mathbb{R}^d)$ the set of all real $d \times d$ matrices such that the real parts of all eigenvalues are positive, i.e., $\mathbf{A} \in M_+(\mathbb{R}^d)$ if and only if $\|\mathrm{e}^{-t\mathbf{A}}\|_2 \to 0$ as $t \to \infty$.

Given $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$, denote by $(\beta_1^\#, \ldots, \beta_d^\#)$ a nonincreasing rearrangement of $|\beta_1|, \ldots, |\beta_d|$. For a vector of tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^d$ not all equal to 0 such that $\lambda_1 \ge \lambda_2 \ge \ldots \ge \lambda_d \ge 0$, we set

$$\|\boldsymbol{\beta}\|_* := \sum_{j=1}^d \lambda_j \beta_j^\#, \quad \boldsymbol{\beta} \in \mathbb{R}^d.$$

Then it is known that $\|\cdot\|_*$ defines a norm on $\mathbb{R}^d$ (see Proposition 1.2 in [6]). In the following, the weights will always be given by

$$\lambda_j = \sqrt{\log\left(\frac{2d}{j}\right)}, \quad j \in \{1, \ldots d\}.$$

For $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$, we set (by a slight abuse of notation) $\|\mathbf{A}\|_* := \|\mathrm{vec}(\mathbf{A})\|_*$, i.e.,

$$\|\mathbf{A}\|_* = \sum_{j=1}^{d_1 d_2} \mathrm{vec}(\mathbf{A})_j^\# \sqrt{\log\left(\frac{2d_1 d_2}{j}\right)}. \tag{C.7}$$

Finally, for stochastic processes $(X_t)_{t \in [0,T]}, (Y_t)_{t \in [0,T]} \in L^2([0,T], \mathrm{d}t)$, we introduce the scalar product

$$\langle X, Y \rangle_{L^2} := \frac{1}{T} \int_0^T X_s^\top Y_s \, \mathrm{d}s.$$

C.1.2 *The Lasso and Slope estimators*

In the following, we assume that a continuous record of observations up to time $T > 0$ of a Lévy-driven OU process $\boldsymbol{X}$ is available, and the goal is to estimate the unknown true drift parameter $\mathbf{A}_0 \in \mathbb{R}^{d \times d}$. Additionally, we assume that the corresponding path of the continuous martingale part $\boldsymbol{X}^{\mathrm{c}} = (X_t^{\mathrm{c}})_{t \ge 0}$ of $\boldsymbol{X}$ and the diffusion parameter $\boldsymbol{\Sigma}$ are known. Extraction of the

continous martingale part from discrete observations of $\boldsymbol{X}$ by employing a truncation approach was discussed in [13] in the context of maximum likelihood estimation.

To begin with our analysis, we introduce the following assumption, which will be in place **implicitly throughout the whole paper**.

($\mathcal{A}_0$) $\mathbf{A}_0 \in M_+(\mathbb{R}^d)$, $\mathbf{C}$ is strictly positive definite, and ν admits a second moment. Additionally, it holds $\pi = \mu$, i.e., $\boldsymbol{X}$ is stationary.

Of course for ($\mathcal{A}_0$) to make sense, an invariant distribution has to exist. It is, however, well known that this is the case if $\mathbf{A}_0 \in M_+(\mathbb{R}^d)$ and $\mathbb{E}[(1 \vee \log(\|Z_1\|))]$ is finite (see Theorems 4.1 and 4.2 in [18] or Proposition 2.2 in [14]).

Under ($\mathcal{A}_0$), we are able to employ the results of [19] where maximum likelihood estimation for general jump diffusions is investigated. As condition **C** of [19] clearly follows from ($\mathcal{A}_0$), we get the following result.

Proposition C.1. *Let* $\mathbf{A} \in \mathbb{R}^{d\times d}, T \geq 0$. *Then,*

$$\frac{\mathrm{d}\mathbb{P}_T^{\mathbf{A}}}{\mathrm{d}\mathbb{P}_T^{\mathbf{0}}} = \exp\left(-\int_0^T (\mathbf{C}^{-1}\mathbf{A}X_{s-})^\top \mathrm{d}X_s^{\mathrm{c}} - \frac{1}{2}\int_0^T (\boldsymbol{\Sigma}^{-1}\mathbf{A}X_{s-})^\top \boldsymbol{\Sigma}^{-1}\mathbf{A}X_{s-}\,\mathrm{d}s\right). \tag{C.8}$$

Proof. From (C.6), we have by the Lévy–Itō decomposition that, under $\mathbb{P}^{\mathbf{A}}$,

$$X_t = \mathrm{e}^{-tA}X_0 + \int_0^t \mathrm{e}^{-(t-s)\mathbf{A}}b^*\,\mathrm{d}s + \int_0^t \mathrm{e}^{-(t-s)\mathbf{A}}\boldsymbol{\Sigma}\,\mathrm{d}W_s + \int_0^t\int_{\mathbb{R}^d} \mathrm{e}^{-(t-s)\mathbf{A}}z\widetilde{N}(\mathrm{d}s,\mathrm{d}z),$$

where everything is given as in Section C.1.1 and

$$b^* := b + \int_{\|z\|>1} z\nu(\mathrm{d}z).$$

Thus, Corollary 4.4.24 in [1] implies that $\mathbb{E}[\sup_{0\leq s\leq t}\|X_s\|^2]$ is finite, since μ and ν admit a second moment by ($\mathcal{A}_0$) and Corollary C.16. Hence, we get

$$\mathbb{P}^{\mathbf{A}}\left(\int_0^T (\boldsymbol{\Sigma}^{-1}\mathbf{A}X_{s-})^\top\boldsymbol{\Sigma}^{-1}\mathbf{A}X_{s-}\,\mathrm{d}s < \infty\right) = \mathbb{P}^{\mathbf{0}}\left(\int_0^T (\boldsymbol{\Sigma}^{-1}\mathbf{A}X_{s-})^\top\boldsymbol{\Sigma}^{-1}\mathbf{A}X_{s-}\,\mathrm{d}s < \infty\right) = 1,$$

where we argued analogously for $\mathbb{P}^{\mathbf{0}}$. This concludes the proof by Theorem 2.1 in [19]. ∎

Given (C.8), we are able to determine the likelihood function and thus can define the Lasso and Slope estimators. For doing so, we set

$$\mathcal{L}_T(\mathbf{A}) := -\frac{1}{T}\log\left(\frac{\mathrm{d}\mathbb{P}_T^{\mathbf{A}}}{\mathrm{d}\mathbb{P}_T^{\mathbf{0}}}\right), \quad \mathbf{A} \in \mathbb{R}^{d\times d}. \tag{C.9}$$

Furthermore, as we do *not* assume $\boldsymbol{\Sigma}$ to be the identity (as in [11] and [9]) or the identity matrix multiplied by some factor (as in [4]), we have to adjust the classical definitions of Lasso and Slope estimators slightly in our setting. We define the Lasso estimator to be given as

$$\widehat{\mathbf{A}}_{\text{lasso}} \in \operatorname{argmin}_{\mathbf{A}\in\mathbb{R}^{d\times d}}\left(\mathcal{L}_T(\mathbf{A}) + \lambda_{\mathrm{L}}\|\boldsymbol{\Sigma}^{-1}\mathbf{A}\|_1\right), \tag{C.10}$$

where $\lambda_L > 0$ is a tuning parameter. For the Slope estimator, we set

$$\widehat{\mathbf{A}}_{\text{slope}} \in \operatorname{argmin}_{\mathbf{A}\in\mathbb{R}^{d\times d}}\Big(\mathcal{L}_T(\mathbf{A}) + \lambda_S\|\boldsymbol{\Sigma}^{-1}\mathbf{A}\|_*\Big), \tag{C.11}$$

where again $\lambda_S > 0$ is a tuning parameter and $\|\cdot\|_*$ is defined in (C.7). Our interest in this estimator is motivated by the fact that, in the classical context of high-dimensional linear regression on the class of s-sparse vectors in $\mathbb{R}^d$, the Slope estimator with suitably chosen tuning parameters achieves the optimal rate $(s/n)\log(d/s)$, n denoting the number of observations, for both the prediction and the ℓ_2 estimation risks under suitable assumptions. As both estimators are defined as a solution of a convex optimization problem, they can be computed efficiently.

C.2 PROBABILITY ESTIMATES FOR THE LASSO AND SLOPE ESTIMATORS

The goal of this section is to provide probability estimates for the performance of Lasso and Slope estimators with tuning parameters not tied to a confidence level. The starting point of both proofs is given by the following auxiliary result.

LEMMA C.2. *Let $h\colon \mathbb{R}^{d\times d} \to \mathbb{R}$ be a convex function, and recall the definition of $\mathcal{L}_T(\cdot)$ in (C.9). If $\widehat{\mathbf{A}}$ is a solution of the minimization problem $\min_{\mathbf{A}\in\mathbb{R}^{d\times d}}(\mathcal{L}_T(\mathbf{A}) + h(\mathbf{A}))$, then $\widehat{\mathbf{A}}$ satisfies for all $\mathbf{A} \in \mathbb{R}^{d\times d}$*

$$\|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}}-\mathbf{A}_0)X\|_{L^2}^2 - \|\boldsymbol{\Sigma}^{-1}(\mathbf{A}-\mathbf{A}_0)X\|_{L^2}^2 \le 2(\langle\boldsymbol{\epsilon}_T, \boldsymbol{\Sigma}^{-1}(\mathbf{A}-\widehat{\mathbf{A}})\rangle_2 + h(\mathbf{A}) - h(\widehat{\mathbf{A}})) - \|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}}-\mathbf{A})X\|_{L^2}^2,$$

where

$$\boldsymbol{\epsilon}_T^\top := \frac{1}{T}\int_0^T X_s\,\mathrm{d}\widetilde{W}_s^\top, \tag{C.12}$$

with $(\widetilde{W}_s)_{s\ge0}$ being a $\mathbb{P}^{\mathbf{A}_0}$-Wiener process.

The proof of Lemma C.2 relies on the convexity of $\mathcal{L}_T$ and h, combined with an application of Girsanov's theorem, and can be found in Appendix C.II. Additionally, the proofs for our main results on the performance of the estimators require deviation inequalities for $\boldsymbol{\epsilon}_T$ and properties resembling the restricted eigenvalue property, which is a classical assumption in the context of linear regression. These results can be found in Section C.3. They are based on the following assumption:

($\mathcal{H}$) There exists a function $H\colon \mathbb{R}^+\times\mathbb{R}^+ \to \mathbb{R}^+$ such that

(i) for any $T, r > 0$, the functions $H(T,\cdot)$ and $H(\cdot,r)$ are non-increasing and such that $\lim_{T\to\infty} H(T,r) = 0$ for all $r > 0$, and

(ii) for any vector $u \in \mathbb{R}^d$ with $\|u\| \le 1$, it holds

$$\forall T, r > 0, \quad \mathbb{P}\Big(|u^\top(\widehat{\mathbf{C}}_T - \mathbf{C}_\infty)u| \ge r\Big) \le H(T,r),$$

where

$$\widehat{\mathbf{C}}_T := \frac{1}{T}\int_0^T X_s X_s^\top\,\mathrm{d}s \quad \text{and} \quad \mathbf{C}_\infty := \int xx^\top\mu(\mathrm{d}x). \tag{C.13}$$

Let $\kappa_{\min} := \lambda_{\min}(\mathbf{C}_\infty)$ and $\kappa_{\max} := \lambda_{\max}(\mathbf{C}_\infty)$. Note that $0 < \kappa_{\min} \leq \kappa_{\max}$ holds because of $(\mathcal{A}_0)$ (see the remark at the end of Appendix C.I). For easing the notation, we also introduce the events

$$Q_T(r) := \left\{ \sup_{\mathbf{B}\in\mathbb{B}_2(1)} |\operatorname{tr}(\mathbf{B}(\widehat{\mathbf{C}}_T - \mathbf{C}_\infty)\mathbf{B}^\top)| \leq r \right\}, \quad r > 0. \tag{C.14}$$

In Proposition C.9, we will see that $(\mathcal{H})$ in fact implies a lower bound on $\mathbb{P}(Q_T(r))$ for any $r > 0$.

It was shown in [11] and [9] in the Gaussian OU case that the restricted eigenvalue property holds with high probability for large enough values of T and thus follows implicitly from the model as soon as assumption $(\mathcal{H})$ is in place. In Section C.4.1, we investigate $(\mathcal{H})$ in more detail by providing sufficient conditions in the Lévy-driven case for $(\mathcal{H})$ to hold and recalling the results in the Gaussian case.

C.2.1 *Main results on the Lasso estimator*

A notable feature of many nonasymptotic bounds for the Lasso estimator available in the literature (see Corollary 1 in [11] or Corollary 4.3 in [9]) is that the confidence level is tied to the tuning parameter λ. In the high-dimensional linear regression model, [4] develop new proof strategies for the Lasso estimator, which in particular allow to derive bounds in probability at *any* level of confidence with the same tuning parameter. We now adapt their findings to the high-dimensional Lévy-driven OU model considered in this paper and show that here, too, there is no need for the confidence level to be linked to the tuning parameter.

Proposition C.3. *Grant Assumption $(\mathcal{H})$. Set $s = \|\boldsymbol{\Sigma}^{-1}\mathbf{A}_0\|_0$, and let $\widehat{\mathbf{A}} = \widehat{\mathbf{A}}_{\text{lasso}}$ be the Lasso estimator* (C.10) *with tuning parameter*

$$\lambda_T \geq 2c_*\sqrt{\frac{\kappa_{\max}}{T}\log\left(\frac{2ed^2}{s}\right)}, \tag{C.15}$$

where c_* is defined in Proposition C.11. Then, for any $\mathbf{A} \in \mathbb{R}^{d\times d}$ satisfying $\|\boldsymbol{\Sigma}^{-1}\mathbf{A}\|_0 \leq s$, the upper bound

$$\|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)X\|_{L^2}^2 + 2\lambda_T\|\boldsymbol{\Sigma}^{-1}(\mathbf{A} - \widehat{\mathbf{A}})\|_1 \leq \|\boldsymbol{\Sigma}^{-1}(\mathbf{A} - \mathbf{A}_0)X\|_{L^2}^2 + \frac{8s\lambda_T^2}{\kappa_{\min}}\left(1 \vee \frac{\log(4\varepsilon_0^{-1})}{s\log(2ed^2/s)}\right)$$

holds with probability of at least

$$1 - \frac{\varepsilon_0}{2} - (21(d \wedge \mathrm{e}))^d H\left(T, \frac{\kappa_{\min}}{6d}\right),$$

for all $\varepsilon_0 \in (0,1)$ and $T > 0$.

Proof. By Propositions C.9 and C.11, it holds

$$\begin{aligned} &1 - \frac{\varepsilon_0}{2} - (21(d \wedge \mathrm{e}))^d H\left(T, \frac{\kappa_{\min}}{6d}\right) \\ &\qquad \leq \mathbb{P}\left(\inf_{\mathbf{B}\in\mathbb{R}^{d\times d}\setminus\{0\}} \frac{\|\mathbf{B}X\|_{L^2}^2}{\|\mathbf{B}\|_2^2} \geq \frac{\kappa_{\min}}{2}, \sup_{\mathbf{B}\in\mathbb{R}^{d\times d}, \mathbf{B}\neq 0} \frac{\langle \boldsymbol{\epsilon}_T, \mathbf{B}\rangle_2}{\|\mathbf{B}\|_S} \leq c_*\sqrt{\frac{\kappa_{\max}}{T}} \right), \end{aligned} \tag{C.16}$$

where $\|\cdot\|_S$ is defined in (C.26). From now on, we assume that the event on the rhs of (C.16) occurs, and we fix $\mathbf{A} \in \mathbb{R}^{d\times d}$. By Lemma C.2, we then have for $h(\cdot) = \lambda_T\|\boldsymbol{\Sigma}^{-1}\cdot\|_1$ and $\mathbf{B} := \mathbf{A} - \widehat{\mathbf{A}}$,

$$\|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)X\|_{L^2}^2 + \lambda_T\|\boldsymbol{\Sigma}^{-1}\mathbf{B}\|_1 \le \|\boldsymbol{\Sigma}^{-1}(\mathbf{A} - \mathbf{A}_0)X\|_{L^2}^2 - \|\boldsymbol{\Sigma}^{-1}\mathbf{B}X\|_{L^2}^2 + \Delta_*, \tag{C.17}$$

where

$$\Delta_* := \lambda_T\|\boldsymbol{\Sigma}^{-1}\mathbf{B}\|_1 + 2\langle \boldsymbol{\epsilon}_T, \boldsymbol{\Sigma}^{-1}\mathbf{B}\rangle_2 + 2\lambda_T\Big(\|\boldsymbol{\Sigma}^{-1}\mathbf{A}\|_1 - \|\boldsymbol{\Sigma}^{-1}\widehat{\mathbf{A}}\|_1\Big).$$

Now note that, due to the Cauchy–Schwarz inequality and equation (2.7) in [4], for any $s \in \{1, \dots, d^2\}$,

$$\begin{aligned}\|\boldsymbol{\Sigma}^{-1}\mathbf{B}\|_* &\le \|\boldsymbol{\Sigma}^{-1}\mathbf{B}\|_2\sqrt{\sum_{i=1}^{s}\log(2d^2/i)} + \sum_{i=s+1}^{d^2}\mathrm{vec}(\boldsymbol{\Sigma}^{-1}\mathbf{B})_i^{\#}\sqrt{\log(2d^2/i)} \\ &\le \sqrt{\log(2ed^2/s)}\left(\sqrt{s}\|\boldsymbol{\Sigma}^{-1}\mathbf{B}\|_2 + \sum_{i=s+1}^{d^2}\mathrm{vec}(\boldsymbol{\Sigma}^{-1}\mathbf{B})_i^{\#}\right) =: F(\boldsymbol{\Sigma}^{-1}\mathbf{B}).\end{aligned} \tag{C.18}$$

Hence, by Lemma A.1 in [4], if $\|\boldsymbol{\Sigma}^{-1}\mathbf{A}\|_0 \le s$, the term Δ_* is bounded from above by

$$\begin{aligned}&2\lambda_T\Big(\frac{1}{2}\|\boldsymbol{\Sigma}^{-1}\mathbf{B}\|_1 + \|\boldsymbol{\Sigma}^{-1}\mathbf{A}\|_1 - \|\boldsymbol{\Sigma}^{-1}\widehat{\mathbf{A}}\|_1\Big) \\ &\qquad + \left(\log\left(\frac{2ed^2}{s}\right)\right)^{-\frac{1}{2}}\lambda_T\Big(F(\boldsymbol{\Sigma}^{-1}\mathbf{B}) \vee \sqrt{\log(4\epsilon_0^{-1})}\|\boldsymbol{\Sigma}^{-1}\mathbf{B}\|_2\Big) \\ &\le \lambda_T\left(3\sqrt{s}\|\boldsymbol{\Sigma}^{-1}\mathbf{B}\|_2 - \sum_{i=s+1}^{d^2}\mathrm{vec}(\mathbf{B})_i^{\#}\right) + \left(\log\left(\frac{2ed^2}{s}\right)\right)^{-\frac{1}{2}}\lambda_T\left(F(\boldsymbol{\Sigma}^{-1}\mathbf{B}) \vee \sqrt{\frac{2\log(4\epsilon_0^{-1})}{\kappa_{\min}}}\|\boldsymbol{\Sigma}^{-1}\mathbf{B}X\|_{L^2}\right),\end{aligned}$$

where we also used (C.16) and $2c_*\sqrt{T^{-1}\kappa_{\max}} \le (\log(2ed^2/s))^{-1/2}\lambda_T$. We continue by examining the two different cases which can occur due to the maximum term. On the one hand,

$$\sqrt{\frac{2\log(4\epsilon_0^{-1})}{\kappa_{\min}}}\|\boldsymbol{\Sigma}^{-1}\mathbf{B}X\|_{L^2} \ge F(\boldsymbol{\Sigma}^{-1}\mathbf{B}) \quad\Longrightarrow\quad \|\boldsymbol{\Sigma}^{-1}\mathbf{B}\|_2 \le \sqrt{\frac{2\log(4\epsilon_0^{-1})}{s\log(2ed^2/s)\kappa_{\min}}}\|\boldsymbol{\Sigma}^{-1}\mathbf{B}X\|_{L^2}.$$

Thus, in this case,

$$\begin{aligned}\Delta_* &\le 3\lambda_T\sqrt{s}\|\boldsymbol{\Sigma}^{-1}\mathbf{B}\|_2 + \lambda_T\sqrt{\frac{2\log(4\epsilon_0^{-1})}{\log(2ed^2/s)\kappa_{\min}}}\|\boldsymbol{\Sigma}^{-1}\mathbf{B}X\|_{L^2} \\ &\le 4\lambda_T\sqrt{\frac{2\log(4\epsilon_0^{-1})}{\log(2ed^2/s)\kappa_{\min}}}\|\boldsymbol{\Sigma}^{-1}\mathbf{B}X\|_{L^2} \\ &\le 8\lambda_T^2\frac{\log(4\epsilon_0^{-1})}{\log(2ed^2/s)\kappa_{\min}} + \|\boldsymbol{\Sigma}^{-1}\mathbf{B}X\|_{L^2}^2.\end{aligned}$$

In the opposite case,

$$\Delta_* \le 4\lambda_T\sqrt{s}\|\boldsymbol{\Sigma}^{-1}\mathbf{B}\|_2$$

$$\leq 4\lambda_T \sqrt{\frac{2s}{\kappa_{\min}}} \|\mathbf{\Sigma}^{-1}\mathbf{B}X\|_{L^2}$$
$$\leq \frac{8s\lambda_T^2}{\kappa_{\min}} + \|\mathbf{\Sigma}^{-1}\mathbf{B}X\|_{L^2}^2.$$

Inserting these results into (C.17) yields

$$\|\mathbf{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)X\|_{L^2}^2 + 2\lambda_T\|\mathbf{\Sigma}^{-1}\mathbf{B}\|_1 \leq \|\mathbf{\Sigma}^{-1}(\mathbf{A} - \mathbf{A}_0)X\|_{L^2}^2 + \frac{8\lambda_T^2}{\kappa_{\min}}\left(s \vee \frac{\log(4\varepsilon_0^{-1})}{\log(2ed^2/s)}\right).$$

■

While the lower bound (C.15) for the specification of the tuning parameter λ_T does not depend on ε_0, as promised, the sparsity s of $\mathbf{\Sigma}^{-1}\mathbf{A}_0$ appears there, which is of course unknown in general. As, however, $\mathbf{\Sigma}$ and $\mathbf{A}_0$ are invertible by assumption, it always holds $s \geq d$. Thus, choosing

$$\lambda_T \geq 2c_*\sqrt{T^{-1}\kappa_{\max}\log(2ed)}$$

implies the conditions in Proposition C.3 to be fulfilled. This specification leads to the same rate for the Lasso estimator as derived in [11] and [9] for Gaussian OU processes. We also find this in the following result, where we apply Proposition C.3 for getting high probability estimates in various norms.

Corollary C.4. *Let everything be given as in Proposition C.3, and let $\varepsilon_1 \in (0,1)$. Then, for*

$$T > T_0(\varepsilon_1) := \inf\left\{T > 0 : (21(d \wedge \mathrm{e}))^d H\left(T, \frac{\kappa_{\min}}{6d}\right) \leq \frac{\varepsilon_1}{2}\right\}, \tag{C.19}$$

the following assertions hold, each with probability larger than $1 - \frac{1}{2}(\varepsilon_0 + \varepsilon_1)$, for all $\varepsilon_0 \in (0,1)$:

(a) $$\|\mathbf{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)X\|_{L^2}^2 \leq \frac{8s\lambda_T^2}{\kappa_{\min}}\left(1 \vee \frac{\log(4\varepsilon_0^{-1})}{s\log(2ed^2/s)}\right),$$

(b) $$\|\mathbf{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)\|_2^2 \leq \frac{16s\lambda_T^2}{\kappa_{\min}^2}\left(1 \vee \frac{\log(4\varepsilon_0^{-1})}{s\log(2ed^2/s)}\right),$$

(c) $$\|\mathbf{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)\|_1 \leq \frac{4s\lambda_T}{\kappa_{\min}}\left(1 \vee \frac{\log(4\varepsilon_0^{-1})}{s\log(2ed^2/s)}\right).$$

Proof. Assertions (a) and (b) follow immediately by applying Proposition C.3 with $\mathbf{A} = \mathbf{A}_0$. For (b), note that (C.16) implies $\kappa_{\min}\|\mathbf{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)\|_2^2 \leq 2\|\mathbf{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)X\|_{L^2}^2$, and hence the assertion follows by (a). ■

C.2.2 *Main results on the Slope estimator*

We now state our main result on the Slope estimator in an analogous form to Proposition C.3.

PROPOSITION C.5. *Set* $s = \|\Sigma^{-1}\mathbf{A}_0\|_0$, *and choose* $c_S \geq c_*\sqrt{\kappa_{\max}}$, *where* c_* *is defined in Proposition C.11. Let* $\widehat{\mathbf{A}} = \widehat{\mathbf{A}}_{\text{slope}}$ *be the Slope estimator* (C.11) *with tuning parameter*

$$\lambda_T := \frac{2c_S}{\sqrt{T}}. \tag{C.20}$$

Then, for any $\mathbf{A} \in \mathbb{R}^{d\times d}$ *satisfying* $\|\Sigma^{-1}\mathbf{A}\|_0 \leq s$, *the inequality*

$$\|\Sigma^{-1}(\widehat{\mathbf{A}}-\mathbf{A}_0)X\|_{L^2}^2 + \frac{2c_S\|\Sigma^{-1}(\widehat{\mathbf{A}}-\mathbf{A})\|_*}{\sqrt{T}} \leq \|\Sigma^{-1}(\mathbf{A}-\mathbf{A}_0)X\|_{L^2}^2 + 32c_S^2\frac{s\log\left(\frac{2ed^2}{s}\right)}{T\kappa_{\min}}\left(1 \vee \frac{\log(4\varepsilon_0^{-1})}{s\log\left(\frac{2ed^2}{s}\right)}\right)$$

holds with probability of at least $1-\varepsilon_0/2-(21(d\wedge \mathrm{e}))^d H\big(T, \frac{\kappa_{\min}}{6d}\big)$, *for all* $\varepsilon_0 \in (0,1)$ *and* $T>0$.

Note that (by the choice of λ_T in (C.20)) the Slope estimator achieves the stated (optimal) rate of convergence even if the sparsity s of $\Sigma^{-1}\mathbf{A}_0$ is not known.

Proof of Proposition C.5. As in the proof of Proposition C.3, we start with inequality (C.16), and we assume that the event appearing in the upper bound holds true. By Lemma C.2, we then get for $h(\cdot) = 2c_S\sqrt{T^{-1}}\|\Sigma^{-1}\cdot\|_*$ and all $\mathbf{A}\in\mathbb{R}^{d\times d}$,

$$\|\Sigma^{-1}(\widehat{\mathbf{A}}-\mathbf{A}_0)X\|_{L^2}^2 - \|\Sigma^{-1}(\mathbf{A}-\mathbf{A}_0)X\|_{L^2}^2 \leq 2(\langle \boldsymbol{\epsilon}_T, \Sigma^{-1}(\mathbf{A}-\widehat{\mathbf{A}})\rangle_F + h(\mathbf{A}) - h(\widehat{\mathbf{A}})) - \|\Sigma^{-1}(\widehat{\mathbf{A}}-\mathbf{A})X\|_{L^2}^2.$$

Thus, for $\mathbf{B} := \mathbf{A}-\widehat{\mathbf{A}}$,

$$2c_S\sqrt{T^{-1}}\|\Sigma^{-1}\mathbf{B}\|_* + \|\Sigma^{-1}(\widehat{\mathbf{A}}-\mathbf{A}_0)X\|_{L^2}^2 \leq \|\Sigma^{-1}(\mathbf{A}-\mathbf{A}_0)X\|_{L^2}^2 - \|\Sigma^{-1}\mathbf{B}X\|_{L^2}^2 + \Delta^*, \tag{C.21}$$

where

$$\Delta^* := 2\langle \boldsymbol{\epsilon}_T, \Sigma^{-1}\mathbf{B}\rangle_F + 4c_S\sqrt{T^{-1}}\left(\|\Sigma^{-1}\mathbf{A}\|_* - \|\Sigma^{-1}\widehat{\mathbf{A}}\|_* + \frac{1}{2}\|\Sigma^{-1}\mathbf{B}\|_*\right).$$

Now, (C.16) and Lemma A.1 in [4] imply, if $\|\Sigma^{-1}\mathbf{A}\|_0 \leq s$, that

$$\begin{aligned}
\Delta^* &\leq 4c_S\sqrt{T^{-1}}\left(\frac{1}{2}\|\Sigma^{-1}\mathbf{B}\|_S + \|\Sigma^{-1}\mathbf{A}\|_* - \|\Sigma^{-1}\widehat{\mathbf{A}}\|_* + \frac{1}{2}\|\Sigma^{-1}\mathbf{B}\|_*\right)\\
&\leq 4c_S\sqrt{T^{-1}}\left(\frac{1}{2}\|\Sigma^{-1}\mathbf{B}\|_S + \frac{3}{2}\sqrt{\sum_{i=1}^{s}\log\left(\frac{2d^2}{i}\right)}\|\Sigma^{-1}\mathbf{B}\|_2 - \frac{1}{2}\sum_{i=s+1}^{d^2}\mathrm{vec}(\Sigma^{-1}\mathbf{B})_i^{\#}\sqrt{\log\left(\frac{2d^2}{i}\right)}\right)\\
&\leq 4c_S\sqrt{T^{-1}}\left(\frac{1}{2}\|\Sigma^{-1}\mathbf{B}\|_S + \frac{3}{2}\sqrt{s\log\left(\frac{2ed^2}{s}\right)}\|\Sigma^{-1}\mathbf{B}\|_2 - \frac{1}{2}\sum_{i=s+1}^{d^2}\mathrm{vec}(\Sigma^{-1}\mathbf{B})_i^{\#}\sqrt{\log\left(\frac{2d^2}{i}\right)}\right),
\end{aligned}$$

where we used equation (2.4) in [4] in the last step. Furthermore, arguing as in the derivation of equation (C.18) in the proof of Proposition C.3 and using (C.16), we arrive at

$$\Delta^* \leq 4c_S\sqrt{T^{-1}}\left(\frac{1}{2}\left(F(\Sigma^{-1}\mathbf{B}) \vee \sqrt{\frac{2\log(4\varepsilon_0^{-1})}{\kappa_{\min}}}\|\Sigma^{-1}\mathbf{B}X\|_{L^2}\right)\right.$$

$$+\frac{3}{2}\sqrt{s\log\left(\frac{2ed^2}{s}\right)}\|\Sigma^{-1}\mathbf{B}\|_2-\frac{1}{2}\sum_{i=s+1}^{d^2}\mathrm{vec}(\Sigma^{-1}\mathbf{B})_i^{\#}\sqrt{\log\left(\frac{2d^2}{i}\right)}\Bigg).$$

Again, we continue by investigating the two different cases related to the maximum term. Firstly,

$$\sqrt{\frac{2\log(4\varepsilon_0^{-1})}{\kappa_{\min}}}\|\Sigma^{-1}\mathbf{B}X\|_{L^2}\geq F(\Sigma^{-1}\mathbf{B})\quad\Longrightarrow\quad\|\Sigma^{-1}\mathbf{B}\|_2\leq\sqrt{\frac{2\log(4\varepsilon_0^{-1})}{s\log(2ed^2/s)\kappa_{\min}}}\|\Sigma^{-1}\mathbf{B}X\|_{L^2},$$

and hence

$$\begin{aligned}\Delta^*&\leq 4c_S\sqrt{T^{-1}}\left(\frac{1}{2}\sqrt{\frac{2\log(4\varepsilon_0^{-1})}{\kappa_{\min}}}\|\Sigma^{-1}\mathbf{B}X\|_{L^2}+\frac{3}{2}\sqrt{s\log\left(\frac{2ed^2}{s}\right)}\|\mathbf{B}\|_2\right)\\&\leq 8c_S\sqrt{\frac{2\log(4\varepsilon_0^{-1})}{T\kappa_{\min}}}\|\Sigma^{-1}\mathbf{B}X\|_{L^2}\\&\leq 32c_S^2\frac{\log(4\varepsilon_0^{-1})}{T\kappa_{\min}}+\|\Sigma^{-1}\mathbf{B}X\|_{L^2}^2.\end{aligned}$$

In the other case, we get

$$\begin{aligned}\Delta^*&\leq 4c_S\sqrt{T^{-1}}\left(\frac{1}{2}F(\Sigma^{-1}\mathbf{B})+\frac{3}{2}\sqrt{s\log\left(\frac{2ed^2}{s}\right)}\|\Sigma^{-1}\mathbf{B}\|_2-\frac{1}{2}\sum_{i=s+1}^{d^2}\mathrm{vec}(\Sigma^{-1}\mathbf{B})_i^{\#}\sqrt{\log\left(\frac{2d^2}{i}\right)}\right)\\&\leq 8c_S\sqrt{T^{-1}s\log\left(\frac{2ed^2}{s}\right)}\|\Sigma^{-1}\mathbf{B}\|_2\\&\leq 8c_S\sqrt{\frac{2s\log\left(\frac{2ed^2}{s}\right)}{T\kappa_{\min}}}\|\Sigma^{-1}\mathbf{B}X\|_{L^2}\\&\leq 32c_S^2\frac{s\log\left(\frac{2ed^2}{s}\right)}{T\kappa_{\min}}+\|\Sigma^{-1}\mathbf{B}X\|_{L^2}^2,\end{aligned}$$

and combining these results with (C.21) completes the proof. ∎

As for the Lasso estimator, this leads to results on various norms.

Corollary C.6. *Let everything be given as in Proposition C.5, and let $\varepsilon_1\in(0,1)$. Then, for $T>T_0(\varepsilon_1)$ and $T_0(\cdot)$ defined as in (C.19), the following assertions hold, each with probability larger than $1-\frac{1}{2}(\varepsilon_0+\varepsilon_1)$, for all $\varepsilon_0\in(0,1)$:*

(a) $$\|\Sigma^{-1}(\widehat{\mathbf{A}}-\mathbf{A}_0)X\|_{L^2}^2\leq 32c_S^2\frac{s\log\left(\frac{2ed^2}{s}\right)}{T\kappa_{\min}}\left(\frac{\log(4\varepsilon_0^{-1})}{s\log\left(\frac{2ed^2}{s}\right)}\vee 1\right),$$

(b) $$\|\Sigma^{-1}(\widehat{\mathbf{A}}-\mathbf{A}_0)\|_2^2\leq 64c_S^2\frac{s\log\left(\frac{2ed^2}{s}\right)}{T\kappa_{\min}^2}\left(\frac{\log(4\varepsilon_0^{-1})}{s\log\left(\frac{2ed^2}{s}\right)}\vee 1\right),$$

(c) $$\|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)\|_* \leq 16 c_S \frac{s \log\left(\frac{2ed^2}{s}\right)}{\sqrt{T}\kappa_{\min}} \left(\frac{\log(4\varepsilon_0^{-1})}{s \log\left(\frac{2ed^2}{s}\right)} \vee 1 \right),$$

(d) $$\|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)\|_1 \leq 16 c_S \frac{s \log\left(\frac{2ed^2}{s}\right)}{\sqrt{T}\kappa_{\min} \log(2)} \left(\frac{\log(4\varepsilon_0^{-1})}{s \log\left(\frac{2ed^2}{s}\right)} \vee 1 \right).$$

Proof. The proof is completely analogous to the proof of Corollary C.4, except for (d), where it suffices to note that $\log(2)\|\mathbf{B}\|_1 \leq \|\mathbf{B}\|_*$ for all $\mathbf{B} \in \mathbb{R}^{d\times d}$. ∎

C.2.3 *Optimality of the convergence rates*

Alongside the extension of the analysis of the Gaussian OU model to the Lévy-driven case, the principal question of determining *rate-optimal* estimators for high-dimensional models of continuous-time processes is in the focus of our study in this paper. To mark out the framework, we start by recalling that Theorem 2 in [11] provides a minimax lower bound of the form

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}\in\Gamma_s} \mathbb{E}_{\mathbf{A}}\left[\|\widehat{\mathbf{A}} - \mathbf{A}\|_2^2 \right] \geq c' ds \log(cd/s)/T,$$

for certain constants $c, c' > 0$, where Γ_s is the set of row-s-sparse matrices and the infimum is taken over all possible estimators of the drift parameter $\mathbf{A}$ in the classical OU model (C.1) with $\boldsymbol{\Sigma} = \mathbb{I}_{d\times d}$. In what follows, we will derive a similar result under the sparsity assumption $\|\mathbf{A}\|_0 \leq s$. As regards compatibility of lower and upper bounds, it is of specific advantage that the probability estimates in the previous subsections apply to any confidence level. In particular, this allows to prove upper bounds in expectation, conditioned on the event $Q_T(\kappa_{\min}/2)$.

COROLLARY C.7. *Grant the assumptions of Proposition C.3 and C.5, respectively, let $\varepsilon_1 \in (0,1)$, and recall the definition of the event $Q_T(\cdot)$ in (C.14). Then, for $T > T_0(2\varepsilon_1)$ and $T_0(\cdot)$ defined as in (C.19),*

$$\mathbb{E}_{\mathbf{A}_0}\left[\frac{\kappa_{\min}}{2} \|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}}_{\text{lasso}} - \mathbf{A}_0)\|_2^2 + 2\lambda_T \|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}}_{\text{lasso}} - \mathbf{A}_0)\|_1 \,\middle|\, Q_T\left(\frac{\kappa_{\min}}{2}\right) \right] \leq \frac{8s\lambda_T^2}{\kappa_{\min}(1-\varepsilon_1)} \left(1 + \frac{2}{\log(2ed^2)}\right),$$

$$\mathbb{E}_{\mathbf{A}_0}\left[\frac{\kappa_{\min}}{2} \|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}}_{\text{slope}} - \mathbf{A}_0)X\|_2^2 + \frac{2c_S}{\log(2)\sqrt{T}} \|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}}_{\text{slope}} - \mathbf{A}_0)\|_1 \,\middle|\, Q_T\left(\frac{\kappa_{\min}}{2}\right) \right] \leq \frac{32c_S^2 s \log(2ed^2/s)}{T\kappa_{\min}(1-\varepsilon_1)} \left(1 + \frac{2}{\log(2ed^2)}\right).$$

The fact that the above bounds are for the *conditional* expectation, which is in contrast to the results for sparse regression in [4], can be justified by us not *assuming* our property of restricted eigenvalue type, but *proving* that it holds with high probability. For this, also note that Proposition C.9 implies that $Q_T(\kappa_{\min}/2)$ is a subset of the event where the restricted eigenvalue type property holds true.

Proof of Corollary C.7. We start by proving the assertion for the Lasso estimator. For now, let $\widehat{\mathbf{A}} = \widehat{\mathbf{A}}_{\text{lasso}}$, and set

$$Y \coloneqq \left(\|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)X\|_{L^2}^2 + 2\lambda_T \|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)\|_1 \right) \frac{\kappa_{\min} \log(2ed^2/s)}{8\lambda_T^2} \, \mathbb{1}_{Q_T(\kappa_{\min}/2)}.$$

Inspection of the proof of Proposition C.3 shows that $Y \leq \log(4\varepsilon_0^{-1})$ holds with probability of at least $1 - \varepsilon_0/2$, for all $0 < \varepsilon_0 \leq \varepsilon_0^* < 1$. Hence, for any $r \geq r^* = \log(4/\varepsilon_0^*)$,

$$\mathbb{P}_{\mathbf{A}_0}(Y > r) \leq 2\mathrm{e}^{-r}$$

and therefore

$$\mathbb{E}_{\mathbf{A}_0}[Y] \leq \int_0^\infty \mathbb{P}_{\mathbf{A}_0}(Y > r) \, \mathrm{d}r \leq r^* + 2\int_{r^*}^\infty \mathrm{e}^{-r} \, \mathrm{d}r \leq r^* + 2.$$

Choosing $\varepsilon_0^* = 4(2ed^2/s)^{-s}$, we thus obtain

$$\begin{aligned} &\mathbb{E}_{\mathbf{A}_0}\left[\left(\|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)X\|_{L^2}^2 + 2\lambda_T \|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)\|_1 \right) \mathbb{1}_{Q_T(\kappa_{\min}/2)} \right] \\ &\qquad \leq \frac{8s\lambda_T^2}{\kappa_{\min}} + \frac{16s\lambda_T^2}{\kappa_{\min} s \log(2ed^2/s)} \leq \frac{8s\lambda_T^2}{\kappa_{\min}} + \frac{16s\lambda_T^2}{\kappa_{\min} \log(2ed^2)}. \end{aligned}$$

Applying Proposition C.9 then yields the assertion for the Lasso estimator. We continue with the proof for the Slope estimator. By an abuse of notation, $\widehat{\mathbf{A}} = \widehat{\mathbf{A}}_{\text{slope}}$ now denotes the Slope estimator defined as in Proposition C.5. Furthermore, let

$$Z \coloneqq \left(\|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)X\|_{L^2}^2 + \frac{2c_S \|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)\|_*}{\sqrt{T}} \right) \frac{T\kappa_{\min}}{32c_S^2} \, \mathbb{1}_{Q_T(\kappa_{\min}/2)}.$$

Analogously to the proof for the Lasso estimator, we have that $Z \leq \log(4\varepsilon_0^{-1})$ holds with probability of at least $1 - \varepsilon_0/2$, for all $0 < \varepsilon_0 \leq \varepsilon_0^* < 1$. Hence, choosing ε_0^* as above,

$$\begin{aligned} &\mathbb{E}_{\mathbf{A}_0}\left[\left(\|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)X\|_{L^2}^2 + \frac{2c_S \|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}} - \mathbf{A}_0)\|_*}{\sqrt{T}} \right) \mathbb{1}_{Q_T(\kappa_{\min}/2)} \right] \\ &\qquad \leq \frac{32c_S^2 s \log(2ed^2/s)}{T\kappa_{\min}} \left(1 + \frac{2}{s\log(2ed^2/s)} \right) \leq \frac{32c_S^2 s \log(2ed^2/s)}{T\kappa_{\min}} \left(1 + \frac{2}{\log(2ed^2)} \right). \end{aligned}$$

Applying Proposition C.9, together with $\log(2)\|\mathbf{B}\|_1 \leq \|\mathbf{B}\|_*$ for all $\mathbf{B} \in \mathbb{R}^{d\times d}$, completes the proof. ∎

For proving a lower bound for estimation of the drift parameter $\mathbf{A}_0$ over the set of s-sparse matrices, belonging to $M_+(\mathbb{R}^d)$, we follow the strategy developed by [4] in the high-dimensional regression setting. By providing a lower bound on the expected value of a general loss function, one in particular also obtains results allowing for comparison with upper bounds in probability as they are stated in Corollary C.4 and C.6, respectively.

THEOREM C.8 (cf. Theorem 7.1 in [4]). *Let $d \geq 4$, $s \geq 2d$, and consider a nondecreasing function $\ell\colon \mathbb{R}_+ \to \mathbb{R}_+$ fulfilling $\ell(0) = 0$ and $\ell \not\equiv 0$. Assume that the Lévy triplet of the BDLP $\mathbf{Z}$ is given by $(0, \mathbb{I}_{d\times d}, 0)$. Then, for $1 \leq p < \infty$, there exist positive constants c, c', depending only on $\ell(\cdot)$, such that*

$$\inf_{\widehat{\mathbf{A}}} \sup_{\mathbf{A}_0 \in M_+(\mathbb{R}^d)\cap\mathbb{B}_0(s)} \mathbb{E}_{\mathbf{A}_0}\left[\ell\left(c\psi_{T,p}^{-1}\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_p\right)\right] \geq c',$$

where the infimum extends over all estimators of $\mathbf{A}_0$ and

$$\psi_{T,p} := s^{1/p}\sqrt{\frac{\log(ed^2/s)}{T}}.$$

Proof. Let r be the largest even number such that $r \leq (s-d)/2$, and let Ω_r be the set of antisymmetric matrices in $\{-1,0,1\}^{d\times d}$ with sparsity exactly equal to r. Then, every matrix in Ω_r is uniquely determined by its upper triangular section, which corresponds to a vector in $\{-1,0,1\}^{d(d-1)/2}$ with $r/2$ non-zero entries. Now since $d \geq 4$ and $s \geq 2d$ imply $d(d-1)/2 \geq 2$ and $1 \leq r/2 \leq (s-d)/4 \leq d(d-1)/4$, Lemma F.1 in [4] entails the existence of a set $\widetilde{\Omega}_r \subset \Omega_r$ such that, for all $\mathbf{B} \neq \mathbf{B}' \in \widetilde{\Omega}_r$,

$$\|\mathbf{B}\|_0 \leq r \leq s-d, \tag{C.22}$$

$$\log\left(|\widetilde{\Omega}_r|\right) \geq cr\log(ed(d-1)/r) \geq cr\log(ed^2/s),$$

$$\|\mathbf{B}-\mathbf{B}'\|_p^p \geq r/8 \geq ((s-d)/2-1)/8 \geq s/64, \quad p \geq 1, \tag{C.23}$$

where $c > 0$ is an absolute constant and we used $d \geq 4$ and $s \geq 2d$ for (C.23). Now, for $w > 0$, set

$$\Omega_w := \left\{\tfrac{1}{2}\mathbb{I}_{d\times d} + w\mathbf{B} : \mathbf{B} \in \widetilde{\Omega}_r\right\}.$$

Note that $\mathrm{i}\mathbf{B}$ is unitarily diagonizable for every $\mathbf{B} \in \widetilde{\Omega}$, because of its antisymmetricity. Hence, for any $\mathbf{A} \in \Omega_w$, there exist a unitary matrix $\mathbf{U}$ and a real diagonal matrix $\mathbf{D}$ such that, for $t > 0$,

$$\|\mathrm{e}^{-t\mathbf{A}}\|_{\mathrm{Sp}} \leq \mathrm{e}^{-\frac{t}{2}}\|\mathbf{U}\mathrm{e}^{-\mathrm{i}tw\mathbf{D}}\mathbf{U}^*\|_{\mathrm{Sp}} \leq \mathrm{e}^{-\frac{t}{2}}\|\mathrm{e}^{-\mathrm{i}tw\mathbf{D}}\|_{\mathrm{Sp}} = \mathrm{e}^{-\frac{t}{2}}.$$

Thus, $\|\mathrm{e}^{-t\mathbf{A}}\|_{\mathrm{Sp}} \to 0$ as $t \to \infty$ holds for any $\mathbf{A} \in \Omega_w$, implying that $\Omega_w \subset M_+(\mathbb{R}^d)$. Furthermore, by (C.22), $\|\mathbf{A}\|_0 \leq s$ holds for all $\mathbf{A} \in \Omega_w$. Lemma 6 in [11] entails that, under $\mathbb{P}_{\mathbf{A}}$,

$$\mathbf{C}_\infty = \mathbb{I}_{d\times d},$$

and (C.22) gives for $\mathbf{A}, \mathbf{A}' \in \Omega_w$ that

$$\|\mathbf{A}-\mathbf{A}'\|_2^2 \leq 4w^2r.$$

For the Kullback–Leibler divergence of the probability measures associated to $\mathbf{A}, \mathbf{A}' \in \Omega_w$, we then get as in the proof of Corollary 3 in [11]

$$\mathrm{KL}(\mathbb{P}_{\mathbf{A}}\|\mathbb{P}_{\mathbf{A}'}) = \frac{T}{2}\operatorname{tr}((\mathbf{A}'-\mathbf{A})\mathbf{C}_\infty(\mathbf{A}'-\mathbf{A})^\top) \leq 2Tw^2r,$$

and (C.23) implies for $p \geq 1$

$$\|\mathbf{A}-\mathbf{A}'\|_p^p \geq sw^p/64.$$

Now, choosing $w_0 > 0$ such that $w_0^2 = cT^{-1}\log(ed^2/s)/25$ and setting $\Omega = \Omega_{w_0}$, it holds for all $\mathbf{A}, \mathbf{A}' \in \Omega$

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_{\mathbf{A}}\|\mathbb{P}_{\mathbf{A}'}) &\leq 2cr\log(ed^2/s)/25 < \log(|\Omega|)/8,\\ \|\mathbf{A}-\mathbf{A}'\|_p^p &\geq s(cT^{-1}\log(ed^2/s))^{p/2}/320,\end{aligned}$$

which completes the proof by applying Theorem 2.7 in [21]. ∎

Using the indicator loss $\ell(u) = \mathbb{1}\{u \geq 1\}$, Theorem C.8 yields, e.g., the following statement for $d \geq 4$ and $s \geq 2d$: For any estimator $\widehat{\mathbf{A}}$, there exists some s-sparse matrix $\mathbf{A}_0 \in M_+(\mathbb{R}^d)$ such that, with $\mathbb{P}^{\mathbf{A}_0}$-probability of at least c_0,

$$\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_2 \geq c_1\sqrt{\frac{s\log(ed^2/s)}{T}},$$

for some constants $c_0, c_1 > 0$. Note that this lower bound matches the upper bound for the Slope estimator of the drift parameter $\mathbf{A}_0$ in the model (C.4) with $\boldsymbol{\Sigma} = \mathrm{Id}_{d\times d}$ which was derived in Corollary C.6(b). The restrictions $d \geq 4$ and $s \geq 2d$ are consequences of the assumption $s \in [1, p/2]$ in Lemma F.1 of [4] and the construction of the hypotheses. More specifically, as we want to apply Lemma 6 in [11] for showing that $\mathbf{C}_\infty$ is identical for all hypotheses, we use a similar construction by antisymmetric matrices as in Lemma 5 of the same reference. However, as the constructed set then needs to contain matrices with sparsity $\leq s - d$, we are in need of a lower bound of the form $s - d \geq cs$ for some constant $c > 0$ which holds for *all* $d \geq 4$. This is reflected in (C.23) in the proof of Theorem C.8, where it can also be seen that the assumption $s \geq 2d$ solves this problem.

C.3 Deviation inequalities

Having presented our main results for the Lasso and Slope estimator, respectively, in the last section, we now give the two central deviation inequalities used in the proofs. In particular, the approach to bounding the stochastic error introduced in Section C.3.2 provides the key to achieving the optimal rate of convergence.

C.3.1 *Property of restricted eigenvalue type*

In previous works on Lasso and Slope estimators, the analysis relied on the so-called restricted eigenvalue property, which in our setting corresponds to the assumption

$$\inf_{\mathbf{B}\in\mathcal{C}} \frac{\|\mathbf{B}X\|_{L^2}^2}{\|\mathbf{B}\|_2^2} \geq c_{\mathrm{REP}},$$

for certain cones $\mathcal{C} \subset \mathbb{R}^{d\times d}$ and a constant $c_{\mathrm{REP}} > 0$. It was discovered in [11] and [9] that this property holds with high probability in the context of Lasso estimation for Gaussian OU processes as soon as assumptions corresponding to ($\mathcal{H}$) are fulfilled (see Theorem 3 in [11] and Theorem 3.3 in [9]). As these findings essentially rely on the discretization procedure presented in Lemma F.2 of [3], it is not surprising that similar results can be obtained in the Lévy-driven case.

PROPOSITION C.9. *For any $T > 0$, it holds*

$$Q_T\left(\frac{\kappa_{\min}}{2}\right) \subset \left\{ \inf_{\mathbf{B}\in\mathbb{R}^{d\times d}\setminus\{0\}} \frac{\|\mathbf{B}X\|_{L^2}^2}{\|\mathbf{B}\|_2^2} \geq \frac{\kappa_{\min}}{2} \right\}$$

and, for any $r > 0$, we have

$$\mathbb{P}(Q_T(r)) \geq 1 - (21(d \wedge \mathrm{e}))^d H\left(T, \frac{r}{3d}\right). \tag{C.24}$$

Proof. First note that

$$\begin{aligned}
\left\{ \inf_{\mathbf{B}\in\mathbb{R}^{d\times d}\setminus\{0\}} \frac{\|\mathbf{B}X\|_{L^2}^2}{\|\mathbf{B}\|_2^2} < \frac{\kappa_{\min}}{2} \right\} &= \left\{ \kappa_{\min} - \inf_{\mathbf{B}\in\mathbb{R}^{d\times d}\setminus\{0\}} \frac{\operatorname{tr}(\mathbf{B}\widehat{\mathbf{C}}_T\mathbf{B}^\top)}{\|\mathbf{B}\|_2^2} > \frac{\kappa_{\min}}{2} \right\} \\
&= \left\{ \inf_{\mathbf{B}\in\mathbb{R}^{d\times d}\setminus\{0\}} \frac{\operatorname{tr}(\mathbf{B}\mathbf{C}_\infty\mathbf{B}^\top)}{\|\mathbf{B}\|_2^2} - \inf_{\mathbf{B}\in\mathbb{R}^{d\times d}\setminus\{0\}} \frac{\operatorname{tr}(\mathbf{B}\widehat{\mathbf{C}}_T\mathbf{B}^\top)}{\|\mathbf{B}\|_2^2} > \frac{\kappa_{\min}}{2} \right\} \\
&\subset \left\{ \sup_{\mathbf{B}\in\mathbb{B}_2(1)} |\operatorname{tr}(\mathbf{B}(\widehat{\mathbf{C}}_T - \mathbf{C}_\infty)\mathbf{B}^\top)| > \frac{\kappa_{\min}}{2} \right\},
\end{aligned}$$

where we used the identities $\|\mathbf{B}X\|_{L^2}^2 = \operatorname{tr}(\mathbf{B}\widehat{\mathbf{C}}_T\mathbf{B}^\top)$ and

$$\operatorname{tr}(\mathbf{B}\mathbf{C}_\infty\mathbf{B}^\top) = \operatorname{vec}(\mathbf{B}^\top)^\top(\mathbb{I}_{d\times d} \otimes \mathbf{C}_\infty)\operatorname{vec}(\mathbf{B}^\top),$$

with $\otimes$ denoting the Kronecker product, combined with the min-max theorem. This proves the first assertion. Since, for $\mathbf{B} \in \mathbb{R}^{d\times d}$,

$$\operatorname{tr}(\mathbf{B}(\widehat{\mathbf{C}}_t - \mathbf{C}_\infty)\mathbf{B}^\top) = \sum_{i=1}^d \mathbf{B}_{i,\cdot}(\widehat{\mathbf{C}}_t - \mathbf{C}_\infty)\mathbf{B}_{i,\cdot}^\top, \tag{C.25}$$

where $\mathbf{B}_{i,\cdot}$ denotes the i-th row of $\mathbf{B}$, assumption ($\mathcal{H}$) now implies together with Lemma 7 in [11] for any $r > 0$

$$\begin{aligned}
\mathbb{P}\left(\sup_{\mathbf{B}\in\mathbb{B}_2(1)} |\operatorname{tr}(\mathbf{B}(\widehat{\mathbf{C}}_T - \mathbf{C}_\infty)\mathbf{B}^\top)| \geq r \right) &\leq \mathbb{P}\left(\sup_{u\in\mathbb{R}^d:\|u\|\leq 1} |u^\top(\widehat{\mathbf{C}}_T - \mathbf{C}_\infty)u| \geq \frac{r}{d} \right) \\
&\leq (21(d \wedge \mathrm{e}))^d H\left(T, \frac{r}{3d}\right),
\end{aligned}$$

i.e., (C.24) holds. ∎

As can be seen in Proposition C.9, we choose $\mathcal{C} = \mathbb{R}^{d\times d} \setminus \{\mathbf{0}\}$. This may seem counterintuitive at first, since the cones used in previous works are much smaller than the whole space. There are two main reasons for our choice. Firstly, we employ Proposition C.9 in Section C.3.2 for obtaining a deviation inequality for the stochastic error term involving $\boldsymbol{\epsilon}_T$ in Lemma C.2. As this deviation inequality must hold for all matrices, we have to choose $\mathcal{C}$ in the specified way. Secondly, as our framework concerns sparsity instead of row-sparsity, it becomes hard to exploit the property (C.25). A good indicator for this is the difference between the threshold time index T_0 in Theorem

3.3 of [9] and Corollary 4 in [11]: In the row-sparse setting, the dominating term wrt sparsity and dimension is of the form $s\log(d/s)$, whereas in the sparse setting the corresponding term is given as $s\log(d)$, which is clearly larger since the sparsity always dominates the row-sparsity. This is in fact a direct consequence of the different concentration results stated in Lemma 6.2 in [9] and Lemma 8 in [11], which can be seen as analogues to Proposition C.9. Recall that [9] assumes the true parameter to be sparse whereas in [11] the parameter is assumed to be row-sparse.

C.3.2 *Bounding the stochastic error*

We now prove a uniform deviation inequality for the stochastic error term involving $\boldsymbol{\epsilon}_T$ in the basic inequality stated in Lemma C.2. As we want to obtain results for the Slope estimator, we are in need of a statement similar to Theorem 4.1 in [4]. However, the proof of said theorem strongly relies on the noise being normally distributed, as it uses as a key argument the classical concentration result for Lipschitz functions of Gaussian random variables (see, e.g., Theorem 5.2.2 in [22]). Since the noise in our case is given by an Itō integral, we are not able to directly employ the same techniques as used in [4]. We overcome this challenge by noting that Proposition C.9 allows us to find a uniform bound for the quadratic variation of the noise term, which holds with high probability. This implies that the noise is sub-Gaussian with high probability, thus enabling us to apply Talagrand's generic chaining device and majorizing measure theorem (see e.g. Chapter 2 in [20] or Section 8.6 in [22]) to return to the Gaussian setting. These findings yield the following important auxiliary result, for which we define the Gaussian width $w(\mathcal{D})$ and radius $\operatorname{rad}(\mathcal{D})$ of a set $\mathcal{D}\subset\mathbb{R}^{d\times d}$ by setting

$$w(\mathcal{D}) := \mathbb{E}\left[\sup_{\mathbf{B}\in\mathcal{D}}\langle \operatorname{vec}(\mathbf{B}), Z\rangle\right] \quad \text{and} \quad \operatorname{rad}(\mathcal{D}) := \sup_{\mathbf{B}\in\mathcal{D}}\|\mathbf{B}\|_2,$$

where $Z\sim\mathcal{N}(0,\mathbb{I}_{d^2\times d^2})$. Recall the definition of $\boldsymbol{\epsilon}_T$ in (C.12).

Lemma C.10. *There exists a universal constant $c_0>0$ such that, for any $\mathcal{D}\subset\mathbb{R}^{d\times d}$, it holds for all $u>0$*

$$\mathbb{P}\left(\sup_{\mathbf{B}\in\mathcal{D}}\langle\boldsymbol{\epsilon}_T,\mathbf{B}\rangle_2\,\mathbb{1}_{Q_T(\kappa_{\max})}\le c_0\sqrt{\frac{12\kappa_{\max}}{T}}(w(\mathcal{D})+u\operatorname{rad}(\mathcal{D}))\right)\ge 1-2\exp(-u^2).$$

Proof. Note first that, by the min-max theorem, for all $\mathbf{B}\in\mathbb{R}^{d\times d}$,

$$\operatorname{tr}(\mathbf{B}\mathbf{C}_\infty\mathbf{B}^\top)=\operatorname{vec}(\mathbf{B}^\top)^\top(\mathbb{I}_{d\times d}\otimes\mathbf{C}_\infty)\operatorname{vec}(\mathbf{B}^\top)\le\kappa_{\max}\|\mathbf{B}\|_2^2,$$

and hence

$$\begin{aligned}
Q_T(\kappa_{\max}) &= \left\{\sup_{\mathbf{B}\in\mathbb{B}_2(1)}\operatorname{tr}(\mathbf{B}(\widehat{\mathbf{C}}_T-\mathbf{C}_\infty)\mathbf{B}^\top)\le\kappa_{\max}\right\}\\
&\subset\left\{\forall\mathbf{B}_1,\mathbf{B}_2\in\mathcal{D}:\operatorname{tr}((\mathbf{B}_1-\mathbf{B}_2)(\widehat{\mathbf{C}}_T-\mathbf{C}_\infty)(\mathbf{B}_1-\mathbf{B}_2)^\top)<\kappa_{\max}\|\mathbf{B}_1-\mathbf{B}_2\|_2^2\right\}\\
&\subset\left\{\forall\mathbf{B}_1,\mathbf{B}_2\in\mathcal{D}:\operatorname{tr}((\mathbf{B}_1-\mathbf{B}_2)\widehat{\mathbf{C}}_T(\mathbf{B}_1-\mathbf{B}_2)^\top)<2\kappa_{\max}\|\mathbf{B}_1-\mathbf{B}_2\|_2^2\right\}.
\end{aligned}$$

Let $\mathbf{B}_1 \neq \mathbf{B}_2 \in \mathcal{D}$ be given. Then, using Bernstein's inequality for continuous martingales, we get

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{\langle \boldsymbol{\epsilon}_T, \mathbf{B}_1 - \mathbf{B}_2\rangle_2^2 \, \mathbb{1}_{Q_T(\kappa_{\max})}}{12T^{-1}\kappa_{\max}\|\mathbf{B}_1 - \mathbf{B}_2\|_2^2}\right)\right] \\
&\quad \leq \int_0^\infty \mathbb{P}\left(\exp\left(\frac{\langle \boldsymbol{\epsilon}_T, \mathbf{B}_1 - \mathbf{B}_2\rangle_2^2 \, \mathbb{1}_{Q_T(\kappa_{\max})}}{12T^{-1}\kappa_{\max}\|\mathbf{B}_1 - \mathbf{B}_2\|_2^2}\right) > u\right) \mathrm{d}u \\
&\quad \leq 1 + \int_1^\infty \mathbb{P}\left(|\langle \boldsymbol{\epsilon}_T, \mathbf{B}_1 - \mathbf{B}_2\rangle| \, \mathbb{1}_{Q_T(\kappa_{\max})} > \sqrt{12T^{-1}\kappa_{\max}\log(u)}\|\mathbf{B}_1 - \mathbf{B}_2\|_2\right) \mathrm{d}u \\
&\quad = 1 + \int_1^\infty \mathbb{P}\left(\left|\int_0^T ((\mathbf{B}_1 - \mathbf{B}_2)X_s)^\top \, \mathrm{d}W_s\right| > \sqrt{12T\kappa_{\max}\log(u)}\|\mathbf{B}_1 - \mathbf{B}_2\|_2, Q_T(\kappa_{\max})\right) \mathrm{d}u \\
&\quad \leq 1 + \int_1^\infty \mathbb{P}\left(\left|\int_0^T ((\mathbf{B}_1 - \mathbf{B}_2)X_s)^\top \, \mathrm{d}W_s\right| > \sqrt{12T\kappa_{\max}\log(u)}\|\mathbf{B}_1 - \mathbf{B}_2\|_2, \right. \\
&\qquad\qquad \left. \int_0^T \|(\mathbf{B}_1 - \mathbf{B}_2)X_s\|^2 \, \mathrm{d}s < 2T\kappa_{\max}\|\mathbf{B}_1 - \mathbf{B}_2\|_2^2\right) \mathrm{d}u \\
&\quad \leq 1 + 2\int_1^\infty u^{-3} \, \mathrm{d}u = 2.
\end{aligned}$$

This shows that $(\langle \boldsymbol{\epsilon}_T, \mathbf{B}\rangle_2 \, \mathbb{1}_{Q_T(\kappa_{\max})})_{\mathbf{B}\in\mathcal{D}}$ is sub-Gaussian in the sense of Definition 2.5.6 in [22], since

$$\|\langle \boldsymbol{\epsilon}_T, \mathbf{B}_1\rangle_2 - \langle \boldsymbol{\epsilon}_T, \mathbf{B}_2\rangle_2\|_{\psi_2}^2 \leq \frac{12}{T}\kappa_{\max}\|\mathbf{B}_1 - \mathbf{B}_2\|_2^2,$$

holds, where $\|Y\|_{\psi_2} := \inf\{t > 0 : \mathbb{E}[\exp(Y^2/t^2) \leq 2]\}$, for any random variable Y. Hence, we can apply Exercise 8.6.5 in [22], which yields for any $\mathcal{D} \subset \mathbb{R}^{d\times d}$ and for all $u > 0$ the asserted inequality. ∎

Combining Lemma C.10 with the concentration property for Lipschitz functions of Gaussian random variables and Proposition E.2 in [4], which allow us to bound the Gaussian width of the relevant set, we arrive at the following proposition.

PROPOSITION C.11. *For $0 < \varepsilon_0 < 1$, set*

$$\|\mathbf{B}\|_S := \|\mathbf{B}\|_* \vee \sqrt{\log(4\varepsilon_0^{-1})}\|\mathbf{B}\|_2, \quad \forall \mathbf{B} \in \mathbb{R}^{d\times d}. \tag{C.26}$$

Then, for all $T > 0$,

$$\mathbb{P}\left(\sup_{\mathbf{B}\in\mathbb{R}^{d\times d}, \mathbf{B}\neq 0} \frac{\langle \boldsymbol{\epsilon}_T, \mathbf{B}\rangle_2}{\|\mathbf{B}\|_S} \, \mathbb{1}_{Q_T(\kappa_{\max})} \leq c_* \sqrt{\frac{\kappa_{\max}}{T}}\right) \geq 1 - \frac{\varepsilon_0}{2},$$

where, for c_0 being the constant from Lemma C.10,

$$c_* := c_0\left(\sqrt{\frac{3\pi}{\log(2)}} + \sqrt{300}\right).$$

Proof. First note that

$$\sup_{\mathbf{B}\in\mathbb{R}^{d\times d},\mathbf{B}\neq 0}\frac{\langle \boldsymbol{\epsilon}_T,\mathbf{B}\rangle_2}{\|\mathbf{B}\|_S}=\sup_{\mathbf{B}\in\mathcal{D}_*}\langle \boldsymbol{\epsilon}_T,\mathbf{B}\rangle_2,\qquad \text{where } \mathcal{D}_*:=\left\{\mathbf{B}\in\mathbb{R}^{d\times d}:\|\mathbf{B}\|_S=1\right\}.$$

Thus, to apply Lemma C.10, we need to bound $w(\mathcal{D}_*)$ and $\mathrm{rad}(\mathcal{D}_*)$. Therefore, let $Z\sim\mathcal{N}(0,\mathbb{I}_{d^2\times d^2})$, and note that the function

$$f\colon\mathbb{R}^{d^2}\to\mathbb{R},\quad v\mapsto f(v):=\sup_{\mathbf{B}\in\mathcal{D}_*}\langle \mathrm{vec}(\mathbf{B}),v\rangle,$$

is Lipschitz continuous with Lipschitz constant $\log(4)^{-1/2}$ wrt the Euclidean distance. Thus, equation (1.4) in [12] gives

$$\begin{aligned}w(\mathcal{D}_*)=\mathbb{E}[f(Z)]&\le\int_0^\infty\mathbb{P}(|f(Z)-\mathrm{Med}(f(Z))|\ge u)\,\mathrm{d}u+\mathrm{Med}(f(Z))\\&\le\int_0^\infty\exp\left(-\frac{\log(4)u^2}{2}\right)\mathrm{d}u+\mathrm{Med}(f(Z))\\&=\sqrt{\frac{\pi}{4\log(2)}}+\mathrm{Med}(f(Z)).\end{aligned}$$

Combining Proposition E.2 in [4] with

$$\begin{aligned}f(Z)&\le\sup_{\mathbf{B}\in\mathbb{R}^{d\times d}:\|\mathbf{B}\|_*\le 1}\langle\mathrm{vec}(\mathbf{B}),Z\rangle\\&=\sup_{\mathbf{B}\in\mathbb{R}^{d\times d}:\|\mathbf{B}\|_*\le 1}\sum_{i=1}^{d^2}\mathrm{vec}(\mathbf{B})_i^{\#}\sqrt{\log(2d^2/i)}\frac{Z_i^{\#}}{\sqrt{\log(2d^2/i)}}\\&\le\max_{i=1,\dots,d^2}\frac{Z_i^{\#}}{\sqrt{\log(2d^2/i)}}\end{aligned}$$

yields $\mathrm{Med}(f(Z))\le 4$, implying that

$$w(\mathcal{D}_*)\le\sqrt{\frac{\pi}{4\log(2)}}+4.$$

Since $\mathrm{rad}(\mathcal{D}_*)\le\log(4/\varepsilon_0)^{-1/2}$ trivially holds, the assertion follows. ∎

C.4 Discussion of assumption (ℋ) and outlook

C.4.1 *Sufficient conditions for assumption (ℋ)*

We first recall the results of [11] and [9] on assumption (ℋ) for the case where the BDLP is given as a standard Wiener process. Moreover, we prove that in the Lévy-driven case (ℋ) is satisfied as soon as the Lévy measure of the BDLP admits a fourth moment.

The Gaussian case As mentioned above, both [11] and [9] assume that $\mathbf{Z}$ is a standard Wiener process, i.e., the characteristic triplet of $\mathbf{Z}$ is given by $(0, \mathbb{I}_{d\times d}, 0)$. In this case, [11] were able to show that $(\mathcal{H})$ holds under assumptions implied by $(\mathcal{A}_0)$ if $\mathbf{A}_0$ is symmetric. This result was achieved by exploiting that symmetricity of $\mathbf{A}_0$ implies μ to fulfill a log-Sobolev inequality, which then yields $(\mathcal{H})$ by Theorem 2.1 of [8]. [9] extended this finding to the general case of possibly non-symmetric $\mathbf{A}_0$, i.e., $\mathbf{A}_0 \in M_+(\mathbb{R}^d)$ already implies $(\mathcal{H})$ in the classical Gaussian case. The proof of this result relies on Malliavin calculus methods, especially Theorem 4.1 in [15]. In both papers, the function H in $(\mathcal{H})$ is of the form

$$H(T, r) = 2\exp(-TH_0(r)), \quad T, r > 0,$$

where H_0 is positive and increasing. For the sake of completeness, we state the findings of [9] below.

PROPOSITION C.12 (cf. Proposition 3.2 in [9]). *Assume that the characteristic triplet of the BDLP $\mathbf{Z}$ is given by $(0, \mathbb{I}_{d\times d}, 0)$. Denote by $\lambda_1, \dots, \lambda_d$ the eigenvalues of $\mathbf{A}_0$, and let $\mathbf{P}_0$ be the matrix such that $\mathbf{A}_0 = \mathbf{P}_0 \operatorname{diag}(\lambda_1, \dots, \lambda_d)\mathbf{P}_0^{-1}$. Then, for all $r > 0$,*

$$\sup_{u\in\mathbb{R}^d : \|u\|=1} \mathbb{P}\Big(|u^\top(\widehat{\mathbf{C}}_T - \mathbf{C}_\infty)u| \geq r\Big) \leq 2\exp(-TH_0(r)),$$

where

$$H_0(r) = \frac{\tau_0 r^2}{8\lambda_{\max}(\mathbf{C}_\infty)p_0(r + \lambda_{\max}(\mathbf{C}_\infty))}, \quad r > 0,$$

with $\tau_0 = \min(\operatorname{Re}(\lambda_i))$ and $p_0 = \|\mathbf{P}_0\|_{\mathrm{Sp}}\|\mathbf{P}_0^{-1}\|_{\mathrm{Sp}}$.

The Lévy-driven case Since the derivation of the results in the previous paragraph strongly relies on the Gaussianity of $\boldsymbol{X}$, achieving similar results in the Lévy-driven setting is a challenging task. However, an application of the stochastic Fubini theorem (similar to the proof of equation (2.17) in [2]), combined with classical martingale results, yields that $(\mathcal{H})$ is fulfilled as soon as the Lévy measure ν of the BDLP admits a fourth moment.

PROPOSITION C.13. *Assume that ν admits a fourth moment. Then, there exists a constant $c > 0$ such that, for all $u \in \mathbb{R}^d$ fulfilling $\|u\| \leq 1$,*

$$\mathbb{P}\Big(|u^\top(\widehat{\mathbf{C}}_T - \mathbf{C}_\infty)u| \geq r\Big) \leq \frac{c}{t(r \wedge r^2)} + \frac{c}{(tr)^2}.$$

In particular, Assumption $(\mathcal{H})$ is fulfilled.

The proof of Proposition C.13, which also contains the explicit value of the constant c, can be found in Appendix C.III.

C.4.2 *Outlook*

Following the pioneering work of [11] and [9] which clarified the statistical foundations of a high-dimensional modelling of the classical OU process, we have extended the investigation to the Lévy-driven case. In particular, this requires finding tools that do not explicitly rely on Gaussian structures.

As usual in high-dimensional statistics, the proof of our main results (Propositions C.3 and C.5) is based on two central elements: On the one hand, we confine ourselves to the study of a benign event, in our context of the form

$$\mathcal{C} := \left\{ \inf_{\mathbf{B}\in\mathbb{R}^{d\times d}\setminus\{0\}} \frac{\|\mathbf{B}X\|_{L^2}^2}{\|\mathbf{B}\|_2^2} > \frac{\kappa_{\min}}{2} \right\} \bigcap \left\{ \sup_{\mathbf{B}\in\mathbb{R}^{d\times d}:\mathbf{B}\neq 0} \frac{\langle \boldsymbol{\epsilon}_T, \mathbf{B}\rangle_2}{\|\mathbf{B}\|_S} \leq c_* \sqrt{\frac{\kappa_{\max}}{T}} \right\}.$$

As becomes clear in the proof of the aforementioned propositions, the investigation on this event is driven by purely deterministic arguments, which can be developed analogously to the high-dimensional linear regression model as it is studied in [4]. It then remains to show that the event $\mathcal{C}$ is of high probability.

With respect to the first sub-event, this amounts to verifying a property of restricted eigenvalue type. Similarly to the Gaussian case, we identified a concentration condition (assumption $(\mathcal{H})$) that can be used to show this. Proposition C.13 stated in the previous subsection gives a concrete criterion for $(\mathcal{H})$ to be fulfilled. This result is obviously weaker than its Gaussian counterpart (Proposition C.12) in the sense of the temporal decay not being exponential but polynomial. The primary influence of this is on the value of the threshold value T_0 specified in (C.19) appearing in Corollaries C.4 and C.6, which increases. Nevertheless, as the main results of this paper are developed in such a way that they only rely on assumption $(\mathcal{H})$ in its general form, it would be easy to implement results implying an exponential decay in the Lévy-driven case to achieve values of T_0 similar to the Gaussian case.

The second sub-event of $\mathcal{C}$ involves both the process $\boldsymbol{\epsilon}_T$ (specified in (C.12)) and the norm $\|\cdot\|_S$ (as introduced in (C.26)). At this point, the main differences with the studies of [11] and [9] do not arise because of the structure of the process, but because of the different statistical approach. In fact, controlling the second sub-event provides the key to removing the additional logarithmic factor in the convergence rate. The derivations in Section C.3.2 are therefore of independent interest. As noted in Remark 4.4 of [9], the development of general high-dimensional diffusion models requires a suitable representation of the likelihood function (given in our case by Proposition C.1) and appropriate techniques for proving concentration phenomena. If these ingredients are combined with our techniques for bounding the stochastic error, estimators (of the Lasso or Slope type) that achieve minimax optimal convergence rates might also be formulated in a general diffusion model.

C.5 Simulation study

In this section, we investigate our theoretical results by applying them to simulated data. For this purpose, we compare the errors of the maximum likelihood, Lasso and Slope estimators in different dimensions. Of course, our results were derived in the setting of continuous observations, but they can easily be transferred to the more realistic framework of discrete observations by discretising the integrals involved. The data will always be generated by an Euler–Maruyama scheme with step size $\delta = 10^{-2}$. We choose this value for δ because Figure 6 in [11] indicates that the quality of estimation does not improve with a smaller step size. Since it is well known that choosing tuning parameters by theoretical results leads to too large values, we select the tuning parameters by cross-validation, with the first 80% of the path acting as the training set and the remainder as the validation set. More precisely, we define a candidate set $\Lambda \subset \mathbb{R}_+$, and

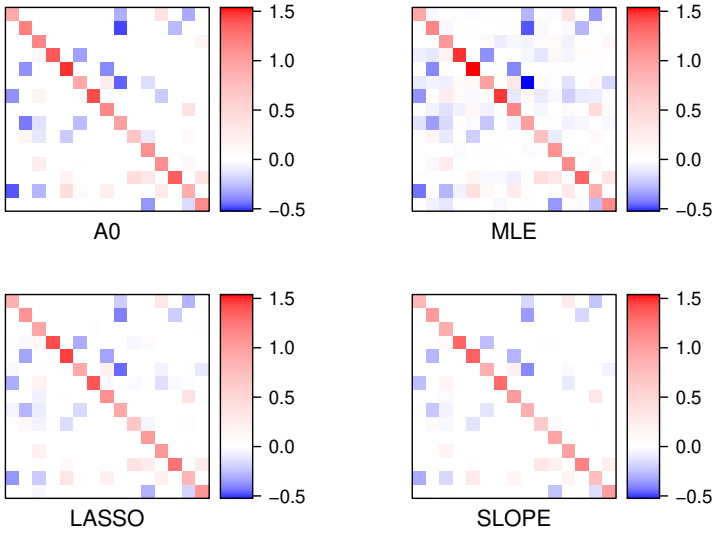

Figure C.5.1: Comparison of the true parameter $\mathbf{A}_0$ with MLE, Lasso and Slope.

for each $\lambda \in \Lambda$, we set

$$\widehat{\mathbf{A}}_\lambda^{\text{lasso}} = \operatorname{argmin}_{\mathbf{A}\in\mathbb{R}^{d\times d}} \mathcal{L}_{[0,0.8T]}(\mathbf{A}) + \lambda\|\mathbf{A}\|_1, \quad \widehat{\mathbf{A}}_\lambda^{\text{slope}} = \operatorname{argmin}_{\mathbf{A}\in\mathbb{R}^{d\times d}} \mathcal{L}_{[0,0.8T]}(\mathbf{A}) + \lambda\|\mathbf{A}\|_*$$

and

$$\widehat{\lambda}^{\text{lasso}} = \operatorname{argmin}_{\lambda\in\Lambda} \frac{\mathcal{L}_{[0.8T,T]}(\widehat{\mathbf{A}}_\lambda^{\text{lasso}})}{\|\widehat{\mathbf{A}}_\lambda^{\text{lasso}}\|_1}, \quad \widehat{\lambda}^{\text{slope}} = \operatorname{argmin}_{\lambda\in\Lambda} \frac{\mathcal{L}_{[0.8T,T]}(\widehat{\mathbf{A}}_\lambda^{\text{slope}})}{\|\widehat{\mathbf{A}}_\lambda^{\text{slope}}\|_*},$$

where $\mathcal{L}_{[0,0.8T]}$ and $\mathcal{L}_{[0.8T,T]}$, respectively, correspond to the negative log-likelihood function computed on the relevant intervals. This then leads to $\widehat{\mathbf{A}}^{\text{lasso}}_{\widehat{\lambda}^{\text{lasso}}}$ and $\widehat{\mathbf{A}}^{\text{slope}}_{\widehat{\lambda}^{\text{slope}}}$ as our final estimators. The candidate set will always be a logarithmic grid with values between 10^{-3} and 10. We choose this particular form of our estimators because it is closer to practice compared to the theoretical definitions in (C.10) and (C.11). For a more in-depth numerical analysis in the Gaussian framework and, in particular, an application to real world financial data, we refer to Section 4 of [11], and for a comparison between Lasso and Dantzig estimators to Section 5 of [9].

In Figure C.5.1, we give a first example of the different estimators compared to the ground truth $\mathbf{A}_0$, which in this case is given as a 15×15 matrix with sparsity ~ 0.2. For comparability, we depict the matrices as heat maps. In this example, we set $T = 300$ and let the matrix $\boldsymbol{\Sigma}$ be generated as a diagonal matrix with entries generated from a uniform distribution with values in $[0, 5]$, and the jumps are given by a composite Poisson process with intensity 10 and Laplace-distributed jump sizes. We choose $\boldsymbol{\Sigma}$ as the diagonal matrix because our results rely on the sparsity of $\boldsymbol{\Sigma}^{-1}\mathbf{A}_0$ and this is the simplest way to preserve the sparsity of $\mathbf{A}_0$.

For more general results, we compare the estimation error in L_1 and Frobenius norm (hereafter referred to as L_2 norm) of the three estimators over 10 iterations for dimensions 10 to 30, with $T = 100$. For each dimension, we generate $\mathbf{A}_0 \in M_+$ with sparsity ~ 0.2 and $\boldsymbol{\Sigma}$ similar to Figure C.5.1, with the only difference that the uniform distribution is now on $[0, 10]$. The jump intensity is given as 5, and the jump sizes are also Laplace distributed. The results of this

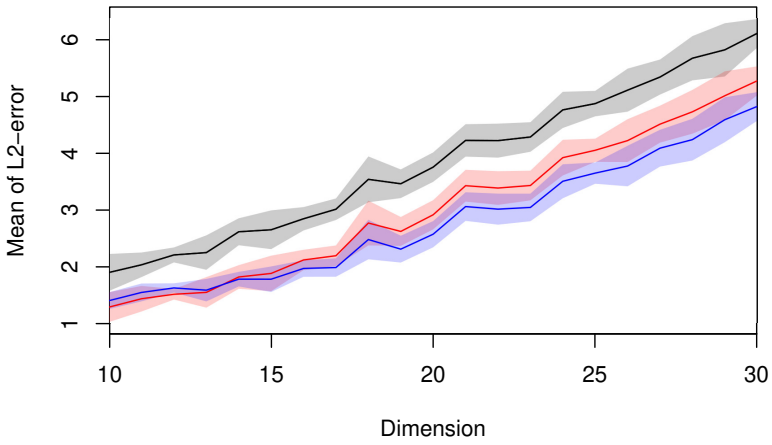

Figure C.5.2: L_2 errors of MLE, Lasso and Slope ± one standard deviation.

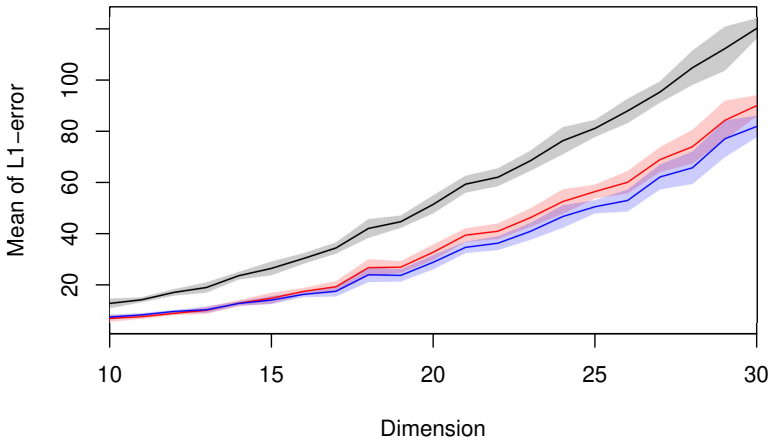

Figure C.5.3: L_1 errors of MLE, Lasso and Slope ± one standard deviation.

simulation study can be seen in Figures C.5.2 and C.5.3. We see that Lasso and Slope constantly outperform the maximum likelihood estimator for both error measures, and that Lasso and Slope behave very similarly, which is in line with our theoretical results. Moreover, the L_2 error grows linearly, while the growth of the L_2 error is of quadratic nature. This also matches our theoretical results.

Appendices

C.I Some results on Lévy processes and Lévy-driven OU processes

We start by presenting some results on Lévy-driven OU processes and Lévy processes, respectively infinitely divisible distributions. For this, recall that a function $g\colon \mathbb{R}^d \to \mathbb{R}$ is called sub-multiplicative if it is nonnegative and there exists a constant $c > 0$ such that

$$g(x+y) \leq cg(x)g(y), \quad \forall x, y \in \mathbb{R}^d.$$

Lemma C.14 (cf. Theorem 25.3 in [17]). *Let $\mathbf{Z} = (Z_t)_{t\geq 0}$ be an $\mathbb{R}^d$-valued Lévy process with Lévy triplet $(b, \mathbf{\Sigma}, \nu)$, and let $g\colon \mathbb{R}^d \to \mathbb{R}$ be a measurable, locally bounded and sub-multiplicative function. Then, $\mathbb{E}[g(Z_t)] < \infty$ holds for all $t > 0$ if and only if $\int_{\|z\|\geq 1} g(z)\nu(\mathrm{d}z) < \infty$.*

In particular, the function $g(x) = (1 \vee \|x\|)^p$ is sub-multiplicative for all $p > 0$ (see Proposition 25.4 in [17]) and thus $\mathbb{E}[\|Z_t\|^p] < \infty$ holds true for all $t > 0$ as soon as $\int_{\|z\|\geq 1} \|z\|^p \nu(\mathrm{d}z) < \infty$ is fulfilled. We continue with the following result, which characterizes the invariant distribution of $\boldsymbol{X}$.

Lemma C.15 (cf. Theorem 4.1 in [18], Proposition 2.2 in [14]). *Assume ($\mathscr{A}_0$). Then, $\boldsymbol{X}$ has a unique invariant distribution μ which is infinitely divisible with characteristic triplet $(b_\mu, \mathbf{C}_\mu, \nu_\mu)$ where*

$$\begin{aligned}
b_\mu &= \mathbf{A}^{-1}b + \int \int_0^\infty \mathrm{e}^{-s\mathbf{A}}z\Big(\mathbb{1}_{\|z\|\leq 1}(\mathrm{e}^{-s\mathbf{A}}z) - \mathbb{1}_{\|z\|\leq 1}(z)\Big)\,\mathrm{d}s\nu(\mathrm{d}z),\\
\mathbf{C}_\mu &= \int_0^\infty \mathrm{e}^{-s\mathbf{A}}\mathbf{C}\mathrm{e}^{-s\mathbf{A}^\top}\,\mathrm{d}s,\\
\nu_\mu(B) &= \int_0^\infty \nu(\mathrm{e}^{s\mathbf{A}}B)\,\mathrm{d}s, \quad \forall B \in \mathcal{B}(\mathbb{R}^d),\\
\mathrm{e}^{s\mathbf{A}}B &:= \big\{y \in \mathbb{R}^d : y = \mathrm{e}^{s\mathbf{A}}x, x \in B\big\}, \quad \forall B \in \mathcal{B}(\mathbb{R}^d).
\end{aligned}$$

Combining these results directly leads to the following corollary.

Corollary C.16. *Consider an $\mathbb{R}^d$-valued Lévy process $\mathbf{Z} = (Z_t)_{t\geq 0}$ with Lévy triplet $(b, \mathbf{\Sigma}, \nu)$, and assume ($\mathscr{A}_0$). Let $p \geq 2$ be given, and suppose that $\int_{\|z\|\geq 1} \|z\|^p \nu(\mathrm{d}z) < \infty$. Then,*

$$\int \|x\|^p \mu(\mathrm{d}x) < \infty.$$

Proof. By Lemmas C.14 and C.15, it suffices to show $\int_{\|z\|\geq 1} \|z\|^p \nu_\infty(\mathrm{d}z) < \infty$. It holds

$$\int_{\|z\|\geq 1} \|z\|^p \nu_\mu(\mathrm{d}z) \leq \int_0^\infty \int \|\mathrm{e}^{-s\mathbf{A}}z\|^p \nu(\mathrm{d}z)\,\mathrm{d}s \leq \int_0^\infty \|\mathrm{e}^{-s\mathbf{A}}\|_2^p\,\mathrm{d}s \int \|z\|^p \nu(\mathrm{d}z) < \infty,$$

since ν is a Lévy measure, $p \geq 2$ and $\mathbf{A} \in M_+(\mathbb{R}^d)$. ∎

Another consequence of Lemma C.15 is that $\kappa_{\min} > 0$ follows from ($\mathscr{A}_0$), since then by assumption the Gaussian part of μ is nontrivial. Hence, the support of μ cannot be contained in a hyperplane of $\mathbb{R}^d$, which would be the case if $\kappa_{\min}$ was equal to 0.

C.II PROOFS FOR SECTION C.2

Proof for Lemma C.2. We adapt the proof of Lemma A.2 in [4]; also cf. the proof of Lemma 3 in [11]. Define the functions f and g by the relations $g(\mathbf{A}) = \mathcal{L}_T(\mathbf{A})$, $f \equiv g + h$. By Proposition C.1, we have that

$$\mathcal{L}_T(\mathbf{A}) = \frac{1}{T}\int_0^T (\mathbf{C}^{-1}\mathbf{A}X_{s-})^\top \,\mathrm{d}X_s^{\mathrm{c}} + \frac{1}{2T}\int_0^T (\boldsymbol{\Sigma}^{-1}\mathbf{A}X_{s-})^\top \boldsymbol{\Sigma}^{-1}\mathbf{A}X_{s-}\,\mathrm{d}s.$$

Note that, under $\mathbb{P}^{\mathbf{0}}$, $X_t^{\mathrm{c}} = \boldsymbol{\Sigma}W_t$, where W is a $\mathbb{P}^{\mathbf{0}}$-Wiener process. Additionally, by Girsanov's theorem,

$$\widetilde{W}_t = W_t + \boldsymbol{\Sigma}^{-1}\mathbf{A}_0\int_0^t X_s\,\mathrm{d}s$$

is a $\mathbb{P}^{\mathbf{A}_0}$-Wiener process. Hence, we can write

$$\begin{aligned}
\mathcal{L}_T(\mathbf{A}) &= \frac{1}{T}\int_0^T (\boldsymbol{\Sigma}^{-1}\mathbf{A}X_s)^\top \,\mathrm{d}\widetilde{W}_s + \frac{1}{2T}\int_0^T (\boldsymbol{\Sigma}^{-1}\mathbf{A}X_s)^\top \boldsymbol{\Sigma}^{-1}\mathbf{A}X_s\,\mathrm{d}s - \frac{1}{T}\int_0^T (\boldsymbol{\Sigma}^{-1}\mathbf{A}X_s)^\top \boldsymbol{\Sigma}^{-1}\mathbf{A}_0X_s\,\mathrm{d}s \\
&= \frac{1}{2T}\operatorname{tr}\left(2\boldsymbol{\Sigma}^{-1}\mathbf{A}\int_0^T X_s\,\mathrm{d}\widetilde{W}_s^\top + \boldsymbol{\Sigma}^{-1}\mathbf{A}\int_0^T X_sX_s^\top\,\mathrm{d}s(\boldsymbol{\Sigma}^{-1}(\mathbf{A}-2\mathbf{A}_0))^\top\right) \\
&= \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{A}\boldsymbol{\epsilon}_T^\top + \frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{A}\widehat{\mathbf{C}}_T(\boldsymbol{\Sigma}^{-1}\mathbf{A})^\top - \boldsymbol{\Sigma}^{-1}\mathbf{A}\widehat{\mathbf{C}}_T(\boldsymbol{\Sigma}^{-1}\mathbf{A}_0)^\top\right) \\
&= \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{A}\boldsymbol{\epsilon}_T^\top + \frac{1}{2}\mathbf{A}\widehat{\mathbf{C}}_T\mathbf{A}^\top\mathbf{C}^{-1} - \mathbf{A}\widehat{\mathbf{C}}_T\mathbf{A}_0^\top\mathbf{C}^{-1}\right),
\end{aligned}$$

where $\boldsymbol{\epsilon}_T$ and $\widehat{\mathbf{C}}_T$ are defined according to (C.12) and (C.13), respectively. The gradient is thus given as $(\boldsymbol{\Sigma}^{-1})^\top\boldsymbol{\epsilon}_T + \mathbf{C}^{-1}(\mathbf{A}-\mathbf{A}_0)\widehat{\mathbf{C}}_T$. Since f is convex, it follows that $\mathbf{0}$ is in the subdifferential of f at $\widehat{\mathbf{A}}$. The Moreau–Rockafellar theorem then gives that there exists $\mathbf{B}$ in the subdifferential of h at $\widehat{\mathbf{A}}$ such that $\mathbf{0} = (\boldsymbol{\Sigma}^{-1})^\top\boldsymbol{\epsilon}_T + \mathbf{C}^{-1}(\widehat{\mathbf{A}}-\mathbf{A}_0)\widehat{\mathbf{C}}_T + \mathbf{B}$. Additionally, $\mathbf{B}$ being in the subdifferential of h at $\widehat{\mathbf{A}}$ implies $\langle \mathbf{B}, \mathbf{A}-\widehat{\mathbf{A}}\rangle_2 \le h(\mathbf{A}) - h(\widehat{\mathbf{A}})$. Consequently,

$$\begin{aligned}
&\|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}}-\mathbf{A}_0)X\|_{L^2}^2 - \|\boldsymbol{\Sigma}^{-1}(\mathbf{A}-\mathbf{A}_0)X\|_{L^2}^2 + \|\boldsymbol{\Sigma}^{-1}(\widehat{\mathbf{A}}-\mathbf{A})X\|_{L^2}^2 \\
&\quad= \langle(\widehat{\mathbf{A}}-\mathbf{A}_0)^\top\mathbf{C}^{-1}(\widehat{\mathbf{A}}-\mathbf{A}_0) - (\mathbf{A}-\mathbf{A}_0)^\top\mathbf{C}^{-1}(\mathbf{A}-\mathbf{A}_0) + (\widehat{\mathbf{A}}-\mathbf{A})^\top\mathbf{C}^{-1}(\widehat{\mathbf{A}}-\mathbf{A}), \widehat{\mathbf{C}}_T\rangle_2 \\
&\quad= 2\langle(\widehat{\mathbf{A}}-\mathbf{A}_0)^\top\mathbf{C}^{-1}(\widehat{\mathbf{A}}-\mathbf{A}), \widehat{\mathbf{C}}_T\rangle_2 \\
&\quad= 2\langle(\widehat{\mathbf{A}}-\mathbf{A})^\top, \widehat{\mathbf{C}}_T(\widehat{\mathbf{A}}-\mathbf{A}_0)^\top\mathbf{C}^{-1}\rangle_2 \\
&\quad= 2\langle(\mathbf{A}-\widehat{\mathbf{A}}), (\boldsymbol{\Sigma}^{-1})^\top\boldsymbol{\epsilon}_T + \mathbf{B}\rangle_2 \\
&\quad\le 2(\langle\boldsymbol{\Sigma}^{-1}(\mathbf{A}-\widehat{\mathbf{A}}), \boldsymbol{\epsilon}_T\rangle_2 + h(\mathbf{A}) - h(\widehat{\mathbf{A}})).
\end{aligned}$$

■

C.III PROOFS FOR SECTION C.4.1

Proof of Proposition C.13. Let $u \in \mathbb{R}^d$ be given such that $\|u\| \le 1$. Recall that, for $s > 0$, X_s is given explicitly as

$$X_s = \mathrm{e}^{-s\mathbf{A}}X_0 + \int_0^s \mathrm{e}^{-(s-r)\mathbf{A}}\,\mathrm{d}Z_r.$$

This implies that

$$(u^\top X_s)^2 = (u^\top \mathrm{e}^{-s\mathbf{A}} X_0)^2 + 2(u^\top \mathrm{e}^{-s\mathbf{A}} X_0)(u^\top Y_s) + (u^\top Y_s)^2,$$

where

$$Y_s := \int_0^s \mathrm{e}^{-(s-r)\mathbf{A}}\,\mathrm{d}Z_r.$$

Now, by the Lévy– Itō decomposition, for all $s > 0$,

$$\begin{aligned} Y_s = \int_0^s \mathrm{e}^{-(s-r)\mathbf{A}} b\,\mathrm{d}r + \int_0^s \mathrm{e}^{-(s-r)\mathbf{A}}\boldsymbol{\Sigma}\,\mathrm{d}W_r + \int_0^s \int_{|z|\geq 1} \mathrm{e}^{-(s-r)\mathbf{A}} z N(\mathrm{d}r,\mathrm{d}z) \\ + \int_0^s \int_{|z|<1} \mathrm{e}^{-(s-r)\mathbf{A}} z \widetilde{N}(\mathrm{d}r,\mathrm{d}z), \end{aligned}$$

which allows us to apply Itō's formula (see e.g. Theorem 4.4.7 in [1]). It gives

$$\begin{aligned} &(u^\top Y_s)^2 \\ &= 2\int_0^s (u^\top Y_{r-}) u^\top \mathrm{e}^{-(s-r)\mathbf{A}} b\,\mathrm{d}r + 2\int_0^s (u^\top Y_{r-}) u^\top \mathrm{e}^{-(s-r)\mathbf{A}}\boldsymbol{\Sigma}\,\mathrm{d}W_r + \int_0^s u^\top \mathrm{e}^{-(s-r)\mathbf{A}}\mathbf{C}\mathrm{e}^{-(s-r)\mathbf{A}^\top} u\,\mathrm{d}r \\ &\quad + \int_0^s \int_{\|z\|\geq 1} (u^\top (Y_{r-} + \mathrm{e}^{-(s-r)\mathbf{A}} z))^2 - (u^\top Y_{r-})^2 N(\mathrm{d}r,\mathrm{d}z) \\ &\quad + \int_0^s \int_{\|z\|<1} (u^\top (Y_{r-} + \mathrm{e}^{-(s-r)\mathbf{A}} z))^2 - (u^\top Y_{r-})^2 \widetilde{N}(\mathrm{d}r,\mathrm{d}z) \\ &\quad + \int_0^s \int_{\|z\|<1} (u^\top (Y_{r-} + \mathrm{e}^{-(s-r)\mathbf{A}} z))^2 - (u^\top Y_{r-})^2 - 2(u^\top Y_{r-}) u^\top \mathrm{e}^{-(s-r)\mathbf{A}} z \nu(\mathrm{d}z)\,\mathrm{d}r \\ &= 2\int_0^s (u^\top Y_{r-}) u^\top \mathrm{e}^{-(s-r)\mathbf{A}} b^*\,\mathrm{d}r + 2\int_0^s (u^\top Y_{r-}) u^\top \mathrm{e}^{-(s-r)\mathbf{A}}\boldsymbol{\Sigma}\,\mathrm{d}W_r + \int_0^s u^\top \mathrm{e}^{-(s-r)\mathbf{A}}\mathbf{C}\mathrm{e}^{-(s-r)\mathbf{A}^\top} u\,\mathrm{d}r \\ &\quad + \int_0^s \int 2(u^\top Y_{r-}) u^\top \mathrm{e}^{-(s-r)\mathbf{A}} z + (u^\top \mathrm{e}^{-(s-r)\mathbf{A}} z)^2 \widetilde{N}(\mathrm{d}r,\mathrm{d}z) + \int_0^s \int (u^\top \mathrm{e}^{-(s-r)\mathbf{A}} z)^2 \nu(\mathrm{d}z)\,\mathrm{d}r, \end{aligned} \tag{C.27}$$

where $b^* = b + \int_{\|z\|\geq 1} z\nu(\mathrm{d}z)$. Stationarity of $\boldsymbol{X}$ implies, for any $s \geq 0$,

$$\begin{aligned} (u^\top X_s)^2 - \int (u^\top x)^2 \mu(\mathrm{d}x) &= (u^\top X_s)^2 - \mathbb{E}[(u^\top X_s)^2] \\ &= u^\top \mathrm{e}^{-s\mathbf{A}} (X_0 X_0^\top - \mathbb{E}[X_0 X_0^\top]) \mathrm{e}^{-s\mathbf{A}^\top} u \\ &\quad + 2u^\top \mathrm{e}^{-s\mathbf{A}} (X_0 Y_s^\top - \mathbb{E}[X_0 Y_s^\top]) u + (u^\top Y_s)^2 - \mathbb{E}[(u^\top Y_s)^2], \end{aligned}$$

and (C.27) gives

$$\begin{aligned} (u^\top Y_s)^2 - \mathbb{E}[(u^\top Y_s)^2] = 2\int_0^s u^\top (Y_{r-} - \mathbb{E}[Y_{r-}]) u^\top \mathrm{e}^{-(s-r)\mathbf{A}} b^*\,\mathrm{d}r + 2\int_0^s (u^\top Y_{r-}) u^\top \mathrm{e}^{-(s-r)\mathbf{A}}\boldsymbol{\Sigma}\,\mathrm{d}W_r \\ + \int_0^s \int 2u^\top Y_{r-} u^\top \mathrm{e}^{-(s-r)\mathbf{A}} z + (u^\top \mathrm{e}^{-(s-r)\mathbf{A}} z)^2 \widetilde{N}(\mathrm{d}r,\mathrm{d}z). \end{aligned}$$

Additionally, the independence of X_0 and $\mathbf{Z}$ leads to

$$\mathbb{E}[X_0 Y_s^\top] = \mathbb{E}[X_0](\mathbf{A}^{-1}(\mathbb{I} - \mathrm{e}^{-s\mathbf{A}})b^*)^\top.$$

Hence,

$$\begin{aligned}
&\int_0^t (u^\top X_s)^2 - \int (u^\top x)^2 \mu(\mathrm{d}x)\,\mathrm{d}s \\
&= \int_0^t u^\top \mathrm{e}^{-s\mathbf{A}}(X_0X_0^\top - \mathbb{E}[X_0X_0^\top])\mathrm{e}^{-s\mathbf{A}^\top}u + 2u^\top \mathrm{e}^{-s\mathbf{A}}(X_0Y_s^\top - \mathbb{E}[X_0](\mathbf{A}^{-1}(\mathbb{I} - \mathrm{e}^{-s\mathbf{A}})b^*)^\top)u \\
&\quad + \int_0^s 2(u^\top Y_{r-})u^\top \mathrm{e}^{-(s-r)\mathbf{A}}\boldsymbol{\Sigma}\,\mathrm{d}W_r + \int_0^s \int 2u^\top Y_{r-}u^\top \mathrm{e}^{-(s-r)\mathbf{A}}z + (u^\top \mathrm{e}^{-(s-r)\mathbf{A}}z)^2 \widetilde{N}(\mathrm{d}r, \mathrm{d}z)\,\mathrm{d}s \\
&= A_t^1 + A_t^2 + M_t^c + M_t^d,
\end{aligned}$$

where

$$\begin{aligned}
A_t^1 &= \int_0^t u^\top \mathrm{e}^{-s\mathbf{A}}(X_0X_0^\top - \mathbb{E}[X_0X_0^\top])\mathrm{e}^{-s\mathbf{A}^\top}u\,\mathrm{d}s, \\
A_t^2 &= \int_0^t 2u^\top \mathrm{e}^{-s\mathbf{A}}(X_0Y_s^\top - \mathbb{E}[X_0](\mathbf{A}^{-1}(\mathbb{I} - \mathrm{e}^{-s\mathbf{A}})b^*)^\top)u\,\mathrm{d}s, \\
M_t^c &= \int_0^t \int_0^s 2(u^\top Y_{r-})u^\top \mathrm{e}^{-(s-r)\mathbf{A}}\boldsymbol{\Sigma}\,\mathrm{d}W_r\,\mathrm{d}s, \\
M_t^d &= \int_0^t \int_0^s \int 2(u^\top Y_{r-})u^\top \mathrm{e}^{-(s-r)\mathbf{A}}z + (u^\top \mathrm{e}^{-(s-r)\mathbf{A}}z)^2 \widetilde{N}(\mathrm{d}r, \mathrm{d}z)\,\mathrm{d}s.
\end{aligned}$$

For the following bound, first note that, since $\mathbf{A}$ is diagonalizable, there exists some matrix V such that

$$\forall s \in \mathbb{R}, \qquad \|\mathrm{e}^{s\mathbf{A}}\|_2^2 \le \alpha^2 \mathrm{e}^{2\max_i(s\,\mathrm{Re}(\lambda_i))},$$

where $\alpha = \|V\|_2\|V^{-1}\|_2\sqrt{d} > 0$ and λ_i, $i = 1, \dots, d$, are the eigenvalues of $\mathbf{A}$. In particular, for $s > 0$ it holds $\|\mathrm{e}^{-s\mathbf{A}}\|_2^2 \le \alpha^2\mathrm{e}^{-2s\beta}$, where $\beta = \min_i \mathrm{Re}(\lambda_i) > 0$ by ($\mathcal{A}_0$). The Itō isometry thus implies

$$\begin{aligned}
\mathbb{E}[\|Y_s\|^2] &= \mathbb{E}\left[\int_0^s \|\mathrm{e}^{-(s-r)\mathbf{A}}\boldsymbol{\Sigma}\|_2^2\,\mathrm{d}r\right] + \mathbb{E}\left[\int_0^s \int \|\mathrm{e}^{-(s-r)\mathbf{A}}z\|^2 N(\mathrm{d}r, \mathrm{d}z)\right] + \|\mathbf{A}^{-1}(\mathbb{I} - \mathrm{e}^{-s\mathbf{A}})b^*\|^2 \\
&\le \alpha^2\|\boldsymbol{\Sigma}\|_2^2 \int_0^s \mathrm{e}^{-2(s-r)\beta}\,\mathrm{d}r + \alpha^2 \int_0^s \mathrm{e}^{-2(s-r)\beta} \int \|z\|^2 \nu(\mathrm{d}z)\,\mathrm{d}r + \alpha^2\|\mathbf{A}^{-1}\|_2^2\|b^*\|^2 \\
&\le (\alpha^2/(2\beta))(\|\boldsymbol{\Sigma}\|_F^2 + \int \|z\|^2\nu(\mathrm{d}z)) + \alpha^2\|\mathbf{A}^{-1}\|_2^2\|b^*\|^2 =: c_1. \qquad \text{(C.28)}
\end{aligned}$$

Now, (C.28) yields

$$\begin{aligned}
\mathbb{E}[|A_t^1|] &\le \int_0^t \alpha^2\mathrm{e}^{-2\beta s}\mathbb{E}[\|X_0X_0^\top - \mathbb{E}[X_0X_0^\top]\|_2]\,\mathrm{d}s \\
&\le \alpha^2/(2\beta)\mathbb{E}[\|X_0X_0^\top - \mathbb{E}[X_0X_0^\top]\|_2] =: c_2
\end{aligned}$$

and

$$\mathbb{E}[|A_t^2|] \le 2\int_0^t \alpha\mathrm{e}^{-s\beta}(\mathbb{E}[\|X_0\|]\sqrt{c_1} + \alpha\|\mathbf{A}^{-1}\|_2\|b^*\|\mathbb{E}[\|X_0\|])\,\mathrm{d}s$$

$$\leq 2(\alpha/\beta)\mathbb{E}[\|X_0\|](\sqrt{c_1} + \alpha\|\mathbf{A}^{-1}\|_2\|b^*\|) =: c_3.$$

Turning our attention to M_t^c and M_t^d, Fubini's theorem for stochastic integrals (see, e.g., Theorem 65 in [16]) gives

$$\begin{aligned} M_t^c &= \int_0^t 2(u^\top Y_{r-})u^\top \mathrm{e}^{r\mathbf{A}} \int_r^t \mathrm{e}^{-s\mathbf{A}}\,\mathrm{d}s\boldsymbol{\Sigma}\,\mathrm{d}W_r \\ &= \int_0^t 2(u^\top Y_{r-})u^\top (\mathbb{I} - \mathrm{e}^{-(t-r)\mathbf{A}})\mathbf{A}^{-1}\boldsymbol{\Sigma}\,\mathrm{d}W_r \end{aligned}$$

and

$$\begin{aligned} M_t^d &= \int_0^t \int \int_r^t \left(2u^\top Y_{r-}u^\top \mathrm{e}^{-(s-r)\mathbf{A}}z + (u^\top \mathrm{e}^{-(s-r)\mathbf{A}}z)^2\right)\mathrm{d}s\widetilde{N}(\mathrm{d}r,\mathrm{d}z) \\ &= \int_0^t \int \left(2u^\top Y_{r-}u^\top (\mathbb{I} - \mathrm{e}^{-(t-r)\mathbf{A}})\mathbf{A}^{-1}z + \int_r^t (u^\top \mathrm{e}^{-(s-r)\mathbf{A}}z)^2\,\mathrm{d}s\right)\widetilde{N}(\mathrm{d}r,\mathrm{d}z). \end{aligned}$$

This, together with (C.28) and the Itō isometry, implies

$$\begin{aligned} \mathbb{E}[|M_t^c|^2] &\leq 4c_1\|\mathbf{A}^{-1}\boldsymbol{\Sigma}\|_2^2 \int_0^t \left(2 + 2\alpha^2\mathrm{e}^{-2(t-r)\beta}\right)\mathrm{d}r \\ &\leq 4c_1\|\mathbf{A}^{-1}\boldsymbol{\Sigma}\|_2^2\left(2t + \frac{\alpha^2}{\beta}\right). \end{aligned}$$

Similarly, we obtain

$$\begin{aligned} \mathbb{E}\left[|M_t^d|^2\right] &\leq 4c_1\|\mathbf{A}^{-1}\|_2^2\left(\int \|z\|^2\nu(\mathrm{d}z)\int_0^t \|(\mathbb{I} - \mathrm{e}^{-(t-r)\mathbf{A}})\|_2^2\,\mathrm{d}r\right) \\ &\quad + \left(\int \|z\|^4\nu(\mathrm{d}z)\int_0^t\left(\int_r^t \|\mathrm{e}^{-(s-r)\mathbf{A}}\|_2^2\,\mathrm{d}s\right)^2\mathrm{d}r\right) \\ &\leq 4c_1\|\mathbf{A}^{-1}\|_2^2\left(\int \|z\|^2\nu(\mathrm{d}z)\left(2t + \frac{\alpha^2}{\beta}\right)\right) + \left(\int \|z\|^4\nu(\mathrm{d}z)\frac{\alpha^4}{4\beta^2}t\right), \end{aligned}$$

which is finite by the assumption of Z_1 admitting a fourth moment. Markov's inequality now implies for any $r > 0$

$$\begin{aligned} &\mathbb{P}\left(|u^\top(\widehat{\mathbf{C}}_T - \mathbf{C}_\infty)u| \geq r\right) \\ &\quad\leq \mathbb{P}\left(|A_t^1| \geq \frac{tr}{4}\right) + \mathbb{P}\left(|A_t^2| \geq \frac{tr}{4}\right) + \mathbb{P}\left(|M_t^c| \geq \frac{tr}{4}\right) + \mathbb{P}\left(|M_t^d| \geq \frac{tr}{4}\right) \\ &\quad\leq \frac{4(c_2+c_3)}{tr} + \frac{64c_1\|\mathbf{A}^{-1}\boldsymbol{\Sigma}\|_2^2\left(2t + \frac{\alpha^2}{\beta}\right)}{(tr)^2} \\ &\qquad + \frac{64c_1\|\mathbf{A}^{-1}\|_2^2\left(\int \|z\|^2\nu(\mathrm{d}z)\left(2t + \frac{\alpha^2}{\beta}\right)\right) + 4\left(\int \|z\|^4\nu(\mathrm{d}z)\frac{\alpha^4}{\beta^2}t\right)}{(tr)^2} \\ &\quad\leq \frac{4(c_2+c_3) + 128c_1\|\mathbf{A}^{-1}\|_2^2(\|\boldsymbol{\Sigma}\|_2^2 + \int \|z\|^2\nu(\mathrm{d}z)) + (2\alpha^2/\beta)^2\int \|z\|^4\nu(\mathrm{d}z)}{t(r \wedge r^2)} \end{aligned}$$

$$+\frac{64c_1\alpha^2\|\mathbf{A}^{-1}\|_2^2(\|\mathbf{\Sigma}\|_2^2+\int\|z\|^2\nu(\mathrm{d}z))}{\beta(tr)^2},$$

which concludes the proof. ■

References

[1] D. Applebaum. *Lévy Processes and Stochastic Calculus*. 2nd ed. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2009.

[2] O. E. Barndorff-Nielsen. "Processes of normal inverse Gaussian type". In: *Finance Stoch.* 2.1 (1997), pp. 41–68.

[3] S. Basu and G. Michailidis. "Regularized estimation in sparse high-dimensional time series models". In: *Ann. Statist.* 43.4 (2015), pp. 1535 –1567.

[4] P. C. Bellec, G. Lecué, and A. B. Tsybakov. "Slope meets Lasso: improved oracle bounds and optimality". In: *Ann. Statist.* 46.6B (2018), pp. 3603–3642.

[5] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. "Simultaneous analysis of Lasso and Dantzig selector". In: *Ann. Statist.* 37.4 (2009), pp. 1705 –1732.

[6] M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candès. "SLOPE—adaptive variable selection via convex optimization". In: *Ann. Appl. Stat.* 9.3 (2015), pp. 1103–1140.

[7] R. Carmona, J. Fouque, and L. Sun. "Mean field games and systemic risk". English (US). In: *Communications in Mathematical Sciences* 13.4 (2015). Publisher Copyright: © 2015 International Press., pp. 911–933.

[8] P. Cattiaux and A. Guillin. "Deviation bounds for additive functionals of Markov processes". In: *ESAIM Probab. Stat.* 12 (2008), pp. 12–29.

[9] G. Ciołek, D. Marushkevych, and M. Podolskij. "On Dantzig and Lasso estimators of the drift in a high dimensional Ornstein–Uhlenbeck model". In: *Electron. J. Stat.* 14.2 (2020), pp. 4395–4420.

[10] J.-P. Fouque and T. Ichiba. "Stability in a Model of Interbank Lending". In: *SIAM J. Finanical Math.* 4.1 (2013), pp. 784–803. eprint: https://doi.org/10.1137/110841096.

[11] S. Gaïffas and G. Matulewicz. "Sparse inference of the drift of a high-dimensional Ornstein–Uhlenbeck process". In: *J. Multivariate Anal.* 169 (2019), pp. 1–20.

[12] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. A Series of Modern Surveys in Mathematics Series. Springer, 1991.

[13] H. Mai. "Efficient maximum likelihood estimation for Lévy-driven Ornstein–Uhlenbeck processes". In: *Bernoulli* 20.2 (2014), pp. 919 –957.

[14] H. Masuda. "On multidimensional Ornstein-Uhlenbeck processes driven by a general Lévy process". In: *Bernoulli* 10.1 (2004), pp. 97–120.

[15] I. Nourdin and F. G. Viens. "Density formula and concentration inequalities with Malliavin calculus". In: *Electron. J. Probab.* 14 (2009), no. 78, 2287–2309.

[16] P. E. Protter. *Stochastic integration and differential equations*. Second. Vol. 21. Applications of Mathematics (New York). Stochastic Modelling and Applied Probability. Springer-Verlag, Berlin, 2004, pp. xiv+415.

[17] K.-i. Sato. *Lévy processes and infinitely divisible distributions*. Cambridge Studies in Advanced Mathematics, Vol. 68. Cambridge University Press, Cambridge, 1999.

[18] K.-i. Sato and M. Yamazato. "Operator-self-decomposable distributions as limit distributions of processes of Ornstein-Uhlenbeck type". In: *Stochastic Process. Appl.* 17.1 (1984), pp. 73–100.

[19] M. Sørensen. "Likelihood methods for diffusions with jumps". In: *Statistical inference in stochastic processes*. Vol. 6. Probab. Pure Appl. Dekker, New York, 1991, pp. 67–105.

[20] M. Talagrand. *Upper and lower bounds for stochastic processes*. Vol. 60. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics. Modern methods and classical problems. Springer, Heidelberg, 2014, pp. xvi+626.

[21] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009, pp. xii+214.

[22] R. Vershynin. *High-dimensional probability*. Vol. 47. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018, pp. xiv+284.