Sharp adaptive drift estimation and Donsker-type theorems for multidimensional ergodic diffusions

Dissertation

**zur Erlangung des Doktorgrades
der Fakultät für Mathematik, Informatik
und Naturwissenschaften
der Universität Hamburg**

vorgelegt
im Fachbereich Mathematik

von
Claudia Strauch
aus Pritzwalk

Hamburg 2013

Als Dissertation angenommen vom Fachbereich
Mathematik der Universität Hamburg

Auf Grund der Gutachten von Prof. Dr. Angelika Rohde
und Prof. Dr. Enno Mammen

Hamburg, den 04.06.2013

Prof. Dr. Ulf Kühn
Leiter des Fachbereichs Mathematik

Für meine Mutter.

Contents

1. Introduction

This first chapter introduces relevant definitions related to diffusion processes in $\mathbb{R}^d$ and notions from nonparametric statistics and empirical process theory. For Section 1.1, the main references are Stroock and Varadhan (1979), Revuz and Yor (1999), Ledoux (2000) and Bakry (2008). Main sources for Section 1.4.1 are Tsybakov (1998), Nussbaum (2006) and Tsybakov (2009), and for Section 1.4.2, we refer to Chapter 14 in Ledoux and Talagrand (1991) and the books of van der Vaart and Wellner (1996) and Dudley (1999), respectively.

1.1. Diffusion processes

Diffusions present a particularly important class of stochastic processes. There is a long standing theoretical interest in these processes as illustrated, for example, by the seminal books by Itō and McKean (1965) and Stroock and Varadhan (1979). In addition, diffusions are of substantial practical importance as they can be used to model many natural phenomena arising in biology, physics, or mathematical finance. Loosely speaking, a diffusion is a continuous Markov process which is characterized by its local infinitesimal drift b and variance a. At a more rigorous level, assume that $a : \mathbb{R}^d \to S_d$ and $b : \mathbb{R}^d \to \mathbb{R}^d$ are Borel measurable maps, with S_d denoting the set of symmetric non-negative definite $d \times d$ real matrices. Let A be the following second-order differential operator associated with (a, b),

$$A := \frac{1}{2}\sum_{i,j=1}^{d} a_{ij}(\cdot)\ \frac{\partial^2}{\partial x^i \partial x^j} + \sum_{i=1}^{d} b^i(\cdot)\ \frac{\partial}{\partial x^i}, \tag{1.1.1}$$

where, for any $g : \mathbb{R}^d \to \mathbb{R}^d$, g^j denotes the j-th component, $j \in \{1, \ldots, d\}$. Let $C_c^\infty(\mathbb{R}^d)$ denote the class of all infinitely differentiable functions on $\mathbb{R}^d$ with compact support.

Definition 1.1.1 (Revuz and Yor (1999), Definition VII.2.1)**.** A Markov process X, defined on a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P}_x)$, with state space $\mathbb{R}^d$ is called a *diffusion process with generator* A if it has continuous paths and if, for any $x \in \mathbb{R}^d$ and any $f \in C_c^\infty(\mathbb{R}^d)$,

$$\mathbb{E}_x f(X_t) = f(x) + \mathbb{E}_x\left(\int_0^t Af(X_u)\,\mathrm{d}u\right),$$

where $\mathbb{E}_x$ denotes expectation under the initial value x at time $t = 0$. The diffusion X thus defined has *diffusion coefficient* a and *drift* b.

Diffusions appear in many different situations, and depending on the context, it might be convenient to emphasize different aspects. Classically, a diffusion is described as a strong Markov process associated with a Markov semigroup with a specific infinitesimal generator. The Itô approach is to view it as a strong solution of an (Itô) stochastic differential equation (SDE), and Stroock and Varadhan advocate the approach to diffusions as solutions of the corresponding martingale problem. The topic of this thesis are two specific questions related to nonparametric estimation theory and weak convergence properties of diffusion processes in $\mathbb{R}^d$, namely

- sharp adaptive estimation of the drift of a diffusion process, given as a solution of an Itô SDE, from continuous-time observations, and

- the formulation of (necessary and sufficient) conditions for uniform central limit theorems for ergodic diffusions, thus providing an extension of the classical central limit theorem for additive functionals of continuous Markov processes.

While the first question is concerned with diffusions arising as solution of some Itô SDE, the second question will be dealt with by means of Markovian semigroup theory. Before we turn to some further explanation of the two questions, let us briefly introduce the respective frameworks and some basic assumptions.

1.1.1. Diffusions and Itô processes

It is beyond the scope of this thesis to treat multidimensional diffusion processes in full generality, and it is also common practice to restrict attention to the class of diffusion processes which can be described by stochastic differential equations. In the sequel, we will deal exclusively with homogeneous diffusions.

Definition 1.1.2 (Stroock and Varadhan (1979), Section 4.3)**.** Let $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P})$ be a filtered probability space, and let $X : \Omega \to \mathbb{R}^d$, $a : \Omega \to S_d$ and $b : \Omega \to \mathbb{R}^d$ be locally bounded progressively measurable functions. If, for each $(t, q) \in [0, \infty) \times \Omega$ and any $f \in C_c^\infty(\mathbb{R}^d)$,

$$f(X_t) - \int_0^t (Af)(X_u)\,\mathrm{d}u$$

is a martingale relative to $(\Omega, \mathscr{F}, \mathscr{F}_t, \mathbb{P})$ for $t \geq 0$, then X is called an homogeneous *Itô process* on $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P})$ with covariance a and drift b, and we write $X \sim \mathscr{I}_d(a, b)$.

Itô processes can be expressed as stochastic integrals. Let W be an $(\mathscr{F}_t)$-Brownian motion, defined on a probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P})$, and let $\sigma : \Omega \to \mathbb{R}^{d\times d}$ and $b : \Omega \to \mathbb{R}^d$ be locally bounded progressively measurable functions. Any process X satisfying

$$X_t - X_0 = \int_0^t \sigma(X_u)\,\mathrm{d}W_u + \int_0^t b(X_u)\,\mathrm{d}u, \qquad t \geq 0, \tag{1.1.2}$$

is an Itô process with covariance $a(\cdot) = \sigma\sigma^t(\cdot)$ and drift $b(\cdot)$; cf., e.g., Theorem 4.3.8 in Stroock and Varadhan (1979) or Proposition VII.2.6 in Revuz and Yor (1999). Conversely, if $X \sim \mathscr{I}_d(a, b)$ on $(\Omega, \mathscr{F}, \mathscr{F}_t, \mathbb{P})$, then there exist a predictable process σ and a Brownian motion W, possibly on an enlargement of the probability space, such that (1.1.2) holds (see Theorems 4.5.1 and 4.5.2 in Stroock and Varadhan (1979) or Theorem VII.2.7 in Revuz and Yor (1999)).

1.1.2. Diffusions and Markovian semigroups

Let $(E, \mathscr{B}(E))$, $E \subseteq \mathbb{R}^d$, be a Borel measurable space, and let $X = (X_t)_{t\geq 0}$ be an E-valued Markov process with invariant probability measure μ. Its transition semigroup is denoted $(P_t)_{t\geq 0}$, acting on the bounded measurable functions f on E by

$$P_t f(x) = \int_E f(y) p_t(x, y)\mathrm{d}y, \qquad x \in E, \tag{1.1.3}$$

where $p_t(\cdot, \cdot)$ are the corresponding transition densities. Denote the center of the semigroup by $\mathbb{B}_0$, that is,

$$\mathbb{B}_0 := \left\{ f \in C_0(\mathbb{R}^d) : \|P_t f - f\|_\infty \to 0 \text{ as } t \to 0 \right\},$$

where $C_0(\mathbb{R}^d)$ is the space of continuous functions $f : \mathbb{R}^d \to \mathbb{R}$ with $f(x) \to 0$ as x approaches the boundary of E. The domain $\mathbb{D}_A$ of the infinitesimal generator A of $(P_t)_{t\geq 0}$ is defined as the

set

$$\left\{f \in \mathbb{B}_0 : \left\|\frac{P_t f - f}{t} - g\right\|_\infty \to 0 \text{ for some } g \in \mathbb{B}_0, \text{ as } t \to 0\right\},$$

and the infinitesimal generator A is given by

$$Af = \lim_{t\to 0} \frac{1}{t}\left(P_t f - f\right).$$

It will be assumed throughout that the transition probabilities $p_t(\cdot,\cdot)$ admit an invariant probability measure μ. In this case, (P_t) defines a contraction semigroup on $L^2(E,\mu) =: L^2(\mu)$ (cf. Bhattacharya (1982)). Slightly abusing notation, the infinitesimal generator of (P_t) will also be denoted by A (although it is actually an extension of A on $L^2(\mu)$), with corresponding domain $\mathbb{D}_A \subset L^2(\mu)$. It can be equipped with the topology given by

$$\|f\|_{\mathbb{D}_A} = \|f\|_{L^2(\mu)} + \|Af\|_{L^2(\mu)}. \tag{1.1.4}$$

The generator A and its domain $\mathbb{D}_A$ completely determine $(P_t)_{t\geq 0}$, that is, there exists a unique semigroup $(P_t)_{t\geq 0}$ of bounded operators on $L^2(\mu)$ satisfying (1.1.3) for all functions of the domain $\mathbb{D}_A$. Conversely, if P_t is a Feller semigroup on $\mathbb{R}^d$ and the corresponding process has continuous paths, then its infinitesimal generator on $C_c^2(\mathbb{R}^d)$ is given by (1.1.1) (cf. Theorem VII.1.13 and the subsequent remark in Revuz and Yor (1999)). We shall not deepen the presentation of Markov semigroups at this point but will introduce further details when needed.

1.1.3. Basic assumptions

Throughout, we consider some diffusion process X with invariant measure μ, given as a solution of (1.1.2) or described by the associated transition semigroup $(P_t)_{t\geq 0}$. The measure μ is related to $(P_t)_{t\geq 0}$ by different properties. Since μ is supposed to be *invariant* with respect to $(P_t)_{t\geq 0}$, it holds for any $f \in L^1(\mu)$,

$$\int P_t f \mathrm{d}\mu = \int f \mathrm{d}\mu.$$

Furthermore, it will be assumed throughout that $(P_t)_{t\geq 0}$ is *ergodic* with respect to μ in the sense that

$$\lim_{t\to\infty} P_t f = \int f \mathrm{d}\mu \qquad \mu\text{-a.e.}$$

Here and throughout the sequel, $\mathbf{P}_\mu(\cdot) := \int \mathbf{P}_x(\cdot)\mu(\mathrm{d}x)$, such that $((X_t)_{t\geq 0}, \mathbf{P}_\mu)$ is a stationary ergodic process.

We now list some additional conditions on μ and (the transition semigroup $(P_t)_{t\geq 0}$ associated with) X which will be important in our investigation.

(D1) (*Symmetry.*) The measure μ is said to be *reversible* with respect to $(P_t)_{t\geq 0}$, or $(P_t)_{t\geq 0}$ is *symmetric* with respect to μ, if for every $f, g \in L^2(\mu)$,

$$\int f P_t g \mathrm{d}\mu = \int g P_t f \mathrm{d}\mu.$$

The generator A of a symmetric diffusion is self-adjoint, and it holds $\int fAf\mathrm{d}\mu \leq 0$.

(D2) (*Poincaré's inequality.*) There exists some constant c_P such that, for any smooth enough function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathrm{Var}_\mu(f) := \int f^2 \mathrm{d}\mu - \left(\int f \mathrm{d}\mu\right)^2 \leq c_P^{-1} \int |\nabla f|^2 \mathrm{d}\mu. \tag{\textbf{PI}}$$

For symmetric diffusions, Poincaré's inequality is commonly referred to as *spectral gap inequality* since (**PI**) is then equivalent to the existence of a spectral gap,

$$\lambda_1 := \sup\{\lambda \geq 0 : E_\lambda - E_0 = 0\} = \frac{1}{c_P} > 0,$$

where $-\int_0^\infty \lambda \, \mathrm{d}E_\lambda$ denotes the spectral decomposition of A, and $c = c_P$ appears to be the smallest possible constant in (**PI**). Furthermore, (**PI**) is equivalent to the exponential decay of P_t to the invariant measure μ in $L^2(\mu)$ (see, e.g., Theorem 1.3 in Bakry et al. (2008)),

$$\forall f \in L^2(\mu), \qquad \operatorname{Var}_\mu(P_t f) \leq \exp\left(-\frac{2t}{c_P}\right) \operatorname{Var}_\mu(f).$$

(D3) (*Bound on the transition densities.*) There exists some $C_0 > 0$ such that, for any $u \geq t > 0$ and for any pair of points $x, y \in \mathbb{R}^d$ satisfying $\|x-y\|^2 \leq u$, we have

$$p_t(x,y) \ \leq \ C_0(t^{-d/2} + u^{3d/2}). \tag{1.1.5}$$

(D4) (*Uniform ellipticity.*) There exists some positive constant $0 < a_* < \infty$ such that the diffusion matrix a satisfies the uniform ellipticity condition,

$$\zeta^t a(y)\zeta \geq a_*\|\zeta\|^2 \text{ for all } \zeta \in \mathbb{R}^d \setminus \{0\} \text{ and for all } y \in \mathbb{R}^d.$$

(D5) (*Boundedness of the invariant density.*) The invariant measure μ is Lebesgue continuous with density ρ which is bounded away from zero and uniformly bounded in $x \in \mathbb{R}^d$.

At times, we require the following more specific upper bound on the invariant density.

(D6) (*Exponential decay of the invariant density.*) The invariant measure μ is Lebesgue continuous with density ρ which satisfies for some positive constants C_1, C_2,

$$\rho(x) \leq C_1 \exp\left(-C_2\|x\|^2\right), \qquad x \in \mathbb{R}^d.$$

We shall comment on the above assumptions in the sequel. For the moment, we merely note that – although the theory for multidimensional diffusion processes in general is less satisfactory than for scalar diffusions, – the above conditions equip us with tools powerful enough to carry out an in-depth analysis. We proceed by describing an important class of diffusion processes satisfying the above conditions.

Example: Kolmogorov diffusions with at most linear growth of the drift

Let $V \in C^2(\mathbb{R}^d)$, and consider the so-called *Kolmogorov process*

$$X_t = X_0 - \int_0^t \nabla V(X_u)\,\mathrm{d}u + W_t, \tag{1.1.6}$$

where W is a d-dimensional Brownian motion and the initial value X_0 is independent of W. If there exist constants $K_1, K_2 \in \mathbb{R}$ such that

$$\forall \ \|x\| \in \mathbb{R}^d, \qquad \langle \nabla V(x), x\rangle \geq -K_1\|x\|^2 - K_2,$$

then the equation (1.1.6) has a unique continuous solution $t \mapsto X_t(\omega)$ defined on $\mathbb{R}_+$ (Theorem 2.2.29 in Royer (2007)). If V is such that $\mathrm{e}^{-2V} \in L^1(\mathbb{R}^d)$, then the probability measure μ defined by $\mathrm{d}\mu(x) = \rho(x)\mathrm{d}x$ with

$$\rho(x) := c^{-1}(V)\exp\left(-2V(x)\right), \qquad \text{for } c(V) := \int_{\mathbb{R}^d} \mathrm{e}^{-2V(u)}\mathrm{d}u < \infty, \tag{1.1.7}$$

is an invariant measure for X. In addition, μ is reversible for X (see, e.g., Lemma 2.2.3 in Royer (2007)). Denoting by N_t the semigroup which acts on the bounded measurable functions f on $\mathbb{R}^d$ as

$$N_t f(x) = \mathbb{E}_x f(X_t) = \int p_t(x,y) f(y)\mathrm{d}y,$$

where X is the solution of (1.1.6), this can also be stated by saying that N_t induces a strongly continuous semigroup of self-adjoint operators on $L^2(\mu)$. Poincaré's inequality holds whenever there exist constants K_3, K_4 such that

$$\forall\ \|x\| \geq K_3, \qquad \langle \nabla V(x), x\rangle \geq K_4 \|x\| \tag{K1}$$

(cf. Bakry et al. (2008), p. 747). If there exists some constant $K_5 > 0$ such that, for any $x \in \mathbb{R}^d$,

$$\|b(x)\| \leq K_5 \left(1 + \|x\|\right), \tag{K2}$$

then Theorem 3.2 in Qian and Zheng (2004) entails that (1.1.5) holds.

1.2. Adaptive estimation

The minimax theory of adaptive estimation of functions in various statistical models has been studied intensively in the last decades. For setting the scene, consider the Gaussian white noise model

$$\mathrm{d}Y(t) = f(t)\mathrm{d}t + \varepsilon\ \mathrm{d}W(t), \qquad t \in [0,1], \tag{1.2.1}$$

where $W(\cdot)$ is a standard Wiener process and $0 < \varepsilon < 1$ is the noise level. A classical issue in nonparametric statistics is to estimate the function f on the interval $[0,1]$ from the observation $(Y(t))_{t\in[0,1]}$, assuming that f belongs to $\mathcal{F}_\nu$, where $\{\mathcal{F}_\nu, \nu \in \mathcal{N}\}$ is a collection of functional classes. To evaluate the quality of an estimator $\widetilde{f}$, introduce its maximal risk for a given loss function $l(\cdot,\cdot)$ as

$$\mathscr{R}_{\varepsilon,\nu}(\widetilde{f}, \psi_\nu) := \sup_{f\in\mathcal{F}_\nu} \mathbb{E}_f\left(\psi_\nu^{-p}(\varepsilon)\ l^p(\widetilde{f}, f)\right), \qquad p > 0 \text{ fixed},$$

where $\mathbb{E}_f$ denotes expectation with respect to the distribution $\mathbb{P}_f$ of the observation satisfying (1.2.1), and $\psi_\nu(\varepsilon)$ is a normalizing factor satisfying $\psi_\nu(\varepsilon) > 0$ and $\psi_\nu(\varepsilon) \to_{\varepsilon\to 0} 0$ for every ν. For the case of the L^2 loss, one obtains the *mean-integrated squared error*

$$\mathbb{E}_f \|\widetilde{f} - f\|^2_{L^2([0,1])} = \mathbb{E}_f \int_0^1 \left(\widetilde{f}(x) - f(x)\right)^2 \mathrm{d}x, \tag{1.2.2}$$

and the goal is to understand the asymptotics as $\varepsilon \to 0$ of the minimax risk

$$\inf_{\widetilde{f}} \sup_{f\in\mathcal{F}_\nu} \mathbb{E}_f\Big(\psi_\nu^{-2}(\varepsilon)\ \|\widetilde{f} - f\|^2_{L^2([0,1])}\Big),$$

the infimum being taken over all estimators $\widetilde{f}$ of f. Typically, $\mathcal{F}_\nu$ is given as

$$\mathcal{F}_\nu = \left\{f : [0,1] \to \mathbb{R} : \eta_\beta(f) \leq L\right\}, \qquad \nu = (\beta, L),$$

where $\eta_\beta(\cdot)$ is a given functional (e.g., a seminorm), $\beta > 0$ is some smoothness parameter, and $L > 0$ is the radius of the ball $\mathcal{F}_\nu$. If ν is known, the functional class $\mathcal{F}_\nu$ is fixed, and in many situations one can construct an estimator $\widehat{f}_\nu$ which achieves the *optimal asymptotic minimax rate of convergence* on $\mathcal{F}_\nu$, that is, there exist some normalizing factor $\psi^*_\nu(\varepsilon)$ and positive constants c, C such that

$$c \leq \liminf_{\varepsilon\to 0} \inf_{\widetilde{f}} \mathscr{R}_{\varepsilon,\nu}(\widetilde{f}, \psi^*_\nu) \leq \limsup_{\varepsilon\to 0} \mathscr{R}_{\varepsilon,\nu}(\widehat{f}_\nu, \psi^*_\nu) \leq C.$$

Denote the set of all optimal rates of convergence $\psi^*_\nu(\cdot)$ for fixed ν by $\{\psi^*_\nu(\cdot)\}$.

Two obvious severe obstacles have stipulated research in nonparametric estimation theory.

(I) Results on the convergence *rate* have only a limited significance and do not provide further insight into the specific nature of the estimation problem at hand. Furthermore, exact results are of avail in order to be able to compare estimators on the level of constants. The interest therefore is in finding an estimator $\widehat{f}_\nu$ satisfying the equality

$$\lim_{\varepsilon\to 0} \inf_{\widetilde{f}} \mathscr{R}_{\varepsilon,\nu}(\widetilde{f}, \psi_\nu^*) = \lim_{\varepsilon\to 0} \mathscr{R}_{\varepsilon,\nu}(\widehat{f}_\nu, \psi_\nu^*).$$

(II) The parameter ν is usually unknown, and due to dependence of the estimator on the parameter, the minimax approach is of limited use in practice. The next step is to construct *optimally rate adaptive estimators* on $\mathcal{N}$, that is, estimators f^* which do not depend on ν and which achieve the optimal rate uniformly over $\nu \in \mathcal{N}$,

$$\limsup_{\varepsilon\to 0} \sup_{\nu\in\mathcal{N}} \mathscr{R}_{\varepsilon,\nu}(f^*, \Psi_\nu) \leq C,$$

where $\Psi_\nu(\varepsilon) = \psi_\nu^*(\varepsilon)$ is any optimal rate of convergence on $\mathcal{F}_\nu$ for the loss l, and C is a positive constant. A priori, it is not clear whether such estimators exist or whether one can only achieve a rate slower than the optimal one when the parameter ν is not known. In cases where rate optimal estimators exist, it is moreover of interest whether there is a *loss of efficiency under adaptation* which is reflected in a worsening of the asymptotic constant.

As concerns (I), the pioneering result in nonparametric estimation theory on *exact minimax asymptotics* is due to Pinsker (1980) who considers filtering problems over ellipsoids in Hilbert space with respect to mean-integrated squared error criteria. There exists a canonical map of the sequence model considered in Pinsker (1980) to the Gaussian white noise model (1.2.1). Given $m \in \mathbb{N}$ and $M > 0$, denote by $\widetilde{W}^{m,2}(M)$ the periodic L^2-Sobolev class of order m and radius M on $[0,1]$, that is,

$$\widetilde{W}^{m,2}(M) := \left\{ f : f = \sum_{j=1}^{\infty} \theta_j \phi_j,\ \sum_{k=1}^{\infty} (2\pi k)^{2m} \left(\theta_{2k}^2 + \theta_{2k+1}^2\right) \leq M \right\},$$

where $(\phi_j)_{j\geq 1}$ is the trigonometric orthonormal basis in $L^2[0,1]$, and $\theta_j := \int \phi_j(t) f(t) \mathrm{d}t$, $j \geq 1$, are the Fourier coefficients of f. Pinsker's theorem states that

$$\lim_{\varepsilon\to 0} \varepsilon^{-\frac{4m}{2m+1}} \inf_{\widetilde{f}} \sup_{f\in\widetilde{W}^{m,2}(M)} \mathbb{E}_f \|\widetilde{f} - f\|_{L^2(\mathbb{R}^d)}^2 = (L(2m+1))^{\frac{1}{2m+1}} \left(\frac{m}{\pi(m+1)}\right)^{\frac{2m}{2m+1}}. \tag{1.2.3}$$

The minimax linear filtering estimate constructed in Pinsker (1980) which attains the exact bound depends on the ellipsoid via the set of the coefficients describing it. Efroimovich and Pinsker (1984) give an algorithm of adaptive estimation which allows to attain the exact bound for periodic Sobolev classes by an estimator which does not depend on the degree of smoothness m and on the bound M. As concerns item (II), it thus has been shown that optimal rate adaptive estimators exist and that there is no loss of efficiency in the "Pinsker case," that is, in the setting of nonparametric function estimation in the Gaussian white noise model over the scale of L^2 Sobolev classes with respect to mean-integrated squared error criteria. Put differently, this means that *asymptotically exact adaptive estimators* on the Sobolev scale for the L^2 risk exist, in the sense of the following

Definition 1.2.1 (Tsybakov (1998), Definition 1)**.** An optimal rate adaptive estimator f_ε^* is called *asymptotically exact adaptive on the scale of classes* $(\mathcal{F}_\nu)_{\nu\in\mathcal{N}}$ *for the loss* l if it satisfies

$$\lim_{\varepsilon\to 0} \inf_{\widetilde{f}} \sup_{\nu\in\mathcal{N}} \mathscr{R}_{\varepsilon,\nu}(\widetilde{f}, \Psi_\nu) = \lim_{\varepsilon\to 0} \sup_{\nu\in\mathcal{N}} \mathscr{R}_{\varepsilon,\nu}(f_\varepsilon^*, \Psi_\nu), \tag{1.2.4}$$

where $\Psi_\nu(\varepsilon) \in \{\psi_\nu^*(\cdot)\}$ for every fixed ν.

One further crucial result on exact asymptotics is due to Korostelev (1993) who studies the behavior of the minimax risk in Gaussian nonparametric regression. Instead of Sobolev-type smoothness assumptions, he considers estimation over the scales of Hölder classes $\mathcal{H}(\beta, L)$, given as

$$\mathcal{H}(\beta, L) := \left\{ f : |f(x) - f(y)| \leq L|x-y|^\beta,\ x, y \in [0,1] \right\}, \qquad \beta \in (0,1],\ L > 0,$$

and the squared L^2 loss in (1.2.2) is substituted with the sup-norm loss. Donoho (1992) extends the result to signal estimation in Gaussian white noise and establishes a relation to the optimal recovery problem. Lepski (1992) shows that adaptive minimax estimators do not exist in the "Korostelev case," that is, for nonparametric estimation over the scale of Hölder classes of smoothness $\beta \in (0,1)$ with respect to sup-norm loss. The loss of efficiency, defined as the maximal over $\nu \in \mathcal{N}$ ratio of the risk of the best adaptive estimator to the minimax nonadaptive risk, is asymptotically strictly greater than one.

Instead of using the L^2 risk defined in (1.2.2) or sup-norm losses, one might be interested in nonparametric estimation of f in some fixed point $x \in [0,1]$, and in this case accuracy of an estimator $\widetilde{f}$ is conveniently measured by the *mean-squared error*

$$\mathbb{E}_f |\widetilde{f}(x) - f(x)|^2.$$

The investigation of the problem of pointwise adaptive estimation in the minimax framework was pioneered by Lepski (1991). In particular, it is shown that there are no optimal rate adaptive estimators on the scale of Hölder classes. The best adaptive estimators are proven to achieve only a rate which is slower than the optimal one in a logarithmic factor. To distinguish them from optimal rate adaptive estimators, such estimators are called *rate adaptive.*

Definition 1.2.2 (Tsybakov (1998), Definition 3)**.** The sequence $\Psi_\nu(\varepsilon)$ is an *adaptive rate of convergence on the scale of classes* $(\mathcal{F}_\nu)_{\nu \in \mathcal{N}}$ *for the loss* l if the following two conditions are satisfied:

- There exists an estimator f^* which is optimal rate adaptive on $(\mathcal{F}_\nu)_{\nu \in \mathcal{N}}$, that is, for $\Psi_\nu(\varepsilon) = \psi_\nu^*(\varepsilon)$, it holds
$$\limsup_{\varepsilon \to 0} \sup_{\nu \in \mathcal{N}} \mathscr{R}_{\varepsilon,\nu}(f^*, \Psi_\nu) \leq C,$$
where $\psi_\nu^*(\varepsilon)$ is any optimal rate of convergence on $\mathcal{F}_\nu$ for the loss l;
- if the rate of convergence $S_\nu(\varepsilon) > 0$ satisfies for some estimator f^{**}
$$\limsup_{\varepsilon \to 0} \sup_{\nu \in \mathcal{N}} \mathscr{R}_{\varepsilon,\nu}(f^{**}, S_\nu) \leq C$$
and if, in addition, there exists some $\nu' \in \mathcal{N}$ such that $S_{\nu'}(\varepsilon)/\Psi_{\nu'}(\varepsilon) \to 0$ as $\varepsilon \to 0$, then there exists $\nu'' \in \mathcal{N}$ such that
$$\frac{S_{\nu'}(\varepsilon)}{\Psi_{\nu'}(\varepsilon)} \cdot \frac{S_{\nu''}(\varepsilon)}{\Psi_{\nu''}(\varepsilon)} \to_{\varepsilon \to 0} +\infty.$$

For the scale of Hölder classes and under the pointwise risk, Lepski and Spokoiny (1997) find the best among all rate adaptive estimators in the sense of exact risk asymptotics. The exact asymptotics of minimax adaptive risks on the scale of L^2 Sobolev classes for estimation under the sup-norm risk and for pointwise estimation are studied in Tsybakov (1998). In both works, the Gaussian white noise model is considered. Further references will be given in the sequel.

There exists indeed a large variety of situations of nonparametric estimation in between and beyond the classical "Pinsker case" and the "Korostelev case" mentioned above. Typical situations are smoothness assumptions of Hölder, Besov or Sobolev type, combined with pointwise, uniform or L^p risk criteria. We already noted that the results may differ substantially when passing from one risk criterion to another, and a similar remark holds, e.g., for the case of pointwise estimation over Hölder classes as compared to pointwise estimation over Sobolev classes. As concerns the choice of the statistical model, it is folklore that estimators proposed for one specific model usually can be modified for other types of observations. It has become accepted to develop results first in a Gaussian white noise model in the spirit of (1.2.1) as it "approximates" common models with discrete observations such as nonparametric regression or nonparametric density estimation. Arguments based on the concept of asymptotic statistical equivalence allow to extend results on exact estimation which are available for one specific model to another equivalent model.

Asymptotic statistical equivalence for the drift estimation experiment. Dalalyan and Reiß (2007) develop asymptotic equivalence of experiments in the Le Cam sense for inference on the drift in multidimensional ergodic Kolmogorov diffusions, given as a strong solution of the SDE

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \mathrm{d}W_t, \qquad X_0 = \xi,\ t \in [0,T]. \tag{1.2.5}$$

Denote by $\Sigma\left(M_1, M_2\right)$ the set of all functions $b = -\nabla V : \mathbb{R}^d \to \mathbb{R}^d$ satisfying for all $x, y \in \mathbb{R}^d$ the following conditions,

$$\|b(x)\| \leq M_1\left(1 + \|x\|\right), \qquad \langle b(x) - b(y), x - y\rangle \leq -M_2\left\|x - y\right\|.$$

The main result of Dalalyan and Reiß (2007) establishes local equivalence of the diffusion experiment with an accompanying Gaussian shift experiment. It is instructive to consider the corresponding models in detail.

Definition 1.2.3. (i) (*Diffusion experiment*) Let $\Sigma_\beta(L, K)$ be the set of functions $b \in \Sigma(K)$ such that all d components b^i of b belong to the isotropic Hölder smoothness class of regularity $\beta > 0$ and radius $L > 0$. Given some fixed $b_0 \in \Sigma_\beta(L, K)$, let

$$\begin{aligned}\Sigma\left(b_0, \varepsilon, \eta, A\right) = \Big\{ b \in \Sigma_\beta(L, K) : |b(x) - b_0(x)| \leq \varepsilon\ \mathbb{1}_A(x),\ x \in \mathbb{R}^d,& \\ |\rho_b(x) - \rho_{b_0}(x)| \leq \eta\ \rho_{b_0}(x),\ x \in A \Big\}.&\end{aligned} \tag{1.2.6}$$

Suppose $\Sigma \subset \Sigma(K)$ for some $K > 0$. For any $T > 0$, let $\mathbb{E}(\Sigma, T)$ be the statistical experiment of observing the diffusion defined by (1.2.5) with $b \in \Sigma$.

(ii) (*Gaussian shift model*) For $b \in L^2\left(\mu_{b_0}\right)$, denote by $\mathbf{Q}_{b,T}$ the Gaussian measure on the measurable space $\left(C(\mathbb{R}^d; \mathbb{R}^d), \mathscr{B}(C(\mathbb{R}^d; \mathbb{R}^d))\right)$, induced by the d-dimensional process Z satisfying

$$\mathrm{d}Z(x) = b(x)\sqrt{\rho_{b_0}(x)}\ \mathrm{d}x + \frac{1}{\sqrt{T}}\ \mathrm{d}B(x), \qquad Z(0) = 0,\ x \in \mathbb{R}^d,$$

where $B(x) = \left(B^1(x), \ldots, B^d(x)\right)$ and $B^1(x), \ldots, B^d(x)$ are independent d-variate Brownian sheets.

Following the methodology developed by Nussbaum (1996), one may determine exact asymptotic constants for the case of bounded loss functions, but this approach is of limited use in our setting. One limitation is due to the fact that the main local equivalence result of Dalalyan and Reiß (2007) on equivalence with the Gaussian shift model holds for regularity

$$\beta > \left(d - 1 + \sqrt{2(d-1)^2 - 1}\right)/2 \sim \left(\frac{1}{2} + \frac{1}{\sqrt{2}}\right) d, \qquad \text{as } d \to \infty,$$

while asymptotic equivalence with heteroskedastic Gaussian regression is established for

$$\beta > \max\left\{\frac{d^2}{4} - 1, \left(d - 2 + \sqrt{(d-2)^2 + 4d^2}\right)/4\right\}.$$

As mentioned by the authors, it is not clear whether for Hölder classes of smaller regularity asymptotic equivalence holds or not. Even if equivalence would fail, the minimax risk asymptotics could possibly still hold in the diffusion model for substantially smaller smoothness index β. This conjecture is supported by the work of Korostelev and Nussbaum on the exact constant of the risk asymptotics in the sup-norm for density estimation. The constant for nonparametric regression (the "Korostolev case") and for signal estimation in Gaussian white noise (Donoho (1992)) was already known before, and a concise argument based on asymptotic equivalence in the Le Cam sense allows to derive the exact constant for density estimation under the additional assumption that the densities are uniformly bounded away from zero and that the smoothness index $\beta > 1/2$ (cf. Korostelev and Nussbaum (1995)). This last restriction is due to the fact that asymptotic equivalence is known to fail for Hölder smoothness $\beta \leq 1/2$. The direct approach in Korostelev and Nussbaum (1999) gives the sup-norm constant for density estimation for all $\beta > 0$ and without any further conditions on the density. The asymptotic equivalence approach to determine exact risk bounds thus clearly has its limitations. We prefer the approach to prove results in the spirit of (1.2.4) by verifying an upper bound and the corresponding lower bound on the risk directly.

1.2.1. Lower bounds

A convenient strategy to prove lower bounds for the maximum risk over a functional class $(\mathcal{F}_\nu)_{\nu\in\mathcal{N}}$ of the form

$$\liminf_{\varepsilon\to 0} \inf_{\widetilde{f}} \sup_{\nu\in\mathcal{N}} \sup_{f\in\mathcal{F}_\nu} \mathbb{E}_f\big(\psi_\nu^{-p}(\varepsilon)\ l^p(\widetilde{f}, f)\big) \geq 1, \qquad p > 0, \tag{1.2.7}$$

consists in bounding it from below by

- the Bayes risk over a parametric subfamily indexed by a continuous parameter, or
- the maximum or the average risk over a finite family of members of the class.

In the sequel, we shall briefly introduce two useful auxiliary results related to these ideas. A concise introduction, many concrete devices for proving lower bounds and references to the literature are given in Chapter 2 of Tsybakov (2009).

Van Trees' inequality. As noted in Tsybakov (2009), van Trees' inequality applies only to squared loss functions, but in this setting, it can lead in some cases to asymptotically exact lower bounds. Gill and Levit (1995) were apparently the first to use a version of van Trees' inequality for estimation of functionals, and Belitser and Levit (1995) show that the inequality can be applied for deriving the Pinsker constant. The general version below is given in Tsybakov (2009), Section 2.7.3. Let $T := [t_1, t_2] \subset \mathbb{R}$ such that $-\infty < t_1 < t_2 < +\infty$, and let $(\mathbb{P}_t)_{t\in T}$ be a family of probability measures on some measurable space $(\mathcal{X}, \mathscr{A})$. Assume that there exists a σ-finite measure ν on $(\mathcal{X}, \mathscr{A})$ such that $\mathbb{P}_t \ll \nu$ for all $t \in T$, introduce a probability distribution with Lebesgue density $\mu(\cdot)$ on T, and, given any estimator $\widehat{t}$, consider the Bayes risk with respect to μ, defined by

$$\mathscr{R}^\mu(\widehat{t}) := \int_T \mathbb{E}_t\big(\widehat{t} - t\big)^2 \mu(t)\mathrm{d}t.$$

Assume that the density $p(x,t)$ is measurable in (x,t) and absolutely continuous in t for almost all x with respect to ν, and that the prior density $\mu \ll \lambda\!\!\lambda$ on T and such that $\mu(t_1) = \mu(t_2) = 0$. Introduce the Fisher informations

$$\mathcal{I}(t) := \int \left(\frac{p'(x,t)}{p(x,t)}\right)^2 p(x,t)\nu(\mathrm{d}x), \qquad \mathcal{J}(\mu) := \int_T \frac{(\mu'(t))^2}{\mu(t)}\mathrm{d}t,$$

and assume that $\mathcal{I}(t) < \infty$ for all $t \in T$, $\int \mathcal{I}(t)\mathrm{d}t < \infty$ and $\mathcal{J}(\mu) < \infty$. In this set-up, van Trees' inequality states that, for any estimator $\widehat{t}(X)$,

$$\int_T \mathbb{E}_t \left(\widehat{t}(X) - t\right)^2 \mu(t)\mathrm{d}t \geq \frac{1}{\int I(t)\mu(t)\mathrm{d}t + \mathcal{J}(\mu)}.$$

Another important auxiliary result. Another tool which will be notedly useful in the sequel was developed by Tsybakov (1998) for proving lower bounds in the Gaussian white noise model (1.2.1). It has been used, e.g., in Butucea (2001), Klemelä and Tsybakov (2001) and Klemelä and Tsybakov (2004) for proving exact lower bounds on the pointwise risk. In order to prove a (sharp) lower bound for estimating functions $f \in \mathcal{F}_\nu$ at some fixed $x_0 \in (0,1)$, $\nu \in \mathcal{N}$ is some smoothness parameter, consider two hypotheses $f = f_0$ and $f = f_1$. Denote by $\mathbb{P}_{f_i}$ the distribution of observations satisfying (1.2.1) for the hypothesis $f = f_i$, and denote by $\mathbb{E}_{f_i}$ the expectation with respect to $\mathbb{P}_{f_i}$, $i = 0,1$. Then, if $f_0 \in \mathcal{F}_{\nu_0}$ and $f_1 \in \mathcal{F}_{\nu_1}$ for some $\nu_0, \nu_1 \in \mathcal{N}$, it holds for any $p > 0$

$$\begin{aligned}
&\inf_{\widetilde{f}} \sup_{\nu\in\mathcal{N}} \sup_{f\in\mathcal{F}_\nu} \mathbb{E}_f \big(\psi_\nu^{-p}(\varepsilon)\, |\widetilde{f}(x_0) - f(x_0)|^p\big) \\
&\qquad \geq \inf_{\widetilde{f}} \max \left\{ \mathbb{E}_{f_0}\big(\psi_{\nu_0}^{-p}(\varepsilon)\, |\widetilde{f}(x_0) - f_0(x_0)|^p\big),\ \mathbb{E}_{f_1}\big(\psi_{\nu_1}^{-p}(\varepsilon)\, |\widetilde{f}(x_0) - f_1(x_0)|^p\big) \right\}.
\end{aligned}$$

For suitably chosen hypothesis f_0 and f_1, this last term is bounded from below by

$$\inf_{\widetilde{f}} \max \left\{ Q^p\ \mathbb{E}_{f_0} D^p(\widetilde{f}, 0),\ \mathbb{E}_{f_1} D^p(\widetilde{f}, 1) \right\},$$

where, for any fixed $\delta \in (0, 1/2)$, $D(u,v) := (1-\delta)|u-v|$, $u, v \in \mathbb{R}$, and $Q := \psi_{\nu_0}/\psi_{\nu_1}$. If $\mathbb{P}_{f_0} \ll \mathbb{P}_{f_1}$ and, for fixed $\tau > 0$ and $\alpha \in (0,1)$,

$$\mathbb{P}_{f_1}\left(\frac{\mathrm{d}\mathbb{P}_{f_0}}{\mathrm{d}\mathbb{P}_{f_1}} \geq \tau\right) \geq 1 - \alpha,$$

then, for any $p > 0$, Theorem 6(i) in Tsybakov (1998) states that

$$\inf_{\widetilde{f}} \max \left\{ Q^p \mathbb{E}_{f_0} D^p(\widetilde{f}, 0), \mathbb{E}_{f_1} D^p(\widetilde{f}, 1) \right\} \geq \frac{(1-\alpha)\tau(Q\delta)^p(1-2\delta)^p}{(1-2\delta)^p + \tau(Q\delta)^p}. \tag{1.2.8}$$

If δ can be chosen such that the right-hand side of (1.2.8) tends to 1 (for $\varepsilon \to 0$), then a result in the spirit of (1.2.7) follows.

1.2.2. Upper bounds

There exists a large variety of methods to construct adaptive estimators, and the principle of unbiased risk estimation is at the bottom of many of them. A different idea is due to Oleg Lepski and is based on implicit bias-variance comparison schemes. For further information, see Tsybakov (2009).

1.3. Weak convergence and empirical processes

In the classical theory of empirical processes based on independent observations, one considers extensions of classical limit results such as law of large numbers, central limit theorem or law of the iterated logarithm. Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables, defined on some probability space $(\Omega, \mathscr{A}, \mathbb{P})$ and taking values in some space S. Denote the corresponding empirical measure by

$$\mathbb{P}_n(\omega) := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}, \qquad \omega \in \Omega,\ n \geq 1.$$

Let $\mathcal{F}$ be a countable class of functions on $(S, \mathscr{B}(S))$ such that $\sup_{f \in \mathcal{F}} |f(x)| < \infty$ for all $x \in S$. The *empirical process* $(\mathbb{G}_n(f))_{f \in \mathcal{F}}$ based on $\mathbb{P}$ and indexed by $\mathcal{F} \subset L^1(\mathbb{P})$ is defined as

$$\mathbb{G}_n(f) := \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \int f \mathrm{d}\mathbb{P} \right), \qquad f \in \mathcal{F},\ n \geq 1. \tag{1.3.1}$$

The object of empirical process theory is to study the properties of the approximation of the expectation of functions by their empirical measures, uniformly in $\mathcal{F}$. A class $\mathcal{F}$ is said to be a (strong) *Glivenko–Cantelli class for* $\mathbb{P}$ if, a.s.,

$$\lim_{n \to \infty} \sup_{f \in \mathcal{F}} \left| \int f (\mathrm{d}\mathbb{P}_n - \mathrm{d}\mathbb{P}) \right| = \lim_{n \to \infty} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \int f \mathrm{d}\mathbb{P} \right| = 0.$$

$\mathcal{F}$ is called a *Donsker class for* $\mathbb{P}$, or $\mathbb{P}$ is said to satisfy a CLT uniformly over $\mathcal{F}$, if

$$\mathbb{G}_n = \sqrt{n} (\mathbb{P}_n - \mathbb{P}) \rightsquigarrow \mathbb{G} \qquad \text{in } \ell^\infty(\mathcal{F}), \tag{1.3.2}$$

where the limit $\mathbb{G}$ is a tight, Borel measurable element in $\ell^\infty(\mathcal{F})$. Here and in what follows, $\rightsquigarrow$ denotes convergence in law of random elements in the generalized sense of Hoffmann-Jørgensen (cf. Dudley (1999), Chapter 3). By the finite-dimensional CLT, the limit process $(\mathbb{G}(f))_{f \in \mathcal{F}}$ must be a centered Gaussian process with covariance given by

$$\mathbb{E}\mathbb{G}(f)\mathbb{G}(g) = \mathbb{P}(f - \mathbb{P}f)(g - \mathbb{P}g) = \mathbb{P}(fg) - \mathbb{P}f\mathbb{P}g.$$

It is well-known that the limit exists as tight map into $\ell^\infty(\mathcal{F})$ if and only if $(\mathbb{G}(f))_{f \in \mathcal{F}}$ admits a version with bounded sample paths which are continuous on $\mathcal{F}$ with respect to the intrinsic semimetric induced by the limiting process $\mathbb{G}$, given by

$$\varpi(f, g) := \|(f - \mathbb{P}f) - (g - \mathbb{P}g)\|_{L^2(\mathbb{P})}, \qquad f, g \in \mathcal{F}.$$

A class $\mathcal{F}$ of measurable functions is called $\mathbb{P}$-*pregaussian* if the tight limit process in (1.3.2) exists. While a $\mathbb{P}$-Donsker class is necessarily $\mathbb{P}$-pregaussian, the reverse assertion does not hold. The first example of a pregaussian class which is not Donsker is attributed to Strassen and Dudley (1969). A class $\mathcal{F}$ is a $\mathbb{P}$-Donsker class if and only if the empirical processes $\mathbb{G}_n$ satisfy an *asymptotic equicontinuity* criterion.

1.4. The problems

1.4.1. Sharp adaptive drift estimation for ergodic diffusions in higher dimension

In the first part of the thesis, we are concerned with nonparametric estimation of the drift of an ergodic stationary diffusion X with invariant measure μ. We study the problem under the following assumptions:

- (*stationarity*) X is strictly stationary, that is, $X_0 \sim \mu$.
- (*continuous observations*) A continuous record of observations $X^T := (X_t)_{0 \le t \le T}$ is available.
- (*long-time asymptotics*) The process X is observed over $[0, T]$, and $T \to \infty$.

Subject to these constraints, we obtain a framework which presents a suitable starting point for providing as precise understanding of the drift estimation problem as possible. In particular, it allows to prove exact results which can be considered as a benchmark, e.g., for the case of discrete observations or finite-sample results.

Global drift estimation

For ease of presentation, let us restrict attention to the case of Kolmogorov diffusions introduced in Section 1.1.3. The relation

$$b(x) = (2\rho(x))^{-1}\nabla\rho(x) \tag{1.4.1}$$

suggests to decompose the drift estimation problem by estimating the invariant density ρ and its gradient separately. The same approach has been used by Dalalyan and Kutoyants (2002) who consider the problem of nonparametric estimation of the derivative of the invariant density and of the drift coefficient for scalar ergodic diffusion processes over weighted L^2 Sobolev classes. The construction of the asymptotically efficient estimator in Dalalyan and Kutoyants (2002) requires the knowledge of the smoothness and the radius of these weighted Sobolev balls. On the basis of these results, Dalalyan (2005) develops an adaptive procedure which does not depend on the characteristics of the Sobolev ball and which is asymptotically minimax simultaneously over a broad scale of Sobolev classes. The adaptive procedure relies on the principle of unbiased risk estimation and is inspired by Cavalier et al. (2002). We already mentioned that a real satisfactory theory only for one-dimensional diffusion processes is available, and the same assertion holds true for the statistical theory. The concept of local time and related tools such as the occupation times formula are frequently used in the works of Dalalyan and Kutoyants (2002) and Dalalyan (2005). One task to be solved in the sequel therefore consists in extending these results to the multidimensional setting and to identify a class of diffusion processes for which exact asymptotics can be described.

Pointwise drift estimation

To the best of our knowledge, exact results for pointwise drift estimation are not available. To construct an adaptive estimator of the drift at some fixed point $x_0 \in \mathbb{R}^d$, we again use the relation (1.4.1) and decompose the problem into the separate questions of estimating the invariant density and its gradient. Klemelä and Tsybakov (2001) consider the problem of estimating a linear functional $T(f)$, where f is an unknown function observed in the multidimensional Gaussian white noise model

$$\mathrm{d}Y(t) = f(t)\mathrm{d}t + \varepsilon\, \mathrm{d}B(t), \qquad t \in \mathbb{R}^d, \tag{1.4.2}$$

where B denotes the standard Brownian sheet in $\mathbb{R}^d$ and $\varepsilon \in (0, 1)$. A connection between sharp adaptation and optimal recovery is established, and asymptotically sharp adaptive estimators on various scales of smoothness classes are proposed. As an application, Klemelä and Tsybakov (2001) consider the problem of estimating the partial derivatives of f under Sobolev smoothness assumptions. In their follow-up article Klemelä and Tsybakov (2004), nonparametric estimation of a multivariate function and its partial derivatives at a fixed point for observations of the Riesz transform in Gaussian white noise is investigated. The Lepski-type adaptation procedure proposed there serves as a source for our drift estimation scheme. One further reference is

Butucea (2001) who considers the problem of exact adaptive estimation of a density f at a fixed point $x_0 \in \mathbb{R}$ on the Sobolev classes. In the density model, the value of the unknown density at the estimation point appears in the exact normalization. This requires to use a preliminary estimator of $f(x_0)$, and the same is to be expected in the adaptive drift estimation procedure.

1.4.2. Donsker theorems for multidimensional diffusions

The second part of the thesis is devoted to the investigation of infinite-dimensional extensions of the CLT for additive functionals, that is, the investigation of so-called *Donsker theorems*. The chapter is based on the preprint "Uniform central limit theorems for multidimensional diffusions" which is a joint work with Angelika Rohde. All results which are taken from the preprint into this chapter are marked.

Let $(X_t)_{t\geq 0}$ be an ergodic diffusion process on $E \subset \mathbb{R}^d$ with invariant measure μ. A strong law of large numbers for additive functionals of $(X_t)_{t\geq 0}$ is available. Precisely, if $X_0 \sim \mu$, then it holds for any $f \in L^1(\mu)$, a.s. and in L^1,

$$\lim_{t\to\infty} \left(\frac{1}{t}\int_0^t f(X_u)\,\mathrm{d}u\right) = \int f \mathrm{d}\mu.$$

Assume that $X_0 \sim \mu$, and consider $0 \not\equiv f \in L^2(\mu)$ with $\int f\mathrm{d}\mu = 0$. The additive functional $\left(\int_0^t f(X_u)\,\mathrm{d}u\right)_{t\geq 0}$ is said to satisfy a central limit theorem (CLT) if

$$\frac{1}{\sqrt{t}}\int_0^t f(X_u)\,\mathrm{d}u \Rightarrow \mathcal{N}(0,1). \tag{1.4.3}$$

Such limit results can be proven under mild assumptions; we refer to Chapter VII in Jacod and Shiryaev (2002) and to Cattiaux et al. (2012) for an account of situations where the CLT for additive functionals of ergodic diffusions holds. The goal in the sequel is to find uniform versions of (1.4.3). Results of Glivenko–Cantelli or Donsker type for one-dimensional diffusions are given, e.g., in van Zanten (2003) and van der Vaart and van Zanten (2005), respectively. Precisely, van Zanten (2003) considers a regular scalar diffusion process X with finite speed measure m and invariant measure μ. The uniform law of large numbers

$$\lim_{t\to\infty}\sup_{f\in\mathcal{F}}\left|\frac{1}{t}\int_0^t f(X_u)\,\mathrm{d}u - \int f\mathrm{d}\mu\right| = 0$$

is shown to hold for any class $\mathcal{F}$ which either has an μ-integrable envelope function or which is bounded in $L^p(\mu)$ for some $p > 1$. The main result of van der Vaart and van Zanten (2005) is that, whenever the limit $\mathbb{G}$ exists as a tight, Borel measurable map, it holds $\mathbb{G}_t \Rightarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$. In contrast with uniform laws of large numbers and uniform CLTs for i.i.d. random variables, no conditions on the "size" of the function class $\mathcal{F}$ in terms of bracketing or covering numbers are required. The results on empirical processes of one-dimensional diffusions are a consequence of a number of asymptotic properties of diffusion local time and therefore restricted to the scalar case.

For some countable function class $\mathcal{F} \subset L^2(\mu)$ of measurable functions, define the *empirical diffusion process* in analogy to (1.3.1) as

$$\mathbb{G}_t(f) := \sqrt{t}\left(\frac{1}{t}\int_0^t f(X_u)\,\mathrm{d}u - \int_E f\mathrm{d}\mu\right), \qquad f \in \mathcal{F},\ t > 0. \tag{1.4.4}$$

The motivation for investigating empirical diffusion processes as in (1.4.4) is two-fold. First, we aim at contributing to the theory of empirical processes of multidimensional diffusions. Various methods have been developed to prove Donsker properties, for example, bracketing conditions, arguments from probability in Banach spaces, Vapnik–Chervonenkis type arguments or random entropy criteria. With a view towards statistical applications, it is of interest to learn which of the classical techniques and tools work well and can be used for establishing Donsker-type theorems in the context of empirical diffusion processes. On a more abstract level, the goal is to establish uniform CLTs (or modified versions thereof) for multidimensional diffusions under necessary and sufficient conditions.

1.5. Outline of the thesis

As measured by their substantial importance and in view of the well-developed statistical theory for scalar diffusions, only few theoretical results on multidimensional diffusions are available. This is at least partially due to the lack of local time in the multidimensional framework. Both the work of Dalalyan (2005) on sharp adaptive drift estimation and the result of van der Vaart and van Zanten (2005) on Donsker theorems for scalar regular diffusions rely on the concept of local time.

The goal of the first part of the thesis (Chapter 2) is to investigate the exact asymptotics of minimax adaptive risks under reasonable smoothness assumptions on the drift and to construct sharp adaptive estimators, that is, adaptive estimators which do not only achieve the best possible rate of convergence but the best asymptotic constant associated to it. Two situations will be considered, namely

- global estimation, that is, estimation with respect to some global distance measure, and
- pointwise estimation, that is, estimation of b in some fixed $x_0 \in \mathbb{R}^d$.

From the outset, it is neither clear whether optimal rate adaptive estimators exist nor how the exact constants look like. We further have to carve out “reasonable” smoothness assumptions on the drift, taking care of its implications on the existence of a solution of the associated Itô SDE, existence of the invariant measure and regularity properties of its invariant density. Dalalyan and Kutoyants (2002) and Dalalyan (2005) advocate the use of a local minimax approach for global estimation of the drift of scalar diffusion processes. We shall explore the implications of this approach in the multidimensional diffusion framework. One reference point is the work of Dalalyan and Reiß (2007) who establish asymptotic statistical equivalence between the diffusion experiment and the Gaussian shift model. Their results hold for Kolmogorov diffusions under rather restrictive Hölder smoothness assumptions on the drift. Exact asymptotic findings for smoothness classes of smaller regularity could serve as an indication that statistical equivalence, at least in some reduced sense, still holds. Leaving aside the case of Kolmogorov diffusions, it is in particular interesting to describe the influence of the diffusion coefficient on the drift estimation problem. We assume that a continuous record of observations $(X_t)_{0\leq t\leq T} =: X^T$ is available, and this implies that the diffusion coefficient $\sigma\sigma^t$ can be identified using the semimartingale quadratic variation of the diffusion X. Although it is identifiable at any point visited by X^T, it is to be expected that the diffusion coefficient still influences the drift estimation procedure.

In the second part of this work (Chapter 3), Donsker-type theorems for empirical processes of multidimensional diffusions are studied. The proof of such results requires verifying asymptotic tightness and convergence of the marginal distributions. We will use results in the spirit of the (functional) CLT due to Bhattacharya (1982) for establishing marginal convergence. Explicit criteria which ensure asymptotic tightness will be given. The results of van der Vaart and van Zanten (2005) reflect the fact that the empirical measure of scalar diffusions is substantially smoother than the empirical measure based on i.i.d. observations. Such a phenomenon is known in a less distinctive degree for multidimensional diffusions which satisfy suitable functional inequalities. In order to establish increased regularity, we study smoothed versions of the empirical diffusion process. This allows in particular to establish uniform CLTs for smoothed empirical processes of Kolmogorov diffusions with very fast bandwidths under only pregaussian conditions, thus verifying the exceptional regularity of empirical processes of multidimensional ergodic diffusions.

A short selective overview of central results obtained in the thesis is given in Chapter 4. In addition, we sketch possible extensions and questions which might be tackled in the future. For the convenience of the reader, some of the auxiliary results which were used in Chapters 2 and 3 are summarized in the Appendix.

Acknowledgements

Danken möchte ich an erster Stelle Frau Professor Dr. Angelika Rohde, von der ich ungemein viel lernen konnte. Es war ein großes Glück, eine solche Betreuerin zu finden, die stets bereit war, ihr Wissen und ihre Ideen mit mir zu teilen, mich in die Welt der Wissenschaft einzuführen und mir dort eine Perspektive zu bieten. Für fachliche Fragen ist sie jederzeit ansprechbar, und ihre unerschrockene Art und Weise, sich mathematischen Herausforderungen zu stellen, hat mir stets Mut gemacht.

Herrn Professor Dr. Enno Mammen danke ich vielmals für sein großes Entgegenkommen und seine Bereitschaft, das Zweitgutachten zu verfassen.

Die Universität Hamburg bot mir die Gelegenheit, an dieser Dissertation zu arbeiten. Mein Dank gilt insbesondere den Kollegen des Fachbereichs Stochastik, die ich vom Anfang bis hin zum Ende meiner Tätigkeit stets als ausgesprochen hilfsbereit und aufgeschlossen erlebt habe.

Von ganzem Herzen danke ich schließlich all jenen in meinem privaten Umfeld, die es in den letzten Jahren nicht immer leicht mit mir hatten. In Zukunft hoffe ich vieles von dem zurückgeben zu können, was Freunde und insbesondere meine Mutter und meine Tante für mich getan haben. Auch wenn ich es in der letzten Zeit nicht immer genügend zum Ausdruck gebracht habe: Ich weiß genau, wem ich etwas zu verdanken habe!

The chapter “Donsker theorems for multidimensional ergodic diffusions” is based on the preprint “Uniform central limit theorems for multidimensional diffusions” (`arXiv:1010.3604`) which arose in joint work with Angelika Rohde. All results in this dissertation which are borrowed from the preprint are marked as take-overs.

2. Sharp adaptive drift estimation for ergodic diffusions in higher dimension

Let X be an ergodic multidimensional diffusion process with drift function b. We consider the problem of adaptive estimation of b. Exact data-driven procedures both for global and pointwise estimation are proposed, attaining the optimal constant under natural smoothness conditions on the drift. Both a connection to the classical Pinsker result on estimation with respect to L^2 risk over Sobolev classes and to the problem of optimal recovery are established. The sharp results in particular allow to evaluate the influence of the diffusion matrix.

2.1. Introduction and overview of the results

The setting. Consider the diffusion process X which is given as a strong solution of the stochastic differential equation

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \qquad X_0 = \xi,\ t \geq 0, \tag{2.1.1}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ are the unknown characteristics, $W = (W_t)_{t\geq 0}$ is a d-dimensional Brownian motion and the initial vector $\xi \in \mathbb{R}^d$ is independent of W. Given a continuous record of observations $(X_t)_{0\leq t\leq T} =: X^T$, $T > 0$, the aim is to estimate the drift vector b. We restrict attention to ergodic diffusions with unique invariant probability measure μ which is absolutely continuous with respect to Lebesgue measure. Denote by ρ the invariant density of μ. The initial value ξ is assumed to follow the invariant law such that the process X is strictly stationary. Throughout, we consider the case of continuous-time observations of the diffusion process, thus laying the foundation and providing benchmarks for the more involved setting of (possibly low-frequency) discrete-time data. In the continuous-time framework, the diffusion matrix $a := \sigma\sigma^t$ is identifiable by means of the semimartingale quadratic variation, and the problem of its estimation does not arise. Nevertheless, it is to be expected that the characteristics of the diffusion coefficient have an impact on the drift estimation problem, and one focus in the sequel will be on identifying the influence of the diffusion matrix.

The estimators. Let $K : \mathbb{R}^d \to \mathbb{R}$ be a fixed function satisfying $\int_{\mathbb{R}^d} K(x)\mathrm{d}x = 1$. Given a bandwidth $h > 0$, the invariant density ρ can be estimated by a kernel estimator of the form

$$\widehat{\rho}_T(x) := \frac{1}{Th^d}\int_0^T K\left(\frac{X_u - x}{h}\right)\mathrm{d}u = \frac{1}{T}\int_0^T K_h(X_u - x)\mathrm{d}u, \qquad x \in \mathbb{R}^d, \tag{2.1.2}$$

where $K_h(\cdot) := h^{-d}K(\cdot/h)$. There exist several approaches to motivate drift estimators. A kernel estimator of b which is suggested by analogy to the model of regression with random design is given by

$$\overline{b}_T(x) := \frac{T^{-1}\int_0^T K_h(X_u - x)\mathrm{d}X_u}{\widehat{\rho}_T(x) \vee \rho_*(x)}, \qquad x \in \mathbb{R}^d, \tag{2.1.3}$$

for some a priori lower bound $\rho_*(x) > 0$ on $\rho(x)$. If such a lower bound ρ_* is not available, an additional carefully chosen term may be included in the denominator for ensuring that it does not vanish too rapidly.

Alternatively, one may use an analytical approach for suggesting a drift estimator. Under specific smoothness conditions on the coefficients of the diffusion X, the j-th component b^j of the drift vector b can be expressed as follows,

$$b^j(x) = (2\rho(x))^{-1} \sum_{k=1}^{d} \partial_k (a_{jk}(x)\rho(x)), \qquad j \in \{1, \ldots, d\},\ x \in \mathbb{R}^d, \tag{2.1.4}$$

where $a := \sigma\sigma^t$ is the diffusion matrix with entries (a_{jk}). This suggests the following kernel estimator for the drift vector $b(\cdot)$,

$$\widehat{b}_T(x) := \frac{T^{-1}h^{-d} \int_0^T a(X_u)\nabla K_h(X_u - x)\mathrm{d}u}{2\left(\widehat{\rho}_T(x) \vee \rho_*(x)\right)}, \qquad x \in \mathbb{R}^d. \tag{2.1.5}$$

Further details on the definition of drift estimators are given in Section 2.2.3.

The problem. The central problem in any kernel-based estimation procedure is the selection of an appropriate bandwidth. Two different approaches to the question of bandwidth selection will be considered, depending on the concrete statement of the problem. First, the drift function b is estimated globally, and the goal is to find estimators $\widetilde{b}_T$ which are optimal with respect to the weighted L^2 risk

$$\mathcal{R}_{\mathrm{glob}}(\widetilde{b}_T, b) := \mathbf{E}_{b,\sigma} \int_{\mathbb{R}^d} \left\|\widetilde{b}_T(x) - b(x)\right\|_2^2 \rho^2(x)\mathrm{d}x, \tag{2.1.6}$$

where $\mathbf{E}_{b,\sigma}$ is the expectation with respect to the invariant measure associated with the coefficients b and σ, and $\|\cdot\|_p$ denotes the p-norm,

$$\|x\|_p := \left(\sum_{i=1}^{d} |x_i|^p\right)^{1/p}, \qquad p \geq 1,\ x \in \mathbb{R}^d.$$

The use of the squared density as a weight function presents one difference to the definition of the mean integrated squared error commonly employed in the classical statistical models. In combination with growth conditions on the drift coefficient, it allows to formulate clear results for drift estimation on $\mathbb{R}^d$. As an alternative to global criteria, one might be interested in estimating the drift in some fixed $x_0 \in \mathbb{R}^d$ in which case the quality of an estimator $\widetilde{b}_T$ is measured by its pointwise risk,

$$\mathcal{R}_{\mathrm{point}}(\widetilde{b}_T, b) := \mathbf{E}_{b,\sigma}\left\|\widetilde{b}_T(x_0) - b(x_0)\right\|_2^2.$$

The goal in both settings is to define minimax adaptive estimators b_T^*, b_T^{**} satisfying

$$\mathcal{R}_{\mathrm{glob}}(b_T^*, b) = \inf_{\widetilde{b}_T} \sup_{b \in \mathcal{B}} \mathcal{R}_{\mathrm{glob}}(\widetilde{b}_T, b) \quad \text{and} \quad \mathcal{R}_{\mathrm{point}}(b_T^{**}, b) = \inf_{\widetilde{b}_T} \sup_{b \in \mathcal{B}} \mathcal{R}_{\mathrm{point}}(\widetilde{b}_T, b),$$

respectively. The infimum is taken over all estimators $\widetilde{b}_T$ of the drift b, and the supremum extends over a given class of functions $\mathcal{B}$. Typically, $\mathcal{B}$ is a class of smooth functions, for example of Sobolev type. In practice, the smoothness parameters usually are unknown, and this necessitates the usage of data-driven estimation procedures. This chapter is devoted to the development of such procedures for estimating the drift vector of SDEs in the Itô sense. Let us point out the following:

- We propose exact adaptive procedures for global and pointwise estimation of the components of the drift vector over L^2 Sobolev classes, and we provide Pinsker-type results and findings on the relation between optimal recovery and nonparametric pointwise estimation in the diffusion framework.
- The exact results apply to a broad class of ergodic reversible diffusion processes which satisfy the Poincaré hypothesis. We do not explicitly use mixing properties by referring to decoupling results but rely on analytical tools related to the Markovian semigroup character of the diffusion.

The agenda. Recall that $b : \mathbb{R}^d \to \mathbb{R}^d$ denotes the drift vector with components b^j, $j \in \{1,\dots,d\}$, and $a := \sigma\sigma^t$ is the diffusion matrix with entries a_{ij}, $i,j \in \{1,\dots,d\}$. Provided that (2.1.4) is satisfied, both numerators of (2.1.3) and (2.1.5) can be considered as consistent estimators of the vector

$$\operatorname{div}(a\rho) = \begin{pmatrix} \sum_{k=1}^d \partial_k(a_{1k}\rho) \\ \vdots \\ \sum_{k=1}^d \partial_k(a_{dk}\rho) \end{pmatrix}.$$

The estimators defined in (2.1.3) and (2.1.5), respectively, thus suggest to decompose the problem of drift estimation into the separate questions of estimating the invariant density ρ and the divergence $\operatorname{div}(a\rho)$. In many cases, the invariant density of the diffusion can be estimated faster than the numerators in (2.1.3) and (2.1.5), so the main focus is on finding exact asymptotics for estimating $\operatorname{div}(a\rho)$.

There is a substantial difference between the problem of minimax adaptive estimation with respect to L^2 risk on the one hand and with respect to pointwise loss on the other hand. The classical result on nonparametric function estimation in the Gaussian white noise model for L^2 risk is due to Efroimovich and Pinsker (1984) who have shown that there is no loss of efficiency for adaptation over the scale of L^2 Sobolev classes. Concerning the question of pointwise adaptive estimation on the scale of L^2 Sobolev classes, Tsybakov (1998) proves that optimal rate adaptive estimators do not exist but that there is a loss of efficiency under adaptation. We recover the same situation for estimating the components of the vector $\operatorname{div}(a\rho)$ under reasonably mild conditions on the diffusion.

(i) As will be explained in Section 2.3.1, the asymptotically efficient divergence estimator in the global minimax sense does not fit well to the problem of drift estimation. We therefore adopt the local minimax approach considered in Dalalyan (2005). Fix $j \in \{1,\dots,d\}$. Given some fixed (sufficiently regular) central function $b_0 : \mathbb{R}^d \to \mathbb{R}^d$, some set $\Sigma^j(\beta) \subset C^\beta(\mathbb{R}^d)$, $\beta \in \mathbb{N}$, and $\delta > 0$, define the neighborhood

$$U_\delta^j(b_0) := \left\{ b : \mathbb{R}^d \to \mathbb{R}^d : b \in \Sigma^j(\beta);\ \forall k \in \{1,\dots,d\},\ \sup_{x\in\mathbb{R}^d} \left|b^k(x) - b_0^k(x)\right| \le \delta \right\}. \quad (2.1.7)$$

Under the assumption that $\operatorname{div}(a_j\rho)$ belongs to the intersection Σ_δ^j of $U_\delta^j(b_0)$ and an (isotropic) Sobolev ball of unknown integer regularity $\beta > 1$ with radius $L > 0$, we find the constant $P_j^*(\sigma,\beta,L)$ such that

$$\liminf_{\delta\to 0} \liminf_{T\to\infty} \inf_{\widehat{g}_T^j} \sup_{\operatorname{div}(a_j\rho)\in\Sigma_\delta^j} T^{\frac{2\beta}{2\beta+d}}\, \mathbf{E}_b \int_{\mathbb{R}^d} \left|\widehat{g}_T^j(x) - \operatorname{div}(a_j\rho)(x)\right|^2 \mathrm{d}x = P_j^*(\sigma,\beta,L), \quad (2.1.8)$$

the infimum being taken over all estimators $\widehat{g}_T^j$ of $\operatorname{div}(a_j\rho)$. The adaptive estimator defined in (2.3.72) in Section 2.3 asymptotically attains the infimum in (2.1.8).

(ii) In the pointwise adaptive procedure, we consider the scale $(\beta, L) \in [\beta_*, \beta_T] \times [L_*, L^*]$ of Sobolev classes $\mathcal{S}(\beta, L)$, for fixed positive numbers β_*, L_*, L^*, and some sequence $\beta_T \to \infty$ slowly enough. We find the normalizing factor $\psi_T(\beta, L; \rho, \sigma)$ which cannot be improved in the rate and in the constant, and we construct an adaptive estimator $\widetilde{g}_T^j$ (cf. (2.5.22)) satisfying this upper bound.

The proof of the results stated in (i) and (ii) mainly relies on developing analogues of classical statistical procedures in the diffusion setting. Due to the complexity of the diffusion model, the extension is not always straightforward. Let us briefly sketch central points of the proof of lower bounds and the construction of adaptive procedures for estimation of the components of the vector $\operatorname{div}(a\rho)$ in the respective settings. For ease of presentation, it is assumed in the sequel that a is constant.

Lower bounds. In the global setting, we reduce the proof of the lower bound for estimating the components $\operatorname{div}(a_j\rho) = \sum_{k=1}^d a_{jk}\ \partial_k\rho,\ j \in \{1,\ldots,d\}$, to proving a lower bound for the Bayes risk over some conveniently chosen parametric subfamily. The Bayes risk is dealt with by means of van Trees' inequality which was already introduced in Section 1.2.1. The idea is classical and has been used before, e.g., in Golubev and Levit (1996) who consider the problem of distribution function estimation from i.i.d. observations. Dalalyan and Kutoyants (2002) modify the approach for proving a lower bound for estimating the derivative of the invariant density of scalar ergodic diffusions. In particular, they exploit the fact that the relation between the invariant density of scalar ergodic diffusions and the drift coefficient is explicit. We start the proof of the lower bound for estimating $\operatorname{div}(a_j\rho)$ by defining a parametric family of invariant densities ρ_θ, motivated by the respective definitions in Golubev and Levit (1996). The drift function is then suitably defined to ensure that $\operatorname{div}(a_j\rho_\theta) = 2b^j\rho_\theta$.

The proof of the lower bound for pointwise estimation of $\operatorname{div}(a_j\rho)$ relies on suitable modifications of the proof of lower bounds in Tsybakov (1998), Butucea (2001), Klemelä and Tsybakov (2001) and Klemelä and Tsybakov (2004). The basic idea is to reduce the proof of a lower bound over Sobolev-type smoothness classes to verifying a lower bound for the risk of two suitably chosen hypotheses. Our definition of the invariant density hypotheses ρ_0 and ρ_1 is motivated by the choice of Butucea (2001) who considers pointwise density estimation over Sobolev classes from i.i.d. observations. The drift vector and the divergence are then defined such that $\operatorname{div}(a_j\rho_i) = 2b_i^j\rho_i,\ i \in \{1,2\}$.

The adaptive estimation procedures. In the global framework, we suggest a multidimensional version of the approach proposed in Dalalyan (2005) for estimating $\operatorname{div}(a\rho)$. It relies on the empirical risk minimization method. The local time techniques used at several places in the scalar case do not extend to the multivariate setting. Assuming that Poincaré's inequality holds, sharp adaptivity can be proven with alternative tools. The adaptive procedure for estimating the components of $\operatorname{div}(a\rho)(x_0)$ pointwise is inspired by the schemes in Klemelä and Tsybakov (2001) and Klemelä and Tsybakov (2004) who use a version of Lepski's method. In contrast to the Gaussian white noise setting considered in these works, the complexity of the diffusion model requires *uniform* exponential inequalities. These can be proven by chaining techniques, starting, e.g., from a Bernstein-type deviation inequality for reversible Markov processes.

We will see below that − under mild auxiliary conditions − the results on exact adaptive estimation of $\operatorname{div}(a_j\rho)$ in (i) and (ii) allow to derive results on exact adaptive component-wise estimation of the drift vector, both in the L^2 risk and in the pointwise case.

Outlook and one first resume. We will establish minimax results for estimating the drift for a class of multidimensional ergodic diffusion processes. Having regard to the fact that the actual smoothness of the drift coefficient is typically unknown, adaptive estimation procedures are proposed. In particular, these schemes are asymptotically exact, that is, they attain the lower bounds and converge with the best possible rate to the best possible constant.

In correspondence to the use of the **weighted** L^2 risk criterion in (2.1.6) for quantifying the global risk of drift estimators, it appears appropriate to consider adaptation over **weighted** Sobolev balls of order $\beta \in \mathbb{N}$. The global adaptive rate of convergence is given as $T^{-2\beta/(2\beta+d)}$. Denoting by $\mathbb{S}_d := 2\pi^{d/2}/\Gamma(d/2)$ the surface of the unit sphere in $\mathbb{R}^d$, where $\Gamma(\cdot)$ is the Gamma function, the asymptotically exact constant for estimating one component of the drift of some Kolmogorov diffusion with $\sigma \equiv \mathbf{Id}_{d\times d}$ is identified as

$$\frac{(L(2\beta+d))^{\frac{d}{2\beta+d}}}{d}\left(\frac{\beta\,\mathbb{S}_d}{(2\pi)^d(\beta+d)}\right)^{\frac{2\beta}{2\beta+d}}, \qquad \beta > 1,\ L > 0. \tag{2.1.9}$$

For $d=1$, the constant in (2.1.9) coincides with Pinsker's constant (see, e.g., Tsybakov (2009), p. 138), and, for $d \geq 1$, it equals the constant obtained in Rigollet and Tsybakov (2007).

For pointwise adaptive estimation of the components of the drift vector, we consider some scale of Sobolev classes $(\Pi(\beta,L))_{(\beta,L)\in\mathcal{B}_T}$, where, for $\beta_* > d/2$ and $0 < L_* < L^* < \infty$,

$$\mathcal{B}_T := \{(\beta,L) : \beta_* \leq \beta < \beta_T,\ L_* \leq L \leq L^*\}, \qquad \beta_T = (\log\log T)^\delta,\ \delta \in (0,1) \text{ fixed.}$$

The exact constant for estimating one component of the drift of Kolmogorov diffusions with respect to the pointwise mean-squared error is given as

$$\frac{L^{\frac{d}{2\beta}}}{\rho(x_0)}\ \frac{2\beta}{d}\left(\frac{d^2\ \rho(x_0)}{\beta(2\beta-d)}\right)^{\frac{\beta-d/2}{2\beta}}\ \mathbb{I}_\beta, \qquad \beta > \frac{d}{2},\ L > 0. \tag{2.1.10}$$

Here, with $B(\cdot,\cdot)$ denoting the Beta function,

$$\mathbb{I}_\beta^2 := \frac{1}{2\beta}B\left(1+\frac{d}{2\beta}, 1-\frac{d}{2\beta}\right)(2\pi)^{-d}\mathbb{S}_d = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\frac{\|\lambda\|^{2\beta}}{(1+\|\lambda\|^{2\beta})^2}\,\mathrm{d}\lambda.$$

For $d=1$, the expression in (2.1.10) coincides with the constant obtained by Butucea (2001) for exact adaptive estimation of some density ρ at the point $x_0 \in \mathbb{R}$ under Sobolev smoothness assumptions from i.i.d. observations. In dimension $d \geq 1$, (2.1.10) equals the constant for (classical) estimation of some density ρ which is to be conjectured from the results obtained in Klemelä and Tsybakov (2004).

The accordance of the constants in the exact asymptotics mentioned above reflects the statistical folklore that different statistical models such as Gaussian white noise, nonparametric regression or density estimation show similar behavior, at least from an asymptotic point of view. Dalalyan and Reiß (2007) prove asymptotic statistical equivalence in the sense of Le Cam for inference on the drift in multidimensional ergodic (Kolmogorov) diffusions. Their results hold only for large enough Hölder smoothness of the drift coefficient which is substantially larger than the lower bound of $d/2$ which would correspond to the results of Brown and Zhang (1998). Our exact asymptotic results on drift estimation indicate that asymptotic equivalence – at least in some reduced sense – also holds under smoothness assumptions less severe than those imposed in Dalalyan and Reiß (2007).

Outline of the chapter. Basic assumptions on the diffusion processes are summarized in Section 2.2. We further recall essential ideas for nonparametric estimation of the drift coefficient and the invariant density of diffusions. The exact adaptive estimation procedures are described in Section 2.3 (global case) and Section 2.5 (pointwise setting). The proofs of lower and of the upper bounds for the adaptive estimation procedure are deferred to Section 2.4 (global setting) and Section 2.6 (pointwise case), respectively.

2.2. Preliminaries

This section subsumes preliminaries and general material from nonparametric statistics for diffusion processes which are needed in the sequel.

2.2.1. General definitions and notation

For $g : \mathbb{R}^d \to \mathbb{R}^d$, denote by g^j its j-th component. For a smooth function $f : \mathbb{R}^d \to \mathbb{R}$, let $\partial_j f := \partial f/\partial x^j$, $\partial^2_{jk} f := \partial^2 f/(\partial x^j \partial x^k)$, and denote its gradient by $\nabla f = (\partial_j f)_j$. For smooth

$g : \mathbb{R}^d \to \mathbb{R}^d$, denote by $\operatorname{div} g = \sum_j \partial_j g^j$ its divergence. The map $a : \mathbb{R}^d \to \mathbb{R}^d \otimes \mathbb{R}^d$ is viewed as $d \times d$ matrices whose rows are denoted by a_j, and $\operatorname{div} a$ denotes the vector $(\operatorname{div} a_j)_j$. Let ϕ_f be the Fourier transform of $f \in L^2(\mathbb{R}^d)$, that is,

$$\phi_f(\lambda) := \int_{\mathbb{R}^d} f(x) \exp(\mathrm{i}\lambda^t x)\mathrm{d}x, \qquad \lambda \in \mathbb{R}^d.$$

For $f \in L^p(\mu) := L^p(\mathbb{R}^d, \mu)$, $p \geq 1$, set $\|f\|_{L^p(\mu)} := (\int_{\mathbb{R}^d} |f|^p \mathrm{d}\mu)^{1/p}$. The Euclidean norm for $x \in \mathbb{R}^d$ is denoted by $\|x\|_2 =: \|x\|$, and $\lfloor a \rfloor$ is the largest integer strictly smaller than $a \in \mathbb{R}$. For any multi-index $\alpha \in \mathbb{N}^d$ and $x \in \mathbb{R}^d$, set $|\alpha| := \sum_{i=1}^d \alpha_i$, $x^\alpha := \prod_{i=1}^d x_i^{\alpha_i}$, and

$$D^\alpha f := \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}, \qquad f \in C^{|\alpha|}(\mathbb{R}^d).$$

For $\beta, L > 0$, the isotropic Hölder class $\mathcal{H}(\beta, L)$ is defined as

$$\mathcal{H}(\beta, L) := \begin{cases} \left\{ f : \mathbb{R}^d \to \mathbb{R} : |f(x) - f(y)| \leq L\|x-y\|^\beta \right\}, & \beta \leq 1, \\ \left\{ f \in C^{\lfloor \beta \rfloor}(\mathbb{R}^d) : \left| f(x) - P_y^{(f)}(x) \right| \leq L\|x-y\|^\beta \right\}, & \beta > 1, \end{cases} \tag{2.2.1}$$

where $P_y^{(f)}$ is the Taylor polynomial of f at the point y up to order $\lfloor \beta \rfloor$. The isotropic Sobolev class $\mathcal{S}(\beta, L)$ is given as

$$\mathcal{S}(\beta, L) := \left\{ f \in L^2(\mathbb{R}^d) : \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \|\lambda\|^{2\beta} |\phi_f(\lambda)|^2 \,\mathrm{d}\lambda \leq L^2 \right\}. \tag{2.2.2}$$

For integer $\beta > 0$, it holds

$$\mathcal{S}(\beta, L) = \left\{ f \in L^2(\mathbb{R}^d) : \sum_{|\alpha| \leq \beta} \int_{\mathbb{R}^d} |D^\alpha f|^2 \leq L^2 \right\},$$

where

$$D^\alpha f(x) = \mathrm{i}^{|\alpha|} \int_{\mathbb{R}^d} \lambda^\alpha \phi_f(\lambda) \,\mathrm{e}^{-\mathrm{i}\lambda^t x} \mathrm{d}\lambda, \qquad x \in \mathbb{R}^d.$$

More generally, for $p, r \in \mathbb{N}$, denote by $W^{p,r}(\mathbb{R}^d)$ the Sobolev class of functions that belong to $L^p(\mathbb{R}^d)$, together with their partial weak derivatives up to order r. For integer $l \geq 1$, a function $K : \mathbb{R}^d \to \mathbb{R}$ will be called kernel of order l if the functions $u \mapsto (u^i)^j K(u)$, $j = 0, 1, \dots, l$, $i = 1, \dots, d$, are integrable and satisfy $\int_{\mathbb{R}^d} K(u)\mathrm{d}u = 1$ and

$$\int_{\mathbb{R}^d} (u^i)^j K(u)\mathrm{d}u = 0, \qquad i = 1, \dots, d, \ j = 1, \dots, l.$$

2.2.2. Diffusion processes: Notation and basic assumptions

Basic assumptions on diffusion processes which allow for an in-depth analysis were already introduced in Section 1.1.3. Some care is required indeed in order to define a class of diffusion processes as broad as possible for which results on exact adaptive estimation of the drift function can be derived.

Remark 2.2.1 (The one-dimensional case). Properties of scalar diffusions are often verified straightforwardly. Indeed, consider some one-dimensional diffusion process satisfying the SDE

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \qquad X_0 = \xi, \ t \geq 0,$$

where W_t is a standard Wiener process on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and the initial value $\xi \in \mathbb{R}$ is independent of W. If

$$\lim_{|x|\to\infty} \int_0^x \frac{b(y)}{\sigma^2(y)}\mathrm{d}y = \infty \quad \text{and} \quad N(b) := \int_{-\infty}^{\infty} \sigma^{-2}(x) \exp\left(2\int_0^x \frac{b(y)}{\sigma^2(y)}\mathrm{d}y\right)\mathrm{d}x < \infty,$$

then the diffusion X is positive recurrent (cf. Chapter IV.2 in Mandl (1968)), and the invariant density is given as

$$\rho_b(x) := \frac{1}{N(b)\sigma^2(x)} \exp\left(2\int_0^x \frac{b(y)}{\sigma^2(y)}\mathrm{d}y\right). \tag{2.2.3}$$

It is immediately verified that

$$\left(\sigma^2(x)\rho_b(x)\right)' = \frac{\mathrm{d}}{\mathrm{d}x}\left\{\frac{1}{N(b)} \exp\left(2\int_0^x \frac{b(y)}{\sigma^2(y)}\mathrm{d}y\right)\right\} = 2b(x)\rho_b(x),$$

that is, (2.1.4) holds.

We proceed with introducing a set of assumptions which will play an important role in our subsequent investigation in the multidimensional framework. For fixed $j \in \{1, \ldots, d\}$ and given some $\beta \in \mathbb{N}$, denote by $\Sigma^j(\beta)$ the set of drift functions satisfying assumptions $(\mathrm{C_0}^j)$–$(\mathrm{C_2})$ below.

$(\mathrm{C_0}^j)$ The SDE

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma\ \mathrm{d}W_t, \tag{2.2.4}$$

where $b \in C^\beta(\mathbb{R}^d)$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is non-degenerate, admits a strong solution with Lebesgue continuous invariant measure $\mathrm{d}\mu(x) = \rho(x)\mathrm{d}x$ which satisfies the relation

$$2b^j\rho = \operatorname{div}(a_j\rho) = \sum_{k=1}^d a_{jk}\ \partial_k\rho. \tag{2.2.5}$$

Here and throughout the sequel, $a = \sigma\sigma^t$ denotes the diffusion matrix associated to the dispersion coefficient σ.

$(\mathrm{C_1})$ For some $c_1 \in (0, \infty]$, it holds

$$\limsup_{\|x\|\to\infty} \left\langle b(x), \frac{x}{\|x\|^2}\right\rangle = -c_1. \tag{2.2.6}$$

$(\mathrm{C_2})$ For some constant $c_2 > 0$, it holds $\|b(x)\| \le c_2(1 + \|x\|)$ for all $x \in \mathbb{R}^d$.

Some comments on the above conditions are in order. The assumption of ergodicity of the diffusion in $(\mathrm{C_0}^j)$ is crucial for our analysis. Classical conditions for ensuring the existence of invariant measures are due to Khasminskii and involve Lyapunov-type functions for the generator of the diffusion. For a concise general treatment of the question of existence of invariant measures, we refer to Section 8.1 in Lorenzi and Bertoldi (2007). The condition $(\mathrm{C_1})$ is related to the existence of Lyapunov functions and ensures integrability of functions with respect to the invariant measure. Together with the growth condition on the drift term in $(\mathrm{C_2})$, it further entails exponential bounds on the invariant density. The following result due to Metafune et al. (2005) will be used frequently in the sequel.

Lemma 2.2.2 (Metafune et al. (2005))**.** *Assume that the diffusion X is ergodic with Lebesgue continuous invariant measure $\mathrm{d}\mu(x) = \rho(x)\mathrm{d}x$, and suppose that X satisfies conditions* $(\mathrm{C_1})$ *and* $(\mathrm{C_2})$*. Then there exist positive constants C_1, C_2 such that the invariant density ρ satisfies the following upper bound,*

$$\rho(x) \le C_1 \exp\left(-C_2\|x\|^2\right), \qquad x \in \mathbb{R}^d. \tag{2.2.7}$$

Proof. The lemma is an immediate consequence of the results of Metafune et al. (2005) who study global regularity properties of invariant measures of divergence-form operators. Their results also hold in our specific framework since we restrict attention to the case of constant, uniformly elliptic diffusion part. Denote by $\lambda_{\max}$ the largest eigenvalue of a. Due to Corollary 2.5 in Metafune et al. (2005), (C_1) implies that $\exp\left(\eta\|x\|^2\right) \in L^1(\mu)$ for $\eta < c_1\left(2\lambda_{\max}\right)^{-1}$. Since $\|b(x)\| \leq c_2(1+\|x\|) \lesssim \exp(\|x\|)$, Theorem 6.1 in Metafune et al. (2005) applies and yields the assertion. Here and subsequently, $\lesssim$ means less or equal up to some constant which does not depend on the variable parameters in the expression. □

Throughout this chapter, $\mathbf{E}_{b,\sigma}$ ($\mathrm{Var}_{b,\sigma}$) denotes expectation (variance) with respect to the invariant measure μ related to the diffusion process solution of (2.2.4). For constant σ, we write $\mathbf{E}_b = \mathbf{E}_{b,\sigma}$ and $\mathrm{Var}_b = \mathrm{Var}_{b,\sigma}$. Analogue to the introductory remarks in Section 1.1.3, denote by $(P_t)_{t\geq 0}$ the transition semigroup of the diffusion $(X_t)_{t\geq 0}$ on $L^2(\mu)$, that is,

$$P_t f(x) = \mathbf{E}_{b,\sigma}\left(f(X_t)|X_0 = x\right), \qquad f \in L^2(\mu).$$

Denote by A the infinitesimal generator associated with the diffusion process satisfying (2.2.4), for every C^2 function f on $\mathbb{R}^d$ defined by

$$Af(x) := \sum_{j=1}^d b^j(x)\partial_j f(x) + \frac{1}{2}\sum_{j,k=1}^d a_{jk}\partial^2_{jk} f(x). \tag{2.2.8}$$

The above conditions are similar to the conditions which are typically imposed for estimating the drift of scalar diffusions; compare in particular assumptions C1 and C2 in Dalalyan (2005). We already noted that the investigation of multidimensional processes is substantially more involved and thus often requires to formulate additional assumptions. This remark applies in particular to the relation (2.2.5).

Remark 2.2.3. The validity of the relation (2.2.5) is related to the reversibility property of diffusion processes. For ease of presentation, we restrict attention to the case of constant diffusion matrix a, but the following exposition extends to the general case of non-constant diffusion part; cf. Section 6.2 in Bovier (2012). Consider some Lebesgue continuous reversible measure, $\mu(\mathrm{d}x) = \exp(F(x))\mathrm{d}x =: \rho(x)\mathrm{d}x$, say. The formal adjoint of A in (2.2.8) is given by

$$A^* g(x) = -\sum_{j=1}^d \partial_j\left(b^j(x)g(x)\right) + \frac{1}{2}\sum_{j,k=1}^d a_{jk}\partial^2_{jk} g(x).$$

The reversibility property of μ implies that

$$A^*(g\exp(F))(x) = \exp(F(x))Ag(x). \tag{2.2.9}$$

On the other hand, straightforward algebra gives

$$\begin{aligned} A^*\left(g\mathrm{e}^F\right)(x) = {} & \mathrm{e}^{F(x)}\frac{1}{2}\sum_{j,k} a_{jk}\partial^2_{jk} g(x) + \mathrm{e}^{F(x)}\sum_{j,k} a_{jk}\partial_j g(x)\partial_k F(x) \\ & + \mathrm{e}^{F(x)}\frac{1}{2}\sum_{j,k} a_{jk} g(x)\left(\partial^2_{jk}F(x) + \partial_j F(x)\partial_k F(x)\right)g(x) \\ & - \mathrm{e}^{F(x)}\sum_j b^j(x)\left(\partial_j F(x)g(x) + \partial_j g(x)\right) - \mathrm{e}^{F(x)}\sum_j \partial_j b^j(x)g(x). \end{aligned}$$

(2.2.9) thus only holds true if, for any $j \in \{1,\ldots,d\}$,

$$2b^j(x)\rho(x) = \sum_{k=1}^d a_{jk}\partial_k\rho(x),$$

that is, (2.2.5) is satisfied.

2.2.3. Estimation and regularity properties of the drift coefficient

We turn to the question of *asymptotically exact adaptive* drift estimation in Sections 2.3 and 2.5. Before doing so, we briefly recall essential ideas for constructing a non-adaptive rate-optimal estimator of the drift of multidimensional ergodic diffusions under Hölder smoothness assumptions.

Motivation of drift estimators. Consider the SDE (2.2.4), and suppose we are interested in estimating $b(x)$ in some fixed $x \in \mathbb{R}^d$. We first deliver justifications of the form of the two kernel estimators suggested in Section 2.1.

The estimator $\overline{b}_T$ in (2.1.3) is motivated by the regression type-structure of the drift in Section 4.2.3 in Prakasa Rao (1999), and the basic idea is as follows: Assume that the diffusion process solution of (2.2.4) is stationary, and consider fixed $\Delta > 0$. Introduce the discrete drift and diffusion coefficients b_Δ and d_Δ, satisfying for any $t > 0$,

$$\begin{aligned} X_{t+\Delta} &= \mathbf{E}\left(X_{t+\Delta} \mid X_t = x\right) \ + \ \sqrt{\operatorname{cov}\left(X_{t+\Delta} \mid X_t = x\right)}\ \varepsilon_{t+\Delta} \\ &= X_t + b_\Delta(X_t) + d_\Delta(X_t)\ \varepsilon_{t+\Delta}, \end{aligned}$$

where $\mathbf{E}\left(\varepsilon_{t+\Delta} | X_s, s \leq t\right) = 0$ and $\mathbf{E}\left(\varepsilon_{t+\Delta}\varepsilon^t_{t+\Delta} | X_s, s \leq t\right) = \mathbf{Id}_{d\times d}$. If the discrete drift coefficient b_Δ is constant in some neighborhood U of x, then, for any $j \in \mathbb{N}$ such that $X_{j\Delta} \in U$,

$$X_{j\Delta+\Delta} - X_{j\Delta} = b_\Delta(x) + d_\Delta(X_{j\Delta})\ \varepsilon_{j\Delta+\Delta}.$$

Letting

$$\mathcal{I} := \left\{ j : X_{j\Delta} \in U, 1 \leq j \leq n \right\},$$

a natural estimator of b_Δ is given by

$$\frac{1}{\operatorname{card} \mathcal{I}} \sum_{j\in\mathcal{I}} \left(X_{j\Delta+\Delta} - X_{j\Delta}\right) = \frac{\sum_{j=1}^n \left(X_{j\Delta+\Delta} - X_{j\Delta}\right)\ \mathbb{1}\left\{X_{j\Delta} \in U\right\}}{\sum_{j=1}^n \mathbb{1}\left\{X_{j\Delta} \in U\right\}}.$$

For non-constant $b_\Delta(\cdot)$, one considers a shrinking neighborhood

$$U = U_n = \left\{ y : (y - x)/h_n \in B(0) \right\},$$

where $B(0)$ is some neighborhood of 0 and $h_n \searrow 0$. Denoting $K(\cdot) := \mathbb{1}\{\cdot \in U\}$, this yields the estimator

$$\overline{b}_{\Delta,n}(x) = \frac{\sum_{j=1}^n \left(X_{j\Delta+\Delta} - X_{j\Delta}\right)\ K\left(\frac{X_{j\Delta}-x}{h}\right)}{\sum_{j=1}^n K\left(\frac{X_{j\Delta}-x}{h}\right)}.$$

The passage to the case of continuous-time observations is obvious and explains the form of the estimator defined in (2.1.3).

An alternative interpretation of this estimator as an approximated MLE is sketched in Section 4.3 in Dalalyan and Kutoyants (2002). For the sake of completeness, we also repeat their arguments. Consider the SDEs

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\ \mathrm{d}W_t, \qquad \mathrm{d}Y_t = b_0(Y_t)\mathrm{d}t + \sigma(Y_t)\ \mathrm{d}W_t,$$

and assume that the law of the initial values is independent of the respective drift coefficients. Similarly to (A.1.3), one obtains the following expression for the log-likelihood ratio under the measure $\mathbb{P}_b$ associated with the drift coefficient b,

$$\log \frac{\mathrm{d}\mathbb{P}_b^{(T)}}{\mathrm{d}\mathbb{P}_{b_0}^{(T)}}(X^T) = \int_0^T (b-b_0)^t(X_u)(\sigma\sigma^t)^-(X_u)\mathrm{d}X_u \\ - \frac{1}{2}\int_0^T (b-b_0)^t(X_u)(\sigma\sigma^t)^-(X_u)(b+b_0)(X_u)\mathrm{d}u \tag{2.2.10}$$

(cf. pp. 296–297 in Liptser and Shiryaev (2001)). To extend the maximum likelihood approach to the nonparametric framework, one might use local approximations of the log-likelihood ratio in (2.2.10) by polynomials of order $n \in \mathbb{N}$. For ease of presentation, we consider the case $d = 1$. The arguments extend to the multidimensional framework with $d \geq 2$ with only formal changes. Given two sufficiently regular kernel functions $Q, P \in L^2(\mathbb{R})$ and some bandwidth sequence $h_T \to_{T\to\infty} 0$, such an approximation is given by

$$L_T^{\mathrm{approx}}(\theta, x, X^T) := \sigma^{-2}(x)\int_0^T \sum_{r=0}^{n-1}\theta_r(X_u-x)^r Q\left(\frac{X_u-x}{h_T}\right)\mathrm{d}X_u \\ - \frac{1}{2\sigma^2(x)}\int_0^T\left(\sum_{r=0}^{n-1}\theta_r(X_u-x)^r\right)^2 P\left(\frac{X_u-x}{h_T}\right)\mathrm{d}u.$$

The maximum likelihood approach then suggests to estimate the value $b(x)$ by the first coordinate of θ^*, where

$$\theta^* = \arg\max_{\theta\in\mathbb{R}^n} L_T^{\mathrm{approx}}(\theta, x, X^T). \tag{2.2.11}$$

To shorten notation, for $r \in \{0, 1, 2\}$, let

$$\mu_r(h, T, x) := \frac{1}{T}\int_0^T \left(\frac{X_u-x}{h}\right)^r P\left(\frac{X_u-x}{h_T}\right)\mathrm{d}u,$$
$$\nu_r(h, T, x) := \frac{1}{T}\int_0^T \left(\frac{X_u-x}{h}\right)^r Q\left(\frac{X_u-x}{h_T}\right)\mathrm{d}X_u.$$

For $n = 1$, the ML approach gives an estimator of the same form as in (2.1.3), namely

$$\widehat{\theta}_T(x) := \frac{\nu_0(h, T, x)}{\mu_0(h, T, x)} = \frac{\int_0^T Q\left(\frac{X_u-x}{h_T}\right)\mathrm{d}X_u}{\int_0^T P\left(\frac{X_u-x}{h_T}\right)\mathrm{d}u}, \qquad x \in \mathbb{R}^d.$$

Spokoiny (2000) suggests local linear smoothers for pointwise estimation of the drift. Estimators of this form arise as the solution of the maximization problem (2.2.11) for $n = 2$ and can be written as

$$\frac{\nu_0(h, T, x)\mu_2(h, T, x) - \nu_1(h, T, x)\mu_1(h, T, x)}{\mu_0(h, T, x)\mu_2(h, T, x) - \mu_1^2(h, T, x)}. \tag{2.2.12}$$

The alternate estimator $\widehat{b}_T$ defined in (2.1.5) is a multidimensional analogue of the drift estimator pioneered by Banon (1978). The author studies real-valued diffusion processes satisfying the Itô SDE

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\ \mathrm{d}W_t, \tag{2.2.13}$$

for uniformly Lipschitzian coefficients b and σ which grow at most linearly. The basic assumption is that the unique solution of (2.2.13) admits a stationary transition density $p_{X_t|X_0=a}$ which fulfills Kolmogorov's forward equation and which tends to a limiting density p. Banon (1978) proposes a nonparametric procedure for pointwise estimation of $b(\cdot)$. The central idea for estimating the drift is to construct estimators of $p(\cdot)$ and its derivative $p'(\cdot)$ and to exploit the relation

$$\frac{1}{2}(\sigma^2 p') = bp. \tag{2.2.14}$$

The estimator $\widehat{b}_T$ in (2.1.5) is based on the relation (2.1.4) which can be viewed as a multi-dimensional generalization of (2.2.14). Given some compactly supported, symmetric kernel K satisfying $\int_{\mathbb{R}^d} K(x)\mathrm{d}x = 1$, denote

$$\widehat{g}^j_{T,h}(x) := \frac{1}{Th^d}\int_0^T \sum_{k=1}^d a_{jk}(X_u)\ (\partial_k K)\left(\frac{X_u - x}{h}\right)\mathrm{d}u. \tag{2.2.15}$$

Integration by parts gives

$$\begin{aligned}\mathbf{E}_b\widehat{g}^j_{T,h}(x) &= h^{-d}\int_{\mathbb{R}^d}\sum_{k=1}^d a_{jk}(y)\ \partial_k K\left(\frac{x-y}{h}\right)\rho(y)\mathrm{d}y\\ &= h^{-d+1}\int_{\mathbb{R}^d} K\left(\frac{x-y}{h}\right)\sum_k \partial_k(a_{jk}(y)\rho(y))\mathrm{d}y\\ &= \int_{\mathbb{R}^d} K(v)\operatorname{div}(a_j\rho)(x-hv)\mathrm{d}v. \end{aligned}\tag{2.2.16}$$

(The boundary terms vanish since K has compact support.) Provided the relation (2.1.4) is valid, it holds

$$\lim_{h\to 0}\mathbf{E}_b\widehat{g}^j_{T,h}(x) = \lim_{h\to 0}\int_{\mathbb{R}^d} K(v)\operatorname{div}(a_j\rho)(x-hv)\mathrm{d}v = \operatorname{div}(a_j\rho)(x) \overset{(2.1.4)}{=} 2b^j(x)\rho(x).$$

Consequently, given some consistent estimator $\widehat{\rho}_T$ of the invariant density ρ, the estimator $\widehat{b}_T$ as defined in (2.1.3) is a reasonable estimator of the drift vector.

The agenda. We work under the standing assumption that the relation (2.1.4) is satisfied. The strategy for analyzing the drift estimation problem is

(i) to derive asymptotic results for estimating $2b^j\rho = \operatorname{div}(a_j\rho)$, and

(ii) to transfer these results to the problem of drift estimation.

At this point, we shall briefly highlight particular challenges to be tackled in the sequel, on the one hand due to the dependence structure of the data, and, secondly, due to our wish to identify the optimal constant appearing in the normalizing factor.

Exemplarily, consider the estimator $\overline{b}_T$ defined in (2.1.3), and denote

$$\overline{g}^j_{T,h}(x) := \frac{2}{T}\int_0^T K_h(X_u - x)\mathrm{d}X^j_u, \qquad j\in\{1,\dots,d\},\ x\in\mathbb{R}^d.$$

The j-th component of the estimator $\overline{b}_T$ defined in (2.1.3) can then be written as

$$\overline{b}^j_T(x) = \overline{b}^j_{T,h}(x) = \frac{\overline{g}^j_{T,h}(x)}{2\left(\widehat{\rho}_T(x)\vee\rho_*(x)\right)}, \qquad j\in\{1,\dots,d\},\ x\in\mathbb{R}^d. \tag{2.2.17}$$

(Recall that $\widehat{\rho}_T$ is an estimator of the invariant density ρ and $\rho_* > 0$ denotes some a priori lower bound such that $\rho(x)\geq\rho_*(x)$, $x\in\mathbb{R}^d$.)

In order to analyze the mean-squared error of $\overline{g}^j_{T,h}$, we use the classical bias-variance decomposition, that is,

$$\mathbf{E}_b\big|\overline{g}^j_{T,h}(x) - \operatorname{div}(a_j\rho)(x)\big|^2 = \big|\mathbf{E}_b\overline{g}^j_{T,h}(x) - \operatorname{div}(a_j\rho)(x)\big|^2 + \operatorname{Var}_b\left(\overline{g}^j_{T,h}(x)\right). \tag{2.2.18}$$

If

$$b^j\rho\in C^\beta(\mathbb{R}^d), \qquad \beta\in\mathbb{N}, \tag{2.2.19}$$

the bias of $\overline{g}^j_{T,h}$ is shown to be of the order h^β in the standard way by choosing a smooth kernel $K:\mathbb{R}^d\to\mathbb{R}$ of order $\beta-1$ satisfying $\int_{\mathbb{R}^d}|K(u)|\|u\|^\beta\mathrm{d}u<\infty$. Indeed, it holds

$$\begin{aligned}\left|\mathbf{E}_b\overline{g}^j_{T,h}(x)-2b^j(x)\rho(x)\right| &= \left|2\int_{\mathbb{R}^d}K(u)(b^j(x-uh)\rho(x-uh)-b^j(x)\rho(x))\mathrm{d}u\right|\\ &\leq 2h^\beta\int_{\mathbb{R}^d}|K(u)|\|u\|^\beta\mathrm{d}u\lesssim h^\beta. \end{aligned}\tag{2.2.20}$$

The investigation of the variance term differs from the case of independent data as there appear additional covariances in the dependent case. Under suitable conditions on the precise dependence mechanism governing the observations however, one can show that there exists some positive constant C such that

$$\operatorname{Var}_b\left(\int_0^T f(X_u)\mathrm{d}u\right)\leq CT\int_{\mathbb{R}^d}f^2(y)\mathrm{d}\mu(y),\qquad f\in L^2(\mu),\ T\geq 0.\tag{2.2.21}$$

In this case, a rough upper bound on the variance term in (2.2.18) is given by

$$\begin{aligned}\operatorname{Var}_b\left(\overline{g}^j_{T,h}(x)\right) &\leq \frac{8}{T^2}\operatorname{Var}_b\left(\int_0^T K_h(X_u-x)b^j(X_u)\mathrm{d}u\right)\\ &\qquad +\frac{8}{T^2}\operatorname{Var}_b\left(\int_0^T K_h(X_u-x)\sum_{r=1}^d\sigma_{jr}\mathrm{d}W^r_u\right)\\ &= \frac{8}{T^2}\operatorname{Var}_b\left(\int_0^T K_h(X_u-x)b^j(X_u)\mathrm{d}u\right)+\frac{8a_{jj}}{T}\int_{\mathbb{R}^d}K_h^2(y-x)\rho(y)\mathrm{d}y\\ &\overset{(2.2.21)}{\leq} \frac{8}{T}\int_{\mathbb{R}^d}K_h^2(y-x)(C(b^j)^2(y)+a_{jj})\rho(y)\mathrm{d}y\\ &= \frac{8}{Th^d}\int_{\mathbb{R}^d}K^2(v)(C(b^j)^2(x+hv)+a_{jj})\rho(x+hv)\mathrm{d}v\sim\frac{1}{Th^d}.\end{aligned}\tag{2.2.22}$$

In view of (2.2.18), this allows to deduce the rate of convergence of estimators of $\operatorname{div}(a_j\rho)$.

As concerns item (ii) in our above agenda, that is, the question of transferring the results to the original question of drift estimation, note that

$$\begin{aligned}|\overline{b}^j_{T,h}-b^j| &\overset{(2.2.17)}{=} \left|\frac{\overline{g}^j_{T,h}}{2(\widehat{\rho}_T\vee\rho_*)}-b^j\right|\\ &= \left|\frac{(\overline{g}^j_{T,h}-2b^j\rho)\rho+2b^j\rho(\rho-\widehat{\rho}_T\vee\rho_*)}{2(\widehat{\rho}_T\vee\rho_*)\rho}\right|\\ &\leq \rho_*^{-1}\left(\left|\frac{1}{2}\overline{g}^j_{T,h}-b^j\rho\right|+\left|b^j\rho_*^{-1}(\widehat{\rho}_T-\rho)\right|\right).\end{aligned}\tag{2.2.23}$$

The above steps have certain shortcomings in view of the goal of extending the results on exact estimation in the classical statistical models to the drift estimation problem. Precisely, they raise the following new issues:

1. The smoothness assumption in (2.2.19) refers to $b^j\rho=\operatorname{div}(a_j\rho)/2$ and appears as a natural starting point for estimating $\operatorname{div}(a_j\rho)$. For the original goal, that is, for nonparametric estimation of the drift coefficient, it is desirable to formulate a result under regularity assumptions on the drift coefficient itself. *Weighted* Sobolev smoothness conditions turn out to give an adequate description of the regularity of the drift.

2. It remains to specify preferably general assumptions on the dependence structure of the data which guarantee that (2.2.21) is satisfied. Given any ergodic stationary diffusion X with invariant measure μ, the study of the variance of integral functionals of X of the form

$$\mathrm{Var}_b\left(\int_0^T f(X_u)\mathrm{d}u\right), \qquad f\in L^2(\mu),$$

requires to study covariances of the form $\mathbf{E}_b\left(f(X_u)f(X_v)\right)$, $u,v\in[0,\infty)$. Using stationarity and the Cauchy–Schwarz inequality, one obtains

$$\begin{aligned}\mathbf{E}_b\left(f(X_u)f(X_v)\right) &= \mathbf{E}_b\left(f(X_0)f(X_{|u-v|})\right)\\ &= \mathbf{E}_b\left(f(X_0)\mathbf{E}_b\left(f(X_{|u-v|})|X_0\right)\right) = \mathbf{E}_b\left(f(X_0)P_{|u-v|}f(X_0)\right)\\ &\le \left(\mathbf{E}_b f^2(X_0)\right)^{1/2}\left(\mathbf{E}_b(P_{|u-v|}f)^2(X_0)\right)^{1/2}. \qquad (2.2.24)\end{aligned}$$

In the scalar diffusion framework, Banon (1978) uses the "condition G2" to evaluate the second term in the last display. Dalalyan and Reiß (2007) consider multidimensional Kolmogorov diffusions, and they bound this quantity by means of the spectral gap inequality.

3. The decomposition in (2.2.23) is too crude in view of the wish to identify exact constants for drift estimation. It involves the inverse of the lower bound which typically is such that $\rho_*^{-1}\gg 1$. A more careful decomposition will be shown to give exact results.

2.2.4. Estimation of the invariant density

The previous section clearly shows that the investigation of the drift estimators suggested there involves us in an analysis of invariant density estimators. It is worth recalling that, in contrast to the classical setting of nonparametric density estimation based on i.i.d. observations, there is a gap between the minimax optimal rates occurring in the scalar and the multidimensional case, respectively. If the invariant density $\rho\in C^\beta(\mathbb{R})$ is continuous in x, the so-called *local-time estimator* $\widehat{\rho}_T^0(x)$ achieves the parametric rate $T^{-1/2}$ for pointwise estimation. It is defined as

$$\widehat{\rho}_T^0(x) := \frac{1}{\sigma^2(x)T}\int_0^T \mathrm{sgn}(x-X_u)\mathrm{d}X_u + \frac{|X_T-x|-|X_0-x|}{\sigma^2(x)T},$$

and the naming is due to the following representation of $\widehat{\rho}_T^0$,

$$\widehat{\rho}_T^0(x) = \frac{2\ell_T(x)}{\sigma^2(x)T}.$$

Here, $\ell_T(x)$ is the local time of the diffusion process X at time T and point x (see, e.g., Section 3.7 in Karatzas and Shreve (1988)). The phenomenon of nonparametric estimation with a parametric rate of convergence is related to the existence of local time and does not transfer to dimension $d\ge 2$.

Note that the drift b appears in the invariant density ρ_b (2.2.3) in an integrated form. In particular, this implies some "increase of regularity" of the invariant density estimation problem as compared to the classical i.i.d. setting. Roughly speaking, the problem of estimating the invariant density of an ergodic diffusion as in (2.2.3) is to be compared with the problem of nonparametric estimation of a distribution function from i.i.d. observations. In turn, the drift estimation problem is similar to nonparametric density estimation in the classical setting.

We now formulate the first auxiliary result in this section. Later on, we shall investigate drift estimators $\widetilde{b}_T$ of the form

$$\widetilde{b}_T(x) = \frac{\widetilde{g}_T(x)}{\widehat{\rho}_T(x)+c_T(x)},$$

where $\widetilde{g}_T$ and $\widehat{\rho}_T$ are estimators of $\operatorname{div}(a\rho)/2 = b\rho$ and the invariant density ρ, respectively. The additional term c_T is included in order to prevent the denominator of $\widetilde{b}_T$ from vanishing. For transferring the results on estimation of $\operatorname{div}(a\rho)$ to the problem of drift estimation, we require exponential inequalities for the risk of $\widehat{\rho}_T$ in order to guarantee integrability of certain functions. Such estimates are given in the following

Lemma 2.2.4. *Consider some ergodic diffusion process X with Lebesgue-continuous invariant measure $\mathrm{d}\mu(x) = \rho(x)\mathrm{d}x$, and assume that X fulfills conditions* (C_1), (C_2) *and the relation (2.2.21).*

1) *If $\rho \in \mathcal{H}(\beta+1, L)$, for some $\beta \in \mathbb{N}$ and $L > 0$, then there exist positive constants K_1, K_2 and some invariant density estimator $\widehat{\rho}_{T,1}$ satisfying, for any $q \geq 1$,*

$$\mathbf{E}_b\big|\widehat{\rho}_{T,1}(x) - \rho(x)\big|^{2q} \leq K_1\ T^{-\frac{2q(\beta+1)}{2(\beta+1)+d}}\ \exp\left(-K_2 q\|x\|\right), \qquad x \in \mathbb{R}^d. \tag{2.2.25}$$

2) *If $\rho \in \mathcal{S}(\beta'+1, L')$, for some $\beta' > d/2$ and $L' > 0$, then there exist positive constants K_1', K_2' and some invariant density estimator $\widehat{\rho}_{T,2}$ such that, for any $q \geq 1$,*

$$\mathbf{E}_b\big|\widehat{\rho}_{T,2}(x) - \rho(x)\big|^{2q} \leq K_1'\ T^{-\frac{q(\beta'+1-d/2)}{\beta'+1}}\ \exp\left(-K_2' q\|x\|\right), \qquad x \in \mathbb{R}^d. \tag{2.2.26}$$

The above remark on the formal correspondence between the problems of estimating the invariant density and the classical distribution function applies to scalar diffusions. For multidimensional diffusions which satisfy the relation $2b\rho = \operatorname{div}(a\rho)$, a similar formal analogy holds true. This explains in particular the setting considered in Lemma 2.2.4: Later on, we will impose smoothness assumptions of order β on the drift function b, and thus it is of interest to investigate the case $\rho \in C^{\boldsymbol{\beta+1}}(\mathbb{R}^d)$.

Proof of Lemma 2.2.4. Let $K : \mathbb{R}^d \to \mathbb{R}$ be some kernel and $h = h_T \searrow 0$ be some bandwidth, both to be specified later, and define

$$\widehat{\rho}_{T,i}(x) := \frac{1}{Th^d}\int_0^T K\left(\frac{X_u - x}{h}\right)\mathrm{d}u, \qquad i \in \{1,2\}. \tag{2.2.27}$$

We use the classical decomposition

$$\big|\widehat{\rho}_{T,i}(x) - \rho(x)\big| \leq \big|\mathbf{E}_b\widehat{\rho}_{T,i}(x) - \rho(x)\big| + \big|\widehat{\rho}_{T,i}(x) - \mathbf{E}_b\widehat{\rho}_{T,i}(x)\big|, \qquad i \in \{1,2\}.$$

For bounding the stochastic error, note that, for $i \in \{1,2\}$,

$$\begin{aligned}\mathbf{E}_b\big|\widehat{\rho}_{T,i}(x) - \mathbf{E}_b\widehat{\rho}_{T,i}(x)\big|^2 \quad &= \quad \frac{1}{T^2h^{2d}}\ \mathrm{Var}_b\left(\int_0^T K\left(\frac{X_u - x}{h}\right)\mathrm{d}u\right)\\ &\overset{(2.2.21)}{\leq} \quad \frac{C}{Th^{2d}}\int_{\mathbb{R}^d} K^2\left(\frac{y-x}{h}\right)\rho(y)\mathrm{d}y.\end{aligned}$$

We now apply the multidimensional version of Theorem 1A in Parzen (1962) (cf. Lemma A.1.4 in the appendix with $m = 1$). Assume that $\sup_{x\in\mathbb{R}^d}|K(x)| < \infty$, $\int K^2(x)\mathrm{d}x < \infty$ and that $\lim_{\|x\|\to\infty}\left(\|x\|K^2(x)\right) = 0$. Then

$$\begin{aligned}&\left|\frac{1}{h^d}\int_{\mathbb{R}^d} K^2\left(\frac{y-x}{h}\right)\rho(y)\mathrm{d}y\right|\\ &\quad\leq \left|\frac{1}{h^d}\int_{\mathbb{R}^d} K^2\left(\frac{y-x}{h}\right)\rho(y)\mathrm{d}y - \rho(x)\|K\|^2_{L^2(\mathbb{R}^d)}\right| + \left|\rho(x)\|K\|^2_{L^2(\mathbb{R}^d)}\right|\\ &\quad\leq \max_{\|u\|\leq\delta}|\rho(u+x) - \rho(u)|\|K\|^2_{L^2(\mathbb{R}^d)} + \frac{1}{\delta}\sup_{\|u\|>\delta h_T^{-d}}\left(\|u\|K^2(u)\right)\int_{\mathbb{R}^d}|\rho(y)|\mathrm{d}y\\ &\qquad\qquad + |\rho(x)|\int_{\|y\|>\delta h_T^{-d}} K^2(y)\mathrm{d}y + \left|\rho(x)\|K\|^2_{L^2(\mathbb{R}^d)}\right|.\end{aligned}$$

Consequently, for $i \in \{1,2\}$,

$$\mathbf{E}_b\big|\widehat{\rho}_{T,i}(x) - \mathbf{E}_b\widehat{\rho}_{T,i}(x)\big|^2 \le \frac{C}{Th^d}\rho(x)\|K\|^2_{L^2(\mathbb{R}^d)}(1+o_T(1)).$$

It remains to treat the bias term.

Case 1). Assume that K is of order β and satisfies $\int_{\mathbb{R}^d}\|y\|^{\beta+1}|K(y)|\mathrm{d}y < \infty$. Denote by $P_x^{(\rho)}$ the Taylor polynomial of ρ at the point x up to order β, and let $R_{x,y}^{(\rho)} := \rho(y) - P_x^{(\rho)}(y)$ be the corresponding remainder term. Since

$$\begin{aligned}\big|\mathbf{E}_b\widehat{\rho}_{T,1}(x) - \rho(x)\big| &= \left|h^{-d}\int_{\mathbb{R}^d}(\rho(y)-\rho(x))\ K\left(\frac{y-x}{h}\right)\mathrm{d}y\right|\\ &= \left|h^{-d}\int_{\mathbb{R}^d}\left(P_x^{(\rho)}(y) - \rho(x) + R_{x,y}^{(\rho)}\right)\ K\left(\frac{y-x}{h}\right)\mathrm{d}y\right|\end{aligned}$$

and K is of order β, it follows

$$\big|\mathbf{E}_b\widehat{\rho}_{T,1}(x) - \rho(x)\big| = \left|h^{-d}\int_{\mathbb{R}^d}R_{x,y}^{(\rho)}\ K\left(\frac{y-x}{h}\right)\mathrm{d}y\right|.$$

The smoothness assumption on ρ entails by the definition of the Hölder class $\mathcal{H}(\beta+1,L)$ (cf. (2.2.1)) that

$$\sup_{\|x-y\|\le h}\left|R_{x,y}^{(\rho)}\right| = \sup_{\|x-y\|\le h}\left|P_x^{(\rho)}(y) - \rho(x)\right| \le Lh^{\beta+1}.$$

Balancing the upper bounds on the bias and the stochastic error suggests to select a bandwidth

$$h = h_T \sim \left(\frac{C\rho(x)}{L^2T}\right)^{\frac{1}{2(\beta+1)+d}}.$$

In combination with the exponential upper bound from Lemma 2.2.2, (2.2.25) follows.

Case 2). Define $\widehat{\rho}_{T,2}$ according to (2.2.27), for the kernel $K = K_{\beta'+1}$ with Fourier transform

$$\phi_{K_{\beta'+1}}(\lambda) = \int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t y}K_{\beta'+1}(y)\mathrm{d}y = \frac{1}{1+\|\lambda\|^{2(\beta'+1)}}, \qquad \lambda \in \mathbb{R}^d.$$

Thus, using Cauchy–Schwarz,

$$\begin{aligned}\big|\mathbf{E}_b\widehat{\rho}_{T,2}(x) - \rho(x)\big| &= \left|h^{-d}\int_{\mathbb{R}^d}K_{\beta'+1}\left(\frac{y-x}{h}\right)\rho(y)\mathrm{d}y - \rho(x)\right|\\ &= (2\pi)^{-d}\left|\int_{\mathbb{R}^d}\phi_\rho(\lambda)\left(\phi_{K_{\beta'+1}}(h\lambda) - 1\right)\mathrm{e}^{-\mathrm{i}\lambda^t x}\mathrm{d}\lambda\right|\\ &= (2\pi)^{-d}\left|\int_{\mathbb{R}^d}\phi_\rho(\lambda)\frac{\|h\lambda\|^{2(\beta'+1)}}{1+\|h\lambda\|^{2(\beta'+1)}}\mathrm{d}\lambda\right|\\ &\le h^{\beta'+1}\left((2\pi)^{-d}\int_{\mathbb{R}^d}|\phi_\rho(\lambda)|^2\,\|\lambda\|^{2(\beta'+1)}\mathrm{d}\lambda\right)^{1/2}\\ &\qquad\times\left((2\pi)^{-d}\int_{\mathbb{R}^d}\frac{\|h\lambda\|^{2(\beta'+1)}}{\left(1+\|h\lambda\|^{2(\beta'+1)}\right)^2}\mathrm{d}\lambda\right)^{1/2}\\ &\le L'h^{\beta'+1-d/2}\ (2\pi)^{-d/2}\left(\int_{\|y\|\le 1}\frac{\mathrm{d}y}{(1+\|y\|^{2\beta'})^2} + \int_{\|y\|>1}\frac{\mathrm{d}y}{\|y\|^{2\beta'}}\right)^{1/2}\\ &\le L''h^{\beta'+1-d/2},\end{aligned}$$

for some positive constant L''. Specifying

$$h = h_T \sim \left(\frac{C\rho(x)}{(L'')^2 T}\right)^{\frac{1}{2(\beta'+1)}}$$

and using again the upper bound from Lemma 2.2.2, we obtain (2.2.26). □

Remark 2.2.5. The rate of convergence for the pointwise mean-squared error (MSE) in Lemma 2.2.4 corresponds to the results for nonparametric density estimation from i.i.d. observations over Hölder classes. For multidimensional Kolmogorov diffusions, Dalalyan and Reiß (2007) prove a substantially better convergence rate for the pointwise MSE, assuming Hölder smoothness of ρ. The improvement is due to a smaller bound on the variance term in the bias-variance decomposition of the MSE of the invariant density estimator $\widehat{\rho}_T$,

$$\mathbf{E}_b \left|\widehat{\rho}_T(x) - \rho(x)\right|^2 = \mathrm{Var}_b\left(\widehat{\rho}_T(x)\right) + \left|\mathbf{E}_b\widehat{\rho}_T(x) - \rho(x)\right|^2.$$

The smaller bound on the variance term is obtained by applying the spectral gap inequality (**SG**). For the purpose of transferring results on invariant density estimation to the question of estimating the drift coefficient, the result in (2.2.25) suffices.

2.3. Global sharp adaptive estimation on L^2 Sobolev classes

For some sufficiently regular $b : \mathbb{R}^d \to \mathbb{R}^d$, consider the diffusion process which is given as a solution of the SDE

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma\ \mathrm{d}W_t,$$

where $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is some constant, non-degenerate dispersion matrix. Given a continuous record of observations $X^T = (X_t)_{0\le t\le T}$, $T > 0$, the aim is to estimate the drift function b and to quantify the quality of an estimator $\widehat{b}_T$ with respect to the weighted global L^2 risk

$$\mathbf{E}_b \int_{\mathbb{R}^d} \left|\widehat{b}^j_T(x) - b^j(x)\right|^2 \rho^2(x)\mathrm{d}x, \qquad j \in \{1,\dots,d\}.$$

The drift estimation problem is a classical issue in statistics of stochastic processes. For an account on this question, we refer to Chapter 4.5 in Kutoyants (2004) and references to the literature therein. The closest points of reference for our investigation of drift estimation with respect to global risk criteria are Dalalyan and Kutoyants (2002) and Dalalyan (2005). Like a majority of the literature on nonparametric statistics of diffusion processes, these works consider the case of scalar diffusions. Since local time techniques and specific properties of one-dimensional diffusions are heavily used, the results do not admit a straightforward extension to the multidimensional setting.

The section is organized as follows. We start by introducing the local minimax framework. Section 2.3.2 contains a lower bound for estimating the components of the vector $\mathrm{div}(a\rho)$ which readily entails a lower bound for estimating the drift function. In part (**I**) of Section 2.3.3, we construct an asymptotically efficient estimator of $\mathrm{div}(a_j\rho)$, $j \in \{1,\dots,d\}$ fixed, which attains the lower bound for the minimax risk. This estimator can be used for defining an asymptotically efficient estimator of the j-th component of the drift vector. We translate the smoothness conditions on $\mathrm{div}(a_j\rho)$ into conditions on the weighted Sobolev regularity of the drift coefficient b. This in particular yields Corollary 2.3.12 which presents the most convenient formulation of our result on asymptotically efficient global drift estimation. The explicit form of the optimal kernel and the optimal bandwidth of the asymptotically efficient estimator of $\mathrm{div}(a_j\rho)$ are obtained by minimizing the maximum of the asymptotically exact upper risk bound over the Sobolev ball of (known) smoothness β and (known) radius L. The optimal kernel and the optimal bandwidth thus depend on β and L. In part (**II**) of Section 2.3.3, we consider the important special case of Kolmogorov diffusions, and we propose an adaptive estimator of the entire vector $\nabla\rho$ which relies on data-driven approximations of the parameters.

2.3.1. The local minimax setting

Recall the definition of $\Sigma^j(\beta)$ introduced in Section 2.2.2, fix some $b_0 \in \Sigma^j(\beta)$, and denote by ρ_0 the invariant density of the diffusion process solution of the SDE

$$\mathrm{d}X_t = b_0(X_t)\mathrm{d}t + \sigma\ \mathrm{d}W_t. \tag{2.3.1}$$

Consider the neighborhood $U_\delta^j(b_0)$ as defined in (2.1.7), that is,

$$U_\delta^j(b_0) := \left\{ b : \mathbb{R}^d \to \mathbb{R}^d : b \in \Sigma^j(\beta);\ \forall k \in \{1,\dots,d\},\ \sup_{x\in\mathbb{R}^d} \left|b^k(x) - b_0^k(x)\right| \leq \delta \right\}.$$

We require some additional smoothness of the central function b_0 and some bound on the transition densities $p_t^{b_0}(\cdot,\cdot)$ of the diffusion process solution of the SDE (2.3.1), namely:

(C_3) There exists some $\tau > 0$ such that $\int_{\mathbb{R}^d} \|\lambda\|^{2\beta+2\tau+2}|\phi_0(\lambda)|^2\mathrm{d}\lambda < \infty$, where ϕ_0 is the Fourier transform of ρ_0;

(C_4) there exist some $\kappa > 0$ and some $q > 1$ such that

$$\sup_{t>\kappa} \mathbf{E}_{b_0}\left| \sup_{y\in\mathbb{R}^d} \left(p_\kappa^{b_0}(X_{t-\kappa}, y)\right)^q \right| < \infty.$$

The above conditions present multidimensional analogues of conditions C3 and C4 in Dalalyan (2005). Condition (C_3) is satisfied whenever the central function b_0 is chosen a bit more regular than the other drift functions in its neighborhood. Sufficient criteria which ensure that (the one-dimensional version of) condition (C_4) is satisfied are given in Section 4.4 in Dalalyan (2005).

For fixed $j \in \{1,\dots,d\}$, define $\Sigma_\delta^j = \Sigma_\delta^j(\beta, L; b_0, \sigma)$ as

$$\Sigma_\delta^j := \Big\{ b \in U_\delta^j(b_0) : \sum_{|\alpha|\leq\beta} \int_{\mathbb{R}^d} \left|D^\alpha\left(\mathrm{div}(a_j\rho) - \mathrm{div}(a_j\rho_0)\right)\right|^2 \leq 4L \Big\}. \tag{2.3.2}$$

For the problem of nonparametric drift estimation, it is natural to impose smoothness conditions on the components b^j of the drift vector, while the results on exact adaptive estimation of the sums $\mathrm{div}(a_j\rho)$ rely on smoothness conditions on $\mathrm{div}(a_j\rho_b) = 2b^j\rho_b$. The local minimax approach allows to formulate results under more natural smoothness conditions imposed on the drift coefficient itself.

Proposition 2.3.1. *Fix $b_0 \in \Sigma^j(\beta)$, $\beta \in \mathbb{N}$, and consider the neighborhood*

$$\overline{U}_\delta^j(b_0) := \Big\{ b : \mathbb{R}^d \to \mathbb{R}^d : b \in \Sigma^j(\beta); \forall k \in \{1,\dots,d\},\ \alpha \in \mathbb{N}^d \text{ with } |\alpha| \leq \beta - 1,$$
$$\sup_{x\in\mathbb{R}^d} \left|D^\alpha b^j(x) - D^\alpha b_0^j(x)\right| \leq \delta \Big\}.$$

Define the Sobolev ball $\overline{\Sigma}_\delta^j = \overline{\Sigma}_\delta^j(\beta, L; b_0, \sigma)$ as

$$\overline{\Sigma}_\delta^j := \Big\{ b \in \overline{U}_\delta^j(b_0) : \sum_{\alpha\in\mathbb{N}^d:|\alpha|\leq\beta} \int_{\mathbb{R}^d} \left|D^\alpha(b^j - b_0^j)\right|^2 \rho_b^2 \leq L \Big\}. \tag{2.3.3}$$

Assume that the central function $b_0(\cdot)$ satisfies for some positive constants M, κ and for all $x \in \mathbb{R}^d$,

$$\max_{\alpha \in \mathbb{N}^d : |\alpha| \leq \beta} \left| D^\alpha b_0^j(x) \right| \leq M \left(1 + \|x\|^\kappa\right). \tag{2.3.4}$$

Then, for any $\beta \in \mathbb{N}$, $L > 0$ and $\delta > 0$, it holds

$$\overline{\Sigma}_\delta^j(\beta, L; b_0, \sigma) \subseteq \Sigma_\delta^j(\beta, L + o_\delta(1); b_0, \sigma).$$

Proof. To shorten notation, denote $Db = D^i b^j$, $D^2 b = D^i(D^j b^k)$, $i, j, k \in \{1, \ldots, d\}$, and so forth. For any $\beta \in \mathbb{N}$, $D^\beta b$ thus denotes partial derivatives of the components of b of order β. Taking into account the form of the invariant density $\rho = \rho_b$ associated with $b \in \overline{U}_\delta^j(b_0)$, one might verify that, for any multi-index $\alpha \in \mathbb{N}^d$ of order $|\alpha| = \beta \in \mathbb{N}$,

$$D^\alpha \rho_b(x) = P\big(b, Db, \ldots, D^{\beta-1} b\big)\ \rho_b(x), \qquad \beta \in \mathbb{N},\ x \in \mathbb{R}^d,$$

where $P(x_0, x_1, \ldots, x_{\beta-1})$ is some real-valued polynomial. For any $j \in \{1, \ldots, d\}$ and any multi-index $\alpha \in \mathbb{N}^d$ of order $|\alpha| \leq \beta$, it thus holds

$$\begin{aligned}
&\left\| D^\alpha \left(\operatorname{div}(a_j \rho_b) - \operatorname{div}(a_j \rho_0)\right) \right\|_{L^2(\mathbb{R}^d)} \\
&\quad = 2 \left\| D^\alpha b^j \rho_b + b^j D^\alpha \rho_b - D^\alpha b_0^j \rho_0 - b_0^j \rho_0 \right\|_{L^2(\mathbb{R}^d)} \\
&\quad \leq 2 \left\| D^\alpha (b^j - b_0^j) \rho_b \right\|_{L^2(\mathbb{R}^d)} + 2 \left\| D^\alpha b_0^j (\rho_b - \rho_0) \right\|_{L^2(\mathbb{R}^d)} \\
&\qquad + 2 \left\| \Big(P\big(b, Db, \ldots, D^{\beta-1} b\big) - P\big(b_0, Db_0, \ldots, D^{\beta-1} b_0\big) \Big) \rho_b \right\|_{L^2(\mathbb{R}^d)} \\
&\qquad + 2 \left\| P\big(b_0, Db_0, \ldots, D^{\beta-1} b_0\big) (\rho_b - \rho_0) \right\|_{L^2(\mathbb{R}^d)}.
\end{aligned} \tag{2.3.5}$$

The pointwise upper bound on the invariant density ρ_b associated to b in (2.2.7) implies that, for any $\gamma' > 0$,

$$\lim_{\delta \to 0} \sup_{b \in \Sigma_\delta} \int_{\mathbb{R}^d} \|x\|^{2\gamma'} \left(\rho_b - \rho_0\right)^2(x) \mathrm{d}x = 0;$$

this can be shown analogously to the proof of Theorem 4 in Dalalyan and Kutoyants (2002). In view of condition (2.3.4) and the decomposition (2.3.5), the assertion follows. $\square$

2.3.2. Lower risk bound for global drift estimation

The lower bound in the following Theorem is proven by a version of van Trees' inequality which was introduced in its general form in Section 1.2.1. It has been applied for proving asymptotically exact lower bounds in different statistical models. Belitser and Levit (1995) study the filtration problem of estimating the mean θ of an observed infinite-dimensional Gaussian vector with independent components. They consider the model

$$Y_k = \theta_k + \varepsilon\ \sigma_k\ \xi_k, \qquad k = 1, 2, \ldots,$$

where the values $\sigma_k \geq 0$ are given, ξ_k are independent standard Gaussian random variables, and $\varepsilon > 0$ is a small noise parameter. The unknown infinite-dimensional parameter of interest, $\theta = (\theta_1, \theta_2, \ldots)$, is assumed to belong to some ellipsoid

$$\Theta = \Theta(Q, (a_k)_{k \in \mathbb{N}}) = \Big\{ \theta : \sum_{k=1}^{\infty} a_k^2 \theta_k^2 \leq Q \Big\}, \qquad Q > 0,$$

where $(a_k)_{k \in \mathbb{N}}$ is a nonnegative sequence tending to infinity. Belitser and Levit (1995) enhance the approach of Pinsker (1980) by describing the *second-order behavior* of the minimax estimators and the quadratic minimax risk for the filtration problem in Gaussian noise. Similarly, Golubev and Levit (1996) study the problem of second-order minimax estimation of distribution functions over $\mathbb{R}$. This question is more involved than the classical first-order minimax density-type estimation problems. The authors give the exact asympotics of the properly normalized *second-order minimax risk* over Sobolev balls in the class of distribution functions. Dalalyan and Kutoyants (2002) use the scheme of the proof of the lower bound in Golubev and Levit (1996) for proving lower bounds for the problem of estimating the derivative ρ_b' of the invariant density ρ_b of a scalar ergodic diffusion with drift b. For proving a lower bound on the minimax risk

$$\inf_{\widehat{g}_T} \sup_{b\in\Sigma} \mathbf{E}_b \int_{\mathbb{R}} \left(\widehat{g}_T(x) - \rho_b'(x)\right)^2 \mathrm{d}x, \tag{2.3.6}$$

where Σ is a Sobolev-type ellipsoid, Dalalyan and Kutoyants (2003) first minorate the minimax risk in (2.3.6) by the minimax risk over a properly chosen parametric family. As the scheme of the proof, the parametrization is inspired by the proof of Golubev and Levit (1996). The constrained minimax risk is then bounded from below by the Bayesian risk plus an additional term which can be shown to be small in order. The next step is to find a lower bound for the Bayesian risk corresponding to a prior distribution, and this is where van Trees' inequality comes in. An appropriate choice of the prior then yields the lower bound.

Our proof of the lower bound in the multidimensional diffusion framework follows the same scheme. In particular, it heavily relies upon the application of some multivariate extension of the van Trees inequality due to Gill and Levit (1995). For any $j \in \{1,\ldots,d\}$, define

$$P_j(\sigma,\beta,L) := \frac{(L(2\beta+d))^{\frac{d}{2\beta+d}}}{d} \left(\frac{a_{jj}\beta\, \mathbb{S}_d}{(2\pi)^d(\beta+d)}\right)^{\frac{2\beta}{2\beta+d}}, \qquad \beta, L > 0. \tag{2.3.7}$$

Theorem 2.3.2. *For any $\mathbb{N} \ni \beta > 1$ and $j \in \{1,\ldots,d\}$, it holds*

$$\liminf_{\delta\to 0} \liminf_{T\to\infty} T^{\frac{2\beta}{2\beta+d}} \inf_{\widetilde{g}_T^j} \sup_{b\in\Sigma_\delta^j} \mathbf{E}_b \int_{\mathbb{R}^d} \left(\widetilde{g}_T^j(x) - \operatorname{div}(a_j\rho)(x)\right)^2 \mathrm{d}x \geq 4\ P_j(\sigma,\beta,L). \tag{2.3.8}$$

For any drift function $b \in \Sigma^j(\beta)$, (2.3.8) readily entails a lower bound for the drift estimation problem.

Theorem 2.3.3. *For any $\mathbb{N} \ni \beta > 1$ and fixed $j \in \{1,\ldots,d\}$, it holds*

$$\liminf_{\delta\to 0} \liminf_{T\to\infty} T^{\frac{2\beta}{2\beta+d}} \inf_{\widehat{b}_T^j} \sup_{b\in\Sigma_\delta^j} \mathbf{E}_b \int_{\mathbb{R}^d} \left|\widehat{b}_T^j(x) - b^j(x)\right|^2 \rho^2(x)\mathrm{d}x \geq P_j(\sigma,\beta,L).$$

The proofs of Theorem 2.3.2 and Theorem 2.3.3 are deferred to Section 2.4.

2.3.3. Upper risk bounds for global drift estimation

The importance of specific conditions on the precise dependence mechanism governing the diffusion process was already mentioned in the introductory remarks in Section 2.2.3. From a statistical point of view, it is particularly interesting to derive *lower bounds* under *comparably strong assumptions* on the dependence of the data as such results show whether prior information on the underlying dependence structure allows to improve the quality of estimation as compared to the case of independent observations. As concerns the proof of *upper risk bounds*, the principal goal is to establish results under *preferably general assumptions*. It is however beyond the scope of the present work to identify the most general framework possible for proving exact upper risk bounds. Instead, we will impose the following condition,

Assumption (**SG**)**.** The carré du champs associated with the infinitesimal generator of the diffusion satisfies the spectral gap inequality, that is, for some constant c_P and any $f \in L^2(\mu)$,

$$\Big\| P_t f - \int_{\mathbb{R}^d} f \mathrm{d}\mu \Big\|_{L^2(\mu)} \leq \mathrm{e}^{-t/c_P} \|f\|_{L^2(\mu)}. \tag{\textbf{SG}}$$

In this case, for any $f \in L^2(\mu)$ and $T > 0$, there exists some constant $C_P > 0$ such that

$$\mathbf{E}_b \left[\Big(\int_0^T f(X_u) \mathrm{d}u \Big)^2 \right] \leq C_P T \|f\|^2_{L^2(\mu)}. \tag{\textbf{SG}'}$$

The proof of (**SG**') is straightforward. Arguing as in Section 2.2.3 and analogous to the proof of Proposition 1 in Dalalyan and Reiß (2007) (also see the proof of Lemma 2.8 in Cattiaux et al. (2012)), it holds

$$\begin{aligned}
\mathbf{E}_b \left[\Big(\int_0^T f(X_u) \mathrm{d}u \Big)^2 \right] &= 2 \int_0^T \int_0^v \mathbf{E}_b \left(f(X_u) f(X_v) \right) \mathrm{d}u \ \mathrm{d}v \\
&= 2 \int_0^T \int_0^v \mathbf{E}_b (f(X_0) \mathbf{E}_b \left(f(X_{v-u}) | X_0 \right)) \mathrm{d}u \ \mathrm{d}v \\
&= 2 \int_0^T \int_0^v \langle f, P_{v-u} f \rangle_{L^2(\mu)} \mathrm{d}u \ \mathrm{d}v \\
&= 2 \int_0^T \int_0^v \langle f, P_u f \rangle_{L^2(\mu)} \mathrm{d}u \ \mathrm{d}v \\
&= 2 \int_0^T (T - w) \langle f, P_w f \rangle_{L^2(\mu)} \mathrm{d}w \\
&\leq 2T \int_0^T \langle f, P_w f \rangle_{L^2(\mu)} \mathrm{d}w.
\end{aligned}$$

The last term is upper-bounded by applying the Cauchy–Schwarz and the spectral gap inequality such that

$$\begin{aligned}
\mathbf{E}_b \left[\Big(\int_0^T f(X_u) \mathrm{d}u \Big)^2 \right] &\leq 2T \int_0^T \|f\|_{L^2(\mu)} \|P_w f\|_{L^2(\mu)} \mathrm{d}w \\
&\leq 2T \|f\|^2_{L^2(\mu)} \int_0^T \mathrm{e}^{-w/c_P} \mathrm{d}w \\
&= 2 c_P T (1 - \mathrm{e}^{-T/c_P}) \|f\|^2_{L^2(\mu)}.
\end{aligned}$$

Dalalyan and Reiß (2007) prove asymptotic statistical equivalence for Kolmogorov diffusions under the following enhanced condition.

Assumption (**SG+**)**.** The carré du champs associated with the infinitesimal generator of the diffusion satisfies (**SG**), and there exists some $C_0 > 0$ such that, for any $u \geq t > 0$ and for any pair of points $x, y \in \mathbb{R}^d$ satisfying $\|x - y\|^2 \leq u$, it holds

$$p_t(x, y) \ \leq \ C_0 (t^{-d/2} + u^{3d/2}). \tag{2.3.9}$$

Under Assumption (**SG+**), Proposition 1 in Dalalyan and Reiß (2007) provides an upper bound on the variance of additive functionals of multidimensional diffusions. This bound in particular allows to prove convergence rates for the pointwise risk of invariant density estimators for d-dimensional ergodic diffusions which are considerably smaller as the standard convergence rates for density estimation based on i.i.d. observations; see also Remark 2.2.5. For the proof of exact results for drift estimation, we require the following slightly modified version of their result.

Lemma 2.3.4 (Proposition 1 in Dalalyan and Reiß (2007))**.** *Grant Assumption* (**SG+**)*, and let* $G:\mathbb{R}^d\to\mathbb{R}$ *be a compactly supported kernel function. Given any bandwidth* $h=h_T>0$*, denote* $G_h(\cdot):=h^{-d}G(\cdot/h)$*. Then there exists some positive constant* C *such that, for any* $y\in\mathbb{R}^d$*,*

$$\mathrm{Var}_b\left(\int_0^T G_h(X_u-y)\mathrm{d}u\right)\le CT\sup_{z\in[y-h,y+h]}\rho(z)\times\begin{cases}\max\left\{1,\left(\log(h^{-4})\right)^2\right\}, & d=2,\\ h^{2-d}, & d\ge 3.\end{cases}$$

Proof. Analogue to the proof of Proposition 1 in Dalalyan and Reiß (2007), taking into account the shift of the kernel function. □

In view of the results of Dalalyan and Reiß (2007) on asymptotic statistical equivalence for the drift estimation problem mentioned in Section 1.2, it is to be expected (and will be shown in the sequel) that upper risk bounds analogue to those in the case of independent observations can be established under the conditions imposed in Lemma 2.3.4. Note however that the results of Dalalyan and Reiß (2007) are established under rather restrictive (Hölder) smoothness assumptions. Precisely, the critical regularity for proving asymptotic equivalence with the Gaussian shift model grows like $(1/2+1/\sqrt{2})d$ as $d\to\infty$. The authors refer to the question whether for Hölder classes of smaller regularity asymptotic equivalence fails as “a challenging open problem.” The upper risk bounds in this section will be derived for (Sobolev-type) smoothness $\beta>1$ which suggests that asymptotic equivalence (at least in a reduced sense) still holds beyond the critical bounds of Dalalyan and Reiß (2007).

(I) Asymptotically efficient global drift estimation: The non-adaptive set-up

For constructing the drift estimator, we proceed stepwise and, similarly to the proof of the lower bound, we start again with considering the question of estimating (the components of) $\mathrm{div}(a\rho)$. A lower bound for this problem has been stated in Theorem 2.3.2, and the next step is to propose an estimator attaining this lower bound. It is instructive to assume first that $\mathrm{div}(a_j\rho)\in\Sigma_\delta^j(\beta,L;b_0,\sigma)$ for *known* values $\beta>0$ and $L>0$.

(A) Component-wise estimation of $\mathrm{div}(a\rho)$**.** To construct an estimator of the j-th component

$$\mathrm{div}(a_j\rho)=\sum_{k=1}^d a_{jk}\partial_k\rho,\qquad j\in\{1,\dots,d\}\text{ fixed},$$

consider the following estimator,

$$\overline{g}_T^j(x):=\overline{g}_T^{j,(\alpha,\beta)}(x):=\frac{2}{T}\int_0^T K_T(x-X_u)\mathrm{d}X_u^j,\qquad x\in\mathbb{R}^d,\tag{2.3.10}$$

where the kernel $K_T(\cdot)$ is chosen such that its Fourier transform ϕ_{K_T} is given as

$$\phi_{K_T}(\lambda):=\int_{\mathbb{R}^d}K_T(x)\mathrm{e}^{\mathrm{i}\lambda^t x}\mathrm{d}x=\left(1-\|\alpha\lambda\|^{\beta+\gamma}\right)_+,\qquad \lambda\in\mathbb{R}^d,\tag{2.3.11}$$

for some $\mathbb{R}_+\ni\alpha=\alpha_T\to 0$ and $\gamma=\gamma_T:=(\log\log T)^{-1}$. The inverse Fourier transform can be applied to define the kernel K_T.

Remark 2.3.5 (Choice of the kernel). The kernel K_T defined via (2.3.11) belongs to the family of Pinsker kernels. Dalalyan and Kutoyants (2002) use the following one-dimensional version of K_T for defining asymptotically efficient estimators of the drift of scalar ergodic diffusions,

$$K_T(x)=\frac{1}{\pi}\int_0^1\left(1-u^{\beta+\gamma}\right)\cos(ux)\mathrm{d}u,\qquad x\in\mathbb{R}.$$

Multidimensional Pinsker kernels are used in Rigollet and Tsybakov (2007) for proving sharp optimality properties of aggregate density estimators over scales of Sobolev classes of densities.

The aim in the sequel is to derive an upper bound on the mean-integrated squared error of the estimator $\overline{g}_T^j$, and this will be done by means of Fourier methods. Using Fubini's theorem for stochastic integrals, the formula for the Fourier transform of a convolution entails for any $\lambda \in \mathbb{R}^d$ that

$$\begin{aligned}\phi_{\overline{g}_T^j}(\lambda) &= \int_{\mathbb{R}^d} \frac{2}{T}\int_0^T K_T(x-X_u)\mathrm{d}X_u^j\ \mathrm{e}^{\mathrm{i}\lambda^t x}\mathrm{d}x\\ &= \frac{2}{T}\int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u}\mathrm{d}X_u^j\ \int_{\mathbb{R}^d} K_T(y)\mathrm{e}^{\mathrm{i}\lambda^t y}\mathrm{d}y \end{aligned}\tag{2.3.12}$$

such that the Fourier transform of $\overline{g}_T^j$ can be written as $\phi_{\overline{g}_T^j} = 2\overline{\phi}_T^j\phi_{K_T}$, where

$$\overline{\phi}_T(\lambda) := \frac{1}{T}\int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u}\mathrm{d}X_u, \qquad \lambda \in \mathbb{R}^d. \tag{2.3.13}$$

The expectation of the stochastic integral vanishes such that, for any $\lambda \in \mathbb{R}^d$,

$$\begin{aligned}2\mathbf{E}_b\overline{\phi}_T(\lambda) &= 2\mathbf{E}_b\left(\frac{1}{T}\int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u}b(X_u)\mathrm{d}u\right) = 2\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t y}b(y)\rho(y)\mathrm{d}y\\ &= \int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t y}\operatorname{div}(a\rho)(y)\mathrm{d}y = \phi_{\operatorname{div}(a\rho)}(\lambda).\end{aligned}$$

In view of the above relations, Plancherel's theorem implies that

$$\begin{aligned}&\mathbf{E}_b\int_{\mathbb{R}^d}\left|\overline{g}_T^j(x)-\operatorname{div}(a_j\rho)(x)\right|^2\mathrm{d}x\\ &= (2\pi)^{-d}\ \mathbf{E}_b\int_{\mathbb{R}^d}\left|\phi_{\overline{g}_T^j}(\lambda)-\phi_{\operatorname{div}(a_j\rho)}(\lambda)\right|^2\mathrm{d}\lambda\\ &= (2\pi)^{-d}\ \mathbf{E}_b\int_{\mathbb{R}^d}\left|2\phi_{K_T}(\lambda)\overline{\phi}_T^j(\lambda)-\phi_{\operatorname{div}(a_j\rho)}(\lambda)\right|^2\mathrm{d}\lambda\\ &= (2\pi)^{-d}\ \mathbf{E}_b\int_{\mathbb{R}^d}\left|2\phi_{K_T}(\lambda)(\overline{\phi}_T^j(\lambda)-\mathbf{E}_b\overline{\phi}_T^j(\lambda))+\phi_{\operatorname{div}(a_j\rho)}(\lambda)(\phi_{K_T}(\lambda)-1)\right|^2\mathrm{d}\lambda\\ &= (2\pi)^{-d}\ \int_{\mathbb{R}^d}\Big(4\left|\phi_{K_T}(\lambda)\right|^2\operatorname{Var}_b\left(\overline{\phi}_T^j(\lambda)\right)+\left|\phi_{K_T}(\lambda)-1\right|^2\left|\phi_{\operatorname{div}(a_j\rho)}(\lambda)\right|^2\Big)\mathrm{d}\lambda. \end{aligned}\tag{2.3.14}$$

Let

$$\zeta_T(\lambda) := \frac{1}{\sqrt{T}}\int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u}\sigma\ \mathrm{d}W_u, \qquad \lambda\in\mathbb{R}^d,\ T>0.$$

The following lemma provides an upper bound on the first part of the integrand appearing in (2.3.14).

Lemma 2.3.6. *Suppose that the diffusion X satisfies Assumption* (**SG+**)*, and consider the function $\overline{\phi}_T(\cdot)$ defined in (2.3.13). Then, for any function $\widetilde{h}_T : \mathbb{R}^d \to \mathbb{R}$ satisfying $|\widetilde{h}_T| \leq 1$,*

$$\|\widetilde{h}_T\|^2_{L^2(\mathbb{R}^d)} = O\Big(T^{\frac{d}{2\beta+d}}\Big) \quad \text{and} \quad \int_{\mathbb{R}^d}\|\lambda\|^{2\beta}\widetilde{h}_T^2(\lambda)\mathrm{d}\lambda = O(T), \tag{2.3.15}$$

it holds

$$\int_{\mathbb{R}^d}\widetilde{h}_T^2(\lambda)\operatorname{Var}_b\left(\overline{\phi}_T^j(\lambda)\right)\mathrm{d}\lambda \leq \frac{1}{T}\int_{\mathbb{R}^d}\widetilde{h}_T^2(\lambda)\mathrm{d}\lambda\ \left(a_{jj}+o_T(1)\right). \tag{2.3.16}$$

The assertion is a multidimensional analogue of the statement in Lemma 1 in Dalalyan and Kutoyants (2002). Their proof uses the occupation times formula and the martingale representation of local time and is therefore restricted to the scalar setting.

Proof. Throughout the proof, $C_1, C_2, \dots$ denote positive constants. The basic idea in our set-up is to regularize by convolution with some compactly supported kernel. Precisely, consider some symmetric kernel G of order $\beta - 1$ with compact support, satisfying

$$\int_{\mathbb{R}^d} \|z\|^\beta G(z)\mathrm{d}z < \infty \tag{2.3.17}$$

and $\int_{\mathbb{R}^d} G(z)\mathrm{d}z = 1$. Denote $G_h(\cdot) := h^{-d} G(\cdot/h)$ for some bandwidth $h = h_T \searrow 0$ to be specified later. Fix $j \in \{1, \dots, d\}$. The definition of $\overline{\phi}_T$ entails that, for any $T > 0,\ \lambda \in \mathbb{R}^d$,

$$\begin{aligned}
&\overline{\phi}_T^j(\lambda) - \mathbf{E}_b \overline{\phi}_T^j(\lambda)\\
&\quad = \frac{1}{T} \int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u} \mathrm{d}X_u^j - \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} b^j(y)\rho(y)\mathrm{d}y\\
&\quad = \frac{1}{\sqrt{T}} \Bigg[\Bigg\{ \frac{1}{\sqrt{T}} \int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u} b^j(X_u)\mathrm{d}u - \sqrt{T} \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} b^j(y)\rho(y)\mathrm{d}y \Bigg\}\\
&\qquad\qquad + \frac{1}{\sqrt{T}} \int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u} \sum_{m=1}^d \sigma_{jm} \mathrm{d}W_u^m \Bigg]\\
&\quad = \frac{1}{\sqrt{T}} \left(m_b(\lambda, X^T) + \zeta_T^j(\lambda) \right),
\end{aligned} \tag{2.3.18}$$

where

$$m_b(\lambda, X^T) \ := \ \frac{1}{\sqrt{T}} \int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u} b^j(X_u)\mathrm{d}u - \sqrt{T} \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} b^j(y)\rho(y)\mathrm{d}y.$$

For $\lambda \in \mathbb{R}^d$, $T > 0$, write

$$m_b(\lambda, X^T) = m_1(\lambda, X^T) + m_2(\lambda, X^T) + m_3(\lambda, X^T),$$

where

$$\begin{aligned}
m_1(\lambda, X^T) &:= \frac{1}{\sqrt{T}} \int_0^T \left(\mathrm{e}^{\mathrm{i}\lambda^t X_u} - \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} G_h(X_u - y)\mathrm{d}y \right) b^j(X_u)\mathrm{d}u,\\
m_2(\lambda, X^T) &:= \sqrt{T} \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} \left(\frac{1}{T} \int_0^T G_h(X_u - y) b^j(X_u)\mathrm{d}u - \int_{\mathbb{R}^d} b^j(z) G_h(z - y)\rho(z)\mathrm{d}z \right) \mathrm{d}y,\\
m_3(\lambda, X^T) &:= \sqrt{T} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} G_h(z - y)\mathrm{d}y - \mathrm{e}^{\mathrm{i}\lambda^t z} \right) b^j(z)\rho(z)\mathrm{d}z.
\end{aligned}$$

Due to Assumption (**SG**) (precisely, its consequence (**SG**')), and by means of Cauchy–Schwarz, it holds

$$\begin{aligned}
&\int_{\mathbb{R}^d} \mathbf{E}_b |m_1(\lambda, X^T)|^2 \widetilde{h}_T^2(\lambda)\mathrm{d}\lambda\\
&\qquad \leq C_P \int_{\mathbb{R}^d} \left\| \left(\mathrm{e}^{\mathrm{i}\lambda^t \cdot} - \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} G_h(y - \cdot)\mathrm{d}y \right) b^j \right\|_{L^2(\mu)}^2 \widetilde{h}_T^2(\lambda)\mathrm{d}\lambda\\
&\qquad \leq C_P \|b^j\|_{L^4(\mu)}^2 \int_{\mathbb{R}^d} \left\| \mathrm{e}^{\mathrm{i}\lambda^t \cdot} - \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} G_h(y - \cdot)\mathrm{d}y \right\|_{L^4(\mu)}^2 \widetilde{h}_T^2(\lambda)\mathrm{d}\lambda.
\end{aligned} \tag{2.3.19}$$

In order to evaluate the integrand in (2.3.19), note first that, for any $\lambda, z \in \mathbb{R}^d$,

$$\begin{aligned}
\left| \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} G_h(y - z)\mathrm{d}y - \mathrm{e}^{\mathrm{i}\lambda^t z} \right| &= \left| h^{-d} \int_{\mathbb{R}^d} \left(\mathrm{e}^{\mathrm{i}\lambda^t y} - \mathrm{e}^{\mathrm{i}\lambda^t z} \right)\ G\left(\frac{y - z}{h} \right) \mathrm{d}y \right|\\
&= \left| \int_{\mathbb{R}^d} \left(\mathrm{e}^{\mathrm{i}\lambda^t (z + hu)} - \mathrm{e}^{\mathrm{i}\lambda^t z} \right)\ G(u)\mathrm{d}u \right|.
\end{aligned}$$

Using a Taylor expansion of the characteristic function $\mathrm{e}^{\mathrm{i}x}$, $x \in \mathbb{R}$, up to order $\beta - 1$ with remainder term around $a \in \mathbb{R}$, one obtains

$$\mathrm{e}^{\mathrm{i}x} = \mathrm{e}^{\mathrm{i}a} + \sum_{k=1}^{\beta-1} \frac{\mathrm{e}^{\mathrm{i}a}}{k!} \left(\mathrm{i}(x-a)\right)^k + \frac{\mathrm{i}^\beta}{(\beta-1)!} \int_a^x (x-t)^{\beta-1} \, \mathrm{e}^{\mathrm{i}t} \mathrm{d}t.$$

Furthermore, for $x \geq a$,

$$\begin{aligned}\left| \frac{\mathrm{i}^\beta}{(\beta-1)!} \int_a^x (x-t)^{\beta-1} \, \mathrm{e}^{\mathrm{i}t} \mathrm{d}t \right| &\leq \frac{1}{(\beta-1)!} \int_a^x \left|(x-t)^{\beta-1}\right| \left|\mathrm{e}^{\mathrm{i}t}\right| \mathrm{d}t \\ &\leq \frac{1}{(\beta-1)!} \int_a^x (x-t)^{\beta-1} \mathrm{d}t = \frac{(x-a)^\beta}{\beta!}.\end{aligned}$$

Consequently, since G is of order $\beta - 1$ and satisfies (2.3.17), there exists some positive constant C_1 such that

$$\begin{aligned}&\left| \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} \, G_h(y-z) \mathrm{d}y - \mathrm{e}^{\mathrm{i}\lambda^t z} \right| \\ &\quad = \Bigg| \int_{\mathbb{R}^d} \sum_{k=1}^{\beta-1} \frac{\mathrm{e}^{\mathrm{i}\lambda^t z}}{k!} \, \Big(\mathrm{i}(\lambda^t(hu))\Big)^k \, G(u) \mathrm{d}u \\ &\qquad\quad + \int_{\mathbb{R}^d} \frac{\mathrm{i}^\beta}{(\beta-1)!} \int_{\lambda^t z}^{\lambda^t(z+hu)} \Big(\lambda^t(z+hu) - t\Big)^{\beta-1} \mathrm{e}^{\mathrm{i}t} \mathrm{d}t \; G(u) \mathrm{d}u \Bigg| \\ &\quad \leq \int_{\mathbb{R}^d} \frac{\left(\lambda^t(hu)\right)^\beta}{\beta!} \, G(u) \mathrm{d}u \leq C_1 \|\lambda\|^\beta h^\beta. \end{aligned} \tag{2.3.20}$$

Taking into account (2.3.19) and since

$$\|b^j\|^2_{L^4(\mu)} = \left(\int_{\mathbb{R}^d} \left(b^j\right)^4 (x) \rho(x) \mathrm{d}x \right)^{1/2} \leq C_2,$$

we obtain

$$\int_{\mathbb{R}^d} \mathbf{E}_b \left|m_1(\lambda, X^T)\right|^2 \widetilde{h}_T^2(\lambda) \mathrm{d}\lambda \leq C_3 h^{2\beta} \int_{\mathbb{R}^d} \|\lambda\|^{2\beta} \widetilde{h}_T^2(\lambda) \mathrm{d}\lambda \leq C_3 h^{2\beta} O(T). \tag{$\mathbf{A}$}$$

We turn to the second term. Denoting

$$\frac{1}{T} \int_0^T G_h(X_u - y) b^j(X_u) \mathrm{d}u =: F(X^T, y), \qquad T > 0, \; y \in \mathbb{R}^d,$$

it holds (since $|\widetilde{h}_T| \leq 1$ and by Plancherel's formula)

$$\begin{aligned}&\int_{\mathbb{R}^d} \mathbf{E}_b \left|m_2(\lambda, X^T)\right|^2 \widetilde{h}_T^2(\lambda) \mathrm{d}\lambda \\ &\quad \leq T \int_{\mathbb{R}^d} \mathbf{E}_b \left| \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} \Big(F(X^T, y) - \mathbf{E}_b F(X^T, y) \Big) \mathrm{d}y \right|^2 \mathrm{d}\lambda \\ &\quad = T \int_{\mathbb{R}^d} \mathbf{E}_b \left| F(X^T, y) - \mathbf{E}_b F(X^T, y) \right|^2 \mathrm{d}y.\end{aligned}$$

Lemma 2.3.4 then entails that

$$\int_{\mathbb{R}^d} \mathbf{E}_b \left|m_2(\lambda, X^T)\right|^2 \widetilde{h}_T^2(\lambda) \mathrm{d}\lambda \leq C_4 h^{2-d} \int_{\mathbb{R}^d} \sup_{z \in [y-h, y+h]} \rho(z) \mathrm{d}y = C_5 h^{2-d}. \tag{$\mathbf{B}$}$$

It remains to consider

$$\int_{\mathbb{R}^d} \mathbf{E}_b\left|m_3(\lambda, X^T)\right|^2 \widetilde{h}_T^2(\lambda)\mathrm{d}\lambda = T\int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t z}\widetilde{F}(\lambda,z)b^j(z)\rho(z)\mathrm{d}z\right|^2 \widetilde{h}_T^2(\lambda)\mathrm{d}\lambda,$$

where

$$\widetilde{F}(\lambda, z) := \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t(y-z)}G_h(z-y)\mathrm{d}y - 1, \qquad \lambda, z\in\mathbb{R}^d.$$

The upper bound (2.3.20) implies that, for any $\lambda, z \in \mathbb{R}^d$,

$$\left|\widetilde{F}(\lambda,z)\right| = \left|h^{-d}\int_{\mathbb{R}^d}\left(\mathrm{e}^{\mathrm{i}\lambda^t(z-y)}-1\right)G\left(\frac{z-y}{h}\right)\mathrm{d}y\right| \leq C_1\|\lambda\|^\beta h^\beta.$$

Consequently, for any $\lambda\in\mathbb{R}^d$,

$$\left|\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t z}\widetilde{F}(\lambda,z)b^j(z)\rho(z)\mathrm{d}z\right|^2 \leq C_6 h^{2\beta}\phi^2_{b^j\rho}(\lambda)\|\lambda\|^{2\beta}.$$

Recall that $b^j\rho = \operatorname{div}(a_j\rho)/2 \in \Sigma_\delta(\beta, L; b_0, \sigma)$. Thus,

$$\int_{\mathbb{R}^d} \mathbf{E}_b\left|m_3(\lambda, X^T)\right|^2 \widetilde{h}_T^2(\lambda)\mathrm{d}\lambda \leq Th^{2\beta}\int_{\mathbb{R}^d}\phi^2_{b^j\rho}(\lambda)\|\lambda\|^{2\beta}\mathrm{d}\lambda \leq C_7 Th^{2\beta}. \tag{\textbf{C}}$$

Balancing the upper bounds in (**A**), (**B**) and (**C**) suggests to select a bandwidth $h = h_T \sim T^{-1/(2(\beta-1)+d)}$, and this choice entails

$$\int_{\mathbb{R}^d} \mathbf{E}_b\left|m_\ell(\lambda, X^T)\right|^2 \widetilde{h}_T^2(\lambda)\mathrm{d}\lambda \lesssim T^{\frac{d-2}{2(\beta-1)+d}} = o\Big(T^{\frac{d}{2\beta+d}}\Big), \qquad \ell\in\{1,2,3\}. \tag{2.3.21}$$

To complete the proof of (2.3.16), note that

$$\begin{aligned}
\operatorname{Var}_b\left(\overline{\phi}_T^j(\lambda)\right) &\overset{(2.3.18)}{=} \frac{1}{T}\,\mathbf{E}_b\left(m_b(\lambda, X^T) + \zeta_T^j(\lambda)\right)^2\\
&= \frac{1}{T}\left(\mathbf{E}_b\left|m_b(\lambda, X^T)\right|^2 + a_{jj}\right)\\
&\leq \frac{1}{T}\,\mathbf{E}_b\left(4\sum_{\ell=1}^3\left|m_\ell(\lambda, X^T)\right|^2 + a_{jj}\right).
\end{aligned}$$

The last line relies on the rough upper bound $(a+b+c)^2 \leq 4(a^2+b^2+c^2)$. Thus,

$$\int_{\mathbb{R}^d}\widetilde{h}_T^2(\lambda)\operatorname{Var}_b\left(\overline{\phi}_T^j(\lambda)\right)\mathrm{d}\lambda \;\leq\; \frac{1}{T}\int_{\mathbb{R}^d}\widetilde{h}_T^2(\lambda)\,\mathbf{E}_b\left(4\sum_{\ell=1}^3\left|m_\ell(\lambda, X^T)\right|^2 + a_{jj}\right)\mathrm{d}\lambda,$$

and since $\|\widetilde{h}_T\|^2_{L^2(\mathbb{R}^d)} = O\Big(T^{\frac{d}{2\beta+d}}\Big)$, the assertion follows from (2.3.21). □

Remark 2.3.7 (Comparison to classical nonparametric density estimation). The approach to use Fourier analysis for investigating the MISE of kernel estimators is classical; see Section 1.3 in Tsybakov (2009). For pointing out the subtlety arising in the multidimensional diffusion framework, let us consider the problem of estimating a probability density $p\in L^2(\mathbb{R}^d)$ from i.i.d. $\mathbb{R}^d$-valued observations $X_1,\ldots,X_n$. Given some symmetric kernel $K\in L^2(\mathbb{R}^d)$, define the kernel density estimator

$$\widehat{p}_n(x) := \frac{1}{nh^d}\sum_{i=1}^n K\left(\frac{x-X_i}{h}\right), \qquad h>0,\ n\geq 1.$$

Let $\widehat{\phi}_n(\lambda) := \frac{1}{n}\sum_{j=1}^n \mathrm{e}^{\mathrm{i}\lambda^t X_j}$, $\lambda \in \mathbb{R}^d$, be the (classical) empirical characteristic function, and denote by $\mathbf{E}_p$ expectation with respect to p. Analogue to the proof of Theorem 1.4 in Tsybakov (2009), the following exact expression of the MISE of $\widehat{p}_n$ can be derived,

$$\begin{aligned}
&\mathbf{E}_p \int_{\mathbb{R}^d} \left(\widehat{p}_n(x) - p(x)\right)^2 \mathrm{d}x \\
&\quad= (2\pi)^{-d}\ \mathbf{E}_p \int_{\mathbb{R}^d} \left|\phi_{\widehat{p}_n}(\lambda) - \phi_p(\lambda)\right|^2 \mathrm{d}\lambda \\
&\quad= (2\pi)^{-d}\ \mathbf{E}_p \int_{\mathbb{R}^d} \left|\widehat{\phi}_n(\lambda)\phi_K(h\lambda) - \phi_p(\lambda)\right|^2 \mathrm{d}\lambda \\
&\quad= (2\pi)^{-d}\ \mathbf{E}_p \int_{\mathbb{R}^d} \left|(\widehat{\phi}_n(\lambda) - \phi_p(\lambda))\phi_K(h\lambda) - (1-\phi_K(h\lambda))\phi_p(\lambda)\right|^2 \mathrm{d}\lambda \\
&\quad= (2\pi)^{-d} \int_{\mathbb{R}^d} \Big(\left|\phi_K(h\lambda)\right|^2 \mathbf{E}_p\left|\widehat{\phi}_n(\lambda) - \phi_p(\lambda)\right|^2 + \left|1-\phi_K(h\lambda)\right|^2 \left|\phi_p(\lambda)\right|^2\Big) \mathrm{d}\lambda.
\end{aligned}$$

Since

$$\begin{aligned}
\mathbf{E}_p\left|\widehat{\phi}_n(\lambda) - \phi_p(\lambda)\right|^2 &= \mathbf{E}_p\left|\widehat{\phi}_n(\lambda)\right|^2 - \left|\phi_p(\lambda)\right|^2 = \mathbf{E}_p\big(\widehat{\phi}_n(\lambda)\phi_n(-\lambda)\big) - \left|\phi_p(\lambda)\right|^2 \\
&= \mathbf{E}_p\Big(\frac{1}{n}\sum_{j,k:k\neq j} \exp\left(\mathrm{i}\lambda^t(X_j - X_k)\right)\Big) + \frac{1}{n} - \left|\phi_p(\lambda)\right|^2 \\
&= \frac{n-1}{n}\ \phi_p(\lambda)\phi_p(-\lambda) + \frac{1}{n} - \left|\phi_p(\lambda)\right|^2 = \frac{1}{n}(1 - \left|\phi_p(\lambda)\right|^2),
\end{aligned}$$

one obtains in the i.i.d. setting the following exact expression for the MISE,

$$\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \left(\left|\phi_K(h\lambda)\right|^2 \frac{1}{n}\left(1 - \left|\phi_p(\lambda)\right|^2\right) + \left|1-\phi_K(h\lambda)\right|^2 \left|\phi_p(\lambda)\right|^2\right) \mathrm{d}\lambda. \tag{2.3.22}$$

While the computation of (2.3.14) in the multidimensional diffusion framework is analogue to the i.i.d. case, the manipulation of this relation to derive an analogue of (2.3.22) is more involved due to dependence of the data.

Letting $\widetilde{h}_T := \phi_{K_T}$, we obviously have

$$\left|\phi_{K_T}(\lambda)\right| = \left(1 - \|\alpha\lambda\|^{\beta+\gamma}\right)_+ \leq 1.$$

Assume for the moment that α is such that the conditions in (2.3.15) are satisfied. In this case, Lemma 2.3.6 can be applied. Plugging (2.3.16) into (2.3.14) then gives

$$\mathbf{E}_b \int_{\mathbb{R}^d} \left|\overline{g}_T^j(x) - \mathrm{div}(a_j\rho)(x)\right|^2 \mathrm{d}x \leq \frac{1}{(2\pi)^d T}\ \widetilde{\Delta}_T^j(\alpha,\beta)\ (1 + o_T(1)), \tag{2.3.23}$$

where

$$\widetilde{\Delta}_T^j(\alpha,\beta) := 4a_{jj}\int_{\mathbb{R}^d} \left|\phi_{K_T}(\lambda)\right|^2 \mathrm{d}\lambda + T\int_{\mathbb{R}^d} \left|\phi_{K_T}(\lambda) - 1\right|^2 \left|\phi_{\mathrm{div}(a_j\rho)}(\lambda)\right|^2 \mathrm{d}\lambda. \tag{2.3.24}$$

We continue with deriving an upper bound on the second term appearing on the RHS of (2.3.24). Note first that, for the kernel K_T defined via (2.3.11), it holds

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left|\phi_{K_T}(\lambda) - 1\right|^2 \left|\phi_{\mathrm{div}(a_j\rho)}(\lambda) - \phi_{\mathrm{div}(a_j\rho_0)}(\lambda)\right|^2 \mathrm{d}\lambda \\
&\quad= \int_{\|\lambda\|>\alpha^{-1}} \left|\phi_{\mathrm{div}(a_j\rho)}(\lambda) - \phi_{\mathrm{div}(a_j\rho_0)}(\lambda)\right|^2 \mathrm{d}\lambda \\
&\qquad\quad+ \int_{\|\lambda\|\leq\alpha^{-1}} \alpha^{2\beta}\|\lambda\|^{2\beta} \left|\phi_{\mathrm{div}(a_j\rho)}(\lambda) - \phi_{\mathrm{div}(a_j\rho_0)}(\lambda)\right|^2 \|\alpha\lambda\|^{2\gamma} \mathrm{d}\lambda \\
&\quad\leq \int_{\mathbb{R}^d} \alpha^{2\beta}\|\lambda\|^{2\beta} \left|\phi_{\mathrm{div}(a_j\rho)}(\lambda) - \phi_{\mathrm{div}(a_j\rho_0)}(\lambda)\right|^2 \mathrm{d}\lambda \\
&\quad\leq (2\pi)^d\ 4L\ \alpha^{2\beta}.
\end{aligned} \tag{2.3.25}$$

The last line follows from the assumption that the components of the vector div $(a\rho - a\rho_0)$ belong to the Sobolev ball Σ_δ of regularity β with radius $4L$ (see the definition in (2.3.2)) such that

$$\int_{\mathbb{R}^d} \|\lambda\|^{2\beta} \big|\phi_{\mathrm{div}(a_j\rho)}(\lambda) - \phi_{\mathrm{div}(a_j\rho_0)}(\lambda)\big|^2 \mathrm{d}\lambda \leq 4L\ (2\pi)^d.$$

Using the same decomposition as in (2.3.25), it can be shown that, for sufficiently large T, there exists some positive constant C such that

$$\begin{aligned}\int_{\mathbb{R}^d} \big|\phi_{K_T}(\lambda) - 1\big|^2 \big|\phi_{\mathrm{div}(a_j\rho_0)}(\lambda)\big|^2 \mathrm{d}\lambda &\leq \int_{\mathbb{R}^d} \|\alpha\lambda\|^{2(\beta+\gamma)} \big|\phi_{\mathrm{div}(a_j\rho_0)}(\lambda)\big|^2 \mathrm{d}\lambda \\ &\leq \alpha^{2(\beta+\gamma)} \int_{\mathbb{R}^d} \|\lambda\|^{2\beta+\tau} \big|\phi_{\mathrm{div}(a_j\rho_0)}(\lambda)\big|^2 \mathrm{d}\lambda \\ &\leq C\alpha^{2(\beta+\gamma)},\end{aligned}$$

where the last line follows from condition (C_3). Summing up, $\widetilde{\Delta}_T^j(\alpha, \beta)$ as defined in (2.3.24) is bounded from above by

$$\Big\{4a_{jj} \int_{\mathbb{R}^d} \big(1 - \|\alpha\lambda\|^\beta\big)_+^2 \mathrm{d}\lambda + 4T(2\pi)^d L\alpha^{2\beta}\Big\}\ (1 + K\alpha^{2\gamma}), \tag{2.3.26}$$

K some positive constant. Later on, α will be chosen such that $\alpha^{2\gamma} = o_T(1)$. The expression in braced brackets is minimized by the bandwidth α_T^j satisfying

$$a_{jj} \int_{\mathbb{R}^d} \|\lambda\|^\beta \big(1 - \|\alpha_T^j\lambda\|^\beta\big)_+ \mathrm{d}\lambda = \big(\alpha_T^j\big)^\beta T(2\pi)^d L,$$

and plugging this relation into (2.3.26) gives

$$\widetilde{\Delta}_T^j(\alpha_T^j, \beta) \leq 4a_{jj} \int_{\mathbb{R}^d} \big(1 - \|\alpha_T^j\lambda\|^\beta\big)_+ \Big\{\big(1 - \|\alpha_T^j\lambda\|^\beta\big)_+ + \|\alpha_T^j\lambda\|^\beta\Big\} \mathrm{d}\lambda. \tag{2.3.27}$$

Recall that $\mathbb{S}_d = 2\pi^{d/2}/\Gamma(d/2)$ denotes the surface of the unit sphere in $\mathbb{R}^d$. Since

$$\begin{aligned}\int_{\mathbb{R}^d} \|\lambda\|^\beta \big(1 - \|\alpha\lambda\|^\beta\big)_+ \mathrm{d}\lambda &= \alpha^{-(\beta+d)} \int_{\mathbb{R}^d} \|\lambda\|^\beta \big(1 - \|\lambda\|^\beta\big)_+ \mathrm{d}\lambda \\ &= \alpha^{-(\beta+d)} \int_0^1 r^\beta (1 - r^\beta) r^{d-1} \mathbb{S}_d \mathrm{d}r \\ &= \alpha^{-(\beta+d)}\ \frac{\beta\mathbb{S}_d}{(\beta+d)(2\beta+d)},\end{aligned}$$

straightforward algebra yields

$$\alpha_T^j = \bigg(\frac{\beta a_{jj}\mathbb{S}_d}{T(2\pi)^d L(\beta+d)(2\beta+d)}\bigg)^{\frac{1}{2\beta+d}}. \tag{2.3.28}$$

In particular, the specific choice $\alpha = \alpha_T^j$ implies that there exist positive constants C_1, C_2 such that

$$\begin{aligned}\int_{\mathbb{R}^d} \|\lambda\|^{2\beta} \big|\phi_{K_T}(\lambda)\big|^2 \mathrm{d}\lambda &= \int_{\mathbb{R}^d} \|\lambda\|^{2\beta} \big(1 - \|\alpha_T^j\lambda\|^\beta\big)_+^2 \mathrm{d}\lambda \\ &= \big(\alpha_T^j\big)^{-2\beta-d} \int_{\|\omega\|\leq 1} \|\omega\|^{2\beta} \big(1 - \|\omega\|^\beta\big)^2 \mathrm{d}\omega \\ &\leq C_1 \big(\alpha_T^j\big)^{-2\beta-d} = C_2 T.\end{aligned}$$

Consequently, Lemma 2.3.6 is applicable indeed. Denote by $\left(g_T^j\right)^* := \overline{g}_T^{j,(\alpha_T^j,\beta)}$ the estimator defined via (2.3.11) and (2.3.28). In view of (2.3.23) and (2.3.27), it holds

$$\begin{aligned}
&\mathbf{E}_b \int_{\mathbb{R}^d} \left|(g_T^j)^*(x) - \operatorname{div}(a_j\rho)(x)\right|^2 \mathrm{d}x \\
&\quad\leq \frac{4a_{jj}}{(2\pi)^d T} \int_{\mathbb{R}^d} (1-\|\alpha_T^j\lambda\|^\beta)_+ \big((1-\|\alpha_T^j\lambda\|^\beta)_+ + \|\alpha_T^j\lambda\|^\beta\big)\mathrm{d}\lambda\ (1+o_T(1)) \\
&\quad= \frac{4a_{jj}}{(2\pi\alpha_T^j)^d T} \int_{\mathbb{R}^d} (1-\|\lambda\|^\beta)\mathrm{d}\lambda\ (1+o_T(1)) \\
&\quad= \frac{4a_{jj}\mathbb{S}_d}{(2\pi\alpha_T^j)^d T} \int_0^1 r^{d-1}(1-r^\beta)\mathrm{d}r\ (1+o_T(1)) \\
&\quad= 4P_j(\sigma,\beta,L)\ T^{-\frac{2\beta}{2\beta+d}}.
\end{aligned} \tag{2.3.29}$$

(B) Estimation of the vector $\operatorname{div}(a\rho)$**.** As an alternative to component-wise estimation of $\operatorname{div}(a\rho)$, consider the following vector-valued estimator,

$$\check{g}_T(x) := \check{g}_T^{\alpha,\beta}(x) := \frac{2}{T}\int_0^T K_T(x-X_u)\mathrm{d}X_u, \qquad x\in\mathbb{R}^d,$$

where the kernel K_T is defined via its Fourier transform

$$\phi_{K_T}(\lambda) = \left(1-\|\alpha\lambda\|^{\beta+\gamma_T}\right)_+, \qquad \lambda\in\mathbb{R}^d,\ \gamma_T = (\log\log T)^{-1}. \tag{2.3.30}$$

Denote

$$\Sigma_\delta := \Big\{ b\in\bigcap_{k=1}^d U_\delta^k(b_0) : \sum_{k=1}^d \sum_{|\alpha|\leq\beta} \int_{\mathbb{R}^d} \left|D^\alpha\left(\operatorname{div}(a_k\rho) - \operatorname{div}(a_k\rho_0)\right)\right|^2 \leq 4Ld\Big\}. \tag{2.3.31}$$

In contrast to the estimators $\overline{g}_T^1,\ldots,\overline{g}_T^d$ whose respective bandwidths are chosen individually, the sequence $\mathbb{R}_+\ni\alpha=\alpha_T\to 0$ for estimating the components of $\operatorname{div}(a\rho)$ is now always the same. To derive an upper bound on the MISE of $\check{g}_T$, note that (2.3.23) implies that

$$\mathbf{E}_b \int_{\mathbb{R}^d} \|\check{g}_T(\lambda) - \operatorname{div}(a\rho)(\lambda)\|^2\,\mathrm{d}\lambda \leq \frac{1}{(2\pi)^d T}\,\widetilde{\Delta}_T(\alpha,\beta)\ (1+o_T(1)),$$

where, with $\|\sigma\|_{S_2}$ denoting the Frobenius norm of the matrix σ,

$$\widetilde{\Delta}_T(\alpha,\beta) := 4\|\sigma\|_{S_2}^2 \int_{\mathbb{R}^d} |\phi_{K_T}(\lambda)|^2\mathrm{d}\lambda + T\int_{\mathbb{R}^d} |\phi_{K_T}(\lambda)-1|^2 \|\phi_{\operatorname{div}(a\rho)}(\lambda)\|^2\mathrm{d}\lambda.$$

It is straightforward to show that the optimal bandwidth α_T^{vec} is characterized by the equation

$$\|\sigma\|_{S_2}^2 \int_{\mathbb{R}^d} \|\lambda\|^\beta \left(1-\|\alpha_T^{\mathrm{vec}}\lambda\|^\beta\right)_+ \mathrm{d}\lambda = (\alpha_T^{\mathrm{vec}})^\beta\, T(2\pi)^d Ld,$$

and, similarly to (2.3.27), one obtains

$$\widetilde{\Delta}_T(\alpha_T^{\mathrm{vec}},\beta) \leq 4\|\sigma\|_{S_2}^2 \int_{\mathbb{R}^d} \left(1-\|\alpha_T^{\mathrm{vec}}\lambda\|^\beta\right)_+ \left\{\left(1-\|\alpha_T^{\mathrm{vec}}\lambda\|^\beta\right)_+ + \|\alpha_T^{\mathrm{vec}}\lambda\|^\beta\right\}\mathrm{d}\lambda.$$

The optimal bandwidth α_T^{vec} is thus identified as

$$\alpha_T^{\mathrm{vec}} = \left(\frac{\beta\|\sigma\|_{S_2}^2\mathbb{S}_d}{T(2\pi)^d Ld(\beta+d)(2\beta+d)}\right)^{\frac{1}{2\beta+d}}. \tag{2.3.32}$$

Denote by $\check{g}_T^* := \check{g}_T^{\alpha_T^{\mathrm{vec}},\beta}$ the estimator defined via (2.3.30) and (2.3.32). Taking into account Theorem 2.3.2, the above reasoning gives

Theorem 2.3.8. *Suppose that Assumption* (**SG**+) *is satisfied.*

(A) *The estimator* $(\overline{g}_T^j)^*$ *defined via (2.3.11) and (2.3.28) satisfies, for any* $\mathbb{N} \ni \beta > 1$,

$$\lim_{\delta\to 0}\lim_{T\to\infty}\sup_{b\in\Sigma_\delta^j} T^{\frac{2\beta}{2\beta+d}}\ \mathbf{E}_b\int_{\mathbb{R}^d}\left|(\overline{g}_T^j)^*(x)-\operatorname{div}(a_j\rho)(x)\right|^2\mathrm{d}x = 4\ P_j(\sigma,\beta,L).$$

(B) *For any* $\beta\in\mathbb{N}$, *the estimator* $\check{g}_T^*$ *defined via (2.3.11) and (2.3.32) satisfies*

$$\lim_{\delta\to 0}\lim_{T\to\infty}\sup_{b\in\Sigma_\delta} T^{\frac{2\beta}{2\beta+d}}\ \mathbf{E}_b\int_{\mathbb{R}^d}\left\|\check{g}_T^*(x)-\operatorname{div}(a\rho)(x)\right\|^2\mathrm{d}x \leq 4\ P(\sigma,\beta,L),$$

where

$$P(\sigma,\beta,L) := (L(2\beta+d))^{\frac{d}{2\beta+d}}\left(\frac{\|\sigma\|_{S_2}^2\beta\ \mathbb{S}_d}{(2\pi)^d d(\beta+d)}\right)^{\frac{2\beta}{2\beta+d}}. \tag{2.3.33}$$

Remark 2.3.9. Part (A) of Theorem 2.3.8 in particular implies that

$$\lim_{T\to\infty}\sum_{j=1}^d\sup_{b\in\Sigma_\delta^j} T^{\frac{2\beta}{2\beta+d}}\ \mathbf{E}_b\int_{\mathbb{R}^d}\left|(\overline{g}_T^j)^*(x)-\operatorname{div}(a_j\rho)(x)\right|^2\mathrm{d}x \leq 4\sum_{j=1}^d P_j(\sigma,\beta,L).$$

Jensen's inequality entails that

$$\begin{aligned}\sum_{j=1}^d P_j(\sigma,\beta,L) &= (L(2\beta+d))^{\frac{d}{2\beta+d}}\left(\frac{\beta\ \mathbb{S}_d}{(2\pi)^d(\beta+d)}\right)^{\frac{2\beta}{2\beta+d}}\quad \frac{1}{d}\sum_{j=1}^d a_{jj}^{\frac{2\beta}{2\beta+d}}\\ &\leq (L(2\beta+d))^{\frac{d}{2\beta+d}}\left(\frac{\beta\ \mathbb{S}_d}{(2\pi)^d(\beta+d)}\right)^{\frac{2\beta}{2\beta+d}}\quad\left(\frac{1}{d}\sum_{j=1}^d a_{jj}\right)^{\frac{2\beta}{2\beta+d}} = P(\sigma,\beta,L).\end{aligned}$$

Notably, this shows that the approach of component-wise estimation is to be preferred to implementing a vector-wise estimator of $\operatorname{div}(a\rho)$. Comparison of the quantities

$$\frac{1}{d}\sum_{j=1}^d a_{jj}^{\frac{2\beta}{2\beta+d}}\quad\text{and}\quad\left(\frac{\|\sigma\|_{S_2}^2}{d}\right)^{\frac{2\beta}{2\beta+d}}$$

further allows to assess the difference between estimating the drift vector component-wise or by one single (vector-valued) function.

Two steps in our development of an asymptotically efficient global non-adaptive drift estimator still have to be carried out, namely

(i) the definition of an asymptotically efficient estimator of the *drift vector* (which will rely on the asymptotically efficient estimator of $\operatorname{div}(a\rho)$), and

(ii) the translation of the smoothness assumption on $\operatorname{div}(a\rho)$ into (weighted) Sobolev smoothness of the drift vector b.

As concerns item (i), note that the estimators (2.1.3) and (2.1.5) suggested in the introduction involve an a priori lower bound $\rho_*(x)>0$ on the invariant density ρ. In the current framework, we are interested in estimating the drift of diffusions with exponentially fast decreasing invariant density over the entire $\mathbb{R}^d$. The invariant density thus is not bounded away from zero. For defining a reasonable drift estimator in such situations where no lower bound is available, one may replace the denominator $\widehat{\rho}_T(x)\vee\rho_*(x)$ by the expression $\widehat{\rho}_T(x)+c_T(x)$, for some additional term $c_T(x)$ included in order to prevent the denominator of the drift estimator from vanishing.

The idea is classical and has been used in the scalar setting, e.g., in Banon (1978), Dalalyan and Kutoyants (2002) and Dalalyan (2005). We restrict attention to the multidimensional case $d \geq 2$.

Theorem 2.3.10. *In the setting of Theorem 2.3.8, define an invariant density estimator $\widehat{\rho}_{T,1}$ according to Lemma 2.2.4, and let*

$$c_T^*(x) := T^{-\frac{\beta+1}{2(\beta+1)+d}} \varepsilon_T(x), \quad \text{for } \varepsilon_T(x) := \exp\left(\sqrt{\log T} - \frac{\|x\|}{\log T}\right), \qquad x \in \mathbb{R}^d.$$

Define the estimator $\left(\overline{g}_T^j\right)^*$ according to (2.3.11) and (2.3.28), and set

$$\left(b_T^j\right)^*(x) := \frac{\left(\overline{g}_T^j\right)^*(x)}{2\left(\widehat{\rho}_{T,1}(x) + c_T^*(x)\right)}, \qquad x \in \mathbb{R}^d. \tag{2.3.34}$$

Then

$$\lim_{\delta \to 0} \lim_{T \to \infty} \sup_{b \in \Sigma_\delta^j} T^{\frac{2\beta}{2\beta+d}} \; \mathbf{E}_b \int_{\mathbb{R}^d} \left|\left(b_T^j\right)^*(x) - b(x)\right|^2 \rho^2(x) \mathrm{d}x = P_j(\sigma, \beta, L). \tag{2.3.35}$$

Remark 2.3.11 (Comparison to classical nonparametric density estimation)*.* In the classical i.i.d. framework, Rigollet and Tsybakov (2007) consider the problem of nonparametric density estimation over Sobolev classes of densities p on $\mathbb{R}^d$, satisfying the smoothness condition

$$\int_{\mathbb{R}^d} \|\lambda\|^{2\beta} \left|\phi_p(\lambda)\right|^2 \mathrm{d}\lambda \leq Q, \qquad Q > 0,\ \beta > 0.$$

Recall that our smoothness assumption for component-wise estimation reads

$$\sum_{|\alpha| \leq \beta} \int_{\mathbb{R}^d} \left|D^\alpha (b^j \rho - b_0^j \rho_0)\right|^2 = (2\pi)^{-d} \int_{\mathbb{R}^d} \|\lambda\|^{2\beta} \left|\phi_{b^j \rho - b_0^j \rho_0}(\lambda)\right|^2 \mathrm{d}\lambda \leq L. \tag{2.3.36}$$

Replacing Q in the definition of the optimal constant for nonparametric density estimation obtained in Rigollet and Tsybakov (2007) with $(2\pi)^d L$ yields the constant

$$\frac{(L(2\beta+d))^{\frac{d}{2\beta+d}}}{d} \left(\frac{\beta \mathbb{S}_d}{(2\pi)^d (\beta+d)}\right)^{\frac{2\beta}{2\beta+d}} = P_j(\mathbf{Id}_{d\times d}, \beta, L) = \frac{P(\mathbf{Id}_{d\times d}, \beta, L)}{d}.$$

We now turn to formulating the result under conditions on the weighted Sobolev regularity of b, that is, we replace (2.3.36) with the criterion in Proposition 2.3.1, namely,

$$\sum_{\alpha \in \mathbb{N}^d : |\alpha| \leq \beta} \int_{\mathbb{R}^d} \left|D^\alpha (b^j(x) - b_0^j(x))\right|^2 \rho_b^2(x) \mathrm{d}x \leq L. \tag{2.3.37}$$

Combining Proposition 2.3.1 and Theorem 2.3.10, we finally obtain

Corollary 2.3.12. *Assume that b_0 satisfies (2.3.4). Then, in the setting of Theorem 2.3.10, it holds*

$$\lim_{\delta \to 0} \lim_{T \to \infty} \sup_{b \in \overline{\Sigma}_\delta^j} T^{\frac{2\beta}{2\beta+d}} \; \mathbf{E}_b \int_{\mathbb{R}^d} \left|\left(b_T^j\right)^*(x) - b(x)\right|^2 \rho^2(x) \mathrm{d}x = P_j(\sigma, \beta, L).$$

The inclusion $\overline{\Sigma}_\delta^j(\beta, L; b_0, \sigma) \subset \Sigma_\delta^j(\beta, L + o_\delta(1); b_0, \sigma)$ which is proven in Proposition 2.3.1 actually presents an important motivation for considering a local minimax framework and for analyzing the behavior of the drift estimator *over a shrinking neighborhood* of the central function b_0. In the classical statistical models of signal recovery in Gaussian white noise, density or regression function estimation from i.i.d. observations, the Pinsker-type bounds are obtained for Sobolev balls centered at the origin, that is, for some centering function $f_0 \equiv 0$. The choice $b_0 \equiv 0$ however is not appropriate in our setup. Even in the one-dimensional setting, additional conditions are needed in order to ensure ergodicity of the associated process. The case of null-recurrent diffusions has to be excluded since the rates of convergence for drift estimation are significantly worse than for ergodic processes; see Section 4.3 in Dalalyan (2005) for further explanation. The situation in higher dimension is even more difficult as the d-dimensional Brownian motion is known to be transient in dimension $d \geq 3$.

(II) Asymptotically efficient global drift estimation: The sharp adaptive procedure

The estimators described in Theorem 2.3.8 depend on the smoothness parameter β and the radius L of the Sobolev ball which are usually not available, and the question considered now is how to choose the parameters optimally in a completely data-driven way. Given the observations $X^T = (X_t)_{0\leq t\leq T}$, the aim is to select the real parameters α and β such that the associated estimator of $\operatorname{div}(a\rho)$ has asymptotically minimax risk. For ease of presentation, we restrict attention to the case of Kolmogorov diffusions, that is, we consider the case $\sigma \equiv \mathbf{Id}_{d\times d}$. In this case, it means no loss of generality to restrict attention to the approach of estimating the entire drift vector by selecting one common bandwidth.

Dalalyan (2005) investigates the global estimation problem of the drift function for a class of scalar ergodic diffusion processes. The unknown drift is supposed to belong to a nonparametric class of smooth functions of unknown order $\beta \geq 1$. The author proves adaptivity of the procedure up to an optimal (Pinsker) constant, for the integrated squared error criterion, weighted by the square of the invariant density. The proposed data-driven procedure of estimating the drift function relies on the *estimated risk minimization method* as it is developed in Cavalier et al. (2002). As explained in Dalalyan (2005), it is not desirable in the adaptive setup to use an estimator which involves a stochastic integral such that an estimator of the form of $\overline{g}_T$ defined in the previous section (cf. (2.3.10)) is not convenient. The reason is that one obtains an anticipative stochastic integral if the bandwidth is chosen in a data-driven way such that it depends on the observations $X^T = (X_t)_{0\leq t\leq T}$. The manipulation of such integrals is rather involved. To circumvent this obstacle, the adaptive estimator of $\nabla\rho$ will have the following form,

$$\widehat{g}_T(x) := \frac{1}{T}\int_0^T \nabla K_T^*(x - X_u)\mathrm{d}u, \qquad T > 0,\ x \in \mathbb{R}^d,$$

for some Pinsker kernel $K_T : \mathbb{R}^d \to \mathbb{R}$. This alternative type of kernel estimator of $\nabla\rho$ has been motivated already in Section 2.2.3; see in particular (2.2.15). Following Cavalier et al. (2002) and Dalalyan (2005), the parameters which characterize the Fourier transform ϕ_{K_T} of K_T (and thus K_T) will be chosen now according to the principle of *unbiased risk estimation.* A heuristic motivation of Mallows' C_p criterion is given in Cavalier et al. (2002) (pp. 846–847). In the current context, the basic idea can be explained as follows.

Similarly to (2.3.24), it will be shown that

$$\mathbf{E}_b \int_{\mathbb{R}^d} \left\|\widehat{g}_T(x) - \nabla\rho(x)\right\|^2 \mathrm{d}x \leq \frac{1}{(2\pi)^d T}\ \Delta_T\big(1, \phi_{K_T}, \phi_\rho\big)\ \left(1 + o_T(1)\right),$$

where, for some positive parameter $\alpha \in \mathbb{R}$ and some function h,

$$\Delta_T(\alpha, h, \phi_\rho) := T\int_{\mathbb{R}^d} |1 - h(\alpha\lambda)|^2 \left\|\phi_{\nabla\rho}(\lambda)\right\|^2 \mathrm{d}\lambda + 4d\int_{\mathbb{R}^d} |h(\alpha\lambda)|^2 \mathrm{d}\lambda. \tag{2.3.38}$$

If the function h is parametrized by some $\beta \in \mathbb{R}$, the optimal choice of the parameters (α, β) are the values (α^*, β^*) which minimize the functional $\Delta_T(\alpha, h_\beta, \phi_\rho)$. The basic idea of the method of *adaptation via unbiased risk estimation* is to replace $\Delta_T(\alpha, h, \phi_\rho)$ (whose minimizers depend on the unknown drift function b) by some good estimator $\widehat{\Delta}_T(h) = \widehat{\Delta}_T(h, X^T)$. To minimize $\Delta_T(\alpha, h_\beta, \phi_\rho)$, or, equivalently,

$$\begin{aligned}\Delta_T(\alpha, h_\beta, \phi_\rho) &- T\int_{\mathbb{R}^d} \|\phi_{\nabla\rho}(\lambda)\|^2 \,\mathrm{d}\lambda \\ &= T\int_{\mathbb{R}^d} \left(h_\beta^2(\alpha\lambda) - 2h_\beta(\alpha\lambda)\right) \|\phi_{\nabla\rho}(\lambda)\|^2 \,\mathrm{d}\lambda + 4d\int_{\mathbb{R}^d} |h_\beta(\alpha\lambda)|^2 \,\mathrm{d}\lambda\end{aligned}$$

with respect to α and β, we thus replace $\|\phi_{\nabla\rho}\|^2$ by some suitable estimator. A classical strategy in such situations is to resort to the "plug-in approach," that is, to estimate $\Delta_T(\alpha, h, \phi_\rho)$ by replacing $\phi_\rho(\lambda)$ with $\widehat{\phi}_T(\lambda)$, where

$$\widehat{\phi}_T(\lambda) := \frac{1}{T}\int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u}\mathrm{d}u, \qquad \lambda \in \mathbb{R}^d, \tag{2.3.39}$$

denotes the empirical characteristic function. It was noted by Dalalyan (2005) in the scalar setting that it is not convenient to use the plug-in approach in the given framework since plug-in estimators of quadratic functionals are heavily biased. Since the same remark applies in the multidimensional situation, we introduce the functional

$$\begin{aligned}\widehat{\Delta}_T(h) &:= T\int_{\mathbb{R}^d} \left(h^2(\lambda) - 2h(\lambda)\right)\left\{\|\lambda\widehat{\phi}_T(\lambda)\|^2 - 4dT^{-1}\right\}\mathrm{d}\lambda + 4d\int_{\mathbb{R}^d} h^2(\lambda)\mathrm{d}\lambda \qquad (2.3.40)\\ &= T\int_{\mathbb{R}^d} \left(h^2(\lambda) - 2h(\lambda)\right)\|\lambda\widehat{\phi}_T(\lambda)\|^2\mathrm{d}\lambda + 8d\int_{\mathbb{R}^d} h(\lambda)\mathrm{d}\lambda\end{aligned}$$

which depends only on the observed path X^T. The definition of $\widehat{\Delta}_T$ is formally justified as follows. Similarly to (2.3.12), one obtains, for any $\lambda \in \mathbb{R}^d$,

$$\begin{aligned}\phi_{\widehat{g}_T}(\lambda) &= \int_{\mathbb{R}^d} \frac{1}{T}\int_0^T \nabla K_T(x - X_u)\mathrm{d}u\; \mathrm{e}^{\mathrm{i}\lambda^t x}\mathrm{d}x \\ &= \frac{1}{T}\int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u}\mathrm{d}u \int_{\mathbb{R}^d} \nabla K_T(y)\mathrm{e}^{\mathrm{i}\lambda^t y}\mathrm{d}y \\ &= (-\mathrm{i})\lambda\widehat{\phi}_T(\lambda)\phi_{K_T}(\lambda),\end{aligned} \tag{2.3.41}$$

where we used that

$$\int_{\mathbb{R}^d} \nabla K_T(y)\mathrm{e}^{\mathrm{i}\lambda^t y}\mathrm{d}y = -\mathrm{i}\int_{\mathbb{R}^d} K_T(y)\lambda\; \mathrm{e}^{\mathrm{i}\lambda^t y}\mathrm{d}y = (-\mathrm{i})\lambda\phi_{K_T}(\lambda).$$

Furthermore, by Plancherel's formula,

$$\begin{aligned}&\mathbf{E}_b\int_{\mathbb{R}^d} \|\widehat{g}_T(x) - \nabla\rho(x)\|^2\mathrm{d}x \\ &\quad= (2\pi)^{-d}\,\mathbf{E}_b\int_{\mathbb{R}^d} \|\phi_{\widehat{g}_T}(\lambda) - \phi_{\nabla\rho}(\lambda)\|^2\mathrm{d}\lambda \\ &\quad\overset{(2.3.41)}{=} (2\pi)^{-d}\,\mathbf{E}_b\int_{\mathbb{R}^d} \|\lambda(\widehat{\phi}_T(\lambda)\phi_{K_T}(\lambda) - \phi_\rho(\lambda))\|^2\mathrm{d}\lambda \\ &\quad= (2\pi)^{-d}\,\mathbf{E}_b\int_{\mathbb{R}^d} \|\lambda\{\phi_{K_T}(\lambda)(\widehat{\phi}_T - \phi_\rho)(\lambda) - \phi_\rho(\lambda)(1 - \phi_{K_T}(\lambda))\}\|^2\mathrm{d}\lambda.\end{aligned} \tag{2.3.42}$$

Comparing this last display to the expression of the MISE of the estimator

$$\overline{g}_T(x) = \frac{1}{T}\int_0^T K_T(x - X_u)\mathrm{d}X_u$$

investigated in Section 2.3.3, two substantial differences appear, namely,

- the function $\overline{\phi}_T(\lambda) = T^{-1}\int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u}\mathrm{d}X_u$ is replaced by the product $\lambda\widehat{\phi}_T(\lambda)$, where $\widehat{\phi}_T(\lambda) = T^{-1}\int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u}\mathrm{d}u$ is the empirical characteristic function, and
- the Fourier transform ϕ_{K_T} of the kernel K_T is random due to the adaptive choice of the bandwidth.

The following lemma shows that these differences do not entail substantial modifications in the subsequent analysis. It can be viewed as some multidimensional analogue of Lemma 1 in Dalalyan (2005). Slightly abusing notation, let

$$\zeta_T(\lambda) := \frac{1}{\sqrt{T}}\int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u}\mathrm{d}W_u. \tag{2.3.43}$$

Lemma 2.3.13. *Assume that $b \in \Sigma_\delta(\beta, L; b_0, \mathbf{Id}_{d\times d})$, and consider the function $\widehat{\phi}_T(\cdot)$ defined in (2.3.39). Then, for any random bounded function $\widetilde{h}_T : \mathbb{R}^d \to \mathbb{R}$ satisfying*

$$|\widetilde{h}_T| \leq 1 \ \textit{a.s.}, \quad \textit{and} \quad \int_{\mathbb{R}^d} \|\lambda\|^{2\beta}\ \mathbf{E}_b \widetilde{h}_T^2(\lambda)\mathrm{d}\lambda = O(T), \tag{2.3.44}$$

it holds, for any $\lambda \in \mathbb{R}^d$,

$$\mathrm{i}\lambda(\widehat{\phi}_T(\lambda) - \phi_\rho(\lambda)) = \frac{1}{\sqrt{T}}\ \Big(m_b(\lambda, X^T) - 2\zeta_T(\lambda)\Big), \tag{2.3.45}$$

where $m_b(\lambda, X^T)$ is a measurable $\mathbb{C}^d$-valued function, satisfying

$$\int_{\mathbb{R}^d} \mathbf{E}_b \big\|m_b(\lambda, X^T)\ \widetilde{h}_T(\lambda)\big\|^2 \mathrm{d}\lambda = o\Big(T^{\frac{d}{2\beta+d}}\Big). \tag{2.3.46}$$

We start with sketching the basic idea of the proof. An important role in the investigation of the asymptotically exact behavior of the drift estimators for scalar diffusions considered in Dalalyan (2005) is taken on by the martingale representation of the local time estimator developed in Kutoyants (1999). Let X be a scalar ergodic diffusion solving the SDE

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \mathrm{d}W_t,$$

and denote its invariant density by ρ. The occupation times formula implies that

$$\frac{1}{T}\int_0^T \mathrm{e}^{\mathrm{i}\lambda X_u}\mathrm{d}u - \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}\lambda y}\rho(y)\mathrm{d}y = \int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}\lambda y}\left(\frac{1}{T}\ \ell_T(y) - \rho(y)\right)\mathrm{d}y, \qquad \lambda \in \mathbb{R}, \tag{2.3.47}$$

where $\ell_T(x)$ is the local time of X at $x \in \mathbb{R}$ and time T. Tanaka's formula states that

$$\ell_T(x) = |X_T - x| + |X_0 - x| - \int_0^T \operatorname{sgn}(X_s - x)\mathrm{d}X_s, \qquad T \geq 0 \tag{2.3.48}$$

(see, e.g., pp. 428–429 in Kallenberg (2002)). In particular, (2.3.48) allows to prove a representation of ℓ_T as a sum of the local martingale $T^{-1}\int_0^T \mathrm{e}^{\mathrm{i}\lambda X_u}\mathrm{d}W_u$ and an asymptotically negligible integral with respect to time.

In the proof of Lemma 2.3.13, we consider an auxiliary estimator which admits a similar representation as the local time $\ell_T(x)$ in (2.3.48). The basic idea in the multidimensional framework is to approximate the empirical characteristic function $\widehat{\phi}_T$ by an auxiliary estimator $\widetilde{\phi}_{T,h}$ of the form

$$\widetilde{\phi}_{T,h}(\lambda) := \frac{1}{T}\int_0^T \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} G_h(X_u - y)\mathrm{d}y\ \mathrm{d}u, \qquad h, T > 0,\ \lambda \in \mathbb{R}^d,$$

where $G_h := h^{-d}G(\cdot/h)$ for some radially symmetric kernel $G : \mathbb{R}^d \to \mathbb{R}$. The approximation error between $\widehat{\phi}_T$ and $\widetilde{\phi}_{T,h}$ vanishes for suitably chosen bandwidths $h \searrow 0$, while the difference $\widetilde{\phi}_{T,h} - \phi_\rho$ admits a representation similar to (2.3.47), namely

$$\widetilde{\phi}_{T,h}(\lambda) - \phi_\rho(\lambda) = \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} \left(\frac{1}{T} \int_0^T G_h(X_u - y)\mathrm{d}u - \rho(y) \right) \mathrm{d}y, \qquad \lambda \in \mathbb{R}^d.$$

In order to obtain a representation involving an integral of the form given in (2.3.43), the next step is to express $\int_0^T G_h(X_u - y)\mathrm{d}u$ via the *Newtonian potential.* Throughout the sequel, the Newtonian potential is denoted by N, that is,

$$N(x) := \begin{cases} (2\pi)^{-1} \log \|x\|, & d = 2, \\ \frac{1}{(2-d)\mathbb{S}_d} \|x\|^{2-d}, & d \geq 3, \end{cases} \qquad x \in \mathbb{R}^d \setminus \{0\}. \tag{2.3.49}$$

The Newtonian potential is the fundamental solution of the Laplace equation (see (2.3.52) below). For

$$\widetilde{N}_h(x) := G_h * (2N)(x) = 2 \int_{\mathbb{R}^d} N(x-y)G_h(y)\mathrm{d}y, \qquad x \in \mathbb{R}^d, \tag{2.3.50}$$

it thus holds

$$\Delta \widetilde{N}_h = \Delta\left(G_h * (2N)\right) = G_h * 2\Delta N = 2G_h.$$

Furthermore, since $\widetilde{N}_h \in C_b^\infty(\mathbb{R}^d)$, Itô's formula gives for $x \in \mathbb{R}^d$, $T > 0$,

$$\widetilde{N}_h(X_T - x) = \widetilde{N}_h(X_0 - x) + \int_0^T \left(\nabla \widetilde{N}_h(X_u - x)\right)^t \mathrm{d}X_u + \int_0^T G_h(X_u - x)\mathrm{d}u. \tag{2.3.51}$$

This last equation in particular allows to derive a decomposition as in (2.3.45).

To verify (2.3.46), results from classical potential theory prove useful. We now briefly collect some of them. For any $d \geq 2$, the Newtonian potential N is locally integrable over $\mathbb{R}^d$, that is, $\int_K |N(x)|\mathrm{d}x < \infty$ for any compact $K \subset \mathbb{R}^d$. Furthermore,

$$\nabla N(x) = \frac{x}{\mathbb{S}_d \|x\|^d} \quad \text{and} \quad \Delta N(x) = 0 \text{ for } x \in \mathbb{R}^d \setminus \{0\}. \tag{2.3.52}$$

Given any compactly supported $f \in C^2(\mathbb{R}^d)$, the function

$$u(x) := (N * f)(x) = \int_{\mathbb{R}^d} N(x-y)f(y)\mathrm{d}y$$

defines a C^2 function satisfying the Poisson equation $\Delta u = f$ in $\mathbb{R}^d$ (see, e.g., Theorem 2.2.1 in Evans (2010)). For later purposes, we note that, for any $j, k \in \{1, \ldots, d\}$ and $x \in \mathbb{R}^d \setminus \{0\}$,

$$|\partial_j N(x)| \leq C\|x\|^{1-d}, \tag{2.3.53}$$

$$|\partial_{jk}^2 N(x)| \leq C\|x\|^{-d}. \tag{2.3.54}$$

The estimate in (2.3.53) in particular implies the absolute convergence of the integral

$$\int_{B(0,1)} |\partial_j N(x)|\mathrm{d}x \leq C \int_{B(0,1)} \|x\|^{1-d}\mathrm{d}x = C' \int_0^1 r^{1-d} r^{d-1}\mathrm{d}r < \infty.$$

The second derivatives $\partial_{jk}^2 N$, $j, k \in \{1, \ldots, d\}$, are not integrable over the unit ball $B(0,1)$. The following result helps to circumvent this obstacle.

Lemma 2.3.14 (Theorem 1.I.7 in Doob (2001))**.** *Let $G : \mathbb{R}^d \to \mathbb{R}$ be a radially symmetric, Lipschitz-continuous bounded kernel function with compact support, and define $\widetilde{N}_h$ according to (2.3.50). Then, for any $j, k \in \{1, \ldots, d\}$, $x \in \mathbb{R}^d$, $h > 0$, it holds*

$$\partial_j \widetilde{N}_h(x) = 2 \int_{\mathbb{R}^d} \partial_j N(x-y) G_h(y) \mathrm{d}y, \tag{2.3.55}$$

$$\partial^2_{jk} \widetilde{N}_h(x) = 2 \int_{\mathbb{R}^d} \partial^2_{jk} N(x-y) \left(G_h(y) - G_h(x)\right) \mathrm{d}y - 2\delta_{jk} G_h(x). \tag{2.3.56}$$

Proof. Under the given assumptions, the assertions follow immediately from Theorem 1.I.7 (pp. 8–10) in Doob (2001). ☐

Note in particular that (2.3.54) and the Lipschitz property of the kernel G imply that the integrand in (2.3.56) for x in a neighborhood of y is majorized by $\|x-y\|^{1-d}$ such that the integral converges absolutely.

Proof of Lemma 2.3.13. Throughout the proof, $C_1, C_2, \ldots$ denote positive constants. For fixed $j \in \{1, \ldots, d\}$, we aim at a representation

$$\mathrm{i}\lambda^j \left(\widehat{\phi}_T(\lambda) - \phi_\rho(\lambda)\right) = \frac{1}{\sqrt{T}} \left(m_b(\lambda, X^T) - \frac{2}{\sqrt{T}} \int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u} \mathrm{d}W_u^j \right), \qquad \lambda \in \mathbb{R}^d, \tag{2.3.57}$$

where the remainder term $m_b(\cdot,\cdot)$ satisfies

$$\int_{\mathbb{R}^d} \mathbf{E}_b \left| m_b(\lambda, X^T)\ \widetilde{h}_T(\lambda)\right|^2 \mathrm{d}\lambda = o\Big(T^{\frac{d}{2\beta+d}}\Big). \tag{2.3.58}$$

We carry out the idea sketched above and derive (2.3.57) by means of approximation with suitable convolutions (which provide a solution of the Poisson equation) and Itô's formula. To account for the factor $\mathrm{i}\lambda^j$ appearing on the LHS of (2.3.57), the definition of the convolution differs from (2.3.50).

Let $G : \mathbb{R}^d \to \mathbb{R}_+$ be a kernel of order $\beta - 1$ with compact support, satisfying $\int_{\mathbb{R}^d} \|z\|^\beta |G(z)| \mathrm{d}z < \infty$ and $\int_{\mathbb{R}^d} G(z)\mathrm{d}z = 1$. For $h = h_T = c_0 T^{-1/(2(\beta-1)+d)}$, c_0 some positive constant, define $G_h(z) := h^{-d} G(z/h)$ and

$$N_h(x) := \partial_j G_h * (2N)(x) = 2 \int_{\mathbb{R}^d} \partial_j G_h(y)\ N(x-y) \mathrm{d}y, \tag{2.3.59}$$

where N denotes the Newtonian potential introduced in (2.3.49). Write

$$\mathrm{i}\lambda^j \left(\widehat{\phi}_T(\lambda) - \phi_\rho(\lambda)\right) = \mathrm{i}\lambda^j \left(\widehat{\phi}_T(\lambda) - \widetilde{\phi}_{T,h}(\lambda)\right) + \mathrm{i}\lambda^j \left(\widetilde{\phi}_{T,h}(\lambda) - \phi_\rho(\lambda)\right),$$

where

$$\widetilde{\phi}_{T,h}(\lambda) = \frac{1}{T} \int_0^T g_h(\lambda, X_u) \mathrm{d}u,$$

for

$$g_h(\lambda, v) := \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} G_h(v-y) \mathrm{d}y, \qquad \lambda, v \in \mathbb{R}^d,\ h > 0.$$

Integration by parts gives

$$\begin{aligned} &\mathrm{i}\lambda^j g_h(\lambda, v) \\ &\quad = \int_{\mathbb{R}^{d-1}} \mathrm{e}^{\mathrm{i}\lambda^t y} G_h(v-y) \mathrm{d}y^1 \ldots \mathrm{d}y^{j-1} \mathrm{d}y^{j+1} \ldots \mathrm{d}y^d \Big|_{|y^j| \to \infty} - \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} \partial_j \left(G_h(v-y)\right) \mathrm{d}y \\ &\quad = - \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} \partial_j \left(G_h(v-y)\right) \mathrm{d}y. \end{aligned} \tag{2.3.60}$$

Applying Itô's formula to $N_h \in C_0^2(\mathbb{R}^d)$, we obtain

$$N_h(X_T - x) = N_h(X_0 - x) + \int_0^T (\nabla N_h)^t (X_u - x)\mathrm{d}X_u + \frac{1}{2}\int_0^T \Delta N_h(X_u - x)\mathrm{d}u. \tag{2.3.61}$$

It then follows

$$\begin{aligned}
&\sqrt{T}\mathrm{i}\lambda^j\left(\widetilde{\phi}_{T,h}(\lambda) - \phi_\rho(\lambda)\right)\\
&\quad= \frac{1}{\sqrt{T}}\int_0^T \mathrm{i}\lambda^j g_h(\lambda, X_u)\mathrm{d}u - \sqrt{T}\mathrm{i}\lambda^j\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t y}\rho(y)\mathrm{d}y\\
&\quad\overset{(2.3.60)}{=} -\frac{1}{\sqrt{T}}\int_0^T\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t y}\partial_j\left(G_h(X_u - y)\right)\mathrm{d}y\ \mathrm{d}u + \sqrt{T}\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t y}\partial_j\rho(y)\mathrm{d}y\\
&\quad\overset{(2.3.61)}{=} -\frac{1}{\sqrt{T}}\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t y}(N_h(X_T - y) - N_h(X_0 - y))\mathrm{d}y \qquad (\mathbf{I})\\
&\qquad + \frac{1}{\sqrt{T}}\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t y}\int_0^T (\nabla N_h)^t (X_u - y)\mathrm{d}X_u\ \mathrm{d}y + 2\sqrt{T}\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t y}b^j(y)\rho(y)\mathrm{d}y. \qquad (\mathbf{II})
\end{aligned}$$

The treatment of the term in $(\mathbf{II})$ turns out to be the central part of the proof. Lemma 2.3.14 implies that

$$\begin{aligned}
\int_0^T \partial_k N_h(X_u - y)\mathrm{d}X_u^k &= \int_0^T \partial_{kj}^2\widetilde{N}_h(X_u - y)\mathrm{d}X_u^k\\
&= 2\int_0^T\int_{\mathbb{R}^d}\partial_{kj}^2 N(X_u - y - z)\left(G_h(z) - G_h(X_u - y)\right)\mathrm{d}z\ \mathrm{d}X_u^k\\
&\qquad -2\cdot\mathbb{1}\{j = k\}\int_0^T G_h(X_u - y)\mathrm{d}X_u^j.
\end{aligned}$$

Therefore, we may write

$$(\mathbf{II}) = (\mathbf{II}_1) + (\mathbf{II}_2) + (\mathbf{II}_3) - \frac{2}{\sqrt{T}}\zeta_T^j(\lambda),$$

for

$$\begin{aligned}
(\mathbf{II}_1) &:= 2\sqrt{T}\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t y}\left(b^j(y)\rho(y) - \frac{1}{T}\int_0^T G_h(X_u - y)b^j(X_u)\mathrm{d}u\right)\mathrm{d}y,\\
(\mathbf{II}_2) &:= \frac{2}{\sqrt{T}}\left(\int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u}\mathrm{d}W_u^j - \int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t y}\int_0^T G_h(X_u - y)\mathrm{d}W_u^j\ \mathrm{d}y\right),\\
(\mathbf{II}_3) &:= \frac{2}{\sqrt{T}}\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t y}\int_0^T\sum_{k=1}^d\int_{\mathbb{R}^d}\partial_{kj}^2 N(X_u - y - z)\left(G_h(z) - G_h(X_u - y)\right)\mathrm{d}z\ \mathrm{d}X_u^k\ \mathrm{d}y.
\end{aligned}$$

In particular, (2.3.57) holds for

$$m_b(\lambda, X^T) := (\mathbf{I}) + (\mathbf{II}_1) + (\mathbf{II}_2) + (\mathbf{II}_3) + \sqrt{T}\mathrm{i}\lambda^j\left(\widehat{\phi}_T(\lambda) - \widetilde{\phi}_{T,h}(\lambda)\right).$$

We turn to verifying (2.3.58). Note first that the assumption that $|\widetilde{h}_T| \leq 1$, Plancherel's formula and stationarity of X entail that

$$\begin{aligned}
\int_{\mathbb{R}^d}\mathbf{E}_b\left|(\mathbf{I})\ \widetilde{h}_T(\lambda)\right|^2\mathrm{d}\lambda &\leq \frac{1}{T}\int_{\mathbb{R}^d}\mathbf{E}_b\left(N_h(X_T - y) - N_h(X_0 - y)\right)^2\mathrm{d}y\\
&\leq \frac{4}{T}\int_{\mathbb{R}^d}\mathbf{E}_b\left(N_h(X_0 - y)\right)^2\mathrm{d}y.
\end{aligned}$$

Since N_h as defined in (2.3.59) can be rewritten as $N_h = \partial_j(G_h * 2N) = \partial_j\widetilde{N}_h$, Lemma 2.3.14 implies that, for any $y \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbf{E}_b\left(N_h(X_0-y)\right)^2 &= \int_{\mathbb{R}^d}\left(\partial_j\widetilde{N}_h(z-y)\right)^2\rho(z)\mathrm{d}z\\
&\overset{(2.3.55)}{=} 4\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d}\partial_jN(z-y-x)G_h(x)\mathrm{d}x\right)^2\rho(z)\mathrm{d}z\\
&= 4\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d}\partial_jN(w-x)G_h(x)\mathrm{d}x\right)^2\rho(w+y)\mathrm{d}w\\
&\le 4\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d}\partial_jN(w-x)G_h(x)\mathrm{d}x\right)^2\mathrm{e}^{-\|w\|^2}\mathrm{d}w\ \mathrm{e}^{-\|y\|^2}.
\end{aligned}$$

Cauchy–Schwarz and the inequality (2.3.53) entail that

$$\begin{aligned}
\left(\int_{\mathbb{R}^d}\partial_jN(w-x)G_h(x)\mathrm{d}x\right)^2 &\le \int_{\mathbb{R}^d}\left(\partial_jN(w-x)\right)^2\mathrm{d}x\int_{\mathbb{R}^d}G_h^2(x)\mathrm{d}x\\
&\le C_1h^{-d}\int_{\mathbb{R}^d}\|w-x\|^{2-d}\mathrm{d}x.
\end{aligned}$$

Consequently,

$$\mathbf{E}_b\left(N_h(X_0-y)\right)^2 \le 4C_1h^{-d}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|w-x\|^{2-d}\mathrm{d}x\mathrm{e}^{-\|w\|^2}\mathrm{d}w\ \mathrm{e}^{-\|y\|^2},$$

and

$$\begin{aligned}
\int_{\mathbb{R}^d}\mathbf{E}_b\left|(\mathbf{I})\ \widetilde{h}_T(\lambda)\right|^2\mathrm{d}\lambda &\le 4C_2h^{-d}T^{-1}\int_{\mathbb{R}^d}\mathrm{e}^{-\|y\|^2}\mathrm{d}y\\
&\sim T^{-\frac{2(\beta-1)}{2(\beta-1)+d}} = o\Big(T^{\frac{d}{2\beta+d}}\Big).
\end{aligned} \qquad (\mathbf{A})$$

We turn to $(\mathbf{II}_1)$. Letting

$$\widetilde{g}_T(y) := \frac{1}{T}\int_0^T G_h(X_u-y)b^j(X_u)\mathrm{d}u, \qquad T>0,\ y\in\mathbb{R}^d, \tag{2.3.62}$$

the condition that $|\widetilde{h}_T| \le 1$, Plancherel's formula and the usual bias-variance decomposition imply that

$$\begin{aligned}
\int_{\mathbb{R}^d}\mathbf{E}_b\left|(\mathbf{II}_1)\ \widetilde{h}_T(\lambda)\right|^2\mathrm{d}\lambda &\le \int_{\mathbb{R}^d}\mathbf{E}_b\left|(\mathbf{II}_1)\right|^2\mathrm{d}\lambda\\
&= 4T\int_{\mathbb{R}^d}\mathbf{E}_b\left(b^j(y)\rho(y)-\widetilde{g}_T(y)\right)^2\mathrm{d}y\\
&= 4T\left(\int_{\mathbb{R}^d}\left(b^j(y)\rho(y)-\mathbf{E}_b\widetilde{g}_T(y)\right)^2\mathrm{d}y + \int_{\mathbb{R}^d}\mathrm{Var}_b\left(\widetilde{g}_T(y)\right)\mathrm{d}y\right).
\end{aligned}$$

The integrated bias is treated as in the case of independent observations. Using Plancherel's formula, one obtains

$$\begin{aligned}
\int_{\mathbb{R}^d}\left(b^j(y)\rho(y)-\mathbf{E}_b\widetilde{g}_T(y)\right)^2\mathrm{d}y &= (2\pi)^{-d}\int_{\mathbb{R}^d}\left(\phi_{b^j\rho}(\lambda)-\phi_{\mathbf{E}_b\widetilde{g}_T}(\lambda)\right)^2\mathrm{d}\lambda\\
&= (2\pi)^{-d}\int_{\mathbb{R}^d}\left|\phi_{b^j\rho}(\lambda)\right|^2\left(1-\phi_G(h\lambda)\right)^2\mathrm{d}\lambda\\
&\le (2\pi)^{-d}\int_{\mathbb{R}^d}\left|\phi_{b^j\rho}(\lambda)\right|^2\|h\lambda\|^{2\beta}\mathrm{d}\lambda\\
&\le C_3\int_{\mathbb{R}^d}\left(\left|\phi_{\partial_j\rho}(\lambda)-\phi_{\partial_j\rho_0}(\lambda)\right|^2+\left|\phi_{\partial_j\rho_0}(\lambda)\right|^2\right)\|h\lambda\|^{2\beta}\mathrm{d}\lambda\\
&\le C_4h^{2\beta}.
\end{aligned}$$

For the integrated variance term, Lemma 2.3.4 gives

$$\int_{\mathbb{R}^d} \mathrm{Var}_b\left(\widetilde{g}_T(y)\right)\mathrm{d}y \leq C_5 T^{-1} h^{2-d}.$$

Thus,

$$\begin{aligned}\int_{\mathbb{R}^d} \mathbf{E}_b\left|(\mathbf{II}_1)\ \widetilde{h}_T(\lambda)\right|^2 \mathrm{d}\lambda &\leq C_6 T\left(h^{2\beta} + T^{-1}h^{2-d}\right)\\ &\sim T^{\frac{d-2}{2(\beta-1)+d}} = o\Big(T^{\frac{d}{2\beta+d}}\Big).\end{aligned} \tag{B}$$

We turn to the term involving $(\mathbf{II}_2)$. It follows from Cauchy–Schwarz that

$$\int_{\mathbb{R}^d} \mathbf{E}_b\left|(\mathbf{II}_2)\ \widetilde{h}_T(\lambda)\right|^2 \mathrm{d}\lambda \leq \int_{\mathbb{R}^d} \left(\mathbf{E}_b|(\mathbf{II}_2)|^4\right)^{1/2} \left(\mathbf{E}_b|\widetilde{h}_T(\lambda)|^4\right)^{1/2}\mathrm{d}\lambda. \tag{2.3.63}$$

Using the Burkholder–Davis–Gundy and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\mathbf{E}_b|(\mathbf{II}_2)|^4 &= \frac{16}{T^2}\ \mathbf{E}_b\left[\left(\int_0^T \left(\mathrm{e}^{\mathrm{i}\lambda^t X_u} - \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y}\ G_h(X_u-y)\mathrm{d}y\right)\mathrm{d}W_u^j\right)^2\right]\\ &\leq C_7 T^{-2}\ \mathbf{E}_b\left(\int_0^T \left(\mathrm{e}^{\mathrm{i}\lambda^t X_u} - \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y}\ G_h(X_u-y)\mathrm{d}y\right)^2 \mathrm{d}u\right)^2\\ &\leq C_7\ \mathbf{E}_b\left[\left(\mathrm{e}^{\mathrm{i}\lambda^t X_0} - \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y}\ G_h(X_0-y)\mathrm{d}y\right)^4\right].\end{aligned}$$

Arguing as in the proof of (2.3.20), it can be shown that, for any $\lambda, z \in \mathbb{R}^d$,

$$\left|\mathrm{e}^{\mathrm{i}\lambda^t z} - \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} G_h(z-y)\mathrm{d}y\right| \leq C_8\|\lambda\|^\beta h^\beta.$$

Thus,

$$\left(\mathbf{E}_b|(\mathbf{II}_2)|^4\right)^{1/2} \leq C_9\|\lambda\|^{2\beta}h^{2\beta},$$

and plugging this into (2.3.63), we obtain

$$\begin{aligned}\int_{\mathbb{R}^d} \mathbf{E}_b\left|(\mathbf{II}_2)\ \widetilde{h}_T(\lambda)\right|^2 \mathrm{d}\lambda &\leq C_9 \int_{\mathbb{R}^d} \|\lambda\|^{2\beta}\ \sqrt{\mathbf{E}_b|\widetilde{h}_T(\lambda)|^4}\mathrm{d}\lambda\ h^{2\beta}\\ &= O\Big(T^{\frac{d-2}{2(\beta-1)+d}}\Big) = o\Big(T^{\frac{d}{2\beta+d}}\Big).\end{aligned} \tag{C}$$

As concerns $(\mathbf{II}_3)$, note that, using once again the estimate $|\widetilde{h}_T| \leq 1$ and Plancherel's formula,

$$\begin{aligned}&\int_{\mathbb{R}^d} \mathbf{E}_b\left|(\mathbf{II}_3)\ \widetilde{h}_T(\lambda)\right|^2 \mathrm{d}\lambda\\ &\quad\leq \frac{4}{T}\int_{\mathbb{R}^d} \mathbf{E}_b\left(\int_0^T \sum_{k=1}^d \int_{\mathbb{R}^d} \partial_{kj}^2 N(X_u-y-z)\left(G_h(z)-G_h(X_u-y)\right)\mathrm{d}z\ \mathrm{d}X_u^k\right)^2 \mathrm{d}y\\ &\quad\leq 8\int_{\mathbb{R}^d} \mathbf{E}_b\left(\sum_{k=1}^d \int_{\mathbb{R}^d} \partial_{kj}^2 N(X_0-y-z)\left(G_h(z)-G_h(X_0-y)\right)\mathrm{d}z\ b^k(X_0)\right)^2 \mathrm{d}y\\ &\qquad + 8\int_{\mathbb{R}^d} \sum_{k=1}^d \mathbf{E}_b\left(\int_{\mathbb{R}^d} \partial_{kj}^2 N(X_0-y-z)\left(G_h(z)-G_h(X_0-y)\right)\mathrm{d}z\right)^2 \mathrm{d}y.\end{aligned}$$

The first term in the last sum satisfies, for any $y \in \mathbb{R}^d$,

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left(\sum_{k=1}^d \int_{\mathbb{R}^d} \partial_{kj}^2 N(x-y-z) \left(G_h(z) - G_h(x-y)\right) \mathrm{d}z \right)^2 \left(b^k(x)\right)^2 \rho(x) \mathrm{d}x \\
&\quad = \int_{\mathbb{R}^d} \left(\sum_{k=1}^d \int_{\mathbb{R}^d} \partial_{kj}^2 N(w-z) \left(G_h(z) - G_h(w)\right) \mathrm{d}z \right)^2 \left(b^k(w+y)\right)^2 \rho(w+y) \mathrm{d}w \\
&\quad \leq C_{10} \int_{\mathbb{R}^d} (1 + \|w+y\|)^2 \exp\left(-\|w+y\|^2\right) \mathrm{d}w \leq C_{11} \|y\| \mathrm{e}^{-\|y\|^2},
\end{aligned}$$

where the last line follows from (C_2) and (2.2.7). Analogous arguments show that

$$\int_{\mathbb{R}^d} \sum_{k=1}^d \mathbf{E}_b \left(\int_{\mathbb{R}^d} \partial_{kj}^2 N(X_0 - y - z) \left(G_h(z) - G_h(X_0 - y)\right) \mathrm{d}z \right)^2 \mathrm{d}y \leq C_{12} \mathrm{e}^{-\|y\|^2}.$$

Thus,

$$\int_{\mathbb{R}^d} \mathbf{E}_b \big| (\mathbf{II}_3)\ \widetilde{h}_T(\lambda) \big|^2 \mathrm{d}\lambda \leq C_{13} \int_{\mathbb{R}^d} \|y\| \mathrm{e}^{-\|y\|^2} \mathrm{d}y \leq C_{14}. \tag{$\mathbf{D}$}$$

We finally consider the term

$$\sqrt{T} \mathrm{i}\lambda^j \big(\widehat{\phi}_T(\lambda) - \widetilde{\phi}_{T,h}(\lambda)\big) = \frac{1}{\sqrt{T}} \mathrm{i}\lambda^j \left(\int_0^T \left(\mathrm{e}^{\mathrm{i}\lambda^t X_u} - \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} G_h(X_u - y) \mathrm{d}y \right) \mathrm{d}u \right).$$

First, the Cauchy–Schwarz inequality yields

$$\begin{aligned}
&\int_{\mathbb{R}^d} \mathbf{E}_b \big| \widetilde{h}_T(\lambda)\ \mathrm{i}\lambda^j \big(\widehat{\phi}_T(\lambda) - \widetilde{\phi}_{T,h}(\lambda)\big|^2 \mathrm{d}\lambda \\
&\qquad \leq \int_{\mathbb{R}^d} \left(\mathbf{E}_b |\widetilde{h}_T(\lambda)|^4 \right)^{1/2} \left(\mathbf{E}_b \big| \mathrm{i}\lambda^j \big(\widehat{\phi}_T(\lambda) - \widetilde{\phi}_{T,h}(\lambda)\big) \big|^4 \right)^{1/2} \mathrm{d}\lambda.
\end{aligned}$$

Hölder's inequality implies that

$$\begin{aligned}
&\mathbf{E}_b \left[\left| \frac{1}{T} \int_0^T \mathrm{i}\lambda^j \left(\mathrm{e}^{\mathrm{i}\lambda^t X_u} - \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} G_h(X_u - y) \mathrm{d}y \right) \mathrm{d}u \right|^4 \right] \\
&\qquad \leq \mathbf{E}_b \left| \mathrm{i}\lambda^j \left(\mathrm{e}^{\mathrm{i}\lambda^t X_0} - \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} G_h(X_0 - y) \mathrm{d}y \right) \right|^4 \leq C_{15} \|\lambda h\|^{4(\beta-1)},
\end{aligned}$$

where the last estimate follows from arguments similar to those in the derivation of (2.3.20). Summing up,

$$\begin{aligned}
&\int_{\mathbb{R}^d} \mathbf{E}_b \big| \widetilde{h}_T(\lambda)\ \mathrm{i}\lambda^j \big(\widehat{\phi}_T(\lambda) - \widetilde{\phi}_{T,h}(\lambda)\big|^2 \mathrm{d}\lambda \\
&\qquad \leq C_{16} \int_{\mathbb{R}^d} \|\lambda\|^{2\beta} \sqrt{\mathbf{E}_b\big(\widetilde{h}_T^4(\lambda)\big)} \mathrm{d}\lambda\ h^{2\beta} = o\Big(T^{\frac{d}{2\beta+d}}\Big).
\end{aligned}$$

In view of ($\mathbf{A}$), ($\mathbf{B}$), ($\mathbf{C}$) and ($\mathbf{D}$), this last assertion completes the proof of (2.3.58). $\square$

Remark 2.3.15 (Comparison to the one-dimensional case). Consider an ergodic scalar diffusion solving the SDE $\mathrm{d}X_t = b(X_t)\mathrm{d}t + \mathrm{d}W_t$, denote by $\rho^{(1)}$ its invariant density, and let

$$\widehat{\phi}_T^{(1)}(\lambda) := \frac{1}{T} \int_0^T \mathrm{e}^{\mathrm{i}\lambda X_u} \mathrm{d}u, \qquad \zeta_T^{(1)}(\lambda) := \frac{1}{\sqrt{T}} \int_0^T \mathrm{e}^{\mathrm{i}\lambda X_u} \mathrm{d}W_u, \qquad \lambda \in \mathbb{R}.$$

Lemma 1 in Dalalyan (2005) states that, for any $\lambda \in \mathbb{R}$,

$$\lambda\Big(\widehat{\phi}_T^{(1)}(\lambda) - \phi_{\rho^{(1)}}(\lambda)\Big) = \frac{1}{\sqrt{T}}\Big(2\mathrm{i}\zeta_T^{(1)}(\lambda) + m_b^{(1)}(\lambda, X^T)\Big), \tag{2.3.64}$$

for some measurable function $m_b^{(1)}(\lambda, X^T)$ satisfying

$$\int_{\mathbb{R}} \mathbf{E}_b \big|m_b^{(1)}(\lambda, X^T)\big|^2 \mathrm{d}\lambda < C. \tag{2.3.65}$$

The above estimate is obviously stronger than the upper bound in (2.3.46). The crux about (2.3.65) however is that it allows to conclude that, under appropriate conditions on the random function $\widehat{h}_T$,

$$\begin{aligned}\frac{1}{2\pi}\ \mathbf{E}_b \int_{\mathbb{R}} |\lambda|^2 \left|(\widehat{\phi}_T^{(1)}(\lambda) - \phi_{\rho^{(1)}}(\lambda))\widehat{h}_T(\lambda) - \phi_{\rho^{(1)}}(\lambda)(1 - \widehat{h}_T(\lambda))\right|^2 \mathrm{d}\lambda \\ \simeq \frac{1}{2\pi}\ \mathbf{E}_b \int_{\mathbb{R}} \left|2\mathrm{i}T^{-1/2}\zeta_T^{(1)}(\lambda)\ \widehat{h}_T(\lambda) + \lambda\phi_{\rho^{(1)}}(\lambda)\left(1 - \widehat{h}_T(\lambda)\right)\right|^2 \mathrm{d}\lambda,\end{aligned}$$

where $a_T \simeq b_T$ means that the ratio $a_T/b_T \to_{T\to\infty} 1$, uniformly in all parameters which are involved in the definitions of these functions. For the purpose of deriving an equivalent relation in the multidimensional framework, the assertion proven in Lemma 2.3.13 suffices.

Similarly to the non-adaptive set-up (cf. (2.3.11)), the optimal estimator is defined via some function

$$h : x \mapsto h_\beta(\alpha x) := \Big(1 - \|\alpha x\|^\beta\Big)_+, \qquad x \in \mathbb{R}^d.$$

For $h_\beta(\alpha\ \cdot)$ thus defined, it obviously holds $0 \le h_\beta(\alpha x) \le 1$. Furthermore, for any $\beta > d/2$, there exist positive constants K_1 and K_2, independent of β, such that

$$K_1\alpha^{-d} \le \|h_\beta(\alpha\cdot)\|^2_{L^2(\mathbb{R}^d)} \le K_2\alpha^{-d}. \tag{2.3.66}$$

Indeed, it holds

$$\begin{aligned}\|h_\beta(\alpha\cdot)\|^2_{L^2(\mathbb{R}^d)} &= \int_{\mathbb{R}^d} \Big(1 - \|\alpha\lambda\|^\beta\Big)_+^2 \mathrm{d}\lambda \\ &= \alpha^{-d} \int_{\|\omega\|\le 1} \Big(1 - \|\omega\|^\beta\Big)^2 \mathrm{d}\omega \\ &= \alpha^{-d}\ \mathbb{S}_d \int_0^1 \Big(1 - u^\beta\Big)^2 u^{d-1}\mathrm{d}u \\ &= \alpha^{-d}\ \mathbb{S}_d \left(\frac{1}{d} - \frac{2}{\beta + d} + \frac{1}{2\beta + d}\right),\end{aligned} \tag{2.3.67}$$

and, for any $\beta > d/2$,

$$\frac{1}{6d} \le \frac{1}{d} - \frac{2}{\beta + d} + \frac{1}{2\beta + d} \le \frac{1}{d}.$$

Since

$$\begin{aligned}\int_{\mathbb{R}^d} \|\lambda\|^{2\beta} h_\beta^2(\alpha\lambda)\mathrm{d}\lambda &= \int_{\mathbb{R}^d} \|\lambda\|^{2\beta} \Big(1 - \|\alpha\lambda\|^\beta\Big)_+^2 \mathrm{d}\lambda \\ &= \alpha^{-2\beta-d} \int_{\|\omega\|\le 1} \|\omega\|^{2\beta} \Big(1 - \|\omega\|^\beta\Big)^2 \mathrm{d}\omega \\ &= \alpha^{-2\beta-d}\ \mathbb{S}_d \int_0^1 u^{2\beta} \Big(1 - u^\beta\Big)^2 u^{d-1}\mathrm{d}u \\ &= \alpha^{-2\beta-d}\ \mathbb{S}_d \left(\frac{1}{2\beta + d} - \frac{2}{3\beta + d} + \frac{1}{4\beta + d}\right),\end{aligned}$$

there also exists some positive constant K_3, not depending on β, such that

$$\int_{\mathbb{R}^d} \|\lambda\|^{2\beta} h_\beta^2(\alpha\lambda)\mathrm{d}\lambda \le K_3 \alpha^{-2\beta-d}. \tag{2.3.68}$$

Let us summarize the adaptive procedure for estimating $\nabla\rho$. Consider the set

$$\mathfrak{H}_T := \left\{ h : x \mapsto h_\beta(\alpha x) := \left(1 - \|\alpha x\|^\beta\right)_+ : \alpha \in \left[T^{-\gamma_T}, (\log T)^{-1}\right], \beta > d/2 \right\}. \tag{2.3.69}$$

Direct numerical minimization over the entire set $\mathfrak{H}_T$ is time consuming such that it is convenient to restrict attention to a finite subset $\mathfrak{H}_T^N$, approximating well the weights in $\mathfrak{H}_T$. For this purpose, we require another auxiliary result which is due to Cavalier et al. (2002). In this article, the authors consider a sequence space model of statistical linear inverse problems, and the aim is to estimate a function f from indirect noisy observations. They provide a nonasymptotic, asymptotically exact oracle inequality which can be applied to obtain sharp minimax adaptive results. Similarly to our framework, the inequality is then used to prove that minimax adaptation on ellipsoids (even in the multivariate anisotropic case) can be achieved by minimization of an unbiased risk estimator without any loss of efficiency with respect to optimal nonadaptive procedures. The arguments of the proof of Lemma 5 in Cavalier et al. (2002) imply that, for any $h \in \mathfrak{H}_T$, there exists some $\widetilde{h} \in \mathfrak{H}_T^N$ such that

$$\widehat{\Delta}_T(\widetilde{h}) \le (1 + \delta_N)\ \widehat{\Delta}_T(h),$$

for δ_N not depending on $h, \widetilde{h}$ and $b \in \Sigma_\delta$ with $\delta_N \to_{N\to\infty} 0$. (For the sake of completeness, we include a rephrased version of their result in Section 2.4.2.)

The adaptive procedure for estimating $\nabla\rho$ starts by considering discrete values $\alpha_i, \beta_j \in \mathfrak{H}_T^N$ and selecting

$$\widehat{h}_T = \arg\min_{h \in \mathfrak{H}_T^N} \widehat{\Delta}_T(h). \tag{2.3.70}$$

Denote the values of α_i and β_j corresponding to $\widehat{h}_T$ by $\widehat{\alpha}_T$ and $\widehat{\beta}_T$, respectively. The kernel $\widehat{K_T}$ is chosen such that

$$\phi_{\widehat{K}_T}(\lambda) = \widehat{h}_T(\lambda) = \left(1 - \|\widehat{\alpha}_T\lambda\|^{\widehat{\beta}_T}\right)_+, \tag{2.3.71}$$

and the adaptive estimator of $\nabla\rho$ finally is defined as

$$\widehat{g}_T(x) := \frac{1}{T}\int_0^T \nabla\widehat{K_T}(x - X_u)\mathrm{d}u. \tag{2.3.72}$$

Theorem 2.3.16. *Assume that b_0 satisfies, for any $j \in \{1, \ldots, d\}$, the conditions* $(\mathrm{C_0}^j)$–(C_4). *Then, for any $\beta > d/2$ and the set Σ_δ introduced in (2.3.31), it holds*

$$\limsup_{T\to\infty} \sup_{b \in \Sigma_\delta} T^{\frac{2\beta}{2\beta+d}}\ \mathbf{E}_b \int_{\mathbb{R}^d} \left\|\widehat{g}_T(x) - \nabla\rho(x)\right\|^2 \mathrm{d}x \le 4\ P(\mathbf{Id}_{d\times d}, \beta, L).$$

The principle idea for implementing a sharp adaptive *estimator of the drift vector* is the same as in the non-adaptive setup. In order to obtain a data-driven correspondent of (2.3.34), the estimators g_T^*, ρ_T^* and the sequence c_T^* need to be replaced by data-dependent counterparts. Since it does not contribute to the understanding of the drift estimation problem, we do not deepen the question of how to estimate ρ and its smoothness in order to obtain a data-driven truncation sequence c_T. The results on adaptive estimation of the derivatives indicate that existing adaptive methods can be suitably modified without any substantial problems under the given assumptions. Let us emphasize that we do *not* need an estimator of ρ which is efficient up to the constant to define an exact estimator of the drift. Note further that the problem of bandwidth selection for invariant density estimators does not arise in the one-dimensional setting, since the invariant density can be estimated with the rate $\sqrt{T}$ by kernel estimators with a "universal" bandwidth $h_T \sim T^{-1/2}$.

As concerns the definition of the term $c_T(\cdot)$, the basic idea is to define it in such a way that it converges faster than the numerator (which achieves the classical nonparametric rate $T^{-\beta/(2\beta+d)}$) but not faster than the invariant density estimator $\widehat{\rho}_T$ and that it is, in addition, bounded away from zero.

2.4. Proofs for Section 2.3

2.4.1. Proof of lower bounds

Proof of Theorem 2.3.2. **Preliminaries: A modified version of van Trees' inequality.** There exist different versions of the van Trees inequality; cf. Section 1.2.1. Below we state a multivariate extension due to Gill and Levit (1995), tailored to our diffusion setting. It will be applied for estimating a real-valued function of a vector parameter.

The multidimensional diffusion framework is as follows: Define the drift vector $b_\theta(x)$ as a function of $x \in \mathbb{R}^d$ and $\theta \in \Theta$, Θ some D-dimensional rectangle, $D \geq 1$. For any value of the unknown parameter θ, the functions $b_\theta(\cdot)$ are such that there exists an ergodic solution of the SDE

$$\mathrm{d}X_t = b_\theta(X_t)\mathrm{d}t + \sigma \, \mathrm{d}W_t \tag{2.4.1}$$

which admits a Lebesgue continuous invariant measure with density $\rho_{b_\theta} =: \rho_\theta$.

Assume that some prior distribution of the form $\mathrm{d}\Lambda(\theta) = p(\theta)\mathrm{d}\lambda\!\!\lambda$ is defined on Θ, denote by $\mathbf{P}_\theta$ ($\mathbf{E}_\theta$) the distribution (expectation) of a continuous record $X^T = (X_t)_{0\leq t\leq T}$ of observations of X solving (2.4.1), and denote by $\mathbb{E}$ expectation with respect to $\mathrm{d}\mathbf{P}_\theta(X^T) \times \mathrm{d}\Lambda(\theta)$. Suppose we are interested in estimating some $\mathbb{R}^s$-valued functional $\psi(\theta)$, $s \in \mathbb{N}$. Later on, we will specify to the case $s = 1$. Partial derivatives with respect to the components of θ are arranged in rows such that $\partial\psi/\partial\theta =: \partial_\theta\psi$ is an $s \times D$ matrix. In the formulation of the lemma, a "smooth" real-valued function $r(\theta)$ is assumed to be absolutely continuous in θ^j for almost all values of the other components of θ, and its partial derivatives $\partial_{\theta^j} r$ are supposed to be measurable in θ.

Lemma 2.4.1 (van Trees inequality; cf. Theorem 1 in Gill and Levit (1995))**.** *In the above framework, assume that the following conditions are satisfied:*

(a) *The Radon–Nikodym density $L(\theta, \theta_0; X^T)$ of $\mathbf{P}_\theta$ with respect to $\mathbf{P}_{\theta_0}$, $\theta_0 \in \Theta$ some fixed value, is smooth;*

(b) *the Fisher information matrix*

$$I(\theta) := \mathbf{E}_\theta\left(\left(\partial_\theta \log L(\theta, \theta_0; X^T)\right)^t \left(\partial_\theta \log L(\theta, \theta_0; X^T)\right)\right), \qquad \theta \in \Theta, \tag{2.4.2}$$

exists and $\sqrt{\operatorname{diag}(I(\theta))}$ is locally integrable in θ;

(c) *the prior density p is smooth, positive on the interior of Θ and zero on its boundary;*

(d) *the components of the functional $\psi(\theta)$ are smooth.*

Under the above conditions, it holds for any estimator $\widehat{\psi}_T = \widehat{\psi}_T(X^T)$ and any constant $s \times d$ matrix C,

$$\mathbb{E}\|\widehat{\psi}_T - \psi(\theta)\|^2 = \int_\Theta \mathbf{E}_\theta\|\widehat{\psi}_T - \psi(\theta)\|^2 p(\theta)\mathrm{d}\theta \geq \frac{\left(\int_\Theta \mathrm{tr}\left(C(\partial_\theta\psi(\theta))^t\right)p(\theta)\mathrm{d}\theta\right)^2}{\int_\Theta \mathrm{tr}\left(CI(\theta)C^t\right)p(\theta)\mathrm{d}\theta + J(p)}, \tag{2.4.3}$$

where $\mathrm{tr}(a) = \sum_{k=1}^d a_{kk}$ *denotes the trace of the matrix a and*

$$J(p) = \int_\Theta \mathrm{tr}\left(\sum_{i,k,l} \partial_{\theta^k}\left(C_{ik}p(\theta)\right)\partial_{\theta^l}\left(C_{il}p(\theta)\right)\right)\frac{1}{p(\theta)}\,\mathrm{d}\theta. \tag{2.4.4}$$

Proof. We include the short and elementary proof which is analogue to the proof of Theorem 1 in Gill and Levit (1995). Let $U := \widehat{\psi}_T - \psi(\theta)$ and

$$V_i := \sum_{j=1}^s C_{ij}\,\frac{\partial_{\theta^j}\left(L(\theta,\theta_0;X^T)p(\theta)\right)}{L(\theta,\theta_0;X^T)p(\theta)}, \qquad i = 1,\dots,d.$$

Integrating by parts, we obtain

$$\begin{aligned}\mathbb{E}(U^tV) &= \int_{\mathbb{R}^d}\int_\Theta \sum_{i=1}^d (\widehat{\psi}_T^i - \psi^i(\theta))\ \sum_{k=1}^s C_{ik}\partial_{\theta^k}(L(\theta,\theta_0;X^T)p(\theta))\mathrm{d}\theta\ \mathrm{d}x\\ &= \int_\Theta\int_{\mathbb{R}^d}\sum_{i=1}^d\sum_{k=1}^s \partial_{\theta^k}\psi^i(\theta)\ C_{ik}L(\theta,\theta_0;X^T)p(\theta)\mathrm{d}x\ \mathrm{d}\theta = \int_\Theta \mathrm{tr}\left(C(\partial_\theta\psi)^t\right)p(\theta)\mathrm{d}\theta.\end{aligned}$$

Furthermore, since $\mathbf{E}_\theta\partial_\theta\left(\log L(\theta,\theta_0;X^T)\right) = 0$,

$$\begin{aligned}\mathbb{E}(V^tV) &= \mathbb{E}\sum_{i,j=1}^d\sum_{k,l=1}^s \frac{C_{ik}(p\ \partial_{\theta^k}L + L\ \partial_{\theta^k}p)}{Lp} \times \frac{C_{jl}(p\ \partial_{\theta^l}L + L\ \partial_{\theta^l}p)}{Lp}\\ &= \mathbb{E}\sum_{i,j=1}^d\sum_{k,l=1}^s C_{ik}\left(\partial_{\theta^k}\log L + \partial_{\theta^k}\log p\right) \times C_{jl}\left(\partial_{\theta^l}\log L + \partial_{\theta^l}\log p\right)\\ &= \int_\Theta \mathrm{tr}\left(CI(\theta)C^t\right)p(\theta)\mathrm{d}\theta = J(p).\end{aligned}$$

Finally, by Cauchy–Schwarz,

$$\mathbb{E}(U^tU) = \mathbb{E}\|\widehat{\psi}_T - \psi(\theta)\|^2 \geq \frac{\left(\mathbb{E}(U^tV)\right)^2}{\mathbb{E}(V^tV)}.$$

□

Outline of the proof. We proceed analogously to the proof of the lower bound for estimating the derivative of the invariant density of a scalar ergodic diffusion in the local minimax setting in Theorem 1 in Dalalyan and Kutoyants (2002). The scheme of the proof is originally due to Golubev and Levit (1996). In our framework, the essential steps are as follows:

- Restriction to the problem of estimating $\mathrm{div}(a_j\rho)$, $j \in \{1,\dots,d\}$ fixed, for b belonging to some properly chosen parametric family essentially concentrated on the set Σ_δ^j defined in (2.3.31), with dimension increasing to infinity as $T \to \infty$;
- minorization of the minimax risk by the Bayesian risk with respect to any prior distribution plus an asymptotically negligible term;

- determination of the least favorable prior and evaluation of the corresponding minimax risk.

Main part of the proof. The first step is to approximate the infinite-dimensional classes Σ_δ^j of components of the drift of unbounded support by appropriately chosen finite-dimensional subclasses with bounded support. To do so, fix some interval $[-A, A]$, with $A = A_T = \log(T+1)$, and define

$$a_r := 2rAT^{-\gamma}, \quad r \in \mathbb{Z}, \qquad \text{where } \gamma := \frac{1}{2\beta + d}.$$

Let

$$M = M_T := \max\left\{r \in \mathbb{Z} : \left[a_r - AT^{-\gamma},\ a_r + AT^{-\gamma}\right] \subset [-A, A]\right\},$$

and subdivide the interval $[-A, A]$ into closed intervals by considering the partition points

$$\begin{aligned} a_{-M} - AT^{-\gamma} < a_{-M} + AT^{-\gamma} = a_{-M+1} - AT^{-\gamma} < \cdots \\ \cdots < a_{M-1} + AT^{-\gamma} = a_M - AT^{-\gamma} < a_M + AT^{-\gamma}. \end{aligned}$$

Given any multi-index $m = (m_1, \ldots, m_d)$ with components

$$m_j \in \{-M, -M+1, \ldots, M-1, M\}, \quad j = 1, \ldots, d,$$

denote by $\mathbb{A}_m$ the sub-cube centered at the vector $a_m := (a_{m_1}, \ldots, a_{m_d})^t$, that is,

$$\mathbb{A}_m := \times_{j=1}^d \left[a_{m_j} - AT^{-\gamma},\ a_{m_j} + AT^{-\gamma}\right].$$

Denote by $\mathcal{P}$ the set of all sub-cubes $\mathbb{A}_m = \mathbb{A}_{m_1,\ldots,m_d}$, and note that

$$\mathcal{P} \subset \times_{j=1}^d [-A, A] =: \mathbb{A}.$$

(I) The parametric subfamily and its properties. Let $\varphi_\ell(x) := \exp(\pi \mathrm{i}\langle x, \ell\rangle)$, for $\ell = (\ell_1, \ldots, \ell_d) \in \mathbb{N}^d$, be the trigonometric basis on $[-1, 1]^d$, and consider the following scaled and shifted versions of the trigonometric basis functions,

$$\phi_{\ell,m}(x) = \sqrt{T^d \gamma A^{-d}}\ \varphi_\ell\Big(T^\gamma A^{-1}(x - a_m)\Big) \prod_{j=1}^d U\Big(A - |x_j - a_{m_j}|T^\gamma\Big),$$

where the function U is $(\beta+1)$-times differentiable, increasing and satisfies $U(x) = 0$ for $x \leq 0$ and $U(x) = 1$ for $x \geq 1$.

Thus, $\prod_{j=1}^d U(A - |x_j - a_{m_j}|T^\gamma)$ is a smooth approximation of $\mathbb{1}\{x \in \mathbb{A}_m\}$. For fixed positive integers M and R which are to be specified later, consider the parametrization

$$\rho_\theta(x) := \rho_0(x) \exp\left(\sum_{|m|_\infty \leq M} \sqrt{\frac{(2AT^{-\gamma})^d}{\rho_0(a_m)}} \sum_{1 \leq |\ell|_\infty \leq R} \int_0^1 \phi_{\ell,m}(tx)\theta_{\ell,m}^t x \mathrm{d}t - n(\theta)\right),$$

where

- ρ_0 is the invariant density associated with the diffusion process solution of the SDE $\mathrm{d}Y_t = b_0(Y_t)\mathrm{d}t + \sigma\ \mathrm{d}W_t$ and, in addition, satisfying $b_0 = a\nabla \log \rho_0$,
- the $\mathbb{R}^d$-valued parameters $\theta_{\ell,m}$ are to be specified later, and
- $n(\theta)$ is a normalizing factor.

Let

$$b_\theta(x) := b_0(x) + ab_\theta^-(x), \tag{2.4.5}$$

where, for $l \in \{1, \dots, d\}$,

$$\begin{aligned}(b_\theta^-)^l(x) := \sum_{|m|_\infty \le M} \sqrt{\frac{(2AT^{-\gamma})^d}{\rho_0(a_m)}} \\ \times \sum_{1 \le |\ell|_\infty \le R} \theta_{\ell,m}^l \left(\phi_{\ell,m}(x) - \int_{\mathbb{R}^d} \phi_{\ell,m}(y) \mathrm{d}y \right) \cdot \mathbb{1}\{l = j\},\end{aligned} \tag{2.4.6}$$

and define

$$g_\theta^l(x) := 2 \sum_{k=1}^d a_{jk} \partial_k \rho_\theta(x) \cdot \mathbb{1}\{l = j\}. \tag{2.4.7}$$

The functions b_θ defined in (2.4.5) are β-times continuously differentiable and coincide with b_0 outside of the cube $\mathbb{A}$. In addition, they satisfy conditions $(\mathrm{C_0}^j)$–(C_2). This implies in particular that the SDE (2.4.1) associated with b_θ admits a strong solution which is ergodic with invariant density ρ_θ.

Define the parametric space Γ_T^j as containing all finite $\mathbb{R}^d$-valued sequences

$$(\theta_{\ell,m})_{1 \le |\ell|_\infty \le R,\ |m|_\infty \le M}$$

such that, for any multi-indices ℓ and m,

$$|\theta_{\ell,m}^j| \le G \sqrt{\varsigma_\ell^j(\varepsilon)},$$

where

$$\varsigma_\ell^j(\varepsilon) = \varsigma_{\ell,T}^j(\varepsilon) := \frac{1}{(2A)^d T^{2\beta\gamma}} \left(\left(\frac{\alpha^j(1-\varepsilon)}{\|\ell\|} \right)^\beta - 1 \right)_+,$$

for

$$\alpha^j := 2A \left(\frac{L(\beta + d)(2\beta + d)}{\beta a_{jj} \mathbb{S}_d (2\pi)^{2\beta}} \right)^{1/(2\beta+d)},$$

$\varsigma_0^j := 0$ and $R = R^j := \lceil \alpha^j \rceil$. The positive numbers G, ε are to be defined later.

The following lemma describes properties of the parametrized densities ρ_θ. In particular, it gives an upper bound on the Fisher information associated with the likelihood ratio $L(\theta, \theta_0; X^T)$, for b_θ defined in (2.4.5). Denote

$$I_{\ell,m}(\theta) := \mathbf{E}_\theta \left(\left(\partial_{\theta_{\ell,m}} \log L(\theta, \theta_0; X^T) \right)^t \left(\partial_{\theta_{\ell,m}} \log L(\theta, \theta_0; X^T) \right) \right), \qquad \theta \in \Gamma_T^j.$$

Plugging in the explicit formula for the likelihood ratio given in (A.1.3), we obtain

$$\begin{aligned}&\mathbf{E}_\theta \left[\left(\partial_{\theta_{\ell,m}^j} \log L(\theta, \theta_0; X^T) \right)^2 \right] \\ &\quad = \mathbf{E}_\theta \Bigg[\Bigg(\partial_{\theta_{\ell,m}^j} \log \rho_\theta(X_0) + \partial_{\theta_{\ell,m}^j} \left(\int_0^T b_\theta(X_u)^t a^{-1} \mathrm{d}X_u \right) \\ &\qquad - \frac{1}{2} \partial_{\theta_{\ell,m}^j} \int_0^T \left(b_\theta(X_u)^t a^{-1} b_\theta(X_u) \right) \mathrm{d}u \Bigg)^2 \Bigg].\end{aligned} \tag{2.4.8}$$

The dependence on the parameter θ in the definition of b_θ is linear such that the differential operator $\partial_{\theta^k_{\ell,m}}$ and the stochastic integral with respect to X in (2.4.8) may be interchanged. Consequently, for any $k \in \{1,\dots,d\}$,

$$\begin{aligned}
&\partial_{\theta^k_{\ell,m}}\left(\int_0^T b_\theta(X_u)^t a^{-1}\mathrm{d}X_u\right) - \frac{1}{2}\int_0^T \partial_{\theta^k_{\ell,m}}\left(b_\theta(X_u)^t a^{-1} b_\theta(X_u)\right)\mathrm{d}u\\
&\quad= \int_0^T \partial_{\theta^k_{\ell,m}}\left(b_\theta(X_u)^t a^{-1}\right)\left(b_\theta(X_u)\mathrm{d}u + \sigma\,\mathrm{d}W_u\right)\\
&\qquad -\frac{1}{2}\int_0^T \left(\partial_{\theta^k_{\ell,m}}\left(b_\theta(X_u)^t\right) a^{-1} b_\theta(X_u) + b_\theta(X_u)^t a^{-1}\partial_{\theta^k_{\ell,m}} b_\theta(X_u)\right)\mathrm{d}u\\
&\quad= \frac{1}{2}\int_0^T \partial_{\theta^k_{\ell,m}}\left(b_\theta(X_u)^t\right) a^{-1} b_\theta(X_u)\mathrm{d}u + \int_0^T \partial_{\theta^k_{\ell,m}}\left(b_\theta(X_u)^t\right)\left(\sigma^{-1}\right)^t \mathrm{d}W_u\\
&\qquad -\frac{1}{2}\int_0^T b_\theta(X_u)^t a^{-1}\partial_{\theta^k_{\ell,m}} b_\theta(X_u)\mathrm{d}u\\
&\quad\overset{(2.4.5)}{=} \frac{1}{2}\int_0^T \partial_{\theta^k_{\ell,m}}\left(b_\theta^-(X_u)^t\right) b_\theta(X_u)\mathrm{d}u + \int_0^T \partial_{\theta^k_{\ell,m}}\left(b_\theta^-(X_u)^t\right)\sigma\,\mathrm{d}W_u\\
&\qquad -\frac{1}{2}\int_0^T b_\theta(X_u)^t \partial_{\theta^k_{\ell,m}} b_\theta^-(X_u)\mathrm{d}u\\
&\quad= \int_0^T \partial_{\theta^k_{\ell,m}}\left(b_\theta^-(X_u)\right)^t \sigma\,\mathrm{d}W_u.
\end{aligned}$$

Denote

$$I^j_{\ell,m}(\theta) := \mathbf{E}_\theta\Big(\partial_{\theta^j_{\ell,m}}\log\rho_\theta(X_0) + \int_0^T \partial_{\theta^j_{\ell,m}}\left(b_\theta^-(X_u)\right)^t \sigma\,\mathrm{d}W_u\Big)^2.$$

Lemma 2.4.2. (a) *For any $x \in \mathbb{A}_m$, $\rho_0(x) = \rho_0(a_m)(1+o_T(1))$, and, for any $x \in \mathbb{R}^d$ and $\theta \in \Gamma^j_T$, $\rho_\theta(x) = \rho_0(x)(1+o_T(1))$, where $o_T(1)$ is uniform in x and θ.*

(b) *For any $\theta_{\ell,m} \in \Gamma^j_T$, it holds*

$$I^j_{\ell,m}(\theta) \le (2A)^d T^{2\beta\gamma} a_{jj}\ (1+o_T(1)), \tag{2.4.9}$$

where $o_T(1)$ again is uniform in θ.

Proof. Taking into account that $b_0 = a\nabla\log\rho_0$, the first assertion in (a) follows immediately from the continuity (and thus boundedness) of b_0 on the subcube $\mathbb{A}_m$. The dimension of the space Γ^j_T increases with speed $\left(A_T T^\gamma\right)^d$ as $T \to \infty$ such that only $CT^\gamma A_T$ components of the parameter vector $\theta \in \Gamma^j_T$ do not vanish. The second assertion in (a) thus holds since each of the non-vanishing components is smaller than $CT^{-\beta\gamma}A_T^\beta$.

We proceed with proving (b). The triangular inequality implies that

$$\sqrt{I^j_{\ell,m}(\theta)} \le \left(\mathbf{E}_\theta\Big(\partial_{\theta^j_{\ell,m}}\log\rho_\theta(X_0)\Big)^2\right)^{1/2} + \left(\mathbf{E}_\theta\Big(\int_0^T \partial_{\theta^j_{\ell,m}}\left(b_\theta^-(X_u)\right)^t\sigma\,\mathrm{d}W_u\Big)^2\right)^{1/2}. \tag{2.4.10}$$

Plugging in the definition of $b_\theta^-(\cdot)$ and using Fubini's theorem and stationarity of X, one obtains

$$\begin{aligned}
\mathbf{E}_\theta\Big(\int_0^T \partial_{\theta^j_{\ell,m}}\left(b_\theta^-(X_u)\right)^t\sigma\,\mathrm{d}W_u\Big)^2 &= \frac{(2AT^{-\gamma})^d}{\rho_0(a_m)}\int_0^T a_{jj}\,\mathbf{E}_\theta\left(\phi^2_{\ell,m}(X_u)\right)\mathrm{d}u\\
&= \frac{(2AT^{-\gamma})^d}{\rho_0(a_m)}\,T a_{jj}\int_{\mathbb{A}_m}\phi^2_{\ell,m}(x)\rho_\theta(x)\mathrm{d}x\\
&= (2A)^d T^{2\beta\gamma} a_{jj}\ (1+o_T(1)).
\end{aligned}$$

The definition of ρ_θ implies that there exists some positive constant C such that

$$\partial_{\theta^j_{\ell,m}}(\log \rho_\theta(x)) \leq C.$$

Consequently, (2.4.10) gives

$$\sqrt{I^j_{\ell,m}(\theta)} \leq C + \Big((2A)^d T^{2\beta\gamma} a_{jj}(1 + o_T(1))\Big)^{1/2},$$

and, since $T^{2\beta\gamma} \to \infty$ and for large enough T, (2.4.9) follows. □

(II) Bayesian framework. Define

$$\mathcal{W}^j_T := \left\{\widehat{g}_T : \mathbf{E}_{b_0} \int_{\mathbb{R}^d} |\widehat{g}_T(x) - \operatorname{div}(a_j\rho_0)(x)|^2 \,\mathrm{d}x < 1\right\}.$$

Theorem 2.3.8 guarantees that there exists an asymptotically efficient estimator $(\overline{g}^j_T)^*$ of $\operatorname{div}(a_j\rho_0)$ such that $\mathcal{W}^j_T$ is not empty. For proving the lower bound, it suffices to restrict attention to the set $\mathcal{W}^j_T$.

The minimax risk is bounded from below by the Bayesian risk. Given any probability distribution Λ with density p on Γ^j_T, denote by $\mathcal{R}_j(\Lambda)$ the Bayes risk with respect to the prior distribution Λ for the problem of estimating $\operatorname{div}(a_j\rho_\theta)$ on the parameter set Γ^j_T, that is,

$$\begin{aligned}\mathcal{R}_j(\Lambda) &:= \inf_{\widehat{g}_T \in \mathcal{W}^j_T} \mathbb{E} \int_{\mathbb{R}^d} |\widehat{g}_T(x) - \operatorname{div}(a_j\rho_\theta)(x)|^2 \,\mathrm{d}x \\ &= \inf_{\widehat{g}_T \in \mathcal{W}^j_T} \int_{\Gamma^j_T} \int_{\mathbb{R}^d} \mathbf{E}_b\, |\widehat{g}_T(x) - \operatorname{div}(a_j\rho_\theta)(x)|^2 \,\mathrm{d}x \,\mathrm{d}\Lambda(\theta).\end{aligned}$$

For any prior Λ on Γ^j_T, it holds

$$\begin{aligned}&\inf_{\widehat{g}_T \in \mathcal{W}^j_T} \sup_{b \in \Sigma^j_\delta} \mathbf{E}_b \int_{\mathbb{R}^d} |\widehat{g}_T(x) - \operatorname{div}(a_j\rho)(x)|^2 \,\mathrm{d}x \\ &\quad\geq \inf_{\widehat{g}_T \in \mathcal{W}^j_T} \sup_{b \in \Sigma_\delta \cap \Gamma^j_T} \mathbf{E}_b \int_{\mathbb{R}^d} |\widehat{g}_T(x) - \operatorname{div}(a_j\rho)(x)|^2 \,\mathrm{d}x \\ &\quad\geq \inf_{\widehat{g}_T \in \mathcal{W}^j_T} \int_{\Gamma^j_T} \mathbf{E}_b \int_{\mathbb{R}^d} |\widehat{g}_T(x) - \operatorname{div}(a_j\rho_\theta)(x)|^2 \,\mathrm{d}x \,\mathrm{d}\Lambda(\theta) - \mathcal{R}^{\mathrm{c}}_j(\Lambda) \\ &\quad= \mathcal{R}_j(\Lambda) - \mathcal{R}^{\mathrm{c}}_j(\Lambda),\end{aligned} \tag{2.4.11}$$

where

$$\mathcal{R}^{\mathrm{c}}_j(\Lambda) := \sup_{\widehat{g}_T \in \mathcal{W}^j_T} \int_{\Gamma^j_T \setminus \Sigma_\delta} \mathbf{E}_b \int_{\mathbb{R}^d} |\widehat{g}_T(x) - \operatorname{div}(a_j\rho_\theta)(x)|^2 \,\mathrm{d}x \,\mathrm{d}\Lambda(\theta). \tag{2.4.12}$$

(III) Reduction to functional estimation. As in Golubev and Levit (1996) and Dalalyan and Kutoyants (2002), the proof of the lower bound is now further reduced to that of estimating a sequence of linear functionals. For bounding the Bayesian risk $\mathcal{R}_j(\Lambda)$ from below, consider the orthonormal sequence

$$e_{\ell,m}(x) := \sqrt{T^{d\gamma}A^{-d}}\ \varphi_\ell(T^\gamma A^{-1}(x - a_m))\ \mathbb{1}\{x \in \mathbb{A}_m\},$$

and let

$$\psi^j_{\ell,m,\theta} := \int_{\mathbb{A}_m} g^j_\theta(x) e_{\ell,m}(x)\mathrm{d}x \quad\text{and}\quad \widehat{\psi}^j_{\ell,m,T} := \int_{\mathbb{A}_m} \widehat{g}_T(x) e_{\ell,m}(x)\mathrm{d}x \tag{2.4.13}$$

be the Fourier coefficients of g^j_θ and the estimator $\widehat{g}_T$ with respect to $e_{\ell,m}$, respectively. By means of Parseval's formula, it follows

$$\begin{aligned}\mathcal{R}_j(\Lambda) &= \inf_{\widehat{g}_T} \mathbb{E} \int_{\mathbb{R}^d} \left|\widehat{g}_T(x) - \operatorname{div}(a_j\rho_\theta)(x)\right|^2 \mathrm{d}x \\ &\geq \inf_{\widehat{g}_T} \mathbb{E} \int_{\mathbb{A}} \left|\widehat{g}_T(x) - \operatorname{div}(a_j\rho_\theta)(x)\right|^2 \mathrm{d}x \\ &\geq \inf_{\widehat{g}_T} \sum_{|m|_\infty \leq M} \sum_{1\leq |\ell|_\infty \leq R} \mathbb{E}\left|\widehat{\psi}^j_{\ell,m,T} - \psi^j_{\ell,m,\theta}\right|^2.\end{aligned}$$

We now apply the version of van Trees' inequality in (2.4.3), letting C the transposed j-th unit vector, that is, $C = (0,\dots,0,1,0,\dots,0) \in \mathbb{R}^{1\times d}$, where all entries apart from the j-th component vanish. It then follows that

$$\mathcal{R}_j(\Lambda) \geq \sum_{|m|_\infty \leq M} \sum_{1\leq |\ell|_\infty \leq R} \frac{\left(\int_{\Gamma^j_T} \partial_{\theta^j_{\ell,m}} \psi_{\ell,m,\theta} p(\theta)\mathrm{d}\theta\right)^2}{\int_{\Gamma^j_T} I^j_{\ell,m}(\theta)p(\theta)\mathrm{d}\theta + J^j_{\ell,m}(\mathrm{d}\Lambda)}, \tag{2.4.14}$$

where $J_{\ell,m}(\mathrm{d}\Lambda)$ is the Fisher information matrix of the prior density p (cf. (2.4.4)). Lemma 2.4.2 yields an upper bound on $\int_{\Gamma^j_T} I_{\ell,m}(\theta)p(\theta)\mathrm{d}\theta$. For evaluating the numerator appearing on the right-hand side of (2.4.14), we use the following

Lemma 2.4.3. *For the Fourier coefficients $\psi^j_{\ell,m,\theta}$ defined in (2.4.13), it holds*

$$\partial_{\theta^j_{\ell,m}} \psi^j_{\ell,m,\theta} = 2a_{jj}\sqrt{\left(2AT^{-\gamma}\right)^d \rho_0(a_m)}\ (1 + o_T(1)), \tag{2.4.15}$$

where $o_T(1)$ goes to zero uniformly in ℓ, m and $\theta \in \Gamma^j_T$.

Proof. Note first that, for any $x \in \mathbb{R}^d$,

$$2\rho_\theta(x) b^j_\theta(x) = 2\rho_\theta(x)\sum_{k=1}^d a_{jk}\partial_k\left(\log\rho_\theta(x)\right) = 2\sum_{k=1}^d a_{jk}\partial_k\rho_\theta(x) = g^j_\theta(x)$$

such that

$$\partial_{\theta^j_{\ell,m}} g^j_\theta(x) = 2\rho_\theta(x)\ \partial_{\theta^j_{\ell,m}} b^j_\theta(x) + 2\partial_{\theta^j_{\ell,m}}\rho_\theta(x)\ b^j_\theta(x).$$

Exploiting the definition of ρ_θ, it can be shown that

$$\sup_{\|x\|\leq A^d}\left(2b^j_\theta(x)\ \partial_{\theta^j_{\ell,m}}\rho_\theta(x)\right) = A^{-2d}\ o_T(1).$$

Thus,

$$\begin{aligned}\partial_{\theta^j_{\ell,m}} g^j_\theta(x) &= 2\sum_{l=1}^d a_{jl}\partial_{\theta^j_{\ell,m}}\left(b^-_\theta\right)^l(x)\rho_\theta(x) + A^{-2d}\ o_T(1) \\ &= 2a_{jj}\rho_\theta(x)\left(\frac{\left(2AT^{-\gamma}\right)^d}{\rho_0(a_m)}\right)^{1/2}\left(\phi_{\ell,m}(x) - \int_{\mathbb{R}^d}\phi_{\ell,m}(y)\rho_0(y)\mathrm{d}y\right)\ (1 + o_T(1)),\end{aligned}$$

and it holds

$$\begin{aligned}\partial_{\theta^j_{\ell,m}}\psi^j_{\ell,m,\theta} &= \int_{\mathbb{A}_m}\partial_{\theta^j_{\ell,m}} g^j_\theta(x)e_{\ell,m}(x)\mathrm{d}x\ (1 + o_T(1)) \\ &= 2a_{jj}\int_{\mathbb{A}_m}\rho_0(x)\left(\frac{\left(2AT^{-\gamma}\right)^d}{\rho_0(a_m)}\right)^{1/2} e^2_{\ell,m}(x)\mathrm{d}x\ (1 + o_T(1)).\end{aligned}$$

In view of Lemma 2.4.2 and since $\int_{\mathbb{A}_m} e^2_{\ell,m}(x)\mathrm{d}x = 1$, the assertion follows. □

(IV) Least favorable priors. The estimates (2.4.14), (2.4.9) and (2.4.15) and entail that there exists some $T_0 = T_0(\varepsilon) \in \mathbb{R}_+$ such that, for any $T \geq T_0$,

$$\begin{aligned}\mathcal{R}_j(\Lambda) &\geq \sum_{|m|_\infty \leq M} \sum_{1 \leq |\ell|_\infty \leq R} \frac{4a_{jj}^2 \left(2AT^{-\gamma}\right)^d \rho_0(a_m)}{(1+\varepsilon)(2AT^{-\gamma})^d a_{jj} T + J^j_{\ell,m}(\mathrm{d}\Lambda)} \left(1 + o_T(1)\right) \\ &= \sum_{|m|_\infty \leq M} \sum_{1 \leq |\ell|_\infty \leq R} \frac{4a_{jj} \left(2AT^{-\gamma}\right)^d \rho_0(a_m)}{(1+\varepsilon)(2A)^d T^{2\beta\gamma} + J^j_{\ell,m}(\mathrm{d}\Lambda) a_{jj}^{-1}} \left(1 + o_T(1)\right). \end{aligned} \tag{2.4.16}$$

Proceeding analogue to Golubev and Levit (1996) and Dalalyan and Kutoyants (2002), we now consider the "asymptotically least favorable prior distribution": The choice of the prior Λ is motivated by the attempt to maximize the Bayesian risk asymptotically. Furthermore, it should be essentially concentrated on the set $\left\{\theta \in \Gamma^j_T : b_\theta(\cdot) \in \Sigma^j_\delta\right\}$. Consider i.i.d. random vectors ξ_k with common density $f(x)$ satisfying

$$\int_{\mathbb{R}^d} x f(x) \mathrm{d}x = 0, \qquad \int_{\mathbb{R}^d} x^2 f(x) \mathrm{d}x = a_{jj}^{-1}, \qquad \max_{j=1,\ldots,d} |\xi^j_k| < G$$

and

$$\int_{\mathbb{R}^d} \left(\partial_{x^j} \log f(x)\right)^2 \mathrm{d}x = (1+\varepsilon) a_{jj}.$$

In particular, the above definitions entail that ε vanishes as G tends to infinity. Define the prior distribution Λ as the product measure

$$\mathrm{d}\Lambda(\theta) = \prod_{|m|_\infty \leq M} \prod_{1 \leq |\ell|_\infty \leq R} \lambda_{\ell,m}(\theta_{\ell,m})\, \mathrm{d}\theta_{\ell,m},$$

where

$$\lambda_{\ell,m}(u) = \lambda_\ell(u) := \left(\varsigma^j_\ell\right)^{-1/2}(\varepsilon)\, f\left(u/\sqrt{\varsigma^j_\ell(\varepsilon)}\right).$$

$J^j_{\ell,m}(\mathrm{d}\Lambda)$ then satisfies

$$J^j_{\ell,m}(\mathrm{d}\Lambda) = \frac{(1+\varepsilon) a_{jj}}{\varsigma^j_\ell(\varepsilon)}. \tag{2.4.17}$$

Convergence of the Riemann sums implies that

$$\sum_{|m|_\infty \leq M} \left(2AT^{-\gamma}\right)^d \rho_0(a_m) = \int_{\mathbb{A}} \rho_0(x) \mathrm{d}x \, \left(1 + o_T(1)\right) = 1 + o_T(1),$$

and plugging this last relation and (2.4.17) into (2.4.16), one obtains

$$\begin{aligned}\mathcal{R}_j(\Lambda) &\geq \frac{4a_{jj}}{1+\varepsilon} \sum_{1 \leq |\ell|_\infty \leq R} \frac{\varsigma^j_\ell(\varepsilon)}{(2A)^d T^{2\beta\gamma} \varsigma^j_\ell(\varepsilon) + 1} \left(1 + o_T(1)\right) \\ &= \frac{4a_{jj}}{(1+\varepsilon)(2A)^d T^{2\beta\gamma}} \sum_{1 \leq |\ell|_\infty \leq R} \left(1 - \left\| \frac{\ell^t}{\alpha^j (1-\varepsilon)} \right\|^\beta\right) \left(1 + o_T(1)\right) \\ &= \frac{4a_{jj}}{(1+\varepsilon)(2A)^d T^{2\beta\gamma}} \int_{\mathbb{A}} \left(1 - \left\| \frac{\lambda}{\alpha^j (1-\varepsilon)} \right\|^\beta\right) \mathrm{d}\lambda \, \left(1 + o_T(1)\right) \\ &= \frac{4a_{jj} \left(\alpha^j\right)^d (1-\varepsilon)^d}{(1+\varepsilon)(2A)^d T^{2\beta\gamma}} \int_{\mathbb{A}} \left(1 - \|\omega\|^\beta\right) \mathrm{d}\omega \, \left(1 + o_T(1)\right) \\ &= \frac{4a_{jj} \mathbb{S}_d \left(\alpha^j\right)^d (1-\varepsilon)^d}{(1+\varepsilon)(2A)^d T^{2\beta\gamma}} \int_0^1 r^{d-1} \left(1 - r^\beta\right) \mathrm{d}r \, \left(1 + o_T(1)\right) \\ &\geq 4a_{jj} \left(\alpha^j\right)^d (1-\varepsilon)^{d+1} \mathbb{S}_d (2A)^{-d} T^{-2\beta\gamma} \frac{\beta}{d(\beta+d)} \left(1 + o_T(1)\right) \\ &= 4(1-\varepsilon)^{d+1} T^{-2\beta\gamma} P_j(\sigma, \beta, L) \left(1 + o_T(1)\right).\end{aligned}$$

Combining (2.4.11) and the last display gives

$$\inf_{\widehat{g}_T} \sup_{b\in\Sigma_\delta} \mathbf{E}_b \int_{\mathbb{R}^d} |\widehat{g}_T(x) - \operatorname{div}(a_j\rho)(x)|^2 \,\mathrm{d}x \\ \geq 4(1-\varepsilon)^{d+1} T^{-2\beta/(2\beta+d)} P_j(\sigma,\beta,L)\ (1+o_T(1)) - \mathcal{R}_j^{\mathrm{c}}(\Lambda). \tag{2.4.18}$$

(V) Evaluation of $\mathcal{R}_j^{\mathrm{c}}(\Lambda)$ and conclusion of the proof. To conclude the proof, it needs to be shown that

$$\mathcal{R}_j^{\mathrm{c}}(\Lambda) = \sup_{\widehat{g}_T\in\mathcal{W}_T^j} \int_{\Gamma_T^j\setminus\Sigma_\delta^j} \mathbf{E}_b \int_{\mathbb{R}^d} |\widehat{g}_T(x) - \operatorname{div}(a_j\rho_\theta)(x)|^2 \,\mathrm{d}x \ \mathrm{d}\Lambda(\theta)$$

tends to zero faster than the first term appearing in (2.4.18). To do so, it is shown first that

$$\mathcal{C}_T^j := \left\{\theta \in \Gamma_T^j : V_T^j(\theta) < L(1-\varepsilon)\right\} \subseteq \left\{\theta : b_\theta \in \Sigma_\delta^j\right\},$$

for

$$V_T^j(\theta) := a_{jj}^2\ (2AT^{-\gamma})^d \sum_{|m|_\infty\leq M} \rho_0(a_m) \sum_{1\leq|\ell|_\infty\leq R} |\theta_{\ell,m}^j|^2\ (\pi T^\gamma A^{-1})^{2\beta}\|\ell\|^{2\beta}.$$

Since $\mathcal{C}_T^j \subseteq \Gamma_T^j$, it only needs to be verified that

$$\sup_{\theta\in\mathcal{C}_T^j} \int_{\mathbb{R}^d} \sum_{|\alpha|\leq\beta} \left|D^\alpha(g_\theta^j(x) - g_0^j(x))\right|^2 \mathrm{d}x \leq 4L.$$

In view of the relations $g_\theta^j = 2b_\theta^j\rho_\theta$ and $g_0^j = 2b_0^j\rho_0$ and exploiting the definitions of b_θ and b_0, it is shown analogously to the proof of Proposition 2.3.1 (also cf. the arguments used in the respective scalar setting on pp. 99–100 in Dalalyan and Kutoyants (2003) and p. 10 in Dalalyan and Kutoyants (2002)) that, for any multi-index $\alpha\in\mathbb{N}^d$ of order $|\alpha| = \beta\in\mathbb{N}$,

$$\begin{aligned}\int_{\mathbb{R}^d} \left|D^\alpha(g_\theta^j(x) - g_0^j(x))\right|^2 \mathrm{d}x &= 4\int_{\mathbb{R}^d} \left|D^\alpha(b_\theta^j(x)\rho_\theta(x) - b_0^j(x)\rho_0(x))\right|^2 \rho_0^2(x)\mathrm{d}x \\ &= 4\int_{\mathbb{A}_m} \left|D^\alpha(b_\theta^j(x) - b_0^j(x))\right|^2 \rho_0^2(x)\mathrm{d}x + o_T(1).\end{aligned}$$

Note that

$$b_\theta^j - b_0^j = \sum_{k=1}^d a_{jk}(b_\theta^-)^k = a_{jj}\ (b_\theta^-)^j.$$

Since all derivatives of U up to order β are bounded and vanish for $x\notin\mathbb{A}_m$, it holds, for any $k\in\{1,\dots,d\}$,

$$\frac{\partial^\beta (b_\theta^-)^j(x)}{\partial (x^k)^\beta} = \sum_{|m|_\infty\leq M} \sqrt{\frac{(2AT^{-\gamma})^d}{\rho_0(a_m)}} \sum_{1\leq|\ell|_\infty\leq R} \theta_{\ell,m}^j\ (\mathrm{i}\pi T^\gamma A^{-1})^\beta\ (\ell^k)^\beta.$$

Consequently, it can be shown that

$$\begin{aligned}&\sum_{|\alpha|\leq\beta} \int_{\mathbb{R}^d} \left|D^\alpha(g_\theta^j(x) - g_0^j(x))\right|^2 \mathrm{d}x \\ &\leq 4a_{jj}^2 \sum_{|m|_\infty\leq M} \frac{(2AT^{-\gamma})^d}{\rho_0(a_m)} \sum_{1\leq|\ell|_\infty\leq R} |\theta_{\ell,m}^j|^2\ (\pi T^\gamma A^{-1})^{2\beta}\ \|\ell\|^{2\beta}\ \rho_0^2(a_m)\ (1+o_T(1)) \\ &= 4V_T^j(\theta)\ (1+o_T(1)).\end{aligned}$$

Thus, for any $\theta \in \mathcal{C}_T^j$,

$$\sum_{|\alpha|\le\beta}\int_{\mathbb{R}^d}\left|D^\alpha(g_\theta^j(x)-g_0^j(x))\right|^2\mathrm{d}x \le 4L(1-\varepsilon)\le 4L.$$

To shorten notation, denote $\alpha_\varepsilon := \alpha^j(1-\varepsilon)$. The definition of $V_T^j(\cdot)$ entails that, for any $\theta \in \Gamma_T^j$,

$$\begin{aligned}
\mathbf{E}_b V_T^j(\theta) &= a_{jj}^2\ \left(2AT^{-\gamma}\right)^d \sum_{|m|_\infty\le M}\rho_0(a_m)\sum_{1\le|\ell|_\infty\le R} a_{jj}^{-1}\varsigma_\ell^j(\varepsilon)\ \left(\pi T^\gamma A^{-1}\right)^{2\beta}\|\ell\|^{2\beta}\\
&= a_{jj}\ \left(\frac{\pi T^\gamma}{A}\right)^{2\beta}\sum_{1\le|\ell|_\infty\le R}\frac{1}{(2A)^d}\left(\left(\frac{\alpha_\varepsilon}{\|\ell\|}\right)^\beta-1\right)\ (1+o_T(1))\\
&= a_{jj}\ \left(\frac{\pi}{A}\right)^{2\beta}\frac{\alpha_\varepsilon^{2\beta+d}}{(2A)^d}\sum_{1\le|\ell|_\infty\le R}\alpha_\varepsilon^{-d}\left(\frac{\|\ell\|}{\alpha_\varepsilon}\right)^{2\beta}\left(\left(\frac{\alpha_\varepsilon}{\|\ell\|}\right)^\beta-1\right)\ (1+o_T(1))\\
&= a_{jj}\ \left(\frac{\pi}{A}\right)^{2\beta}\frac{\alpha_\varepsilon^{2\beta+d}}{(2A)^d}\mathbb{S}_d\int_0^1 r^{d-1}\left(r^\beta-r^{2\beta}\right)\mathrm{d}r\ (1+o_T(1))\\
&= a_{jj}\ \left(\frac{2\pi}{2A}\right)^{2\beta}\frac{(2A)^{2\beta+d}}{(2A)^d}\ \frac{L(\beta+d)(2\beta+d)}{\beta\ a_{jj}\mathbb{S}_d(2\pi)^{2\beta}}\ \frac{\mathbb{S}_d\ \beta}{(\beta+d)(2\beta+d)}\ (1-\varepsilon)^{2\beta+d}(1+o_T(1))\\
&= L(1-\varepsilon)^{2\beta+d}(1+o_T(1)).
\end{aligned}$$

Hoeffding's inequality implies that, for some positive constant C,

$$\Lambda(\theta\notin\mathcal{C}_T^j)\le\Lambda\left(V_T^j(\theta)-\mathbf{E}_bV_T^j(\theta)\ge L\varepsilon(1-\varepsilon)\right)\le\exp\left(-\frac{L^2\varepsilon^2(1-\varepsilon)^2}{CT^{-2\beta\gamma}}\right),$$

that is,

$$\Lambda(\theta\notin\mathcal{C}_T^j)=o(T^{-1}).$$

Since

$$\mathcal{R}_j^{\mathrm{c}}(\Lambda)\le\left(8\sup_{\theta\in\Gamma_T^j}\int_{\mathbb{R}^d}\left|g_\theta^j(x)\right|^2\mathrm{d}x+2\right)\Lambda(b_\theta\notin\Sigma_\delta)$$

and $\int_{\mathbb{R}^d}\left|g_\theta^j(x)\right|^2\mathrm{d}x$ is bounded uniformly for $\theta\in\Gamma_T^j$, we obtain $\mathcal{R}_j^{\mathrm{c}}(\Lambda)\le o(T^{-1})$. Plugging the upper bound on $\mathcal{R}_j^{\mathrm{c}}(\Lambda)$ into (2.4.18), we obtain

$$\inf_{\widehat{g}_T}\sup_{b\in\Sigma_\delta^j}\mathbf{E}_b\int_{\mathbb{R}^d}\left|\widehat{g}_T(x)-\operatorname{div}(a_j\rho)(x)\right|^2\mathrm{d}x\ge 4(1-\varepsilon)^{d+1}T^{-2\beta/(2\beta+d)}P_j(\sigma,\beta,L)\ (1+o_T(1))$$

such that

$$\liminf_{\delta\to0}\liminf_{T\to\infty}T^{\frac{2\beta}{2\beta+d}}\inf_{\widetilde{g}_T^j}\sup_{b\in\Sigma_\delta^j}\mathbf{E}_b\int_{\mathbb{R}^d}\left|\widetilde{g}_T^j(x)-\operatorname{div}(a_j\rho)(x)\right|^2\mathrm{d}x\ge 4P_j(\sigma,\beta,L)\ (1-\varepsilon)^{d+1}.$$

Since $\varepsilon>0$ can be chosen arbitrarily small, the assertion follows.

$\square$

Proof of Theorem 2.3.3. Under the given assumptions, it holds $2b^j\rho=\operatorname{div}(a_j\rho)$. For any estimator $\widehat{b}_T^j$ of the j-th component of the drift $b\in U_\delta(b_0)$, denote $\widehat{g}_T^j := 2\widehat{b}_T^j\widehat{\rho}_{T,1}$, where the invariant density estimator $\widehat{\rho}_{T,1}(\cdot)$ is defined according to Lemma 2.2.4 and satisfies

$$\mathbf{E}_b\left|\widehat{\rho}_{T,1}(x)-\rho(x)\right|^2\le K_1\ T^{-\frac{2(\beta+1)}{2(\beta+1)+d}}\ \exp\left(-K_2\|x\|\right),\qquad x\in\mathbb{R}^d.$$

The inverse triangle inequality then entails that

$$\begin{aligned}
&\mathbf{E}_b \int_{\mathbb{R}^d} \left|\widehat{b}_T^j(x) - b^j(x)\right|^2 \rho^2(x)\mathrm{d}x \\
&\quad = \mathbf{E}_b \int_{\mathbb{R}^d} \left|\widehat{b}_T^j(x)\rho(x) - \widehat{b}_T^j(x)\widehat{\rho}_T(x) + \frac{1}{2}\widehat{g}_T^j(x) - \frac{1}{2}\operatorname{div}(a_j\rho)(x)\right|^2 \mathrm{d}x \\
&\quad \geq \left(\sqrt{\frac{1}{4}\mathbf{E}_b \int_{\mathbb{R}^d} \left|\widehat{g}_T^j(x) - \operatorname{div}(a_j\rho)(x)\right|^2 \mathrm{d}x} - \sqrt{\mathbf{E}_b \int_{\mathbb{R}^d} \left|\widehat{b}_T^j(x)\right|^2 (\rho(x) - \widehat{\rho}_{T,1}(x))^2\,\mathrm{d}x}\right)^2.
\end{aligned}$$

The growth condition on the drift term in (C_2) implies that, for any $x \in \mathbb{R}^d$ and T large enough,

$$\sup_{b \in U_\delta(b_0)} \left|b^j(x)\right| \leq c_2(1 + \|x\|) \lesssim \log T \exp\left(\frac{\|x\|}{\log T}\right).$$

If the estimator $\widehat{b}_T^j$ satisfies

$$\left|\widehat{b}_T^j(x)\right| \leq \log T \exp\left(\frac{\|x\|}{\log T}\right), \qquad x \in \mathbb{R}^d, \tag{2.4.19}$$

the risk of $\widehat{b}_T^j \wedge \log T \exp(\|x\|/\log T)$ is smaller than the risk of $\widehat{b}_T^j$ such that the lower bound needs to be proven only for estimators $\widehat{b}_T^j$ fulfilling (2.4.19). In this case, there exists some positive constant C such that

$$\begin{aligned}
\mathbf{E}_b \int_{\mathbb{R}^d} \left|\widehat{b}_T^j(x)\right|^2 (\rho(x) - \widehat{\rho}_T(x))^2\,\mathrm{d}x &\leq K_1 \int_{\mathbb{R}^d} (\log T)^2 \exp\left(\frac{2\|x\|}{\log T} - K_2\|x\|\right)\mathrm{d}x\ T^{-\frac{2(\beta+1)}{2(\beta+1)+d}} \\
&\leq C(\log T)^2\ T^{-\frac{2(\beta+1)}{2(\beta+1)+d}}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\liminf_{T\to\infty} T^{\frac{\beta}{2\beta+d}} \left(\inf_{\widehat{b}_T^j} \sup_{b \in \Sigma_\delta^j} \mathbf{E}_b \int_{\mathbb{R}^d} \left|\widehat{b}_T^j(x) - b^j(x)\right|^2 \rho^2(x)\mathrm{d}x\right)^{1/2} \\
&\quad \geq \liminf_{T\to\infty} \left(T^{\frac{2\beta}{2\beta+d}} \inf_{\widehat{g}_T^j} \sup_{b \in \Sigma_\delta^j} \mathbf{E}_b \int_{\mathbb{R}^d} \left|\widehat{g}_T^j(x) - \operatorname{div}(a_j\rho)(x)\right|^2 \mathrm{d}x/4 \right. \\
&\qquad\qquad\qquad \left. - C(\log T)^2\ T^{-\frac{d}{(2(\beta+1)+d)2(\beta+1)}}\right)^{1/2} \\
&\quad \geq \sqrt{P_j(\sigma,\beta,L)},
\end{aligned}$$

where the last line follows from Theorem 2.3.2. □

2.4.2. Proof of upper bounds

Proof of Theorem 2.3.10. We now show that

$$(b_T^j)^*(x) = \frac{(\overline{g}_T^j)^*(x)}{2\left(\widehat{\rho}_{T,1}(x) + c_T^*(x)\right)}, \qquad x \in \mathbb{R}^d,$$

is an asymptotically efficient estimator of the j-th component of the drift vector. Throughout the proof, $C_1, C_2, \ldots$ denote positive constants. Recall that

$$c_T^*(x) := T^{-\frac{\beta+1}{2(\beta+1)+d}}\ \varepsilon_T(x), \quad \text{for } \varepsilon_T(x) := \exp\left(\sqrt{\log T} - \frac{\|x\|}{\log T}\right), \qquad x \in \mathbb{R}^d,$$

and consider the event

$$\mathcal{T}_T(x) := \{\omega : \rho(x) - \widehat{\rho}_{T,1}(x) < c_T^*(x)\} = \{\omega : \rho(x) < \overline{\rho}_T(x)\}, \qquad x \in \mathbb{R}^d,$$

where $\overline{\rho}_T := \widehat{\rho}_{T,1} + c_T^*$. Recall once again that Lemma 2.2.4 implies that the invariant density estimator $\widehat{\rho}_{T,1}(\cdot)$ satisfies, for any $q \geq 1$,

$$\mathbf{E}_b \left|\widehat{\rho}_{T,1}(x) - \rho(x)\right|^{2q} \leq K_1\ T^{-\frac{2q(\beta+1)}{2(\beta+1)+d}}\ \exp\left(-K_2 q\|x\|\right), \qquad x \in \mathbb{R}^d. \tag{2.4.20}$$

Chebychev's inequality entails that, for T large enough,

$$\begin{aligned}\int_{\mathbb{R}^d} \mathbf{P}_b(\mathcal{T}_T^{\mathrm{c}}(x))\mathrm{d}x &= \int_{\mathbb{R}^d} \mathbf{P}_b(\rho(x) - \widehat{\rho}_{T,1}(x) \geq c_T^*(x))\mathrm{d}x\\ &\leq \int_{\mathbb{R}^d} \left(c_T^*(x)\right)^{-2q}\ \mathbf{E}_b\left|\widehat{\rho}_{T,1}(x) - \rho(x)\right|^{2q}\mathrm{d}x\\ &\leq K_1\ T^{\frac{2q(\beta+1)}{2(\beta+1)+d}} \int_{\mathbb{R}^d} \varepsilon_T^{-2q}(x)\exp\left(-K_2 q\|x\|\right)\mathrm{d}x\ T^{-\frac{2q(\beta+1)}{2(\beta+1)+d}}\\ &\leq C_1\ \exp\big(-2q\sqrt{\log T}\big).\end{aligned}$$

Thus,

$$\begin{aligned}&\int_{\mathbb{R}^d} \mathbf{E}_b \left(\left|(b_T^j)^*(x) - b^j(x)\right|^2 \rho^2(x)\mathbb{1}\left\{\mathcal{T}_T^{\mathrm{c}}(x)\right\}\right)\mathrm{d}x\\ &\quad\leq 2\int_{\mathbb{R}^d} \mathbf{E}_b\left(\left|(b_T^j)^*(x)\right|^2 \rho^2(x)\mathbb{1}\left\{\mathcal{T}_T^{\mathrm{c}}(x)\right\}\right)\mathrm{d}x + 2\int_{\mathbb{R}^d} |b^j(x)|^2\rho^2(x)\ \mathbf{P}_b(\mathcal{T}_T^{\mathrm{c}}(x))\mathrm{d}x\\ &\quad\leq 2\int_{\mathbb{R}^d}\left(\mathbf{E}_b\left|(b_T^j)^*(x)\right|^4\right)^{1/2}\rho^2(x)\left(\mathbf{P}_b(\mathcal{T}_T^{\mathrm{c}}(x))\right)^{1/2}\mathrm{d}x + C_1\exp\left(-2q\sqrt{\log T}\right).\end{aligned} \tag{2.4.21}$$

Recall that $(\overline{g}_T^j)^*(x) = 2T^{-1}\int_0^T K_T^*(x - X_u)\mathrm{d}X_u^j$, where K_T^* is defined via (2.3.11) and (2.3.32). Since $(a+b)^4 \leq 8a^4 + 8b^4$, it holds for the fourth moment of $(\overline{g}_T^j)^*$,

$$\begin{aligned}\mathbf{E}_b\left|(\overline{g}_T^j)^*(x)\right|^4 &\leq 128T^{-4}\mathbf{E}_b\left|\int_0^T K_T^*(x - X_u)b^j(X_u)\mathrm{d}u\right|^4\\ &\qquad + 128T^{-4}\mathbf{E}_b\left|\int_0^T K_T^*(x - X_u)\sum_{k=1}^d \sigma_{jk}\mathrm{d}W_u^k\right|^4.\end{aligned}$$

As concerns the first summand, Hölder's inequality implies that

$$T^{-4}\mathbf{E}_b\left|\int_0^T K_T^*(x - X_u)b^j(X_u)\mathrm{d}u\right|^4 \leq \mathbf{E}_b\Big(\left(K_T^*(x - X_0)\right)^4\left|b^j(X_0)\right|^4\Big).$$

The second term is upper-bounded by means of the Burkholder–Davis–Gundy inequality and Cauchy–Schwarz such that

$$\begin{aligned}\mathbf{E}_b\left|\int_0^T K_T^*(x - X_u)\sum_{k=1}^d \sigma_{jk}\mathrm{d}W_u^k\right|^4 &\leq C_2\ \mathbf{E}_b\left(\int_0^T a_{jj}\ \left|K_T^*(x - X_u)\right|^2\mathrm{d}u\right)^2\\ &\leq C_2\ a_{jj}^2\ T^2\mathbf{E}_b\left(\left(K_T^*(x - X_0)\right)^4\right).\end{aligned} \tag{2.4.22}$$

The Fourier inversion formula entails that

$$\begin{aligned}\sup_{y\in\mathbb{R}^d}|K_T(y)| &\leq C_3\int_{\mathbb{R}^d}\left(1 - \|\alpha\lambda\|^{\beta+\gamma}\right)_+\mathrm{d}\lambda\\ &\leq C_4\alpha^{-d}\int_{\mathbb{R}^d}\left(1 - \|\omega\|^{\beta+\gamma}\right)_+^2\mathrm{d}\omega \leq C_5\alpha^{-d}.\end{aligned}$$

Thus,

$$\sup_{y\in\mathbb{R}^d} |K_T^*(y)| \leq C_6\ T^{d/(2\beta+d)}. \tag{2.4.23}$$

Taking into account (2.4.23) and since the moments of b^j are uniformly bounded on Σ_δ^j (recall the at-most-linear growth of b), it holds for any $x\in\mathbb{R}^d$ and any $b\in\Sigma_\delta^j$,

$$\mathbf{E}_b\big|(\overline{g}_T^j)^*(x)\big|^4 \leq C_7\ T^\kappa, \quad \text{for } \kappa := \frac{4d}{2\beta+d}.$$

Consequently,

$$\begin{aligned}\int_{\mathbb{R}^d} \Big(\mathbf{E}_b (b_T^j)^*(x)\Big)^{1/2} \rho^2(x) \Big(\mathbf{P}_b\left(\mathcal{T}_T^{\mathrm{c}}(x)\right)\Big)^{1/2} \mathrm{d}x \\ &\leq C_8\ T^{\kappa/2} \int_{\mathbb{R}^d} \left(c_T^*(x)\right)^{-2} \rho^2(x)\mathrm{d}x\ \exp\left(-q\sqrt{\log T}\right) \\ &\leq C_8\ T^{\kappa/2+\frac{2(\beta+1)}{2(\beta+1)+d}} \exp\left(-q\sqrt{\log T}\right). \end{aligned}\tag{2.4.24}$$

Choosing

$$q = \left(\frac{\kappa}{2} + \frac{2(\beta+1)}{2(\beta+1)+d} + 1\right)\ \sqrt{\log T},$$

it follows from (2.4.21) and (2.4.24) that

$$\int_{\mathbb{R}^d} \mathbf{E}_b \left(\big|(b_T^j)^*(x) - b^j(x)\big|^2 \rho^2(x) \mathbb{1}\left\{\mathcal{T}_T^{\mathrm{c}}(x)\right\}\right) \leq C_9\ T^{-1}.$$

On the event $\mathcal{T}_T$, it holds $\rho < \overline{\rho}_T$, and, using relation (2.1.4), one obtains

$$\begin{aligned}&\int_{\mathbb{R}^d} \mathbf{E}_b \left(\big|(b_T^j)^*(x) - b^j(x)\big|^2 \rho^2(x) \mathbb{1}\left\{\mathcal{T}_T(x)\right\}\right) \mathrm{d}x \\ &\leq \Bigg\{\bigg(\int_{\mathbb{R}^d} \mathbf{E}_b \left(\big|(b_T^j)^*(x) - b^j(x)\rho(x)/\overline{\rho}_T(x)\big|^2 \rho^2(x) \mathbb{1}\left\{\mathcal{T}_T(x)\right\}\right) \mathrm{d}x\bigg)^{1/2} \\ &\qquad + \bigg(\int_{\mathbb{R}^d} \mathbf{E}_b \left(\big|b^j(x)\rho(x)/\overline{\rho}_T(x) - b^j(x)\big|^2 \rho^2(x) \mathbb{1}\left\{\mathcal{T}_T(x)\right\}\right) \mathrm{d}x\bigg)^{1/2}\Bigg\}^2 \\ &\leq \left\{\sqrt{\frac{1}{4}\int_{\mathbb{R}^d} \mathbf{E}_b \left|(g_T^j)^*(x) - \operatorname{div}(a_j\rho)(x)\right|^2 \mathrm{d}x} + \sqrt{\mathbf{E}_b \int_{\mathbb{R}^d} |b^j(x)|^2 \left(\rho(x) - \overline{\rho}_T(x)\right)^2 \mathrm{d}x}\right\}^2 \\ &=: \left(\sqrt{\mathbf{T}_1} + \sqrt{\mathbf{T}_2}\right)^2. \end{aligned}\tag{2.4.25}$$

(2.4.20) entails that

$$\begin{aligned}\mathbf{E}_b\big|\rho(x) - \overline{\rho}_T(x)\big|^2 &\leq 2\mathbf{E}_b\big|\widehat{\rho}_{T,1}(x) - \rho(x)\big|^2 \mathrm{d}x + 2\left(c_T^*(x)\right)^2 \\ &\leq 2K_1 T^{-\frac{2(\beta+1)}{2(\beta+1)+d}} \left\{\exp\left(-K_2\|x\|\right) + \exp\left(2\sqrt{\log T} - \frac{2\|x\|}{\log T}\right)\right\},\end{aligned}$$

and in view of the growth condition on the drift term, it follows that, for any $b\in\Sigma_\delta^j$,

$$\mathbf{T}_2 \leq T^{-\frac{2(\beta+1)}{2(\beta+1)+d}} \exp\left(2\sqrt{\log T}\right).$$

Thus, for any $\delta>0$,

$$\begin{aligned}&\lim_{T\to\infty} \sup_{b\in\Sigma_\delta^j} T^{\frac{2\beta}{2\beta+d}}\ \mathbf{E}_b \int_{\mathbb{R}^d} \big|(b_T^j)^*(x) - b^j(x)\big|^2\ \rho^2(x)\mathrm{d}x \\ &\qquad \leq \lim_{T\to\infty} \sup_{b\in\Sigma_\delta^j} \left\{T^{\frac{2\beta}{2\beta+d}} \left(\frac{1}{4}\mathbf{E}_b \int_{\mathbb{R}^d} \big|(\overline{g}_T^j)^*(x) - \operatorname{div}(a_j\rho)(x)\big|^2 \mathrm{d}x\right)^{1/2} + o_T(1)\right\}^2 \\ &\qquad \leq P_j(\sigma,\beta,L),\end{aligned}$$

that is, the upper bound part of relation (2.3.35) holds true. □

Proof of Theorem 2.3.16. The principle scheme of the proof of Theorem 2.3.16 is analogue to the proof of Theorem 1 in Dalalyan (2005). For the sake of completeness and since not all arguments carry over unmodified, we explicate the complete proof in the multidimensional framework. Before coming to the details, let us briefly sketch the basic idea for verifying sharp adaptivity of the estimator $\widehat{g}_T$ defined in (2.3.72).

In the proof below, we refer to the result on asymptotically efficient estimation of $\operatorname{div}(a\rho)$. It was shown in Section 2.3.3 that the asymptotically efficient estimator g_T^* of $\nabla\rho$ satisfies

$$\mathbf{E}_b \int_{\mathbb{R}^d} \left\|g_T^*(x) - \nabla\rho(x)\right\|^2 \mathrm{d}x \le \frac{1}{(2\pi)^d T}\ \widetilde{\Delta}_T(\alpha_T^*, \beta)\ \ (1 + o_T(1))$$

(cf. (2.3.23)), where

$$\widetilde{\Delta}_T(\alpha_T^*, \beta) := T \int_{\mathbb{R}^d} (1 - h_T^*(\lambda))^2 \left\|\phi_{\nabla\rho}(\lambda)\right\|^2 \mathrm{d}\lambda + 4d \int_{\mathbb{R}^d} (h_T^*(\lambda))^2 \,\mathrm{d}\lambda,$$

for α_T^* defined in (2.3.28) and

$$h_T^*(\lambda) := \left(1 - \|\alpha_T^*\lambda\|^\beta\right)_+ . \tag{2.4.26}$$

Recall the definition of the functionals Δ_T and $\widehat{\Delta}_T$ in Section 2.3.3,

$$\begin{aligned} \Delta_T(\alpha, h, \phi_\rho) &= T \int_{\mathbb{R}^d} |1 - h(\alpha\lambda)|^2 \left\|\phi_{\nabla\rho}(\lambda)\right\|^2 \mathrm{d}\lambda + 4d \int_{\mathbb{R}^d} |h(\alpha\lambda)|^2 \mathrm{d}\lambda, \\ \widehat{\Delta}_T(h) &= T \int_{\mathbb{R}^d} \left(h^2(\lambda) - 2h(\lambda)\right) \left\|\lambda\widehat{\phi}_T(\lambda)\right\|^2 \mathrm{d}\lambda + 8d \int_{\mathbb{R}^d} h(\lambda)\mathrm{d}\lambda, \end{aligned}$$

and note that

$$\begin{aligned} \Delta_T(1, h_T^*, \phi_\rho) &= T \int_{\mathbb{R}^d} |1 - h_T^*(\lambda)|^2 \left\|\phi_{\nabla\rho}(\lambda)\right\|^2 \mathrm{d}\lambda + 4d \int_{\mathbb{R}^d} |h_T^*(\lambda)|^2 \mathrm{d}\lambda, \\ &= \widetilde{\Delta}_T(\alpha_T^*, \beta). \end{aligned} \tag{2.4.27}$$

The basic idea for the proof of Theorem 2.3.16 is to show that $\widehat{g}_T$ as defined in (2.3.72) satisfies

$$\begin{aligned} \mathbf{E}_b \int_{\mathbb{R}^d} \left\|\widehat{g}_T(x) - \nabla\rho(x)\right\|^2 \mathrm{d}x &\le \frac{1}{(2\pi)^d T}\ \mathbf{E}_b \Delta_T(1, \widehat{h}_T, \phi_\rho)\ \ (1 + o_T(1)) \\ &= \frac{1}{(2\pi)^d T}\ \mathbf{E}_b \widehat{\Delta}_T(\widehat{h}_T)\ \ (1 + o_T(1)) . \end{aligned} \tag{2.4.28}$$

Provided that

$$\mathbf{E}_b \widehat{\Delta}_T(\widehat{h}_T) \le \mathbf{E}_b \widehat{\Delta}_T(h_T^*),$$

(2.4.28) implies in particular that

$$\begin{aligned} \mathbf{E}_b \int_{\mathbb{R}^d} \left\|\widehat{g}_T(x) - \nabla\rho(x)\right\|^2 \mathrm{d}x \quad &\le \quad \frac{1}{(2\pi)^d T}\ \Delta_T(1, h_T^*, \phi_\rho)\ \ (1 + o_T(1)) \\ &\overset{(2.4.27)}{=} \frac{1}{(2\pi)^d T}\ \widetilde{\Delta}_T(\alpha_T^*, \beta)\ \ (1 + o_T(1)) . \end{aligned}$$

The asymptotic behavior of the supremum of (a suitably normalized version of) $\widetilde{\Delta}_T(\alpha_T^*, \beta)$ over the Sobolev ball Σ_δ has been analyzed already in Section 2.3.3.

Preliminaries. We start with giving two auxiliary results which will be important in the sequel. The first lemma contains multidimensional analogues of Lemma 2 and Lemma 3 in Dalalyan (2005) and will be useful later for verifying (2.4.28). In particular, the upper bounds in (2.4.29) and (2.4.31) allow to eliminate the dependence on the random term

$$\zeta_T(\lambda)=\frac{1}{\sqrt{T}}\int_0^T \mathrm{e}^{\mathrm{i}\lambda^t X_u}\mathrm{d}W_u$$

(cf. the expression obtained in Lemma 2.3.13).

Lemma 2.4.4. (a) *Let $\widetilde{f}$ be some random bounded $\mathbb{R}^d$-valued function, taking values in some finite set $\{f_1(\cdot),\ldots,f_N(\cdot)\}$ of functions $f_i\colon\mathbb{R}^d\to\mathbb{R}^d$. Then there exists some constant C, depending only on β,L,b_0, such that*

$$\mathbf{E}_b\left|\int_{\mathbb{R}^d}\zeta_T^t(\lambda)\widetilde{f}(\lambda)\mathrm{d}\lambda\right|\leq C\left(N\mathbf{E}_b\int_{\mathbb{R}^d}\|\widetilde{f}(\lambda)\|^2\mathrm{d}\lambda\right)^{1/2}. \tag{2.4.29}$$

(b) *Assume that $\widetilde{h}$ is some random real-valued function, taking values in the finite set of functions $\{h_1(\cdot),\ldots,h_N(\cdot)\}$. If*

$$\|h_i\|^2_{L^2(\mathbb{R}^d)}\leq T,\qquad i=1,\ldots,N, \tag{2.4.30}$$

then, for $\varepsilon_T=T^{1/\sqrt{\log T}}$ and some constant $C=C(\beta,L,b_0)$,

$$\sup_{b\in\Sigma_\delta}\mathbf{E}_b\int_{\mathbb{R}^d}\widetilde{h}(\lambda)\left(\|\zeta_T(\lambda)\|^2-\|\sigma\|^2_{S_2}\right)\mathrm{d}\lambda\leq C\left(N\varepsilon_T\mathbf{E}_b\int_{\mathbb{R}^d}\widetilde{h}^2(\lambda)\mathrm{d}\lambda\right)^{1/2}. \tag{2.4.31}$$

Proof. Throughout the proof, C,C',C'' denote positive constants whose value may change from line to line. Part (a) is proven analogously to Lemma 2 in Dalalyan (2005), p. 254. First note that, for

$$z_{\widetilde{f}}:=\left(\int_{\mathbb{R}^d}\|\widetilde{f}(\lambda)\|^2\mathrm{d}\lambda\right)^{-1/2}\int_{\mathbb{R}^d}\zeta_T^t(\lambda)\widetilde{f}(\lambda)\mathrm{d}\lambda,$$

it holds by Cauchy–Schwarz

$$\begin{aligned}\mathbf{E}_b\left|\int_{\mathbb{R}^d}\zeta_T^t(\lambda)\widetilde{f}(\lambda)\mathrm{d}\lambda\right|&=\mathbf{E}_b\left(\left(\int_{\mathbb{R}^d}\|\widetilde{f}(\lambda)\|^2\mathrm{d}\lambda\right)^{1/2}|z_{\widetilde{f}}|\right)\\&\leq\left(\mathbf{E}_b\int_{\mathbb{R}^d}\|\widetilde{f}(\lambda)\|^2\mathrm{d}\lambda\right)^{1/2}\left(\sum_{j=1}^N\mathbf{E}_b|z_{f_j}|^2\right)^{1/2}.\end{aligned} \tag{2.4.32}$$

Recall that $\zeta_T(\lambda)=T^{-1/2}\int_0^T\mathrm{e}^{\mathrm{i}\lambda^tX_u}\mathrm{d}W_u$. Thus, for any $j\in\{1,\ldots,N\}$,

$$\begin{aligned}\mathbf{E}_b\left|\int_{\mathbb{R}^d}\zeta_T^t(\lambda)f_j(\lambda)\mathrm{d}\lambda\right|^2&=T^{-1}\ \mathbf{E}_b\left|\int_{\mathbb{R}^d}f_j^t(\lambda)\int_0^T\mathrm{e}^{\mathrm{i}\lambda^tX_u}\mathrm{d}W_u\ \mathrm{d}\lambda\right|^2\\&=T^{-1}\ \mathbf{E}_b\left|\int_0^T\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^tX_u}f_j^t(\lambda)\mathrm{d}\lambda\ \mathrm{d}W_u\right|^2\\&=\sum_{k=1}^d\mathbf{E}_b\left|\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^tX_0}f_j^k(\lambda)\mathrm{d}\lambda\right|^2=\sum_{k=1}^d\mathbf{E}_b\left|\phi_{f_j^k}(X_0)\right|^2\\&\leq\sup_{b\in\Sigma_\delta}\sup_{x\in\mathbb{R}^d}\rho_b(x)\sum_{k=1}^d\int_{\mathbb{R}^d}\left|\phi_{f_j^k}(z)\right|^2\mathrm{d}z\\&=\sup_{b\in\Sigma_\delta}\sup_{x\in\mathbb{R}^d}\rho_b(x)\int_{\mathbb{R}^d}\|f_j(\lambda)\|^2\,\mathrm{d}\lambda.\end{aligned}$$

Consequently, for $C := \sup_{b\in\Sigma_\delta}\sup_{x\in\mathbb{R}^d}\rho_b(x)$ and for any $j\in\{1,\dots,N\}$,

$$\mathbf{E}_b\big|z_{f_j}\big|^2 \le C\Big(\int_{\mathbb{R}^d}\|f_j(\lambda)\|^2\,\mathrm{d}\lambda\Big)^{-1}\int_{\mathbb{R}^d}\|f_j(\lambda)\|^2\,\mathrm{d}\lambda = C.$$

In view of (2.4.32), the assertion (2.4.29) follows.

As regards the proof of part (b), the main idea in the scalar setting is to express the square of the scalar analogue of ζ_T via some linear stochastic integral and to apply the same method as in the proof of (a). The strategy also works in the multidimensional framework as will be sketched in the sequel. Define the $\mathbb{R}^d$-valued martingales

$$\mathrm{M}_u(\lambda) := \int_0^u \cos\left(\lambda^t X_s\right)\mathrm{d}W_s, \qquad \mathrm{N}_u(\lambda) := \int_0^u \sin\left(\lambda^t X_s\right)\mathrm{d}W_s, \quad u>0, \tag{2.4.33}$$

such that, for any $u>0$, $\lambda\in\mathbb{R}^d$,

$$\zeta_u(\lambda) = \frac{1}{\sqrt{u}}\int_0^u \mathrm{e}^{\mathrm{i}\lambda^t X_s}\mathrm{d}W_s = \frac{1}{\sqrt{u}}\big(\mathrm{M}_u(\lambda)+\mathrm{i}\,\mathrm{N}_u(\lambda)\big).$$

Applying Itô's formula to the martingales M and N and the function $f(x) := \|x\|^2 = \sum_{k=1}^d (x^k)^2$, it follows

$$\begin{aligned}\|\mathrm{M}_T(\lambda)\|^2 &= \|\mathrm{M}_0(\lambda)\|^2 + \sum_{i=1}^d 2\int_0^T \mathrm{M}_u^i(\lambda)\mathrm{d}\,\mathrm{M}_u^i + \sum_{i=1}^d\int_0^T \mathrm{d}\langle \mathrm{M}^i(\lambda),\mathrm{M}^j(\lambda)\rangle_u\\ &= 2\int_0^T\cos\left(\lambda^t X_u\right)\mathrm{M}_u(\lambda)\mathrm{d}W_u + d\int_0^T\cos^2\left(\lambda^t X_u\right)\mathrm{d}u\end{aligned}$$

and

$$\|\mathrm{N}_T(\lambda)\|^2 = 2\int_0^T\sin\left(\lambda^t X_u\right)\mathrm{N}_u(\lambda)\mathrm{d}W_u + d\int_0^T\sin^2\left(\lambda^t X_u\right)\mathrm{d}u.$$

Consequently,

$$\|\zeta_T(\lambda)\|^2 = \frac{1}{T}\left(\|\mathrm{M}_T(\lambda)\|^2+\|\mathrm{N}_T(\lambda)\|^2\right) = \frac{2}{T}\int_0^T Y_u(\lambda)\mathrm{d}W_u + d, \tag{2.4.34}$$

where

$$\begin{aligned}Y_u(\lambda) &= \cos\left(\lambda^t X_u\right)\int_0^u\cos\left(\lambda^t X_s\right)\mathrm{d}W_s + \sin\left(\lambda^t X_u\right)\int_0^u\sin\left(\lambda^t X_s\right)\mathrm{d}W_s\\ &= \mathrm{Re}\,\mathrm{e}^{\mathrm{i}\lambda^t X_u}\int_0^u \mathrm{e}^{-\mathrm{i}\lambda^t X_s}\mathrm{d}W_s = \mathrm{Re}\,\mathrm{e}^{\mathrm{i}\lambda^t X_u}\sqrt{u}\zeta_u(\lambda).\end{aligned}$$

Thus,

$$\begin{aligned}\mathbf{E}_b\Big|\int_{\mathbb{R}^d}\widetilde{h}(\lambda)\left(\|\zeta_T(\lambda)\|^2-d\right)\mathrm{d}\lambda\Big|^2 &\overset{(2.4.34)}{=} \frac{4}{T^2}\,\mathbf{E}_b\Big|\int_{\mathbb{R}^d}\widetilde{h}(\lambda)\int_0^T Y_u(\lambda)\mathrm{d}W_u\,\mathrm{d}\lambda\Big|^2\\ &= \frac{4}{T^2}\,\mathbf{E}_b\Big|\int_0^T\int_{\mathbb{R}^d}Y_u(\lambda)\widetilde{h}(\lambda)\mathrm{d}\lambda\,\mathrm{d}W_u\Big|^2\\ &= \frac{4}{T^2}\,\mathbf{E}_b\int_0^T\Big\|\int_{\mathbb{R}^d}Y_u(\lambda)\widetilde{h}(\lambda)\mathrm{d}\lambda\Big\|^2\mathrm{d}u\\ &= 4U_b(\widetilde{h}),\end{aligned} \tag{2.4.35}$$

where

$$U_b(\widetilde{h}) := \frac{1}{T^2}\int_0^T u\,\mathbf{E}_b\Big\|\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t X_u}\zeta_u(\lambda)\widetilde{h}(\lambda)\mathrm{d}\lambda\Big\|^2\mathrm{d}u.$$

For

$$z_{\widetilde{h}} := \left(\int_{\mathbb{R}^d} \widetilde{h}^2(\lambda)\mathrm{d}\lambda\right)^{-1/2} \int_{\mathbb{R}^d} \widetilde{h}(\lambda) \left(\|\zeta_T(\lambda)\|^2 - d\right) \mathrm{d}\lambda,$$

Cauchy–Schwarz implies that

$$\begin{aligned}\mathbf{E}_b \int_{\mathbb{R}^d} \widetilde{h}(\lambda) \left(\|\zeta_T(\lambda)\|^2 - d\right) \mathrm{d}\lambda &= \mathbf{E}_b \left(\left(\int_{\mathbb{R}^d} \widetilde{h}^2(\lambda)\mathrm{d}\lambda\right)^{1/2} z_{\widetilde{h}}\right) \\ &\leq \left(\mathbf{E}_b \int_{\mathbb{R}^d} \widetilde{h}^2(\lambda)\mathrm{d}\lambda\right)^{1/2} \left(\sum_{i=1}^N \mathbf{E}_b \left|z_{h_i}\right|^2\right)^{1/2}.\end{aligned}$$

Since the above estimates hold uniformly in $b \in \Sigma_\delta$, (2.4.31) is proven once we have shown that, for any $b \in \Sigma_\delta$ and $i \in \{1,\ldots,N\}$,

$$\mathbf{E}_b\left|z_{h_i}\right|^2 \overset{(2.4.35)}{=} 4\bigg(\int_{\mathbb{R}^d} h_i^2(\lambda)\mathrm{d}\lambda\bigg)^{-1} U_b(h_i) \leq C'\varepsilon_T. \tag{2.4.36}$$

The subsequent manipulations are nearly analogue to the steps in the proof of Lemma 3 in Dalalyan (2005). For the sake of completeness, we sketch the entire line of reasoning. We will repeatedly use the estimate

$$\alpha_i^{-2d} \overset{(2.3.66)}{\leq} \left(\int_{\mathbb{R}^d} h_i^2(\lambda)\mathrm{d}\lambda\right)^2 = \|h_i\|^4_{L^2(\mathbb{R}^d)} \overset{(2.4.30)}{\leq} \|h_i\|^2_{L^2(\mathbb{R}^d)} T.$$

Denote by κ the constant introduced in condition (C_4), and assume that T is large enough to ensure $T > \kappa \vee \kappa^2$. Given any $h_i \in \mathfrak{H}_T^N$, write

$$U_b(h_i) = \frac{1}{T^2}\left(\int_0^\kappa + \int_\kappa^T\right) u\ \mathbf{E}_b \left\|\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t X_u}\zeta_u(\lambda)h_i(\lambda)\mathrm{d}\lambda\right\|^2 \mathrm{d}u.$$

Note that, for any $k \in \{1,\ldots,d\}$,

$$\begin{aligned}&\int_0^\kappa u\ \mathbf{E}_b \left|\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t X_u}\zeta_u^k(\lambda)h_i(\lambda)\mathrm{d}\lambda\right|^2 \mathrm{d}u \\ &\qquad = \int_0^\kappa u\ \mathbf{E}_b \left|\frac{1}{\sqrt{u}}\int_0^u \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t (X_u+X_s)} h_i(\lambda)\mathrm{d}\lambda\ \mathrm{d}W_s^k\right|^2 \mathrm{d}u \\ &\qquad \leq \kappa^2 \left(\int_{\mathbb{R}^d} h_i(\lambda)\mathrm{d}\lambda\right)^2 = \kappa^2 \left(\int_{\|\lambda\|\leq \alpha_i^{-d}} h_i(\lambda)\mathrm{d}\lambda\right)^2 \\ &\qquad \leq C\kappa^2\alpha_i^{-2d} \leq C'\kappa^2\|h_i\|^2_{L^2(\mathbb{R}^d)}T\end{aligned}$$

and

$$\begin{aligned}&\int_\kappa^T \mathbf{E}_b \left|\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t X_u} h_i(\lambda) \int_{u-\kappa}^u \mathrm{e}^{\mathrm{i}\lambda^t X_s}\mathrm{d}W_s^k\ \mathrm{d}\lambda\right|^2 \mathrm{d}u \\ &\qquad \leq \int_\kappa^T \mathbf{E}_b \left|\int_{\|\lambda\|\leq\alpha_i^{-d}} \int_{u-\kappa}^u \mathrm{e}^{\mathrm{i}\lambda^t X_s}\mathrm{d}W_s^k\ \mathrm{d}\lambda\right|^2 \mathrm{d}u \\ &\qquad \leq C\alpha_i^{-2d}\kappa T \leq C'\kappa\|h_i\|^2_{L^2(\mathbb{R}^d)}T.\end{aligned}$$

Thus,

$$\begin{aligned}U_b(h_i) &= \frac{1}{T^2}\int_0^T u\ \mathbf{E}_b \left\|\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t X_u}\zeta_u(\lambda)h_i(\lambda)\mathrm{d}\lambda\right\|^2 \mathrm{d}u \\ &\leq C\|h_i\|^2_{L^2(\mathbb{R}^d)} + T^{-2}\widetilde{U}_b(h_i),\end{aligned} \tag{2.4.37}$$

for

$$\widetilde{U}_b(h) := \int_\kappa^T \mathbf{E}_b \left\| \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t X_u} h(\lambda) \int_0^{u-\kappa} \mathrm{e}^{\mathrm{i}\lambda^t X_s} \mathrm{d}W_s \,\mathrm{d}\lambda \right\|^2 \mathrm{d}u.$$

Plugging (2.4.37) into (2.4.36) yields

$$\mathbf{E}_b |z_{h_i}|^2 \le \|h_i\|^2_{L^2(\mathbb{R}^d)} \left(C\|h_i\|^2_{L^2(\mathbb{R}^d)} + T^{-2}\widetilde{U}_b(h_i) \right),$$

and the proof of (2.4.36) is reduced to verifying that there exists some constant C such that

$$\widetilde{U}_b(h_i) \le C\varepsilon_T T^2 \|h_i\|^2_{L^2(\mathbb{R}^d)}. \tag{2.4.38}$$

Taking into account (C_4) and denoting

$$\begin{aligned} \eta(X^{u-\kappa}, \lambda) &:= \int_0^{u-\kappa} \mathrm{e}^{\mathrm{i}\lambda^t X_s} \mathrm{d}W_s, & \lambda \in \mathbb{R}^d, \\ Q(X^{u-\kappa}, y) &:= \left| \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} \eta(X^{u-\kappa}, \lambda) h_i(\lambda) \mathrm{d}\lambda \right|, & y \in \mathbb{R}^d, \end{aligned}$$

it can be shown that

$$\begin{aligned} \widetilde{U}_b(h_i) &= \int_\kappa^T \mathbf{E}_b \left(\int_{\mathbb{R}^d} \left\| \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} \eta(X^{u-\kappa}, \lambda) h_i(\lambda) \mathrm{d}\lambda \right\|^2 p_\kappa(X_{u-\kappa}, y) \mathrm{d}y \right) \mathrm{d}u \\ &\le \int_\kappa^T \left(D_1(u) + T^{-2}(D_2(u))^{1/2} \right) \mathrm{d}u, \end{aligned}$$

where

$$D_1(u) := \varepsilon_T \mathbf{E}_b \left(\sup p_\kappa^{b_0}(X_{u-\kappa}, y) \int_{\mathbb{R}^d} \left\| \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t y} h_i(\lambda) \eta(X^{u-\kappa}, \lambda) \mathrm{d}\lambda \right\|^2 \mathrm{d}y \right)$$

and

$$D_2(u) := \mathbf{E}_b \left(\int_{\mathbb{R}^d} Q^4(X^{u-\kappa}, y) p_\kappa^b(X_{u-\kappa}, y) \mathrm{d}y \right).$$

For D_1, Parseval's identity and Hölder's inequality imply that

$$\begin{aligned} D_1(u) &= (2\pi)^d \varepsilon_T \mathbf{E}_b \left(\sup p_\kappa^{b_0}(X_{u-\kappa}, y) \int_{\mathbb{R}^d} \|h_i(\lambda)\eta(X^{u-\kappa}, \lambda)\mathrm{d}\lambda\|^2 \mathrm{d}\lambda \right) \\ &\le C\varepsilon_T (T-\kappa) \left(\mathbf{E}_b \left(\sup p_\kappa^{b_0}(X_{u-\kappa}, y) \right)^q \right)^{1/q} \int_{\mathbb{R}^d} h_i^2(\lambda) \mathrm{d}\lambda \\ &\le C' \varepsilon_T (T-\kappa) \|h_i\|^2_{L^2(\mathbb{R}^d)}. \end{aligned}$$

For D_2, it holds by means of the same arguments as in the scalar setting

$$D_2(u) \le C\|h_i\|^4_{L^2(\mathbb{R}^d)} \alpha_i^{-2d} \le C' \|h_i\|^8_{L^2(\mathbb{R}^d)} \le C'' u^2 T^2 \|h_i\|^4_{L^2(\mathbb{R}^d)}$$

such that (2.4.38) is satisfied. □

The second auxiliary result is due to Cavalier et al. (2002) and was already mentioned in Section 2.3.3. It allows to restrict attention to minimizing the functional $\widehat{\Delta}_T(\cdot)$ over a finite-dimensional subset $\mathfrak{H}_T^N \subset \mathfrak{H}_T$.

Lemma 2.4.5. *For $\mathfrak{H}_T$ defined in (2.3.69), there exists a finite set $\mathfrak{H}_T^N \subset \mathfrak{H}_T$ of cardinality N such that, for any $h \in \mathfrak{H}_T$, there exists a filter $h' \in \mathfrak{H}_T^N$ such that*

$$\widehat{\Delta}_T(h') \le (1+\delta_N)\ \widehat{\Delta}_T(h),$$

where $\delta_N \to_{N\to\infty} 0$ and does not depend on $h, h', b \in \Sigma_\delta$.

Main part of the proof. Note first that, for any $\beta > d/2$ and T sufficiently large,

$$\int_{\mathbb{R}^d} \|\lambda\|^{2\beta} \widehat{h}_T^2(\lambda) \mathrm{d}\lambda \overset{(2.3.68)}{\leq} K_3 \ \alpha_i^{-(2\beta+d)} \leq K_3 \ T^{\frac{2\beta+d}{\log\log T}} \leq K_3 \ T.$$

Thus, Lemma 2.3.13 applies and gives

$$\lambda(\widehat{\phi}_T(\lambda) - \phi_\rho(\lambda)) = \frac{2\mathrm{i}}{\sqrt{T}} \zeta_T(\lambda) - \frac{\mathrm{i}}{\sqrt{T}} m_b(\lambda, X^T),$$

where $m_b(\lambda, X^T)$ satisfies

$$\int_{\mathbb{R}^d} \mathbf{E}_b \big\| m_b(\lambda, X^T) \ \widetilde{h}_T(\lambda) \big\|^2 \mathrm{d}\lambda = o\Big(T^{\frac{d}{2\beta+d}}\Big).$$

Taking into account (2.3.42), it thus follows

$$\begin{aligned}
&\mathbf{E}_b \int_{\mathbb{R}^d} \big\| \widehat{g}_T(x) - \nabla\rho(x) \big\|^2 \mathrm{d}x \\
&\quad = (2\pi)^{-d} \mathbf{E}_b \int_{\mathbb{R}^d} \Big\| \lambda\Big(\widehat{h}_T(\lambda) (\widehat{\phi}_T(\lambda) - \phi_\rho(\lambda)) - \phi_\rho(\lambda)(1 - \widehat{h}_T(\lambda)) \Big) \Big\|^2 \mathrm{d}\lambda \\
&\quad = (2\pi)^{-d} \mathbf{E}_b \int_{\mathbb{R}^d} \left\| \frac{2\mathrm{i}}{\sqrt{T}} \zeta_T(\lambda) \widehat{h}_T(\lambda) - \frac{\mathrm{i}}{\sqrt{T}} m_b(\lambda, X^T) \widehat{h}_T(\lambda) - (1 - \widehat{h}_T(\lambda)) \lambda \phi_\rho(\lambda) \right\|^2 \mathrm{d}\lambda \\
&\quad = (2\pi)^{-d} \mathbf{E}_b \int_{\mathbb{R}^d} \left\| \frac{2\mathrm{i}}{\sqrt{T}} \zeta_T(\lambda) \widehat{h}_T(\lambda) - (1 - \widehat{h}_T(\lambda)) \lambda \phi_\rho(\lambda) \right\|^2 \mathrm{d}\lambda \\
&\qquad + \frac{1}{(2\pi)^d T} \ \mathbf{E}_b \int_{\mathbb{R}^d} \big\| m_b(\lambda, X^T) \ \widetilde{h}_T(\lambda) \big\|^2 \mathrm{d}\lambda \\
&\qquad + \frac{2}{(2\pi)^d T} \ \mathbf{E}_b \int_{\mathbb{R}^d} \big| 2\zeta_T^t(\lambda) \widetilde{h}_T^2(\lambda)(1 - \widetilde{h}_T(\lambda)) m_b(\lambda, X^T) \lambda \phi_\rho(\lambda) \big| \mathrm{d}\lambda \\
&\quad = \frac{4}{(2\pi)^d T} \ \mathbf{E}_b \int_{\mathbb{R}^d} \Big\| \zeta_T(\lambda) \widehat{h}_T(\lambda) \Big\|^2 \mathrm{d}\lambda + (2\pi)^{-d} \mathbf{E}_b \int_{\mathbb{R}^d} \Big\| \Big(1 - \widehat{h}_T(\lambda)\Big) \lambda \phi_\rho(\lambda) \Big\|^2 \mathrm{d}\lambda \\
&\qquad - \frac{4\mathrm{i}}{(2\pi)^d \sqrt{T}} \ \mathbf{E}_b \int_{\mathbb{R}^d} \Big| \widehat{h}_T(\lambda) \Big(1 - \widehat{h}_T(\lambda)\Big) \zeta_T^t(\lambda) \lambda \phi_\rho(\lambda) \Big| \mathrm{d}\lambda + o\Big(T^{-\frac{2\beta}{2\beta+d}}\Big)
\end{aligned}$$

such that

$$\begin{aligned}
&\mathbf{E}_b \int_{\mathbb{R}^d} \big\| \widehat{g}_T(x) - \nabla\rho(x) \big\|^2 \mathrm{d}x \\
&\qquad = \frac{1}{(2\pi)^d T} \Big(\mathbf{E}_b \Delta_T(1, \widehat{h}_T, \phi_\rho) + A_1 - \operatorname{Im} A_2 \Big) + o\Big(T^{-\frac{2\beta}{2\beta+d}}\Big),
\end{aligned} \tag{2.4.39}$$

where

$$\begin{aligned}
A_1 &:= 4\mathbf{E}_b \int_{\mathbb{R}^d} \widehat{h}_T^2(\lambda) \left(\|\zeta_T(\lambda)\|^2 - d \right) \mathrm{d}\lambda, \\
A_2 &:= 4\sqrt{T} \mathbf{E}_b \int_{\mathbb{R}^d} \widehat{h}_T(\lambda) \left(1 - \widehat{h}_T(\lambda)\right) \zeta_T^t(\lambda) \ \lambda \phi_\rho(\lambda) \mathrm{d}\lambda.
\end{aligned}$$

The definition of $\widehat{\Delta}_T$ (cf. (2.3.40)) entails that

$$
\begin{aligned}
\widehat{\Delta}_T(h) &= 8d\int_{\mathbb{R}^d} h(\lambda)\mathrm{d}\lambda + T\int_{\mathbb{R}^d} \left(h^2(\lambda)-2h(\lambda)\right)\|\lambda\phi_\rho(\lambda)\|^2\,\mathrm{d}\lambda \\
&\quad + T\int_{\mathbb{R}^d}\left(h^2(\lambda)-2h(\lambda)\right)\left\|\lambda(\widehat{\phi}_T(\lambda)-\phi_\rho(\lambda))\right\|^2\mathrm{d}\lambda \\
&\quad + 2T\int_{\mathbb{R}^d}\left(h^2(\lambda)-2h(\lambda)\right)\left|\lambda^t\lambda\phi_\rho(\lambda)(\widehat{\phi}_T(\lambda)-\phi_\rho(\lambda))\right|\mathrm{d}\lambda \\
&= \Delta_T(1,h,\phi_\rho) - T\int_{\mathbb{R}^d}\|\phi_{\nabla\rho}(\lambda)\|^2\,\mathrm{d}\lambda \\
&\quad + T\int_{\mathbb{R}^d}\left(h^2(\lambda)-2h(\lambda)\right)\left\|\frac{2\mathrm{i}}{\sqrt{T}}\zeta_T(\lambda)-\frac{\mathrm{i}}{\sqrt{T}}m_b(\lambda,X^T)\mathrm{d}\lambda\right\|\mathrm{d}\lambda \\
&\quad + 2T\int_{\mathbb{R}^d}\left(h^2(\lambda)-2h(\lambda)\right)\left|\lambda^t\phi_\rho(\lambda)\left(\frac{2\mathrm{i}}{\sqrt{T}}\zeta_T(\lambda)-\frac{\mathrm{i}}{\sqrt{T}}m_b(\lambda,X^T)\right)\right|\mathrm{d}\lambda.
\end{aligned}
$$

Define

$$
\begin{aligned}
A_3 &:= 4\sqrt{T}\mathbf{E}_b\int_{\mathbb{R}^d}\left(\widehat{h}_T^2(\lambda)-2\widehat{h}_T(\lambda)\right)\phi_\rho(-\lambda)\zeta_T^t(\lambda)\lambda\mathrm{d}\lambda \\
&= 4\sqrt{T}\mathbf{E}_b\int_{\mathbb{R}^d}\left(1-\widehat{h}_T(\lambda)\right)^2\phi_\rho(-\lambda)\zeta_T^t(\lambda)\lambda\mathrm{d}\lambda, \qquad (2.4.40)\\
A_4 &:= 4\mathbf{E}_b\int_{\mathbb{R}^d}\left(\widehat{h}_T^2(\lambda)-2\widehat{h}_T(\lambda)\right)\left(\|\zeta_T(\lambda)\|^2-d\right)\mathrm{d}\lambda \\
&= 4\mathbf{E}_b\int_{\mathbb{R}^d}\left(1-\widehat{h}_T(\lambda)\right)^2\left(\sum_{j=1}^d\left|\zeta_T^j(\lambda)\right|^2-d\right)\mathrm{d}\lambda, \qquad (2.4.41)
\end{aligned}
$$

where the representations in (2.4.40) and (2.4.41) hold since, for any $k\in\{1,\dots,d\}$,

$$
\mathbf{E}_b\left(\zeta_T^k(\lambda)\right)=0 \quad\text{and}\quad \mathbf{E}_b\left(\|\zeta_T(\lambda)\|^2-d\right)=0.
$$

Consequently,

$$
\begin{aligned}
\mathbf{E}_b\widehat{\Delta}_T(\widehat{h}_T) = \mathbf{E}_b\Delta_T(1,\widehat{h}_T,\phi_\rho) - T\int_{\mathbb{R}^d}\|\phi_{\nabla\rho}(\lambda)\|^2\,\mathrm{d}\lambda& \\
+\operatorname{Im}A_3 + A_4 + o\left(T^{\frac{d}{2\beta+d}}\right).& \qquad (2.4.42)
\end{aligned}
$$

The terms A_1 to A_4 are now shown to be asymptotically smaller than $\mathbf{E}_b\Delta_T(1,\widehat{h}_T,\phi_\rho)$. Indeed, by means of Lemma 2.4.4(a), one obtains

$$
\begin{aligned}
A_2\vee A_3 &\le C_1\sqrt{T}\left(N\mathbf{E}_b\int_{\mathbb{R}^d}\left(1-\widehat{h}_T(\lambda)\right)^2\|\lambda\phi_\rho(\lambda)\|^2\,\mathrm{d}\lambda\right)^{1/2} \\
&\le C_2\sqrt{N}\left(\mathbf{E}_b\Delta_T(1,\widehat{h}_T,\phi_\rho)\right)^{1/2}.
\end{aligned}
$$

Part (b) of Lemma 2.4.4 implies for $\varepsilon_T := T^{1/\sqrt{\log T}}$ that

$$
\begin{aligned}
A_1\vee A_4 &\le 4\mathbf{E}_b\int_{\mathbb{R}^d}\widehat{h}_T(\lambda)\left(\|\zeta_T(\lambda)\|^2-d\right)\mathrm{d}\lambda \\
&\le C_3\left(N\varepsilon_T\mathbf{E}_b\int_{\mathbb{R}^d}\widehat{h}_T^2(\lambda)\mathrm{d}\lambda\right)^{1/2} \\
&\le C_4\varepsilon_T\left(\mathbf{E}_b\Delta_T(1,\widehat{h}_T,\phi_\rho)\right)^{1/2}.
\end{aligned}
$$

The relation (2.4.39) can thus be rewritten for some constant $C > 0$ as

$$\mathbf{E}_b \int_{\mathbb{R}^d} \|\widehat{g}_T(x) - \nabla\rho(x)\|^2 \mathrm{d}x \le \frac{1}{(2\pi)^d T}\left(\left(\mathbf{E}_b \Delta_T(1, \widehat{h}_T, \phi_\rho)\right)^{1/2} + C\varepsilon_T\right)^2 + o\Big(T^{-\frac{2\beta}{2\beta+d}}\Big).$$

The value $\widehat{h}_T$ minimizes the functional $\widehat{\Delta}_T(\cdot)$ over $\mathfrak{H}_T^N$ such that, by means of Lemma 2.4.5,

$$\mathbf{E}_b \widehat{\Delta}_T(\widehat{h}_T) = \min_{h \in \mathfrak{H}_T^N} \mathbf{E}_b \widehat{\Delta}_T(h) \le (1+\delta_N) \min_{h \in \mathfrak{H}_T} \mathbf{E}_b \widehat{\Delta}_T(h) \le (1+\delta_N) \mathbf{E}_b \widehat{\Delta}_T(h_T^*),$$

where h_T^* is defined as in (2.4.26). Consequently,

$$\mathbf{E}_b \widehat{\Delta}_T(\widehat{h}_T) + T\int_{\mathbb{R}^d} \|\phi_{\nabla\rho}(\lambda)\|^2 \,\mathrm{d}\lambda \ge \left(\left(\mathbf{E}_b \Delta_T(1, \widehat{h}_T, \phi_\rho)\right)^{1/2} - C\varepsilon_T\right)^2 + o\Big(T^{\frac{d}{2\beta+d}}\Big)$$

and

$$\mathbf{E}_b \widehat{\Delta}_T(h_T^*) + T\int_{\mathbb{R}^d} \|\phi_{\nabla\rho}(\lambda)\|^2 \,\mathrm{d}\lambda \le \left(\left(\Delta_T(1, \widehat{h}_T, \phi_\rho)\right)^{1/2} + C\varepsilon_T\right)^2 + o\Big(T^{\frac{d}{2\beta+d}}\Big).$$

Summing up,

$$\mathbf{E}_b \int_{\mathbb{R}^d} \|\widehat{g}_T(x) - \nabla\rho(x)\|^2 \mathrm{d}x \le \frac{1}{(2\pi)^d T}\left(\left(\Delta_T(1, h_T^*, \phi_\rho)\right)^{1/2} + C\varepsilon_T\right)^2 + o\Big(T^{-\frac{2\beta}{2\beta+d}}\Big).$$

In particular,

$$\left(T^{\frac{2\beta}{2\beta+d}}\, \mathbf{E}_b \int_{\mathbb{R}^d} \|\widehat{g}_T(x) - \nabla\rho(x)\|^2 \mathrm{d}x\right)^{1/2}$$

is asymptotically bounded by the square-root of the following expression,

$$(2\pi)^{-d} T^{-d/(2\beta+d)} \Delta_T(1, h_T^*, \phi_\rho) \overset{(2.4.27)}{=} (2\pi)^{-d} T^{-d/(2\beta+d)} \widetilde{\Delta}_T(\alpha_T^*, \beta), \tag{2.4.43}$$

plus residual terms of order $T^{1/\sqrt{\log T} - d/(4\beta+2d)}$ and $o_T(1)$. It was already shown in Section 2.3.3 that the supremum of (2.4.43) over the Sobolev ball Σ_δ defined in (2.3.31) tends to the constant $4P(\sigma, \beta, L)$ (see in particular (2.3.23) and (2.3.29)). This observation completes the proof. □

2.5. Pointwise sharp adaptive estimation on L^2 Sobolev balls

Our aim in the sequel is to analyze the asymptotically exact behavior of the pointwise risk for estimating the drift of a diffusion which is given as a solution of the stochastic differential equation

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma\, \mathrm{d}W_t, \qquad X_0 = \xi,\ t \in [0, T], \tag{2.5.1}$$

where W is a d-dimensional standard Wiener process and the initial value $\xi \in \mathbb{R}^d$ is independent of W. Analogous to the investigation of the global drift estimation problem, we exploit the representation of the drift at a point x_0 as an algebraic function of the diffusion matrix $a = \sigma\sigma^t$, the invariant density ρ and its partial derivatives, for $j \in \{1, \dots, d\}$ given as

$$b^j(x_0) = \frac{\operatorname{div}(a_j\rho)(x_0)}{2\rho(x_0)} = \frac{\sum_{k=1}^d a_{jk}\partial_k\rho(x_0)}{2\rho(x_0)}, \qquad x_0 \in \mathbb{R}^d. \tag{2.5.2}$$

The remaining section is organized as follows: Section 2.5.1 introduces the framework and basic assumptions and contains a short review of articles which are directly related to our investigation of pointwise drift estimation: Spokoiny (2000) considers pointwise estimation of the drift of scalar diffusions and suggests a locally linear smoother with data-driven bandwidth choice. His procedure is based on Lepski's method and provides a conditional non-asymptotic result. We suggest *asymptotically exact* adaptive drift estimators, and the adaptation scheme we propose is inspired by the procedures in Klemelä and Tsybakov (2001) and Klemelä and Tsybakov (2004) who modify Lepski's approach in order to perform exact adaptive estimation of linear functionals over L^2 Sobolev classes.

As done in the previous section, we start with considering estimation of $2b^j\rho = \operatorname{div}(a_j\rho)$, $j \in \{1,\dots,d\}$, assuming that $\operatorname{div}(a_j\rho)$ belongs to some Sobolev class of regularity $\beta \in \mathcal{I}$, $\mathcal{I}$ some given interval of the form $[\beta_*, \beta_T]$, where the upper bound β_T depends on T and satisfies $\beta_T \to_{T\to\infty} \infty$ (slowly enough). Similar arguments as in the global framework allow to infer results for estimating the components of the drift vector. Section 2.5.2 contains lower bounds for pointwise estimation of $\operatorname{div}(a_j\rho)$ and b^j, and an adaptive procedure for estimating $\operatorname{div}(a_j\rho)$ which asymptotically attains the respective infimum is introduced in Section 2.5.3. Section 2.5.4 addresses the approach of vector-wise estimation of $\operatorname{div}(a\rho)$, that is, all components $\operatorname{div}(a_j\rho)$ are estimated with one common bandwidth. Some short discussion of our findings and central auxiliary results for the proof of upper and lower bounds are given in Section 2.5.5. The great majority of the proofs is deferred to Section 2.6.

2.5.1. Introduction and motivation

The problem of *pointwise* drift estimation is interesting both from the practical and the theoretical point of view. In view of the interpretation of the drift function as the instantaneous mean of the diffusion (cf. Section 2.2.3), the value of $b(x)$, given the observation $(X_t)_{0\le t\le T}$, can be taken as the value of b at time $T+\delta$, for $\delta > 0$ small. As concerns the theoretical problem of pointwise *adaptive* drift estimation, it is to be expected from the results on pointwise adaptation in the classical models that the results on the speed of convergence in the adaptive case differ significantly from the global estimation setup investigated in the previous section. It is well-known for the classical statistical models and has been verified for drift estimation in multidimensional ergodic diffusion models in Section 2.3 that adaptive estimation of some unknown function *for integrated losses* is possible without loss in the rate of convergence and without loss of efficiency. In other settings however, estimation without payment for adaptation is not possible. One typical example is the problem of estimating a function f at a given point x_0. It was shown in Lepski (1990) that adaptive pointwise estimation leads to a *nearly* minimax rate, which is worse than a minimax one within an extra logarithmic factor. Lepski (1990) investigates the classical Gaussian white noise model, and subsequently similar results have been obtained in other classical statistical models.

In direct relation to the present work, Spokoiny (2000) considers the problem of pointwise adaptive drift estimation and develops a locally linear smoother with data-driven bandwidth choice. The use of local polynomial smoothing methods in nonparametric statistics is classical and goes back to Katkovnik (1985) and Tsybakov (1986). The form of the locally linear smoothers suggested in Spokoiny (2000) was already motivated in Section 2.2.3 (cf. (2.2.12)). His method is derived in a scalar setting but generalizes to the multidimensional framework. The focus of Spokoiny (2000) clearly differs from ours. He provides *nonasymptotic* results which do not require stationarity, ergodicity or mixing properties of the observed diffusion process. His adaptive procedure is based on some deviation inequality which describes the behavior of an estimator $\widetilde{b}_h(\cdot)$, h some bandwidth, restricted to some random set $\mathcal{A}_h$. In principle, an unconditional asymptotic risk bound for the risk of the estimators $\widetilde{b}_h(\cdot)$ can be derived from this deviation inequality; cf. Remark 3.4 in Spokoiny (2000). Theorem 4.1 in Spokoiny (2000) describes properties of the adaptive estimator $\widetilde{b}(x)$, restricted to the set $\mathcal{A}^* = \bigcap_{h\in\mathscr{H}}\mathcal{A}_h$, for a fixed family $\mathscr{H}$ of bandwidths h. Assuming that the drift is twice continuously differentiable, the proposed method is fully adaptive and rate optimal up to a $\log\log$ factor. The logarithmic payment for adaptation – which has been shown to be inevitable for the case of power loss functions $\ell(x) = |x|^p$, $p > 0$, in the classical statistical models such as Gaussian white noise, density estimation or nonparametric regression – is reduced to a $\log\log$ factor. This improvement is due to the fact that the adaptive choice of the bandwidth h is restricted to those h for which the observed path $(X_t)_{0\leq t\leq T}$ belongs to the random set $\mathcal{A}_h$. As a consequence, the cardinality H of the collection of estimators the adaptive estimator is chosen from satisfies $H \sim \log T$, and the adaptive choice of the bandwidth then leads to a worsening in the accuracy by factor $\log H$ at some power.

We have already seen in Section 2.3 that the definition of asymptotically *sharp adaptive* estimators does not only require a data-dependent choice of the smoothing parameter but also a data-driven selection of the kernel. Klemelä and Tsybakov (2001) (subsequently, abbreviated as [KT01]) show that exact adaptive estimators of linear functionals can be constructed by means of a Lepski-type selection procedure over families of estimators with optimal recovery kernels. They consider the Gaussian white noise model

$$\mathrm{d}Y_\varepsilon(x) = f(x)\mathrm{d}x + \varepsilon\ \mathrm{d}B(x), \qquad x \in \mathbb{R}^d, \tag{2.5.3}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is an unknown function to be estimated, B is the standard Brownian sheet in $\mathbb{R}^d$ and $0 < \varepsilon < 1$ is a small noise parameter. [KT01] suggest sharp adaptive estimators of f and its partial derivatives on scales of Sobolev classes.

The framework for pointwise estimation of the components of the drift vector is similar to the global setting. Let $\beta > d/2$, and define the Sobolev seminorm $\eta_\beta(\cdot)$ by

$$\eta_\beta(f) := \left(\frac{1}{(2\pi)^d}\int_{\mathbb{R}^d}\|\lambda\|^{2\beta}\left|\phi_f(\lambda)\right|^2\mathrm{d}\lambda\right)^{1/2}, \qquad f \in L^2(\mathbb{R}^d). \tag{2.5.4}$$

The following condition is central for our subsequent investigation.

$(\mathring{\mathrm{C}}_0{}^j)$ The SDE

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma\ \mathrm{d}W_t,$$

where $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is a constant non-degenerate dispersion matrix, admits a strong solution with Lebesgue continuous measure $\mathrm{d}\mu(x) = \rho(x)\mathrm{d}x$. The invariant density ρ satisfies the relation

$$2b^j\rho = \mathrm{div}(a_j\rho) = \sum_{k=1}^d a_{jk}\partial_k\rho.$$

Furthermore, for some $\beta > d/2$ and some $L' > 0$, it holds $\eta^2_{\beta+1}(\rho) \leq (L')^2$.

Denote by $\Pi^j(\beta)$ the set of measurable drift functions $b : \mathbb{R}^d \to \mathbb{R}^d$ satisfying $(\mathring{\mathrm{C}}_0{}^j)$ and conditions (C_1) and (C_2) introduced in Section 2.2.2. Given any $\beta > d/2$ and $L > 0$, let

$$\Pi^j(\beta, L) = \Pi^j_\epsilon(\beta, L; \rho, \sigma) := \left\{b \in \Pi^j(\beta) : \eta_\beta(\mathrm{div}(a_j\rho)) \leq 2L,\ \rho(x_0) \geq \epsilon\right\}, \tag{2.5.5}$$

where $\epsilon > 0$ is an a priori lower bound on $\rho(x_0)$.

Remark 2.5.1. The explicit knowledge of an a priori lower bound is not necessarily required for defining an exact adaptive estimator of $\mathrm{div}(a_j\rho)$. As noted by Butucea (2001), one should exclude the case of an invariant density ρ with $\rho(x_0) \to 0$ too fast as $T \to \infty$. To do so, one may assume that $\rho(x_0) \geq \rho^*_T$, where ρ^*_T is a sequence of positive real numbers such that $\lim_{T\to\infty}\rho^*_T = 0$ and $\liminf_{T\to\infty}(\rho^*_T \log T) > 0$. Using the same arguments as Butucea (2001), our results can be transferred to this setting where one considers estimation over classes of the form

$$\left\{b \in \Pi^j(\beta) : \eta_\beta(\mathrm{div}(a_j\rho)) \leq 2L,\ \rho(x_0) \geq \rho^*_T\right\}.$$

Remark 2.5.2. In order to get a rough idea of how to define an asymptotically exact procedure for estimating the drift coefficient, one might exploit the fact that different statistical models often exhibit a comparable asymptotic behavior. Spokoiny (2000) remarks that his results on pointwise *drift estimation* are similar to those that can be obtained in the *regression context.* [KT01] consider the *Gaussian white noise* model (2.5.3), but they also indicate extensions of their procedure to other types of observations, and they conjecture that the modified estimators have sharp optimality properties. In dimension $d = 1$, Butucea (2001) investigates the problem of estimating a density f from i.i.d. observations at a fixed point x_0 on Sobolev classes and thus extends the pointwise adaptive result of Tsybakov (1998) for the Gaussian white noise model to the problem of density estimation. The exact asymptotical constant for the minimax adaptive risk is identified as

$$L^{\frac{1}{2\beta}}\ (2\beta)\ \left(\frac{f(x_0)}{\beta(2\beta-1)}\right)^{\frac{\beta-1/2}{2\beta}} \frac{1}{2\pi}\int_{\mathbb{R}} \frac{|u|^{2\beta}}{\left(1+|u|^{2\beta}\right)^2}\,\mathrm{d}u, \qquad \beta > \frac{1}{2},\ L > 0, \tag{2.5.6}$$

and it is noted that this constant is "obviously similar" to the constant found in the Gaussian white noise model in Tsybakov (1998). (In fact, one obtains the constant for the Gaussian white noise model by multiplying the expression in (2.5.6) with $2^{(\beta-1/2)/2\beta}$.) This similarity can be explained by means of the results of Nussbaum (1996) who establishes asymptotic equivalence of the experiments given by the observations

$$y_i,\ i = 1, \dots, n, \qquad \text{i.i.d. with density } f,$$

and, for some standard Wiener process W_t on the unit interval,

$$\mathrm{d}y(t) = \sqrt{f(t)}\,\mathrm{d}t + \frac{1}{2\sqrt{n}}\,\mathrm{d}W_t, \qquad t \in [0,1].$$

It is mentioned in Klemelä and Tsybakov (2004) (subsequently, abbreviated as [KT04]), too, that the proposed adaptive procedure for estimating a multivariate function whose Riesz transform is observed in Gaussian white noise can be adapted to the (classical) nonparametric density or regression estimation problem (see their Remark 4 on p. 450). We shall transform this procedure for the purpose of adaptive drift estimation. The results of Dalalyan and Reiß (2007) on asymptotic statistical equivalence mentioned in Section 1.2 give a description about the relation between the Gaussian shift model and the drift estimation problem. Loosely speaking, the transfer of the procedures due to Klemelä and Tsybakov to the problem of estimating the j-th component of the drift requires to define estimators $\widehat{\rho}_T$ of the invariant density ρ and to implement their procedure with appropriately modified noise level.

2.5.2. Lower bound for pointwise estimation

It has been shown by Lepski (1990) and Brown and Low (1996) that estimators which are optimally rate adaptive with respect to the pointwise risk over the scale of Hölder classes do not exist. Tsybakov (1998) proves an analogous result for adaptation over the scale of Sobolev classes. The above articles consider the Gaussian white noise model. Although our findings are to some extend analogous and the principle ideas of the proof basically rely on techniques developed in the classical framework, the analysis involves some subtleties. It therefore appears instructive to start with considering the bare problem of finding a lower bound on the *pointwise rate of convergence* for estimating the function $\operatorname{div}(a_j\rho)$, $j \in \{1, \dots, d\}$ fixed, assuming that $b \in \Pi^j(\beta, L)$, for some $\beta > d/2$, $L > 0$.

Lemma 2.5.3. *Fix $j \in \{1, \ldots, d\}$. For any $\beta > d/2$, $L > 0$, there exists some positive constant $c < \infty$ such that, for any $x_0 \in \mathbb{R}^d$ fixed,*

$$\liminf_{T\to\infty} \inf_{\widehat{g}_T} \sup_{b\in\Pi^j(\beta,L)} T^{\frac{\beta-d/2}{\beta}} \mathbf{E}_b\left|\widehat{g}_T(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|^2 \geq c, \tag{2.5.7}$$

where the infimum is taken over all estimators $\widehat{g}_T$ of $\operatorname{div}(a_j\rho) = \sum_{k=1}^d a_{jk}\partial_k\rho$.

The proof of Lemma 2.5.3 is given in Section 2.6.

Remark 2.5.4 (The renormalization argument). The role of the notion of *renormalization* for determining optimal rates of convergence for estimating functionals over various smoothness classes has been investigated in Donoho and Low (1992). Precisely, they consider the problem of recovering a linear functional of an unknown function $f : \mathbb{R}^d \to \mathbb{R}$, assuming that f belongs to a certain convex class of functions. In order to calculate optimal rates of convergence, they suggest a procedure which consists in identifying the renormalization exponent s in relations of the form $J(af(b\ \cdot)) = ab^sJ(f(\cdot))$, for three homogeneous functionals J associated with the estimation problem at hand. The rate of convergence is then obtained as a combination of the three exponents.

Note that the Fourier transform ϕ_f is a homogeneous functional satisfying, for $a, b > 0$,

$$\phi_{af(b\ \cdot)}(\lambda) = \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t x} af(bx)\mathrm{d}x = ab^{-d}\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t b^{-1}y} f(y)\mathrm{d}y = ab^{-d}\ \phi_f(b^{-1}\lambda), \quad \lambda \in \mathbb{R}^d.$$

This implies in particular that

$$\begin{aligned}\eta_\beta(af(b\ \cdot)) &= \left((2\pi)^{-d}\int_{\mathbb{R}^d} \|\lambda\|^{2\beta}\left|\phi_{af(b\ \cdot)}(\lambda)\right|^2\mathrm{d}\lambda\right)^{1/2}\\ &= \left((2\pi)^{-d}a^2b^{-d+2\beta}\int_{\mathbb{R}^d} \|\lambda\|^{2\beta}\left|\phi_f(\lambda)\right|^2\mathrm{d}\lambda\right)^{1/2} = ab^{\beta-d/2}\ \eta_\beta(f),\end{aligned}$$

that is, the Sobolev seminorm η_β is homogeneous with dilation exponent $s = \beta - d/2$.

The renormalization argument implies that the optimal pointwise rate for estimating a density $f : \mathbb{R}^d \to \mathbb{R}$ from n i.i.d. observations is given by $n^{-(\beta-d/2)/(2\beta)}$, provided that f satisfies, for fixed $\beta > d/2$ and $L > 0$, the smoothness assumption $\eta_\beta(f) \leq L$. Very likely, the result for pointwise estimation of $\operatorname{div}(a_j\rho)$ stated in Lemma 2.5.3 can also be proven by renormalization arguments in the spirit of Donoho and Low (1992). Instead, we decided to include a much longer, constructive proof of the lower bound which is based on Theorem 2.2 in Tsybakov (2009). A suitable modification of the hypotheses defined in the proof will be applied later for verifying the *exact* adaptive lower bound. For what it's worth, a comparison of both proofs nicely illustrates the additional efforts required to prove exact results.

We now state the exact lower bound for estimating $\operatorname{div}(a_j\rho)$ adaptively, assuming that $b \in \Pi^j(\beta, L)$ for some $\beta \in [\beta_*, \infty)$ and $L \in [L_*, L^*]$. Here, $\beta_* \in (d/2, \infty)$ and $0 < L_* < L^* < \infty$ are fixed values. For any $\beta > d/2$, $L > 0$, let

$$\kappa = \kappa(\beta) := \frac{\beta - \frac{d}{2}}{2\beta}, \qquad \psi_{T,\beta} := \left(\frac{\log T}{T}\right)^{\kappa(\beta)}. \tag{2.5.8}$$

Theorem 2.5.5. *Fix $\beta_* > d/2$ and $\delta \in (0,1)$, denote*

$$\mathcal{B}_T := [\beta_*, \beta_T] \times [L_*, L^*], \quad \textit{for } \beta_T := (\log\log T)^\delta, \tag{2.5.9}$$

and let

$$C_j(\beta, L; \rho, \sigma) := 2L^{\frac{d}{2\beta}}\ \frac{2\beta}{d}\left(\frac{d^2 a_{jj}\rho(x_0)}{\beta(2\beta-d)}\right)^{\frac{\beta-d/2}{2\beta}} \mathbb{I}_\beta. \tag{2.5.10}$$

Then, for any $x_0 \in \mathbb{R}^d$ and $j \in \{1, \dots, d\}$ fixed,

$$\liminf_{T\to\infty} \inf_{\widehat{g}_T^j} \sup_{(\beta,L)\in\mathcal{B}_T} \sup_{b\in\Pi^j(\beta,L)} \left(\psi_{T,\beta}\ C_j(\beta, L; \rho, \sigma)\right)^{-2}\ \mathbf{E}_b\left|\widehat{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|^2 \geq 1, \quad (2.5.11)$$

where the infimum is taken over all estimators $\widehat{g}_T^j$ of $\operatorname{div}(a_j\rho)$*.*

Remark 2.5.6. Theorem 2.5.5 implies that there exists *no smaller constant* than $C_j(\beta, L; \rho, \sigma)$ when the rate of convergence is chosen equal to $\psi_{T,\beta}$. To complete the lower bound, it still has to be shown that $\psi_{T,\beta}$ is the *adaptive rate of convergence* on the scale of Sobolev classes in the sense of Definition 3 in Tsybakov (1998) (see Definition 1.2.2 in Section 1.2). A more complete result, showing that the normalization cannot be improved in the rate (and, by the above result, not even in the constant), is stated in Theorem 2.5.10 below.

The basic (and classical) idea of the proof of Theorem 2.5.5 is to reduce the proof of the lower bound in (2.5.11) to proving a lower bound on the risk of two hypotheses which are chosen to be distant enough. A lower bound on the latter risk is deduced by means of Theorem 6(i) in Tsybakov (1998) as it was also done by Butucea (2001), [KT01] and [KT04]. The verification of the conditions of Tsybakov (1998)'s result in the current diffusion framework requires tools which differ from those used in the references mentioned above. Denoting by $\mathbb{P}_0$ and $\mathbb{P}_1$ the probability measures associated to the two different hypotheses, it needs to be shown that, for some fixed τ and for any $\alpha \in (0, 1/2)$,

$$\mathbb{P}_1\left(\frac{\mathrm{d}\mathbb{P}_0}{\mathrm{d}\mathbb{P}_1} \geq \tau\right) \geq 1 - \alpha. \quad (2.5.12)$$

In the Gaussian white noise framework considered in [KT01] and [KT04], (2.5.12) is verified directly for suitably chosen hypotheses due to the Gaussian nature of the model. For nonparametric density estimation from i.i.d. observations, Butucea (2001) uses Lyapunov's CLT. The lower bound in (2.5.11) is proven by defining distant enough hypotheses g_0, g_1, motivated by the proof of the lower bound for classical density estimation in Butucea (2001). In particular, the hypotheses are chosen such that the Radon–Nikodym density of the associated diffusion processes is well-defined. The condition (2.5.12) is then verified by means of the martingale CLT.

Theorem 2.5.7 (Exact constant for pointwise adaptive estimation of the components of the drift vector: Lower bound)**.** *Define $\mathcal{B}_T$ as in Theorem 2.5.5, let*

$$\widetilde{C}_j(\beta, L; \rho, \sigma) := \frac{L^{\frac{d}{2\beta}}}{\rho(x_0)}\ \frac{2\beta}{d}\left(\frac{d^2 a_{jj}\rho(x_0)}{\beta(2\beta - d)}\right)^{\frac{\beta - d/2}{2\beta}}\ \mathbb{I}_\beta.$$

Then, for any $x_0 \in \mathbb{R}^d$ and $j \in \{1, \dots, d\}$ fixed,

$$\liminf_{T\to\infty} \inf_{\widehat{b}_T^j} \sup_{(\beta,L)\in\mathcal{B}_T} \sup_{b\in\Pi^j(\beta,L)} \left(\psi_{T,\beta}\ \widetilde{C}_j(\beta, L; \rho, \sigma)\right)^{-2}\ \mathbf{E}_b\left|\widehat{b}_T^j(x_0) - b^j(x_0)\right|^2 \geq 1,$$

where the infimum is taken over all estimators $\widehat{b}_T^j$ of the j-th component b^j of the drift vector.

The proofs of Theorem 2.5.5 and Theorem 2.5.7 are deferred to Section 2.6.

2.5.3. Construction of sharp adaptive estimators

To define an asymptotically exact adaptive estimator of the drift coefficient, the estimation problem is again decomposed into estimating the divergences $\operatorname{div}(a_j\rho)$ and the invariant density

ρ appearing in the denominator (cf. (2.5.2)) separately. We will use a two-staged procedure in the spirit of Lepski's method, constructing first a collection of admissible estimators and selecting then an estimator with minimal variance among them.

Many of the procedures proposed in the literature deal with scalar estimation problems. The extension of (Lepski-type) methods based on bias-variance comparison schemes to a higher-dimensional setting may cause some difficulties since monotonicity of the bias and variance with respect to the bandwidth do not hold in general in the multidimensional case. Throughout this section, we restrict attention to an isotropic framework where this problem does not arise.

To define adaptive estimators of the components of $\operatorname{div}(a\rho)(x_0)$, we act similar to [KT04]. The advantage of their estimation scheme as compared to the one proposed in the preceding work [KT01] is that it allows for a wider range of values of the smoothness parameter β. While [KT01] work under the assumption that $\beta \in [\beta_*, \beta^*]$ for some known *fixed* β^*, this condition is dropped in [KT04] by replacing β^* with an upper bound β_T, with $\beta_T \to \infty$ slowly enough. The construction of the estimator thus does not depend on the upper bound β_T, and the upper bound does not appear in the expression for the exact asymptotic constant.

We proceed by introducing another central assumption on the diffusion process X which is required for proving sharp adaptivity of the proposed estimation scheme.

***Assumption* (BI).** Consider the ergodic diffusion process solution of the SDE

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma\, \mathrm{d}W_t, \tag{2.5.13}$$

and denote the invariant measure by μ. Then, for any bounded measurable function $f \in L^2(\mu)$ and for any $r, T > 0$ and $j \in \{1, \ldots, d\}$ fixed, there exists some positive constant C_B such that

$$\begin{aligned}\mathbf{P}_b\left(\left|\frac{1}{T}\int_0^T f(X_u)b^j(X_u)\mathrm{d}u - \int_{\mathbb{R}^d} f(y)b^j(y)\mathrm{d}\mu(y)\right| > r\right)&\\ \leq 2\exp\left(-\frac{Tr^2}{2C_B(\varsigma(f) + r\|f\|_\infty)}\right)&,\end{aligned} \tag{\textbf{BI}}$$

where

$$\varsigma^2(g) := \lim_{T\to\infty} \frac{1}{T}\operatorname{Var}_{\mathbf{P}_b}\left(\int_0^T g(X_u)\mathrm{d}u\right), \quad g \in L^2(\mu),$$

denotes the asymptotic variance appearing in the CLT.

The Bernstein-type deviation inequality (**BI**) in particular allows to prove uniform deviation inequalities which are crucial tools for verifying sharp upper bounds on the pointwise squared risk of the adaptive estimators. For a more detailed discussion of the importance of the above assumption, we refer to Section 2.5.5. Notably, we state sufficient conditions on the diffusion X for Assumption (**BI**) to hold, and we derive a uniform exponential inequality from (**BI**) (cf. Lemma 2.5.13).

For implementing the adaptive procedure, consider a sufficiently fine grid $\mathcal{G} = \mathcal{G}_T$ on the interval $[\beta_*, \beta_T]$, with $\beta_T \to \infty$. It is defined as

$$\mathcal{G} = \mathcal{G}_T := \{\beta_1, \ldots, \beta_m\},$$

where $\beta_* < \beta_1 < \ldots < \beta_m = \beta_T$. Assume that there exist $k_2 > k_1 > 0$ and $\delta_1 \geq \delta > 1$ such that

$$k_1(\log T)^{-\delta_1} \leq \beta_{i+1} - \beta_i \leq k_2(\log T)^{-\delta}, \quad i = 0, 1, \ldots, m-1, \tag{2.5.14}$$

and set $\beta_0 := \beta_* - d/2$.

As in the case of density estimation (cf. Butucea (2001)), the optimal bandwidth for estimating $\operatorname{div}(a_j\rho)$ (which is not available in practice) involves the unknown value of the invariant density ρ at $x_0 \in \mathbb{R}^d$. The adaptive procedure for estimating $\operatorname{div}(a_j\rho)$ therefore starts with a preliminary estimator $\widehat{\varrho}_T(x_0)$ of the value $\rho(x_0)$.

(0) **Definition of the preliminary density estimator.** Define

$$\widehat{\varrho}_T(x_0) := \frac{1}{Th_T^d}\int_0^T Q\left(\frac{X_u - x_0}{h_T}\right)\mathrm{d}u, \tag{2.5.15}$$

where Q is a bounded positive kernel satisfying $\int_{\mathbb{R}^d}\|u\||Q(u)|\mathrm{d}u < \infty$, and the bandwidth $h_T > 0$ is such that

$$\lim_{T\to\infty} h_T = 0, \qquad \lim_{T\to\infty} Th_T^d = \infty, \qquad \lim_{T\to\infty} Th_T^{2d}(\log T)^{-3} = \infty,$$

and, for some $\alpha_0 \in (0,1/2)$, $\limsup_{T\to\infty} h_T^d T^{\alpha_0} < \infty$. Recall that ϵ denotes an a priori lower bound on $\rho(x_0)$, and let

$$\widehat{\rho}_T(x_0) := \max\left\{\widehat{\varrho}_T(x_0), \epsilon\right\}. \tag{2.5.16}$$

(I) **Main part of the procedure: Adaptive estimation of** $\operatorname{div}(a_j\rho)$**.** For fixed $j \in \{1,\ldots,d\}$, we now describe the procedure for defining an adaptive estimator of the j-th component of the vector $\operatorname{div}(a\rho)$. Recall that $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is the dispersion matrix and $a = \sigma\sigma^t$ denotes the associated diffusion coefficient. The adaptive estimator will be selected among the family of estimators $\widehat{g}^j_{T,\beta}(x_0)$, defined as

$$\widehat{g}^j_{T,\beta}(x_0) := \frac{2}{T(\widehat{h}^j_{T,\beta})^d}\int_0^T K_\beta\left(\frac{X_u - x_0}{\widehat{h}^j_{T,\beta}}\right)\mathrm{d}X^j_u, \tag{2.5.17}$$

where

$$\widehat{h}^j_{T,\beta} := \left(\frac{4d\widehat{\rho}_T(x_0)a_{jj}\log T}{\beta T}\right)^{1/(2\beta)}. \tag{2.5.18}$$

Remark 2.5.8. The integral in (2.5.17) can be interpreted as a stochastic integral of some measurable function of the path $X^T = (X_s)_{0\le s\le T}$ and some fixed value h such that we consider some function $G(X^T, h)$. Inserting any choice of h (in particular, $\widehat{h}^j_{T,\beta}$ as defined in (2.5.18)), we obtain a well-defined measurable object. The technical difficulties in the analysis of the stochastic integral due to the adaptive choice of the bandwidth $\widehat{h}^j_{T,\beta}$ are circumvented by investigating estimators of the form

$$\frac{2}{Th^d}\int_0^T K_\beta\left(\frac{X_u - x_0}{h}\right)\mathrm{d}X^j_u,$$

where h is deterministic and belongs to some specified set of bandwidths which are in certain sense "close" to the original random bandwidth.

As in [KT01], the kernel K_β is obtained as a renormalized version of the basic kernel

$$\begin{aligned}\widetilde{K}_\beta(x) &:= (2\pi)^{-d}\int_{\mathbb{R}^d}\left(1+\|\lambda\|^{2\beta}\right)^{-1}\exp(\mathrm{i}\lambda^t x)\mathrm{d}\lambda \\ &= (2\pi)^{-d}\|x\|^{1-d/2}\int_0^\infty \frac{t^{d/2}}{1+t^{2\beta}}\, J_{(d-1)/2}(t\|x\|)\mathrm{d}t,\end{aligned} \tag{2.5.19}$$

where J_n denotes the ordinary Bessel function of order n, that is,

$$J_n(x) = \frac{1}{\pi}\int_0^\pi \cos(n\lambda - x\sin\lambda)\mathrm{d}\lambda = \frac{1}{2\pi}\int_{-\pi}^\pi \mathrm{e}^{-\mathrm{i}(n\lambda - x\sin\lambda)}\mathrm{d}\lambda, \quad x \in \mathbb{R}.$$

Define
$$\mathsf{b} = \mathsf{b}(\beta) := \left(\frac{2\beta - d}{d}\right)^{1/(2\beta)}, \tag{2.5.20}$$
and let
$$K_\beta(x) := \mathsf{b}^d \widetilde{K}_\beta(\mathsf{b}x). \tag{2.5.21}$$
Finally, introduce the thresholding sequence
$$\widehat{\eta}^j_{T,\beta} = \left(\frac{4d\widehat{\rho}_T(x_0)a_{jj}\log T}{\beta T}\right)^{\frac{\beta-d/2}{2\beta}} \|K_\beta\|_{L^2(\mathbb{R}^d)}.$$
The adaptive estimator $\widetilde{g}^j_T$ is defined as
$$\widetilde{g}^j_T(x_0) := \widehat{g}_{T,\widehat{\beta}^j_T}(x_0), \tag{2.5.22}$$
where
$$\widehat{\beta}^j_T := \max\left\{\beta \in \mathcal{G}_T : |\widehat{g}^j_{T,\gamma}(x_0) - \widehat{g}^j_{T,\beta}(x_0)| \le \widehat{\eta}^j_{T,\gamma}\ \forall\, \gamma \in \mathcal{G}_T,\ \gamma \le \beta\right\}. \tag{2.5.23}$$
We continue with the main result on adaptive pointwise estimation of
$$\operatorname{div}(a_j\rho) = \sum_{k=1}^d a_{jk}\partial_k\rho, \quad j \in \{1,\ldots,d\}.$$
Recall the definition of the constant $C_j(\beta, L; \rho, \sigma)$ in (2.5.10).

Theorem 2.5.9. *Grant Assumptions* (**BI**) *and* (**SG**+). *Then, for fixed $\beta_* > d/2$, $0 < L_* < L^* < \infty$, $\delta_2 \in (0,1)$ and for*
$$\mathcal{B}_T := [\beta_*, \beta_T] \times [L_*, L^*], \qquad \text{where } \beta_T := (\log\log T)^{\delta_2},$$
the adaptive estimator $\widetilde{g}^j_T$ defined according to (2.5.22) satisfies, for any $x_0 \in \mathbb{R}^d$,
$$\limsup_{T\to\infty} \sup_{(\beta,L)\in\mathcal{B}_T} \sup_{b\in\Pi^j(\beta,L)} \left(\psi_{T,\beta}\ C_j(\beta,L;\rho,\sigma)\right)^{-2} \mathbf{E}_b\left|\widetilde{g}^j_T(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|^2 \le 1. \tag{2.5.24}$$

One obtains an adaptive drift estimator by defining a suitable invariant density estimator $\widetilde{\rho}_T$ and setting
$$\widetilde{b}_T(x_0) := \frac{\widetilde{g}_T(x_0)}{2\left(\widetilde{\rho}_T(x_0) \vee \epsilon\right)}, \qquad x_0 \in \mathbb{R}^d,$$
where $\epsilon > 0$ is the a priori lower bound on $\rho(x_0)$. Similarly to the proof of Theorem 2.3.10, the exact constant appearing in the upper bound for the pointwise squared risk of $\widetilde{b}^j_T(x_0)$, assuming that $b \in \widetilde{\Pi}_j(\beta, L)$, is identified as
$$\frac{C_j(\beta,L;\rho,\sigma)}{2\rho(x_0)} = \widetilde{C}_j(\beta,L;\rho,\sigma).$$

We now specialize to the important case of Kolmogorov diffusions. Taking into account Lemma 2.5.3, Theorem 2.5.5 and Theorem 2.5.9, we are ready to prove a result on sharp adaptive estimation of the j-th partial derivative of the invariant density, $j \in \{1,\ldots,d\}$ fixed, at a fixed point. To enlighten notation, let
$$C_{\mathrm{Kol}}(\beta,L;\rho) := C_j(\beta,L;\rho,\mathbf{Id}_{d\times d}) = 2L^{\frac{d}{2\beta}}\ \frac{2\beta}{d}\left(\frac{d^2\rho(x_0)}{\beta(2\beta-d)}\right)^{\frac{\beta-d/2}{2\beta}} \mathbb{I}_\beta, \tag{2.5.25}$$

where $\beta > d/2$, $L > 0$ and $x_0 \in \mathbb{R}^d$. We further introduce the maximal risk of an estimator $\breve{g}_T$ of $\partial_j \rho$, for $\beta > d/2$, $L > 0$, $T > 0$ and fixed $x_0 \in \mathbb{R}^d$ defined as

$$\mathscr{R}_{T,\beta,L}(\breve{g}_T) := \sup_{b \in \Pi_\epsilon^j(\beta,L;\rho,\mathbf{Id}_{d\times d})} \mathbf{E}_b \left|\breve{g}_T(x_0) - \partial_j \rho(x_0)\right|^2.$$

Theorem 2.5.10. *Consider the SDE*

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \mathrm{d}W_t,$$

and assume that Assumptions (**BI**) *and* (**SG**+) *are satisfied. Fix $\beta_* > d/2$, $\delta_2 \in (0,1)$, denote $\beta_T := (\log\log T)^{\delta_2}$, and let $\mathcal{B}_T := [\beta_*, \beta_T] \times [L_*, L^*]$. Then the following holds true:*

(a) *For any $x_0 \in \mathbb{R}^d$ and for $C_{\mathrm{Kol}}(\beta, L; \rho)$ defined in (2.5.25), the estimator $\widetilde{g}_T^j$ defined according to (2.5.22) is sharp adaptive.*

(b) *If there exists an estimator $\breve{g}_T$ such that, for some $\beta_0 \geq \beta_*$, $L > 0$,*

$$\limsup_{T\to\infty} \left(\psi_{T,\beta_0}\ C_{\mathrm{Kol}}(\beta_0, L; \rho)\right)^{-2}\ \mathscr{R}_{T,\beta_0,L}(\breve{g}_T) < 1, \tag{2.5.26}$$

then there exists $\beta_0' > \beta_0$ such that

$$\frac{\mathscr{R}_{T,\beta_0',L}(\breve{g}_T)}{\mathscr{R}_{T,\beta_0',L}(\widetilde{g}_T^j)} \geq \Psi_T\ \frac{\mathscr{R}_{T,\beta_0,L}(\widetilde{g}_T^j)}{\mathscr{R}_{T,\beta_0,L}(\breve{g}_T)}, \tag{2.5.27}$$

where $\Psi_T \to_{T\to\infty} \infty$. In particular, for any fixed $\beta \geq \beta_*$, $L > 0$,

$$\limsup_{T\to\infty} \left(\psi_{T,\beta}\ C_{\mathrm{Kol}}(\beta, L; \rho)\right)^{-2}\ \mathscr{R}_{T,\beta,L}(\widetilde{g}_T^j(x_0)) = 1.$$

The statement in part (b) is to be interpreted in the sense that, whenever there exists an estimator $\breve{g}_T$ which performs better than the estimator $\widetilde{g}_T^j$ at least for one smoothness degree β_0, there exists another smoothness factor β_0' for which there is much greater loss of $\breve{g}_T$. The assertion (and its respective proof) are to be compared with Theorem 2 in [KT04].

2.5.4. Extension: Estimation of $\operatorname{div}(a\rho)$ with common bandwidth choice for all components

The bandwidth $\widehat{h}_{T,\beta}^j$ and the threshold $\widehat{\eta}_{T,\beta}^j$ which are used for defining the adaptive estimators of $\operatorname{div}(a_j\rho)$ introduced in the previous section depend on specific entries of the diffusion matrix a and need to be chosen individually for any $j \in \{1, \ldots, d\}$.

Similarly to the global setting, we shall now briefly address the question of *vector-wise* estimation of the divergences (as a pre-stage for estimation of the drift vector). Given any vector-valued function $f : \mathbb{R}^d \to \mathbb{R}^d$ with components $f^j \in L^2(\mathbb{R}^d)$, $j \in \{1, \ldots, d\}$, and $\beta > d/2$, denote

$$\widetilde{\eta}_\beta(f) := \left(\frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \|\lambda\|^{2\beta} \sum_{j=1}^d \left|\phi_{f^j}(\lambda)\right|^2 \mathrm{d}\lambda\right)^{1/2}. \tag{2.5.28}$$

Recall the definition of the sets $\Pi^k(\beta)$ of measurable drift functions $b : \mathbb{R}^d \to \mathbb{R}^d$, $k \in \{1, \ldots, d\}$, as they were introduced in Section 2.5.1. For any $\beta > d/2$ and $L > 0$, define

$$\Pi(\beta, L) = \Pi_\epsilon(\beta, L; \rho, \sigma) := \left\{ b \in \bigcap_{k=1}^d \Pi^k(\beta) : \widetilde{\eta}_\beta^2(\operatorname{div}(a\rho)) \leq 4L^2 d,\ \rho(x_0) \geq \epsilon \right\}.$$

We proceed with describing the announced adaptive procedure for estimating the entire vector $\operatorname{div}(a\rho)$ at once. As in the case of component-wise estimation, the scheme starts with defining a preliminary density estimator $\widehat{\rho}_T$ according to (2.5.15) and (2.5.16). The adaptive estimator of $\operatorname{div}(a\rho)(x_0)$ is selected among the family of estimators $\widehat{g}_{T,\beta}(x_0)$, defined as

$$\widehat{g}_{T,\beta}(x_0) := \frac{2}{T\widehat{h}_{T,\beta}^d}\int_0^T K_\beta\left(\frac{X_u - x_0}{\widehat{h}_{T,\beta}}\right)\mathrm{d}X_u,$$

where

- the kernel
$$K_\beta(x) := \mathsf{b}^d\widetilde{K}_\beta(\mathsf{b}x), \quad \text{for} \quad \mathsf{b} = \mathsf{b}(\beta) := \left(\frac{2\beta - d}{d}\right)^{1/(2\beta)},$$
coincides with the kernel which is used for constructing the estimators of the components of $\operatorname{div}(a\rho)$,
- the bandwidth $\widehat{h}_{T,\beta}$ is defined as
$$\widehat{h}_{T,\beta} := \left(\frac{4\widehat{\rho}_T(x_0)\|\sigma\|_{S_2}^2\log T}{\beta T}\right)^{1/(2\beta)},$$
and
- the thresholding sequence $\widehat{\eta}_{T,\beta}$ is taken as
$$\widehat{\eta}_{T,\beta} := \widehat{h}_{T,\beta}^{\beta - d/2}\|K_\beta\|_{L^2(\mathbb{R}^d)}.$$

The adaptive estimator $\widetilde{g}_T$ of $\operatorname{div}(a\rho)$ is finally defined as

$$\widetilde{g}_T(x_0) := \widehat{g}_{T,\widehat{\beta}_T}(x_0), \tag{2.5.29}$$

where

$$\widehat{\beta}_T := \max\left\{\beta \in \mathcal{G}_T : \|\widehat{g}_{T,\gamma}(x_0) - \widehat{g}_{T,\beta}(x_0)\| \le \sqrt{d}\;\widehat{\eta}_{T,\gamma}\ \forall\, \gamma \in \mathcal{G}_T,\ \gamma \le \beta\right\}. \tag{2.5.30}$$

A slight modification of the proof of Theorem 2.5.9 allows to formulate an upper bound on the mean-squared error of the adaptive estimator $\widetilde{g}_T$. For any $\beta > d/2$, $L > 0$, denote

$$C(\beta, L; \rho, \sigma) := 2L^{d/(2\beta)}\frac{2\beta}{\sqrt{d}}\left(\frac{d\|\sigma\|_{S_2}^2\rho(x_0)}{\beta(2\beta - d)}\right)^{\frac{\beta - d/2}{2\beta}}\mathbb{I}_\beta. \tag{2.5.31}$$

Theorem 2.5.11. *Grant Assumptions* (**BI**) *and* (**SG+**). *Then, for fixed* $\beta_* > d/2$, $0 < L_* < L^* < \infty$, $\delta_2 \in (0,1)$ *and for*

$$\mathcal{B}_T := [\beta_*, \beta_T] \times [L_*, L^*], \qquad \text{where } \beta_T := (\log\log T)^{\delta_2},$$

the adaptive estimator $\widetilde{g}_T$ *defined according to (2.5.29) satisfies, for any* $x_0 \in \mathbb{R}^d$,

$$\limsup_{T\to\infty}\sup_{(\beta,L)\in\mathcal{B}_T}\sup_{b\in\Pi(\beta,L)}\left(\psi_{T,\beta}\;C(\beta, L; \rho, \sigma)\right)^{-2}\mathbf{E}_b\left\|\widetilde{g}_T(x_0) - \operatorname{div}(a\rho)(x_0)\right\|^2 \le 1. \tag{2.5.32}$$

For the proof of Theorem 2.5.11, we refer to Section 2.6.2.

2.5.5. Discussion and central ingredients of the proof of Theorem 2.5.9

Let us first arrange the constants identified in the previous sections and relate it to known results on asymptotically exact adaptive estimation with respect to pointwise risk over the scale of Sobolev classes. Tsybakov (1998) considers the problem of nonparametric function estimation in the Gaussian white noise model (in the one-dimensional case), assuming that the unknown function belongs to some Sobolev class with unknown regularity parameter. The question of density estimation at a fixed point $x_0 \in \mathbb{R}$ is investigated by Butucea (2001). Since the variance of the proposed kernel estimator is proportional to the value of the unknown density f at x_0, the value $f(x_0)$ appears in the exact normalization. [KT04] deal with nonparametric estimation of a multivariate function and its partial derivatives at a fixed point when the Riesz transform is observed in Gaussian white noise. In particular, [KT04] find the exact constant for nonparametric estimation of a function $f : \mathbb{R}^d \to \mathbb{R}$, observed in Gaussian white noise and satisfying $\eta_\beta^2(f) \le L^2$. In combination with the results of Butucea (2001) on classical density estimation (in dimension $d = 1$), the exact constant involving the value of the unknown density $f(x_0)$ to be estimated is then identified as

$$L^{\frac{d}{2\beta}} \ \frac{2\beta}{d} \ \left(\frac{d^2 \ f(x_0)}{\beta(2\beta - d)}\right)^{\frac{\beta-d/2}{2\beta}} \quad \mathbb{I}_\beta. \tag{2.5.33}$$

For the case of Kolmogorov diffusions ($\sigma \equiv \mathbf{Id}_{d\times d}$), $\widetilde{C}_j(\beta, L; \rho, \mathbf{Id}_{d\times d})$ coincides with the constant in (2.5.33). The more general problem of estimating $\operatorname{div}(a_j\rho)$, $j \in \{1, \dots, d\}$, further depends on the form of the diffusion coefficient. This dependence is reflected by the appearance of the j-th diagonal entry of the diffusion matrix $a = \sigma\sigma^t$ in the optimal constants $\widetilde{C}_j(\beta, L; \rho, \sigma)$ (for estimation of the drift component b^j) and $C_j(\beta, L; \rho, \sigma)$ (for estimating $\operatorname{div}(a_j\rho)$). Comparing the latter value to the constant appearing in the upper bound which is found for the vector-valued adaptive estimator $\widetilde{g}_T$ defined in Section 2.5.4, we obtain by means of Jensen's inequality

$$\begin{aligned}
\sum_{j=1}^d C_j^2(\beta, L; \rho, \sigma) &= \left(L^{\frac{d}{2\beta}} \ \frac{4\beta}{d} \left(\frac{d^2\rho(x_0)}{\beta(2\beta - d)}\right)^{\frac{\beta-d/2}{2\beta}} \quad \mathbb{I}_\beta\right)^2 \quad \frac{1}{d}\sum_{j=1}^d a_{jj}^{\frac{\beta-d/2}{\beta}} \\
&\le \left(L^{\frac{d}{2\beta}} \ \frac{4\beta}{d} \left(\frac{d^2\rho(x_0)}{\beta(2\beta - d)}\right)^{\frac{\beta-d/2}{2\beta}} \quad \mathbb{I}_\beta\right)^2 \ \left(\frac{1}{d}\sum_{j=1}^d a_{jj}\right)^{\frac{\beta-d/2}{\beta}} \qquad = C^2(\beta, L; \rho, \sigma).
\end{aligned}$$

This again shows that the approach of component-wise estimation in principle is to be preferred to implementing one single vector-valued estimator.

The above estimation *procedure* is inspired by the method of [KT01] and [KT04], and the *problem* of pointwise drift estimation considered in this section was investigated before by Spokoiny (2000). We shall now briefly point out differences to these works, and we will introduce some important auxiliary results required for the proof of the upper bounds in (2.5.24) and (2.5.32).

The optimal recovery kernel and some of its properties. Spokoiny (2000) suggests a data-driven choice of the bandwidth for defining estimators of the drift of a diffusion at some fixed point $x_0 \in \mathbb{R}$. His non-asymptotic results apply to arbitrary smooth, nonnegative kernel functions which are bounded by 1 and vanish outside of $[-1, 1]$. In contrast, the asymptotically exact procedures suggested in Butucea (2001), [KT01] and [KT04] specify the type of kernel involved in the definition of the estimators. It is conjectured by [KT01] (see their Remark 3) that, for attaining *optimal* estimators, not only the smoothing parameter but also the kernel function should be chosen in a data-driven way.

The kernel K_β which is used for defining the estimators $\widehat{g}^j_{T,\beta}$ (see (2.5.17)) is given as a solution of an optimal recovery problem. Let us briefly recall the general framework. Consider some functional T which is assumed to be linear on

$$\mathcal{D} \subset \{f : \mathbb{R}^d \to \mathbb{R},\ \zeta_\beta(f) < \infty\},$$

where $\zeta_\beta(\cdot)$ denotes a seminorm, and suppose that there exists some $r \geq 0$ such that

$$\forall\ a \geq 0,\ b > 0,\ f \in \mathcal{D}, \qquad T(af(b\ \cdot)) = ab^r T(f(\cdot)).$$

In addition, the modulus of continuity is supposed to be well-defined, that is, for all $\beta > r$, $L > 0$, $\varepsilon > 0$,

$$\omega_{\beta,L}(\varepsilon) := \sup\left\{T(f) : \int_{\mathbb{R}^d} f^2(x)\mathrm{d}\lambda\!\!\lambda(x) \leq \varepsilon,\ \zeta_\beta(f) \leq L\right\} < \infty.$$

Under the given assumptions on the functional T, there exist

- a function $g_{\beta,L,\varepsilon}$ such that
$$T\left(g_{\beta,L,\varepsilon}\right) = \omega_{\beta,L}(\varepsilon), \tag{2.5.34}$$
and
- some function $K_\beta \in L^2(\mathbb{R}^d)$, solving the optimal recovery problem
$$\sup_{\substack{\zeta_\beta(f)\leq 1,\\ \|f-g\|_{L^2(\mathbb{R}^d)}\leq 1}} \left|\int_{\mathbb{R}^d} K_\beta g - T(f)\right| = \inf_K \sup_{\substack{\zeta_\beta(f)\leq 1,\\ \|f-g\|_{L^2(\mathbb{R}^d)}\leq 1}} \left|\int_{\mathbb{R}^d} Kg - T(f)\right| =: E(\beta)$$
and, at the same time, satisfying
$$E(\beta) = T\left(g_{\beta,1,1}\right) = \sup_{\zeta_\beta(f)\leq 1} \left|\int_{\mathbb{R}^d} K_\beta f - T(f)\right| + \|K_\beta\|_{L^2(\mathbb{R}^d)}. \tag{2.5.35}$$

For the corresponding references, we refer to [KT01], p. 1571. For the Sobolev seminorm, the extremal function $g_{\beta,1,1}$ in dimension $d = 1$ is given in Taikov (1968). We use results of [KT01] who consider the multidimensional case $d \geq 1$.

Lemma 2.5.12. *Let $\beta > d/2$, and, for $\widetilde{K}_\beta(\cdot)$ and $\mathsf{b} = \mathsf{b}(\beta)$ defined according to (2.5.19) and (2.5.20), respectively, let*

$$\begin{aligned} K^*_\beta(x) &:= \mathbb{I}_\beta^{-1}\mathsf{b}^{-\beta+d/2}\ \widetilde{K}_\beta(\mathsf{b}x) \qquad (2.5.36)\\ &= \left(\frac{1}{2\beta}\ B\left(1+\frac{d}{2\beta}, 1-\frac{d}{2\beta}\right)(2\pi)^{-d}\mathbb{S}_d\right)^{-1/2} \mathsf{b}^{-\beta+d/2}\ \widetilde{K}_\beta(\mathsf{b}x). \end{aligned}$$

(a) (cf. Proposition 1 in [KT01]) *Consider the Sobolev seminorm η_β defined in (2.5.4). For the functional $T(f) = f(0)$, the extremal function $g_{\beta,1,1}$ is given as $g_{\beta,1,1} = K^*_\beta$. Furthermore,*
$$\widetilde{K}_\beta(0) = (2\pi)^{-d}\int_{\mathbb{R}^d}\left(1+\|\lambda\|^{2\beta}\right)^{-1}\mathrm{d}\lambda = \frac{2\beta}{d}\ \mathbb{I}^2_\beta, \tag{2.5.37}$$
and, for $K_\beta(x) = \mathsf{b}^d\widetilde{K}_\beta(\mathsf{b}x)$,
$$\|K_\beta\|_{L^2(\mathbb{R}^d)} = \mathbb{I}_\beta\left(\frac{2\beta-d}{d}\right)^{\frac{\beta+d/2}{2\beta}} = \mathbb{I}_\beta\ \mathsf{b}^{\beta+d/2}. \tag{2.5.38}$$
In particular,
$$g_{\beta,1,1}(0) = \|K_\beta\|_{L^2(\mathbb{R}^d)} + \mathsf{b}_{\beta,\beta}, \tag{2.5.39}$$
where
$$\mathsf{b}_{\beta,\beta} := (\mathsf{b}(\beta))^{d/2-\widetilde{\beta}}\left((2\pi)^{-d}\int_{\mathbb{R}^d}\frac{\|\lambda\|^{2\beta}}{\left(1+\|\lambda\|^{2\beta}\right)^2}\mathrm{d}\lambda\right)^{1/2}.$$

(b) *For fixed $\delta \in (0,1)$, there exists some compactly supported modification $\overline{K}_\beta$ of K^*_β which enjoys the following properties,*

(k1) $\|\overline{K}_\beta\|_{L^2(\mathbb{R}^d)} \leq 1-\delta/2$,

(k2) $\eta_\beta(\overline{K}_\beta) \leq 1-\delta/2$, *and*

(k3) $(1-\delta/2)K^*_\beta(0) \leq \overline{K}_\beta(0) \leq K^*_\beta(0)$.

Proof. For the proof, we refer to the proofs of Proposition 1 in [KT01] and of Lemma 10 Tsybakov (1998). □

The relation (2.5.35) is crucial for the proof of the upper bound of the adaptive procedures in [KT01] and [KT04]. In particular, it allows to prove a decomposition of the normalizing factor into its bias and variance components (cf. Lemma 2(ii) in [KT01] and Section 4.1.2 in [KT04]). A respective decomposition of the normalizing factor $\psi_{T,\beta}\ C_j(\beta, L; \rho, \sigma)$ which appears in the upper bound (2.5.24) in Theorem 2.5.9 is given in Lemma 2.6.6 below.

Exponential bounds for the stochastic error. It was already noted in connection with the proof of the lower bound for pointwise estimation of the components of $\operatorname{div}(a\rho)$ that the complexity of the ergodic diffusion model requires a rather involved analysis. The same remark applies to the proof of exact adaptivity of the procedure for which exponential bounds on the stochastic error are needed. In the Gaussian white noise framework, the derivation of such exponential bounds is straightforward due to the Gaussian nature of the model. Indeed, letting $K_{\beta,h} := h^{-d}K_\beta(\cdot/h)$ and denoting by $Z_\beta := \varepsilon \int_{\mathbb{R}^d} K(x)\mathrm{d}B(x)$ the stochastic part of the kernel estimator $T_{\beta,\varepsilon} := \int_{\mathbb{R}^d} K_{\beta,h}(x)\mathrm{d}Y_\varepsilon(x)$ in the Gaussian white noise model (2.5.3) considered in [KT04], it holds for any $u > 0$,

$$\mathbf{E}\big(|Z_\beta|^2\ \mathbb{1}\{|Z_\beta| \geq u\}\big) \leq 2(\sigma_\beta^2 + u^2)\ \exp\left(-\frac{u^2}{2\sigma_\beta^2}\right),$$

where σ_β is the standard deviation of the linear estimator $T_{\beta,\varepsilon}$.

An additional complication arises in the (classical) pointwise density estimation problem considered in Butucea (2001). The value of the unknown density f at the estimation point x_0 appears in the exact constant since the variance is proportional to $f(x_0)$, and the optimal density estimators involve an estimator of this value. Consequently, one has to bound *uniform* risks of the adaptive estimator. To do so, Butucea (2001) uses the classical Bernstein inequality and a uniform exponential inequality due to van de Geer (2000). The Bernstein inequality for sequences of i.i.d. random variables (cf., e.g., Lemma 2.2.9 in van der Vaart and Wellner (1996)) is of substantial importance in probability theory and its applications. In particular, it provides a description of the tail behavior of stochastic processes which can be used to initiate chaining procedures and to extend deviation inequalities to *uniform* deviation inequalities.

Similarly to the pointwise density estimation problem, the bandwidths used for defining the estimators in our selection procedure for estimating $\operatorname{div}(a_j\rho)$ involve an estimator $\widehat{\rho}_T(x_0)$ of the (unknown) value of the invariant density at x_0. The basic idea is to consider random events on which the bandwidths $\widehat{h}^j_{T,\beta}$, $j \in \{1,\ldots,d\}$, as defined in (2.5.18) are close to their deterministic counterparts

$$\left(\frac{4d\rho(x_0)a_{jj}\log T}{\beta T}\right)^{1/(2\beta)}, \qquad j \in \{1,\ldots,d\},$$

and to show that the probability of the complement of these events is asymptotically negligible. Consequently, one also has to derive risk bounds which hold uniformly over deterministic counterparts of the above random events.

The stochastic error of the estimators in the diffusion model involves a stochastic integral which requires a slightly different approach for proving uniform exponential inequalities. We shall decompose the stochastic integral into the sum of an additive functional of the diffusion and a continuous martingale. Diverse works on generalizations of the classical Bernstein inequality to (additive functionals of) Markov processes or to the case of weakly dependent observations exist. For a recent account on this question, we refer to Section 1.1 in Gao et al. (2010). Combined with chaining arguments and conditions on the size of function classes in terms of bracketing numbers, (**BI**), together with Bernstein's inequality for continuous martingales, allows to derive a uniform exponential inequality which is comparable to Theorem 5.11 in van de Geer (2000). Denote by $N_{[\,]}(\varepsilon,\mathcal{F},L^2(\mu))$ the ε-entropy with bracketing, that is, the smallest number of ε-brackets (in $L^2(\mu)$) which are required to cover $\mathcal{F}$ (cf. van der Vaart and Wellner (1996), Definition 2.1.6).

Lemma 2.5.13. (a) *Consider the ergodic diffusion process solution of the SDE (2.5.13), assume that X is symmetric with respect to μ, and suppose that the associated carré du champs satisfies Poincaré's inequality. Let $f \in L^2(\mu)$ be some bounded measurable function, fix $j \in \{1,\ldots,d\}$, and assume that there exists some positive constant B such that*

$$\max\left\{\sup_{x\in\operatorname{supp}(f)}|b^j(x)|,\ \sup_{x\in\operatorname{supp}(f^2)}|b^j(x)|^2\right\} \leq B. \tag{2.5.40}$$

Then Assumption (**BI**) *is satisfied. Furthermore, it holds, for any $T, r > 0$,*

$$\mathbf{P}_b\left(\left|\frac{1}{T}\int_0^T f(X_u)b^j(X_u)\mathrm{d}u - \int_{\mathbb{R}^d} f(y)b^j(y)\mathrm{d}\mu(y)\right| > r\right) \leq 2\exp\left(-\frac{T}{4Bc_P}\ \min\left\{\frac{r^2}{\|f\|^2_{L^2(\mu)}}, \frac{r}{\|f\|_\infty}\right\}\right).$$

(b) *Let $\mathcal{F} \subset L^2(\mu)$ be some class of measurable functions $f : \mathbb{R}^d \to \mathbb{R}$, and assume that, for some positive constants K and M, it holds*

$$\sup_{f\in\mathcal{F}}\|f\|_\infty \leq K, \qquad \sup_{f\in\mathcal{F}}\|f\|_{L^2(\mu)} \leq M.$$

Grant Assumptions (**BI**) *and* (**SG**). *Then, for arbitrary $T > 0$ and any positive r satisfying, for some positive constants K_1 and K_2,*

$$\frac{K_1}{\sqrt{T}}\int_0^1 \max\left\{\sqrt{\log N_{[\,]}(\varepsilon,\mathcal{F},L^2(\mu))}, 1\right\}\mathrm{d}\varepsilon\ \leq\ r\ \leq\ \frac{K_2\ M^2}{K},$$

there exist some positive constants C_1 and C_2 such that

$$\mathbf{P}_b\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{T}\int_0^T f(X_u)\mathrm{d}X_u^j - \int_{\mathbb{R}^d} f(y)b^j(y)\mathrm{d}\mu(y)\right| > r\right) \leq C_1\exp\left(-\frac{C_2Tr^2}{M^2}\right). \tag{BI+}$$

The proof of Lemma 2.5.13 is given in Section 2.6.2.

Remark 2.5.14 (Exponential inequalities for the Euclidean norm of vector-valued additive functionals). For the proof of the upper bound in Theorem 2.5.11, we require exponential inequalities for the Euclidean norm of the stochastic error of vector-valued estimators. In the setting of Lemma 2.5.13, such inequalities follow immediately by applying (**BI**) component-wise. Grant Assumptions (**BI**) and (**SG**). Given any $\mathcal{F} \subset L^2(\mu)$ satisfying the assumptions of Lemma

2.5.13(b), it holds for any fixed $r, T > 0$ and some positive constants $C_B(k)$, $k = 1, \ldots, d$,

$$
\begin{aligned}
\mathbf{P}_b\bigg(\bigg\|\frac{1}{T}\int_0^T & f(X_u)b(X_u)\mathrm{d}u - \int_{\mathbb{R}^d} f(y)b(y)\mathrm{d}\mu(y)\bigg\| > r\bigg) \\
&= \mathbf{P}_b\left(\sum_{k=1}^d \left|\frac{1}{T}\int_0^T f(X_u)b^k(X_u)\mathrm{d}u - \int_{\mathbb{R}^d} f(y)b^k(y)\mathrm{d}\mu(y)\right|^2 > r^2\right) \\
&\leq \sum_{k=1}^d \mathbf{P}_b\left(\left|\frac{1}{T}\int_0^T f(X_u)b^k(X_u)\mathrm{d}u - \int_{\mathbb{R}^d} f(y)b^k(y)\mathrm{d}\mu(y)\right| > \frac{r}{\sqrt{d}}\right) \\
&\leq 2\sum_{k=1}^d \exp\left(-\frac{Tr^2/d}{2C_B(k)(c_P M^2 + Kr/\sqrt{d})}\right) = 2d\exp\left(-\frac{Tr^2}{C_B'(M^2+Kr)}\right),
\end{aligned}
$$

for some positive constant C_B' depending only on B, σ, c_P and d.

2.6. Proofs for Section 2.5

2.6.1. Proof of lower bounds

Proof of Lemma 2.5.3. The lower bound in (2.5.7) is proven by bounding it from below by the average error by means of (the Kullback version of) Theorem 2.2 in Tsybakov (2009). For $x_0 \in \mathbb{R}^d$ fixed and functions $f, g : \mathbb{R}^d \to \mathbb{R}$, set $d(f,g) := |f(x_0) - g(x_0)|$. Fix $\beta > d/2$, $L > 0$, and denote

$$
\overline{\psi}_{T,\beta} := T^{-\frac{\beta - d/2}{2\beta}}, \qquad \alpha := \left(\sum_{k=1}^d a_{jk}^2\right)^{1/2}.
$$

Using the standard general reduction scheme (cf. pp. 79–80 in Tsybakov (2009)), the proof of the lower bound in (2.5.7) reduces to verifying that

$$
\inf_{\widehat{g}_T} \max_{g \in \{g_0, g_1\}} \mathbf{P}_g\Big(|\widehat{g}_T(x_0) - g(x_0)| \geq A\,\overline{\psi}_{T,\beta}\Big) \geq c' > 0,
$$

where

- the infimum extends over all estimators $\widehat{g}_T$ of $g \in \Pi^j(\beta, L)$,
- the hypotheses g_0, g_1 satisfy for some positive constant A the relation

$$
d(g_0, g_1) \geq 2A\,\overline{\psi}_{T,\beta}, \tag{2.6.1}
$$

- $\mathbf{P}_{g_0}, \mathbf{P}_{g_1}$ are the probability measures associated with g_0 and g_1, respectively, and
- c' is independent of T.

(I) Construction of the hypotheses. Consider some positive density function

$$
\rho \in C_c^\infty(\mathbb{R}^d) \cap \mathcal{S}(\beta + 1, L/\alpha). \tag{2.6.2}
$$

For small $\delta \in (0, 1/2)$ fixed, define

$$
\rho_0(x) := \delta^{1/(\beta+3/2)}\,\rho\left(x\delta^{1/(\beta+3/2)}\right), \tag{2.6.3}
$$

$$
g_0(x) := 2\sum_{k=1}^d a_{jk}\partial_k \rho_0(x). \tag{2.6.4}
$$

Let $K : \mathbb{R}^d \to [0,\infty)$ be some compactly supported kernel function satisfying $\eta_{\beta+1}(K) \leq \alpha^{-1}$. For $h_{T,\beta} := c_0 T^{-1/(2\beta)}$, where c_0 is some positive constant to be specified later, set

$$g_{T,\beta}(x) := (1-\delta)\ L\ h_{T,\beta}^{\beta+1-d/2}\ K\left(\frac{x-x_0}{h_{T,\beta}}\right). \tag{2.6.5}$$

Define

$$\rho_1(x) = \rho_{T,1}(x) := \rho_0(x)\Big(1 - \int_{\mathbb{R}^d} g_{T,\beta}(y)\mathrm{d}y\Big) + g_{T,\beta}(x),$$

and consider the hypothesis

$$\begin{aligned} g_1(x) = g_{T,1}(x) &:= 2\sum_{k=1}^d a_{jk}\ \partial_k\rho_1(x) \qquad (2.6.6)\\ &= \left(1 - \int_{\mathbb{R}^d} g_{T,\beta}(y)\mathrm{d}y\right)\ 2\sum_{k=1}^d a_{jk}\partial_k\rho_0(x) + 2\sum_{k=1}^d a_{jk}\partial_k g_{T,\beta}(x)\\ &= \left(1 - \int_{\mathbb{R}^d} g_{T,\beta}(y)\mathrm{d}y\right)\ g_0(x) + 2\sum_{k=1}^d a_{jk}\partial_k g_{T,\beta}(x). \qquad (2.6.7)\end{aligned}$$

Define the vector-valued functions b_0, b_1 as

$$b_0 := a\nabla\log\rho_0, \qquad b_1 = b_{T,1} := a\nabla\log\rho_{T,1}. \tag{2.6.8}$$

(II) Properties of the hypotheses. By the very definition of the respective quantities, it holds

$$2b_i^j\rho_i = 2\sum_{k=1}^d a_{jk}\partial_k(\log\rho_i)\rho_i = 2\sum_{k=1}^d a_{jk}\partial_k\rho_i = g_i, \qquad i\in\{0,1\}.$$

The Fourier transform of ρ_0 at the point $\lambda\in\mathbb{R}^d$ is given by

$$\begin{aligned}\phi_{\rho_0}(\lambda) &= \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t x}\rho_0(x)\mathrm{d}x = \int_{\mathbb{R}^d}\exp\left(\mathrm{i}\lambda^t y\delta^{-1/(\beta+3/2)}\right)\rho(y)\mathrm{d}y\\ &= \phi_\rho\left(\lambda\delta^{-1/(\beta+3/2)}\right), \qquad (2.6.9)\end{aligned}$$

and, using integration by parts, one obtains

$$\begin{aligned}\phi_{g_0}(\lambda) &= 2\sum_{k=1}^d a_{jk}\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t x}\partial_k\rho_0(x)\mathrm{d}x = -2\mathrm{i}\sum_{k=1}^d a_{jk}\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t x}\lambda_k\rho_0(x)\mathrm{d}x\\ &= -2\mathrm{i}\sum_{k=1}^d a_{jk}\lambda_k\phi_{\rho_0}(\lambda). \qquad (2.6.10)\end{aligned}$$

Thus, applying Cauchy–Schwarz,

$$\begin{aligned}\eta_\beta^2(g_0) &\overset{(2.5.4)}{=} (2\pi)^{-d}\int_{\mathbb{R}^d}\|\lambda\|^{2\beta}|\phi_{g_0}(\lambda)|^2\mathrm{d}\lambda\\ &\overset{(2.6.10)}{=} 4(2\pi)^{-d}\int_{\mathbb{R}^d}\|\lambda\|^{2\beta}\Big|\sum_{k=1}^d a_{jk}\lambda_k\phi_{\rho_0}(\lambda)\Big|^2\mathrm{d}\lambda\\ &\leq 4(2\pi)^{-d}\alpha^2\int_{\mathbb{R}^d}\|\lambda\|^{2(\beta+1)}\left|\phi_\rho(\lambda\delta^{-1/(\beta+3/2)})\right|^2\mathrm{d}\lambda\\ &= 4\delta^2\alpha^2(2\pi)^{-d}\int_{\mathbb{R}^d}\|\lambda\|^{2(\beta+1)}|\phi_\rho(\lambda)|^2\mathrm{d}\lambda\\ &\overset{(2.5.4)}{=} 4\delta^2\ \alpha^2\ \eta_{\beta+1}^2(\rho) \overset{(2.6.2)}{\leq} 4(L\delta)^2 \leq 4L^2. \qquad (2.6.11)\end{aligned}$$

Furthermore,

$$\begin{aligned}\int_{\mathbb{R}^d} g_{T,\beta}(y)\mathrm{d}y &= L\ (1-\delta)\ h_{T,\beta}^{\beta+1-d/2}\ \int_{\mathbb{R}^d} K\left(\frac{y-x_0}{h_{T,\beta}}\right)\mathrm{d}y \\ &= L\ (1-\delta)\ h_{T,\beta}^{\beta+1+d/2}\int_{\mathbb{R}^d} K(u)\mathrm{d}u = O\Big(T^{-\frac{\beta+1+d/2}{2\beta}}\Big). \end{aligned} \tag{2.6.12}$$

For any $\lambda \in \mathbb{R}^d$, the Fourier transform of $\sum_{k=1}^d a_{jk}\partial_k g_{T,\beta}$ can be written as

$$\begin{aligned}\phi_{\sum_{k=1}^d a_{jk}\partial_k g_{T,\beta}}(\lambda) &= \sum_{k=1}^d a_{jk}\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t x}\partial_k g_{T,\beta}(x)\ \mathrm{d}x \\ &= -\mathrm{i}\ \sum_{k=1}^d a_{jk}\lambda_k \int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t x}\ g_{T,\beta}(x)\ \mathrm{d}x \\ &= -\mathrm{i}L\ (1-\delta)\ h_{T,\beta}^{\beta+1-d/2}\ \sum_{k=1}^d a_{jk}\lambda_k\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t x}K\left(\frac{x-x_0}{h_{T,\beta}}\right)\ \mathrm{d}x \\ &= -\mathrm{i}L\ (1-\delta)\ h_{T,\beta}^{\beta+1+d/2}\ \mathrm{e}^{\mathrm{i}\lambda^t x_0}\sum_{k=1}^d a_{jk}\lambda_k\int_{\mathbb{R}^d}\mathrm{e}^{\mathrm{i}\lambda^t z h_{T,\beta}}K(z)\mathrm{d}z \\ &= -\mathrm{i}L\ (1-\delta)\ h_{T,\beta}^{\beta+1+d/2}\ \mathrm{e}^{\mathrm{i}\lambda^t x_0}\sum_{k=1}^d a_{jk}\lambda_k\phi_K(h_{T,\beta}\lambda)\end{aligned}$$

such that

$$\begin{aligned}\eta_\beta^2\Big(\sum_{k=1}^d a_{jk}\partial_k g_{T,\beta}\Big) &= (2\pi)^{-d}\int_{\mathbb{R}^d}\|\lambda\|^{2\beta}|\phi_{\sum_{k=1}^d a_{jk}\partial_k g_{T,\beta}}(\lambda)\big|^2\mathrm{d}\lambda \\ &= (2\pi)^{-d}\int_{\mathbb{R}^d}\|\lambda\|^{2\beta}\big|\ \mathrm{i}L\ (1-\delta)\ h_{T,\beta}^{\beta+1+d/2}\mathrm{e}^{\mathrm{i}\lambda^t x_0}\ \sum_{k=1}^d a_{jk}\lambda_k\phi_K(h_{T,\beta}\lambda)\big|^2\mathrm{d}\lambda \\ &\le L^2\ (1-\delta)^2\ \alpha^2\ (2\pi)^{-d}\int_{\mathbb{R}^d}\|\lambda\|^{2(\beta+1)}\ |\phi_K(\lambda)|^2\ \mathrm{d}\lambda \\ &\le L^2\ (1-\delta)^2\ \alpha^2\ \eta_{\beta+1}^2(K)\le L^2\ (1-\delta)^2.\end{aligned}$$

Thus,

$$\begin{aligned}\eta_\beta(g_{T,1}) &\le \left(1-\int_{\mathbb{R}^d} g_{T,\beta}(y)\mathrm{d}y\right)\ \eta_\beta(g_0)+2\eta_\beta\Big(\sum_{k=1}^d a_{jk}\partial_k g_{T,\beta}\Big) \\ &\le 2L\delta+2L(1-\delta)=2L.\end{aligned} \tag{2.6.13}$$

Furthermore,

$$\begin{aligned}d\,(g_0,g_{T,1}) &= |g_0(x_0)-g_{T,1}(x_0)| \\ &\ge \left|\Big|2\sum_{k=1}^d a_{jk}\partial_k\rho_0(x_0)\int_{\mathbb{R}^d} g_{T,\beta}(y)\mathrm{d}y\Big|-\Big|2\sum_{k=1}^d a_{jk}\partial_k g_{T,\beta}(x_0)\Big|\right|.\end{aligned}$$

In view of (2.6.12) and since

$$\sum_{k=1}^d a_{jk}\partial_k g_{T,\beta}(x_0) = L\ (1-\delta)\ h_{T,\beta}^{\beta-d/2}\sum_{k=1}^d a_{jk}\ \partial_k K(0) = C_1\ \overline{\psi}_{T,\beta},$$

for $C_1 := c_0^{\beta-d/2}\ L\ (1-\delta)\sum_{k=1}^d a_{jk}\ \partial_k K(0)$, it follows that there exists some positive constant A such that (2.6.1) is satisfied.

(III) Application of Theorem 2.2 in Tsybakov (2009). For $b_0, b_{T,1}$ defined in (2.6.8), consider the SDEs

$$\mathrm{d}X_t = b_0(X_t)\mathrm{d}t + \sigma \ \mathrm{d}W_t, \qquad \mathrm{d}Y_t = b_{T,1}(Y_t)\mathrm{d}t + \sigma \ \mathrm{d}W_t,$$

and assume that the initial values X_0, Y_0 follow the respective invariant measures $\mathrm{d}\mu_0 := \rho_0 \mathrm{d}\lambda\!\!\lambda$ and $\mathrm{d}\mu_{T,1} := \rho_{T,1}\mathrm{d}\lambda\!\!\lambda$. Denote by $\mathbf{E}_i = \mathbf{E}_{b_i}$ expectation under the measure $\mathbf{P}_i = \mathbf{P}_{b_i}$ associated with the hypothesis $b = b_i$, $i \in \{0,1\}$. Letting $X^T := (X_t)_{0\le t\le T}$, the log-likelihood ratio $\log\left(\mathrm{d}\mathbf{P}_0/\mathrm{d}\mathbf{P}_1\right)(X^T)$ under $\mathbf{P}_0$ is given by

$$\begin{aligned}\frac{1}{2}\int_0^T &(b_0 - b_{T,1})^t(X_u)a^{-1}(b_0 - b_{T,1})(X_u)\mathrm{d}u\\ &+ \int_0^T \left(\sigma^{-1}(b_0 - b_{T,1})\right)^t (X_u)\mathrm{d}W_u + \log\left(\frac{\rho_0}{\rho_{T,1}}(X_0)\right)\end{aligned}$$

such that the Kullback–Leibler divergence $\mathrm{KL}(\mathbf{P}_0, \mathbf{P}_1)$ between the probability measures $\mathbf{P}_0$ and $\mathbf{P}_1$ associated to b_0 and $b_{T,1}$, respectively, can be written as

$$\begin{aligned}\mathrm{KL}(\mathbf{P}_0, \mathbf{P}_1) &:= \mathbf{E}_0\left(\log\frac{\mathrm{d}\mathbf{P}_0}{\mathrm{d}\mathbf{P}_1}(X^T)\right)\\ &= \mathbf{E}_0\left(\log\frac{\rho_0}{\rho_{T,1}}(X_0)\right) + \frac{T}{2}\ \mathbf{E}_0\left\|\sigma^{-1}(b_0 - b_{T,1})(X_0)\right\|^2\\ &= \mathbf{E}_0\left(\log\frac{\rho_0}{\rho_{T,1}}(X_0)\right) + \frac{T}{2}\ \mathbf{E}_0\left\|\sigma^t\nabla\left(\log\frac{\rho_0}{\rho_{T,1}}(X_0)\right)\right\|^2.\end{aligned}$$

The assertion (2.5.7) will be proven by means of the following

Lemma 2.6.1 (Theorem 2.2 in Tsybakov (2009), p. 90)**.** *Let $\mathbf{P}_0, \mathbf{P}_1$ be two probability measures on some measurable space $(\mathcal{X}, \mathscr{A})$. If $\mathbf{P}_0 \ll \mathbf{P}_1$ and*

$$\mathrm{KL}(\mathbf{P}_0, \mathbf{P}_1) = \int_{\mathcal{X}} \log\frac{\mathrm{d}\mathbf{P}_0}{\mathrm{d}\mathbf{P}_1}\ \mathrm{d}\mathbf{P}_0 \le \alpha < \infty,$$

then

$$\inf_{\Psi}\max_{j\in\{0,1\}} \mathbf{P}_j(\Psi \neq j) \ge \max\left\{\mathrm{e}^{-\alpha}/2, (1-\sqrt{\alpha/2})/2\right\}.$$

Plugging in the specific definitions of ρ_0 and $\rho_{T,1}$, it follows

$$\log\frac{\rho_0}{\rho_{T,1}} = \log\left(1 - \frac{g_{T,\beta}}{\rho_{T,1}}\right) - \log\left(1 - \int_{\mathbb{R}^d} g_{T,\beta}(y)\mathrm{d}y\right). \tag{2.6.14}$$

Note that, for any $k \in \{1, \ldots, d\}$,

$$\begin{aligned}\left|\partial_k \log\left(1 - \frac{g_{T,\beta}}{\rho_{T,1}}\right)\right| &= \left|\frac{\rho_{T,1}\ \partial_k g_{T,\beta} - g_{T,\beta}\ \partial_k\rho_{T,1}}{\rho_{T,1}\ (\rho_{T,1} - g_{T,\beta})}\right|\\ &= \left(1 - \int_{\mathbb{R}^d} g_{T,\beta}(y)\mathrm{d}y\right)\left|\frac{\rho_0\ \partial_k g_{T,\beta} - g_{T,\beta}\ \partial_k\rho_0}{\rho_{T,1}\ \rho_0\,(1 - \int_{\mathbb{R}^d} g_{T,\beta}(y)\mathrm{d}y)}\right|\\ &= \left|\frac{\partial_k g_{T,\beta} - g_{T,\beta}\ \partial_k(\log\rho_0)}{\rho_{T,1}}\right|.\end{aligned} \tag{2.6.15}$$

Consequently,

$$
\begin{aligned}
&\mathbf{E}_0\Big\|\sigma^t\nabla\log\frac{\rho_{T,0}}{\rho_{T,1}}(X_0)\Big\|^2\\
\overset{(2.6.14)}{=}\quad & \mathbf{E}_0\sum_{k=1}^d\left(\sum_{l=1}^d\sigma_{lk}\partial_l\left(\log\left(1-\frac{g_{T,\beta}}{\rho_{T,1}}\right)(X_0)-\log\left(1-\int_{\mathbb{R}^d}g_{T,\beta}(y)\mathrm{d}y\right)\right)\right)^2\\
\leq\quad & \mathbf{E}_0\sum_{k=1}^d\left(\sum_{l=1}^d\sigma_{lk}^2\right)\left(\sum_{m=1}^d\left(\partial_m\log\left(1-\frac{g_{T,\beta}}{\rho_{T,1}}\right)(X_0)\right)^2\right)\\
\overset{(2.6.15)}{=}\quad & \|\sigma\|_{S_2}^2\mathbf{E}_0\sum_{m=1}^d\left(\frac{g_{T,\beta}\partial_m\left(\log\rho_{T,0}\right)}{\rho_{T,1}}-\frac{\partial_m g_{T,\beta}}{\rho_{T,1}}\right)^2\\
\leq\quad & 2\|\sigma\|_{S_2}^2\left\{\mathbf{E}_0\left(\frac{g_{T,\beta}^2(X_0)}{\rho_{T,1}^2(X_0)}\|\nabla(\log\rho_{T,0})(X_0)\|^2\right)+\mathbf{E}_0\left(\frac{\|\nabla g_{T,\beta}(X_0)\|^2}{\rho_{T,1}^2(X_0)}\right)\right\}.
\end{aligned}
$$

The definition of $g_{T,\beta}$ entails that

$$
\begin{aligned}
\mathbf{E}_0\left(\frac{\|\nabla g_{T,\beta}(X_0)\|^2}{\rho_{T,1}^2(X_0)}\right) \quad &= \quad \int_{\mathbb{R}^d}\|\nabla g_{T,\beta}(y)\|^2\,\frac{\rho_{T,0}(y)}{\rho_{T,1}^2(y)}\mathrm{d}y\\
&\overset{(2.6.5)}{=}\quad L^2\,(1-\delta)^2\,h_{T,\beta}^{2(\beta+1)-d}\int_{\mathbb{R}^d}\sum_{i=1}^d\left(\partial_iK\left(\frac{y-x_0}{h_{T,\beta}}\right)\right)^2\frac{\rho_0(y)}{\rho_{T,1}^2(y)}\mathrm{d}y\\
&=\quad L^2\,(1-\delta)^2\,h_{T,\beta}^{2\beta-d}\int_{\mathbb{R}^d}\sum_{i=1}^d(\partial_iK)^2\left(\frac{y-x_0}{h_{T,\beta}}\right)\frac{\rho_0(y)}{\rho_{T,1}^2(y)}\mathrm{d}y\\
&\overset{(A.1.4)}{=}\quad L^2\,(1-\delta)^2\,h_{T,\beta}^{2\beta}\,\frac{\rho_0(x_0)}{\rho_{T,1}^2(x_0)}\int_{\mathbb{R}^d}\sum_{i=1}^d(\partial_iK)^2(y)\mathrm{d}y\,\left(1+o_T(1)\right)\\
&\leq\quad L^2\,(1-\delta^2)^2\,h_{T,\beta}^{2\beta}\,\rho_0^{-1}(x_0)\,\|\nabla K\|_{L^2(\mathbb{R}^d)}^2\,\left(1+o_T(1)\right), \qquad (2.6.16)
\end{aligned}
$$

where we used a modification of Bochner's lemma (Lemma A.1.4 in Section A.1). (2.6.16) follows from the fact that, for T large enough,

$$
\frac{\rho_0(x_0)}{\rho_{T,1}(x_0)}\leq\frac{\rho_0(x_0)}{\rho_0(x_0)\,(1-\int_{\mathbb{R}^d}g_{T,\beta}(y)\mathrm{d}y)}\leq\sqrt{1+\delta/2}.
$$

It is shown by similar arguments that

$$
\mathbf{E}_0\left(\frac{g_{T,\beta}^2(X_0)}{\rho_{T,1}^2(X_0)}\|\nabla(\log\rho_0)(X_0)\|^2\right)\quad=\quad\int_{\mathbb{R}^d}\frac{g_{T,\beta}^2(y)}{\rho_{T,1}^2(y)}\sum_{m=1}^d\partial_m(\log\rho_0(y))^2\rho_0(y)\mathrm{d}y=O\left(h_{T,\beta}^{2\beta+2}\right).
$$

This implies that there exists some positive constant α such that

$$
\begin{aligned}
\mathrm{KL}(\mathbf{P}_0,\mathbf{P}_1)\quad&=\quad\mathbf{E}_0\log\frac{\rho_0}{\rho_{T,1}}(X_0)+\frac{T}{2}\mathbf{E}_0\Big\|\sigma^{-1}(b_0-b_{T,1})(X_0)\Big\|^2\\
&\leq\quad\alpha Th_{T,\beta}^{2\beta}=\alpha<\infty.
\end{aligned}
$$

Lemma 2.6.1 then gives

$$
\begin{aligned}
\inf_{\widehat{g}_T}\sup_{b\in\Pi^j(\beta,L)}\mathbf{P}_b\left(\left|\widehat{g}_T(x_0)-g(x_0)\right|\geq A\,\overline{\psi}_{T,\beta}\right)&\geq\inf_{\widehat{g}_T}\max_{\ell=0,1}\mathbf{P}_\ell\left(\left|\widehat{g}_T(x_0)-g_\ell(x_0)\right|\geq A\,\overline{\psi}_{T,\beta}\right)\\
&\geq\max\left\{\mathrm{e}^{-\alpha}/4,(1-\sqrt{\alpha/2})/2\right\}.
\end{aligned}
$$

This yields the assertion. □

Proof of Theorem 2.5.5. The proof is similar to the preceding proof of Lemma 2.5.3 but relies on an application of Theorem 6.1 in Tsybakov (1998) instead of Theorem 2.2 in Tsybakov (2009). To shorten notation, let $\psi^j_{\beta,L} := \psi_{T,\beta}\, C_j(\beta, L; \rho, \sigma)$, for $\psi_{T,\beta}$ and $C_j(\beta, L; \rho, \sigma)$ defined in (2.5.8) and (2.5.10), respectively.

(I) Construction of the hypotheses. Denote $\alpha_j := \left(\sum_{k=1}^d a_{jk}^2\right)^{1/2}$. Let $L \in [L_*, L^*]$, consider some positive density function

$$\rho \in C_c^\infty(\mathbb{R}^d) \cap \mathcal{S}(\beta_T + 1, 2L/\alpha_j) \cap \mathcal{S}(\beta_* + 1, L/\alpha_j), \tag{2.6.17}$$

fix $\delta \in (0, 1/2)$, and assume that ρ is such that the function

$$\rho_{T,0}(x) := \delta^{1/(\beta_*+3/2)}\ \rho\left(x\delta^{1/(\beta_*+3/2)}\right), \qquad x \in \mathbb{R}^d,$$

satisfies $\rho_{T,0}(x_0) \geq \epsilon$, and, for

$$b_{T,0}(x) := a/2\nabla \log \rho_{T,0}(x), \qquad x \in \mathbb{R}^d,$$

the SDE $\mathrm{d}X_t = b_{T,0}(X_t)\mathrm{d}t + \sigma \mathrm{d}W_t$, $t \geq 0$, admits a strong solution with Lebesgue continuous invariant measure and invariant density $\rho_{T,0}$. Define further

$$g^j_{T,0}(x) := \sum_{k=1}^d a_{jk}\partial_k \rho_{T,0}(x).$$

For $\beta > d/2$, consider $\widetilde{K}_\beta$ and $\mathsf{b} = \mathsf{b}(\beta)$ as defined in (2.5.19) and (2.5.20), respectively, and let

$$K^*_\beta(x) := \mathbb{I}_\beta^{-1}\mathsf{b}^{-\beta+d/2}\ \widetilde{K}_\beta(\mathsf{b}x). \tag{2.6.18}$$

Lemma 2.5.12 implies that

$$K^*_\beta(0) = \mathbb{I}_\beta^{-1}\mathsf{b}^{-\beta+d/2}\ \widetilde{K}_\beta(0) = \frac{2\beta}{d}\left(\frac{d}{2\beta - d}\right)^{\frac{\beta - d/2}{2\beta}}\ \mathbb{I}_\beta \tag{2.6.19}$$

and

$$\|K^*_\beta\|_{L^2(\mathbb{R}^d)} = \mathbb{I}_\beta^{-1}\mathsf{b}^{-\beta+d/2}\left(\int_{\mathbb{R}^d}\widetilde{K}^2_\beta(\mathsf{b}x)\mathrm{d}x\right)^{1/2} = 1.$$

Denote by $\overline{K}_\beta$ the compactly supported modification of K^*_β from Lemma 2.5.12 satisfying (k1), (k2), and (k3). Define the function $g_{T,\beta_*}: \mathbb{R}^d \to \mathbb{R}$ such that, for any $k \in \{1, \ldots, d\}$,

$$\partial_k g_{T,\beta_*}(x) = 2La_{jj}^{-1}(h^j_{T,\beta_*})^{\beta_*-d/2}\ \overline{K}_{\beta_*}\left(\frac{x - x_0}{h^j_{T,\beta_*}}\right)\ \mathbb{1}\{k = j\}, \tag{2.6.20}$$

where

$$h^j_{T,\beta_*} := \left(\frac{d\rho_{T,0}(x_0)a_{jj}\log T}{\beta_* L^2 T}\right)^{1/(2\beta_*)}. \tag{2.6.21}$$

Let

$$\rho_{T,1}(x) := \rho_{T,0}(x)\Big(1 - \int_{\mathbb{R}^d} g_{T,\beta_*}(y)\mathrm{d}y\Big) + g_{T,\beta_*}(x),$$

and consider the hypothesis $g^j_{T,1}$, defined as

$$\begin{aligned} g^j_{T,1}(x) \quad &:= \quad \sum_{k=1}^d a_{jk}\ \partial_k\rho_{T,1}(x) \\ &= \quad \left(1 - \int_{\mathbb{R}^d} g_{T,\beta_*}(y)\mathrm{d}y\right)\ g^j_{T,0}(x) + \sum_{k=1}^d a_{jk}\partial_k g_{T,\beta_*}(x). \end{aligned} \tag{2.6.22}$$

The function $b_{T,1} : \mathbb{R}^d \to \mathbb{R}^d$ is taken as

$$b_{T,1}(x) := a/2\ \nabla(\log \rho_{T,1}(x)). \tag{2.6.23}$$

(II) Properties of the hypotheses. First of all, the above definitions of the hypotheses imply that

$$2b^j_{T,i}\rho_{T,i} = \sum_{k=1}^d a_{jk}\partial_k(\log \rho_{T,i})\rho_{T,i} = \sum_{k=1}^d a_{jk}\partial_k \rho_{T,i} = g^j_{T,i} \qquad i \in \{0,1\}.$$

Furthermore, for T large enough,

$$\frac{\rho_{T,0}(x_0)}{\rho_{T,1}(x_0)} \leq \frac{\rho_{T,0}(x_0)}{\rho_{T,0}(x_0)\,(1 - \int_{\mathbb{R}^d} g_{T,\beta_*}(y)\mathrm{d}y)} \leq 1 + \delta/2. \tag{2.6.24}$$

Similarly to the proof of Lemma 2.5.3, it holds

$$\begin{aligned}
\eta^2_{\beta_T}(g^j_{T,0}) &\overset{(2.5.4)}{=} (2\pi)^{-d}\int_{\mathbb{R}^d} \|\lambda\|^{2\beta_T}|\phi_{g^j_{T,0}}(\lambda)|^2\mathrm{d}\lambda \\
&= (2\pi)^{-d}\int_{\mathbb{R}^d} \|\lambda\|^{2\beta_T}\Big|\sum_{k=1}^d a_{jk}\lambda_k\phi_{\rho_{T,0}}(\lambda)\Big|^2\mathrm{d}\lambda \\
&\leq \alpha_j^2\ (2\pi)^{-d}\int_{\mathbb{R}^d} \|\lambda\|^{2(\beta_T+1)}\Big|\phi_\rho\Big(\lambda\delta^{-1/(\beta+3/2)}\Big)\Big|^2\mathrm{d}\lambda \\
&= \alpha_j^2\ \delta^2(2\pi)^{-d}\int_{\mathbb{R}^d} \|\lambda\|^{2(\beta_T+1)}|\phi_\rho(\lambda)|^2\mathrm{d}\lambda \\
&= \alpha_j^2\ \delta^2\ \eta^2_{\beta_T+1}(\rho) \overset{(2.6.17)}{\leq} (2L\ \delta)^2,
\end{aligned}$$

such that $g^j_{T,0} \in \mathcal{S}(\beta_T, 2L)$. Analogously,

$$\eta^2_{\beta_*}(g^j_{T,0}) \leq \alpha_j^2\ \delta^2\ \eta^2_{\beta_*+1}(\rho) \overset{(2.6.17)}{\leq} (L\delta)^2. \tag{2.6.25}$$

For any $\lambda \in \mathbb{R}^d$, the Fourier transform of $\sum_{k=1}^d a_{jk}\partial_j g_{T,\beta_*}$ is given by

$$\begin{aligned}
\phi_{\sum_{k=1}^d a_{jk}\partial_j g_{T,\beta_*}}(\lambda) &= \phi_{a_{jj}\partial_j g_{T,\beta_*}}(\lambda) \\
&= 2L(h^j_{T,\beta^*})^{\beta_*-d/2}\int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\lambda^t x}\overline{K}_{\beta_*}\left(\frac{x-x_0}{h^j_{T,\beta_*}}\right)\mathrm{d}x \\
&= 2L(h^j_{T,\beta^*})^{\beta_*-d/2}\mathrm{e}^{\mathrm{i}\lambda^t x_0}\phi_{\overline{K}_{\beta_*}}(h^j_{T,\beta_*}\lambda).
\end{aligned}$$

Consequently, taking into account property (k2),

$$\begin{aligned}
\eta^2_{\beta_*}\left(\sum_{k=1}^d a_{jk}\partial_k g_{T,\beta_*}\right) &= (2\pi)^{-d}\int_{\mathbb{R}^d} \|\lambda\|^{2\beta_*}|\phi_{a_{jj}\partial_j g_{T,\beta_*}}(\lambda)|^2\mathrm{d}\lambda \\
&\leq 4L^2(h^j_{T,\beta^*})^{2\beta_*+d}(2\pi)^{-d}\int_{\mathbb{R}^d} \|\lambda\|^{2\beta_*}|\phi_{\overline{K}_{\beta_*}}(h^j_{T,\beta_*}\lambda)|^2\mathrm{d}\lambda \\
&\leq (2L\,(1-\delta/2))^2.
\end{aligned} \tag{2.6.26}$$

For $g^j_{T,1}$, it thus follows from (2.6.25) and (2.6.26),

$$\begin{aligned}
\eta_{\beta_*}(g^j_{T,1}) &\overset{(2.6.22)}{\leq} \left(1 - \int_{\mathbb{R}^d} g_{T,\beta_*}(y)\mathrm{d}y\right)\eta_{\beta_*}(g^j_{T,0}) + \eta_{\beta_*}\left(2\sum_{k=1}^d a_{jk}\partial_k g_{T,\beta_*}\right) \\
&\leq L\delta + 2L\,(1-\delta/2) = 2L.
\end{aligned}$$

(III) A VERSION OF THEOREM 6(i) IN TSYBAKOV (1998). The central ingredient of the proof is a special case of Theorem 6(i) in Tsybakov (1998) whose original formulation is given in Section 1.2.1. It will be applied in the following situation: Denote by $\mathbf{E}_i = \mathbf{E}_{b_i}$ expectation under the measure $\mathbf{P}_i = \mathbf{P}_{b_{T,i}}$ associated with the hypothesis $b = b_{T,i}$, $i \in \{0,1\}$, and note that

$$\begin{aligned}
&\inf_{\widehat{g}_T^j} \sup_{(\beta,L)\in\mathcal{B}_T} \sup_{b\in\Pi^j(\beta,L)} \left(\psi^j_{\beta,L}\right)^{-2} \mathbf{E}_b\left|\widehat{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|^2 \\
&\quad\geq \inf_{\widehat{g}_T^j} \max\Bigg\{ \sup_{b\in\Pi^j(\beta_T,L)} \left(\psi^j_{\beta_T,L}\right)^{-2} \mathbf{E}_b\left|\widehat{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|^2, \\
&\qquad\qquad \sup_{b\in\Pi^j(\beta_*,L)} \left(\psi^j_{\beta_*,L}\right)^{-2} \mathbf{E}_b\left|\widehat{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|^2 \Bigg\} \\
&\quad\geq \inf_{\widehat{g}_T^j} \max\left\{ \mathbf{E}_0\left(\left(\psi^j_{\beta_T,L}\right)^{-2}\left|\widehat{g}_T^j(x_0) - g^j_{T,0}(x_0)\right|^2\right), \mathbf{E}_1\left(\left(\psi^j_{\beta_*,L}\right)^{-2}\left|\widehat{g}_T^j(x_0) - g^j_{T,1}(x_0)\right|^2\right)\right\} \\
&\quad= \inf_{\widehat{T}_T^j} \max\left\{ \mathbf{E}_0\left|Q_T^j\widehat{T}_T^j\right|^2, \mathbf{E}_1\left|\widehat{T}_T^j - \theta_1^j\right|^2\right\}, \qquad (2.6.27)
\end{aligned}$$

where

$$Q_T^j := \psi^j_{\beta_*,L}\left(\psi^j_{\beta_T,L}\right)^{-1}, \qquad \widehat{T}_T^j := \left(\psi^j_{\beta_*,L}\right)^{-1}\left(\widehat{g}_T^j(x_0) - g^j_{T,0}(x_0)\right), \qquad (2.6.28)$$

and

$$\begin{aligned}
\theta_1^j &:= \left(\psi^j_{\beta_*,L}\right)^{-1}\left(g^j_{T,1}(x_0) - g^j_{T,0}(x_0)\right) \qquad (2.6.29)\\
&= \left(\psi^j_{\beta_*,L}\right)^{-1}\sum_{k=1}^d a_{jk}\left(\partial_k g_{T,\beta_*}(x_0) - \partial_k\rho_{T,0}(x_0)\int_{\mathbb{R}^d} g_{T,\beta_*}(y)\mathrm{d}y\right).
\end{aligned}$$

Lemma 2.6.2 (Theorem 6(i) in Tsybakov (1998)). *Consider Q_T^j, $\widehat{T}_T^j$ and θ_1^j defined in (2.6.28) and (2.6.29), respectively, and assume that $\theta_1^j \in \mathbb{R}$ satisfies*

$$|\theta_1^j| \geq 1-\delta. \qquad (\mathbf{A1})$$

If $\mathbf{P}_0, \mathbf{P}_1$ are such that $\mathbf{P}_0 \ll \mathbf{P}_1$ and, for $\tau > 0$ and $\alpha \in (0,1)$ fixed,

$$\mathbf{P}_1\left(\frac{\mathrm{d}\mathbf{P}_0}{\mathrm{d}\mathbf{P}_1} \geq \tau\right) \geq 1-\alpha, \qquad (\mathbf{A2})$$

then

$$\inf_{\widehat{T}_T^j} \max\left\{\mathbf{E}_0\left|Q_T^j\widehat{T}_T^j\right|^2, \mathbf{E}_1\left|\widehat{T}_T^j - \theta_1^j\right|^2\right\} \geq \frac{(1-\alpha)\tau(1-2\delta)^2\left(Q_T^j\delta\right)^2}{(1-2\delta)^2 + \tau\left(Q_T^j\delta\right)^2},$$

where the infimum is taken over all $\widehat{T}_T^j$ as defined in (2.6.28).

Proof. The proof is completely along the lines of the proof of Theorem 6(i) in Tsybakov (1998). □

We proceed with verifying (**A1**) and (**A2**). Note first that

$$\begin{aligned}
\Big|\sum_{k=1}^d a_{jk}\partial_k g_{T,\beta_*}(x_0)\Big| &= 2Lh_{T,\beta_*}^{\beta_*-d/2}\overline{K}_{\beta_*}(0)\\
&\overset{\text{(k3)}}{\geq} 2Lh_{T,\beta_*}^{\beta_*-d/2}\,(1-\delta/2)\;K^*_{\beta_*}(0)\\
&\overset{(2.5.37)}{=} (1-\delta/2)\,2L\left(\frac{d^2a_{jj}\rho_{T,0}(x_0)\log T}{\beta_*(2\beta_*-d)L^2T}\right)^{\frac{\beta_*-d/2}{2\beta_*}}\frac{2\beta_*}{d}\;\mathbb{I}_{\beta_*}\\
&= (1-\delta/2)\,2L^{\frac{d}{2\beta_*}}\left(\frac{d^2a_{jj}\rho_{T,0}(x_0)}{\beta_*(2\beta_*-d)}\right)^{\frac{\beta_*-d/2}{2\beta_*}}\frac{2\beta_*}{d}\;\mathbb{I}_{\beta_*}\;\left(\frac{\log T}{T}\right)^{\frac{\beta_*-d/2}{2\beta_*}}\\
&= (1-\delta/2)\;C_j(\beta_*,L;\rho_{T,0},\sigma)\;\psi_{T,\beta_*} = (1-\delta/2)\;\psi^j_{\beta_*,L}.
\end{aligned}$$

Since, for T large enough,

$$(\psi^j_{\beta_*,L})^{-1}\int_{\mathbb{R}^d} g_{T,\beta_*}(y)\mathrm{d}y \leq \frac{\delta}{2},$$

this implies

$$\begin{aligned}
|\theta^j_1| &= (\psi^j_{\beta_*,L})^{-1}|g^j_{T,1}(x_0)-g^j_{T,0}(x_0)|\\
&\geq (\psi^j_{\beta_*,L})^{-1}\,\Big|\sum_{k=1}^d a_{jk}\Big(\partial_k g_{T,\beta_*}(x_0)-\partial_k\rho_{T,0}(x_0)\int_{\mathbb{R}^d} g_{T,\beta_*}(y)\mathrm{d}y\Big)\Big|\\
&\geq 1-\delta. \qquad (2.6.30)
\end{aligned}$$

In order to verify (**A2**), note that the specifications on pp. 296–297 in Liptser and Shiryaev (2001) (also see Remark A.1.2 in Section A.1) imply that the likelihood ratio under $\mathbf{P}_1$ is given by

$$\begin{aligned}
\frac{\mathrm{d}\mathbf{P}_0}{\mathrm{d}\mathbf{P}_1}(Y^T) &= \frac{\rho_{T,0}}{\rho_{T,1}}(Y_0)\exp\Bigg(-\frac12\int_0^T(b_{T,0}-b_{T,1})^t(Y_u)a^{-1}(b_{T,0}+b_{T,1})(Y_u)\mathrm{d}u\\
&\qquad\qquad + \int_0^T(b_{T,0}-b_{T,1})^t(Y_u)a^{-1}\mathrm{d}Y_u\Bigg)\\
&= \frac{\rho_{T,0}}{\rho_{T,1}}(Y_0)\exp\Bigg(-\frac12\int_0^T(b_{T,0}-b_{T,1})^t(Y_u)a^{-1}(b_{T,0}-b_{T,1})(Y_u)\mathrm{d}u \qquad (2.6.31)\\
&\qquad\qquad + \int_0^T\left(\sigma^{-1}\left(b_{T,0}-b_{T,1}\right)\right)^t(Y_u)\mathrm{d}W_u\Bigg).
\end{aligned}$$

To proceed, set

$$M_t := \int_0^t\left(\sigma^{-1}\left(b_{T,0}-b_{T,1}\right)\right)^t(Y_u)\mathrm{d}W_u, \qquad t\geq 0,$$

denote

$$g(x) := (b_{T,0}-b_{T,1})^t(x)\;a^{-1}\;(b_{T,0}-b_{T,1})(x), \qquad x\in\mathbb{R}^d,$$

and consider the following stationary sequence of random variables,

$$Z_k := \int_{(k-1)t}^{kt} g(Y_u)\mathrm{d}u, \qquad k\geq 1.$$

Since $g \in L^1(\mathbf{P}_1)$, it follows from the ergodic theorem (see, e.g., Theorem 59.1 in Gnedenko (1968)) that, for any $t > 0$,

$$\frac{1}{n}\sum_{k=1}^{n} Z_k = \frac{1}{n}\int_0^{nt} g(Y_u)\mathrm{d}u = \frac{1}{n}\langle M\rangle_{nt} \xrightarrow{\text{a.s.}}_{n\to\infty} tc,$$

where

$$c := \mathbf{E}_1\left((b_{T,0}-b_{T,1})^t(Y_0)a^{-1}(b_{T,0}-b_{T,1})(Y_0)\right) = \mathbf{E}_1\left\|\sigma^{-1}(b_{T,0}-b_{T,1})(Y_0)\right\|^2. \tag{2.6.32}$$

In particular, this implies by means of the martingale CLT (Lemma A.1.3 in Section A.1) that, for some standard Brownian motion W,

$$\frac{M_{nt}}{\sqrt{n}} \xrightarrow{\mathbf{P}_1}_{n\to\infty} \sqrt{c}\, W_t. \tag{2.6.33}$$

Denoting by $[s]$ the integer part of s and considering an arbitrary sequence $\gamma(s) \to_{s\to\infty} 0$, it holds

$$\gamma(s)\int_{[s]}^{s} g(Y_u)\mathrm{d}u \xrightarrow{\mathbf{P}_1}_{s\to\infty} 0.$$

Choosing $t \equiv 1$ in (2.6.33) and passing to the continuous-time case, we obtain

$$\frac{M_T}{\sqrt{T}} \xrightarrow{\mathbf{P}_1}_{T\to\infty} Z \sim \mathcal{N}(0,c).$$

Similarly to (2.6.15), the definition of the hypotheses $b_{T,0}$ and $b_{T,1}$ entails that

$$\begin{aligned}
\sigma^{-1}\left(b_{T,0}-b_{T,1}\right) &= \frac{1}{2}\sigma^t\nabla\left(\log\frac{\rho_{T,0}}{\rho_{T,1}}\right) = \frac{1}{2}\left(1-\frac{g_{T,\beta_*}}{\rho_{T,1}}\right)\ \sigma^t\nabla\left(1-\frac{g_{T,\beta_*}}{\rho_{T,1}}\right)\\
&= \frac{g_{T,\beta_*}\sigma^t\nabla\rho_{T,1}-\rho_{T,1}\sigma^t\nabla g_{T,\beta_*}}{2\rho_{T,1}\left(\rho_{T,1}-g_{T,\beta_*}\right)}\\
&= \frac{\left(1-\int_{\mathbb{R}^d} g_{T,\beta_*}(y)\mathrm{d}y\right)g_{T,\beta_*}\sigma^t\nabla\rho_{T,0}+\left(g_{T,\beta_*}-\rho_{T,1}\right)\sigma^t\nabla g_{T,\beta_*}}{2\rho_{T,1}\left(\rho_{T,1}-g_{T,\beta_*}\right)}\\
&= \frac{g_{T,\beta_*}\sigma^t\nabla\rho_{T,0}+\rho_{T,0}\sigma^t\nabla g_{T,\beta_*}}{2\rho_{T,1}\rho_{T,0}}\\
&= \left(2\rho_{T,1}\right)^{-1}\ \left(g_{T,\beta_*}\sigma^t\nabla\left(\log\rho_{T,0}\right)+\sigma^t\nabla g_{T,\beta_*}\right).
\end{aligned}$$

The definition of g_{T,β_*} further implies that

$$\begin{aligned}
\mathbf{E}_1\left(\frac{\left\|\sigma^t\nabla g_{T,\beta_*}(Y_0)\right\|^2}{4\rho_{T,1}^2(Y_0)}\right) &= \int_{\mathbb{R}^d}\frac{\left\|\sigma^t\nabla g_{T,\beta_*}(y)\right\|^2}{4\rho_{T,1}(y)}\mathrm{d}y\\
&= a_{jj}\int_{\mathbb{R}^d}\frac{\left(\partial_j g_{T,\beta_*}(y)\right)^2}{4\rho_{T,1}(y)}\mathrm{d}y\\
&\overset{(2.6.20)}{=} a_{jj}^{-1}L^2\ \left(h_{T,\beta_*}^j\right)^{2\beta_*-d}\int_{\mathbb{R}^d}\overline{K}_{\beta_*}^2\left(\frac{y-x_0}{h_{T,\beta_*}^j}\right)\frac{\mathrm{d}y}{4\rho_{T,1}(y)}\\
&\overset{(A.1.4)}{=} a_{jj}^{-1}L^2\ \left(h_{T,\beta_*}^j\right)^{2\beta_*}\int_{\mathbb{R}^d}\overline{K}_{\beta_*}^2(y)\mathrm{d}y\ \frac{(1+o_T(1))}{\rho_{T,1}(x_0)}\\
&\overset{(\text{k1})}{\le} (1-\delta/2)^2\, a_{jj}^{-1}L^2\ \left(h_{T,\beta_*}^j\right)^{2\beta_*}\frac{(1+o_T(1))}{\rho_{T,1}(x_0)}.
\end{aligned}$$

Thus, plugging in the definition of h^j_{T,β_*} (see (2.6.21)),

$$\begin{aligned}\mathbf{E}_1\left(\frac{\left\|\sigma^t\nabla g_{T,\beta_*}(Y_0)\right\|^2}{4\rho^2_{T,1}(Y_0)}\right) &\leq (1-\delta/2)^2\, L^2\left(\frac{d\rho_{T,0}(x_0)a_{jj}\log T}{\beta_* L^2 T}\right)\frac{(1+o_T(1))}{a_{jj}\rho_{T,1}(x_0)}\\ &\overset{(2.6.24)}{\leq} \left(1-\delta^2/4\right)^2\frac{d}{\beta_*}\,\frac{\log T}{T}\,(1+o_T(1)).\end{aligned}$$

It can be shown by analogous arguments that the terms

$$\mathbf{E}_1\left(\frac{g^2_{T,\beta_*}(Y_0)\|\sigma^t\nabla(\log\rho_{T,0})(Y_0)\|^2}{4\rho^2_{T,1}(Y_0)}\right)$$

and

$$\mathbf{E}_1\left(\frac{g_{T,\beta_*}(Y_0)\left(\nabla(\log\rho_{T,0})(Y_0)\right)^t a\nabla g_{T,\beta_*}(Y_0)}{2\rho^2_{T,1}(Y_0)}\right)$$

are asymptotically negligible. Thus, for c defined in (2.6.32) and whenever δ is small and T is large enough,

$$c\leq\left(1-\delta^2/4\right)^2\frac{d}{\beta_*}\,\frac{\log T}{T}\,(1+o_T(1)).$$

Consequently, for

$$\tau:=\exp\left(-\frac{(1-\delta^2/4)\,d\log T}{2\beta_*}\right),$$

it holds a.s.

$$\frac{\log\tau-\log\rho_{T,0}(Y_0)+\log\rho_{T,1}(Y_0)+\frac{1}{2}\langle M\rangle_T}{\sqrt{Tc}}\leq-\frac{\delta^2}{8}\sqrt{\frac{d\log T}{\beta_*}}+o_T(1)\to-\infty.$$

The verification of (**A2**) is accomplished by means of a tightness argument. Consider some sequence of probability measures $(\mathbf{P}_n)_{n\geq1}$ on some measurable space, converging weakly to some probability measure $\mathbf{P}$. Tightness of $\mathbf{P}_n$ implies that, for any sequence $\gamma_n\to-\infty$,

$$\lim_{m\to\infty}\max\left\{\mathbf{P}\left((-\infty,\gamma_m)\right),\sup_{n\in\mathbb{N}}\mathbf{P}_n\left((-\infty,\gamma_m)\right)\right\}=0.$$

Thus, $\lim_{m\to\infty}\sup_{n\in\mathbb{N}}\mathbf{P}_n\left((-\infty,\gamma_m)\right)=0$ and $\lim_{m\to\infty}\inf_{n\in\mathbb{N}}\mathbf{P}_n\left((-\infty,\gamma_m)\right)=1$. In particular, $\lim_{m\to\infty}\mathbf{P}_m\left((\gamma_m,\infty)\right)=1$. In the current framework, this last assertion implies that, for

$$\gamma_T:=\frac{\log\tau-\log\rho_{T,0}(Y_0)+\log\rho_{T,1}(Y_0)+\frac{1}{2}\langle M\rangle_T}{\sqrt{Tc}}\to-\infty,$$

it holds

$$\begin{aligned}\lim_{T\to\infty}\mathbf{P}_1\left(\frac{\mathrm{d}\mathbf{P}_0}{\mathrm{d}\mathbf{P}_1}\geq\tau\right) &\overset{(2.6.31)}{=} \lim_{T\to\infty}\mathbf{P}_1\left(\frac{\rho_{T,0}}{\rho_{T,1}}(Y_0)\exp\left(M_T-\frac{1}{2}\langle M\rangle_T\right)\geq\tau\right)\\ &= \lim_{T\to\infty}\mathbf{P}_1\left(\frac{M_T}{\sqrt{Tc}}\geq\gamma_T\right)=1.\end{aligned}$$

For large enough T and fixed $\tau>0$ (where $\delta\in(0,1)$ can be chosen arbitrarily small), we thus obtain

$$\mathbf{P}_1\left(\frac{\mathrm{d}\mathbf{P}_0}{\mathrm{d}\mathbf{P}_1}\geq\tau\right)\geq1-\delta. \tag{2.6.34}$$

(VI) Completion of the proof. In view of (2.6.30) and (2.6.34), Lemma 2.6.2 gives

$$\begin{aligned}\inf_{\widehat{g}_T^j}\sup_{(\beta,L)\in\mathcal{B}_T}\sup_{b\in\Pi^j(\beta,L)}&\mathbf{E}_b\big|\widehat{g}_T^j(x_0)-\operatorname{div}(a_j\rho)(x_0)\big|^2\big(\psi^j_{\beta,L}\big)^{-2}\\ &\overset{(2.6.27)}{\geq}\inf_{\widehat{T}_T^j}\max\left\{\mathbf{E}_0\big|Q_T^j\widehat{T}_T^j\big|^2,\mathbf{E}_1\big|\widehat{T}_T^j-\theta_1^j\big|^2\right\}\\ &\geq\frac{(1-\delta)\tau(1-2\delta)^2\big(Q_T^j\delta\big)^2}{(1-2\delta)^2+\tau\big(Q_T^j\delta\big)^2}.\end{aligned}$$

Since, for $C_j:=C_j(\beta_*,L;\rho_{T,0},\sigma)/C_j(\beta_T,L;\rho_{T,0},\sigma)$,

$$Q_T^j=\frac{\psi^j_{\beta_*,L}}{\psi^j_{\beta_T,L}}=C_j\exp\left(-\frac{d}{4\beta_*\beta_T}\left(\beta_*-\beta_T\right)\left(\log T-\log\log T\right)\right),$$

we have

$$\begin{aligned}\tau\big(Q_T^j\big)^2&=C_j^2\exp\left(\left(-\frac{(1-\delta^2/4)\,d}{2\beta_*}-\frac{d\,(\beta_*-\beta_T)}{2\beta_*\beta_T}\right)\log T\right)\times\exp\left(\frac{d\,(\beta_*-\beta_T)}{2\beta_*\beta_T}\log\log T\right)\\ &=C_j^2\exp\left(\frac{d\,(\delta^2\beta_T/4-\beta_*)}{2\beta_*\beta_T}\log T\right)\times\exp\left(\frac{d\,(\beta_*-\beta_T)}{2\beta_*\beta_T}\log\log T\right).\end{aligned}$$

As $T\to\infty$, $\tau\big(Q_T^j\big)^2\to\infty$. Choosing $\delta>1/A$ for A large enough to ensure $\delta<1/2$, it holds

$$\frac{(1-\delta)\tau(1-2\delta)^2\big(Q_T^j\delta\big)^2}{(1-2\delta)^2+\tau\big(Q_T^j\delta\big)^2}=\frac{(1-\delta)(1-2\delta)^2\delta^2}{\frac{(1-2\delta)^2}{\tau(Q_T^j)^2}+\delta^2}\to_{T\to\infty}(1-\delta)(1-2\delta)^2.$$

Taking now $A\to\infty$, the assertion follows. $\square$

Proof of Theorem 2.5.7. The lower bound for estimating the j-th component of the drift is derived from the lower bound on estimating $\operatorname{div}(a_j\rho)$ similarly to the proof of Theorem 2.3.3. For any estimator $\widetilde{b}_T^j$ of the j-th component of the drift $b\in\Pi^j(\beta,L)$, $j\in\{1,\ldots,d\}$ fixed, denote $\widetilde{g}_T^j:=2\widetilde{b}_T^j\,\widehat{\rho}_{T,2}$, where the invariant density estimator $\widehat{\rho}_{T,2}$ is defined according to Lemma 2.2.4 and satisfies

$$\mathbf{E}_b\big|\widehat{\rho}_{T,2}(x_0)-\rho(x_0)\big|^2\leq K_1'\,T^{-\frac{\beta+1-d/2}{\beta+1}}\,\exp\left(-K_2'\|x_0\|\right).$$

The inverse triangle inequality entails that

$$\begin{aligned}&\left(\psi_{T,\beta}\,\frac{C_j(\beta,L;\rho,\sigma)}{2\rho(x_0)}\right)^{-2}\mathbf{E}_b\big|\widetilde{b}_T^j(x_0)-b^j(x_0)\big|^2\\ &\quad=4\big(\psi_{T,\beta}\,C_j(\beta,L;\rho,\sigma)\big)^{-2}\\ &\qquad\times\mathbf{E}_b\Big|\widetilde{b}_T^j(x_0)\rho(x_0)-\widetilde{b}_T^j(x_0)\widehat{\rho}_{T,2}(x_0)+\frac{1}{2}\widetilde{g}_T^j(x_0)-\frac{1}{2}\operatorname{div}(a_j\rho)(x_0)\Big|^2\\ &\quad\geq4\big(\psi_{T,\beta}\,C_j(\beta,L;\rho,\sigma)\big)^{-2}\\ &\qquad\times\left(\sqrt{\frac{1}{4}\mathbf{E}_b\big|\widetilde{g}_T^j(x_0)-\operatorname{div}(a_j\rho)(x_0)\big|^2}-\sqrt{\mathbf{E}_b\left(\big|\widetilde{b}_T^j(x_0)\big|^2\big(\rho(x_0)-\widehat{\rho}_{T,2}(x_0)\big)^2\right)}\right)^2.\end{aligned}$$

As in the proof of Theorem 2.3.3, we may assume without loss of generality that

$$\big|\widetilde{b}_T^j(x_0)\big|\leq\log T\exp\left(\frac{\|x_0\|}{\log T}\right).$$

Consequently,

$$\mathbf{E}_b\left(\left|\widetilde{b}_T^j(x_0)\right|^2\left(\rho(x_0)-\widehat{\rho}_{T,2}(x_0)\right)^2\right) \leq K_1'(\log T)^2 \exp\left(\frac{2\|x_0\|}{\log T} - K_2'\|x_0\|\right)\ T^{-\frac{\beta+1-d/2}{\beta+1}}.$$

Since $\psi_{T,\beta}^{-2}\ T^{-\frac{\beta+1-d/2}{\beta+1}} \to 0$ and in view of (2.5.11), the assertion follows. $\square$

2.6.2. Proof of upper bounds

Proof of Lemma 2.5.13. (a) Letting, for $r, T > 0$,

$$\mathsf{p}(r) := \mathbf{P}_b\left(\left|\frac{1}{T}\int_0^T\left(f(X_u)b^j(X_u) - \int_{\mathbb{R}^d} f(y)b^j(y)\mathrm{d}\mu(y)\right)\mathrm{d}u\right| > r\right), \tag{2.6.35}$$

Theorem 1.1 in Lezaud (2001) (also see Proposition 2.5 in Gao et al. (2010)) implies that

$$\mathsf{p}(r) \leq 2\exp\left(-\frac{Tr^2}{2\left(\varsigma^2(fb^j) + c_P\|fb^j\|_\infty r\right)}\right). \tag{2.6.36}$$

Using the spectral gap assumption, we get, for any $T>0$, $g \in L^2(\mu)$,

$$\frac{1}{T}\operatorname{Var}_{\mathbf{P}_b}\left(\int_0^T g(X_u)\mathrm{d}u\right) \leq 2\int_0^T \left\langle P_t g, g\right\rangle_\mu \mathrm{d}t \leq 2\|g\|_{L^2(\mu)}^2 \int_0^T \mathrm{e}^{-2t/c_P}\mathrm{d}t \leq c_P\|g\|_{L^2(\mu)}^2.$$

Consequently, in view of (2.5.40),

$$\varsigma^2(fb^j) \leq c_P\|fb^j\|_{L^2(\mu)}^2 \leq c_P B\|f\|_{L^2(\mu)}^2$$

and $\|fb^j\|_\infty \leq B\|f\|_\infty$. Plugging these estimates into (2.6.36), and considering separately the cases

$$r < \frac{\|f\|_{L^2(\mu)}^2}{\|f\|_\infty} \quad \text{and} \quad r \geq \frac{\|f\|_{L^2(\mu)}^2}{\|f\|_\infty},$$

we obtain

$$\begin{aligned}\mathsf{p}(r) &\leq 2\exp\left(-\frac{Tr^2}{2c_P B\ \left(\|f\|_{L^2(\mu)}^2 + \|f\|_\infty r\right)}\right)\\ &\leq 2\exp\left(-\frac{T}{4c_P B}\ \min\left\{\frac{r^2}{\|f\|_{L^2(\mu)}^2}, \frac{r}{\|f\|_\infty}\right\}\right).\end{aligned}$$

(b) Under the given assumptions, Bernstein's inequality for continuous martingales can be used to show that there exists some constant $\widetilde{C}_B$ such that, for any $r, T > 0$,

$$\begin{aligned}\mathsf{q}(r) := \mathbf{P}_b\left(\left|\frac{1}{T}\int_0^T f(X_u)\mathrm{d}X_u^j - \int_{\mathbb{R}^d} f(y)b^j(y)\mathrm{d}\mu(y)\right| > r\right) \qquad\qquad (2.6.37)\\ \leq 2\exp\left(-\frac{Tr^2}{2\widetilde{C}_B\left(\|f\|_{L^2(\mu)}^2 + \|f\|_\infty\right)}\right).\end{aligned}$$

To see this, write $\mathsf{q}(r) \leq \mathsf{p}(r/2) + \mathsf{p}'(r/2)$, for $\mathsf{p}(\cdot)$ introduced in (2.6.35) and

$$\mathsf{p}'(r) := \mathbf{P}_b\left(\left|\frac{1}{T}\int_0^T f(X_u)\sum_{k=1}^d \sigma_{jk}\ \mathrm{d}W_u^k\right| > r\right), \qquad r > 0.$$

Letting

$$M_t(f) := \int_0^t f(X_u) \sum_{k=1}^d \sigma_{jk} \, \mathrm{d}W_u^k, \quad t \geq 0, \tag{2.6.38}$$

and denoting by $\langle M \rangle_\cdot$ the quadratic variation of the martingale M, Bernstein's inequality for continuous martingales (see p. 154 in Revuz and Yor (1999)) gives

$$\begin{aligned} \mathsf{p}'(r/2) &= \mathbf{P}_b\Big(|M_T(f)| > Tr/2; \langle M(f)\rangle_T \leq Tr\|f\|_\infty/2\Big) + \mathbf{P}_b\Big(\langle M(f)\rangle_T > Tr\|f\|_\infty/2\Big) \\ &\leq 2\exp\left(-\frac{Tr}{4\|f\|_\infty}\right) + \underbrace{\mathbf{P}_b\left(T^{-1}\int_0^T f^2(X_u)\mathrm{d}u > a_{jj}^{-1} r\|f\|_\infty/2\right)}_{=:\mathsf{p}''}. \end{aligned} \tag{2.6.39}$$

Theorem 1.1 in Lezaud (2001) then can be used to show that

$$\mathsf{p}'' \leq \exp\left(-\frac{Tr^2}{8c_P a_{jj}(a_{jj}\;\|f\|_{L^2(\mu)}^2 + \|f\|_\infty/2)}\right).$$

The inequality (2.6.37) now follows for $\widetilde{C}_B := 4\max\left\{2c_P B, 2c_P a_{jj}^2, c_P a_{jj}, 1\right\}$. In view of (2.6.37), a uniform exponential inequality in the spirit of Theorem 5.11 in van de Geer (2000) is available. Indeed, Theorem 5.11 in van de Geer (2000) appears as a special case of the uniform inequality for martingales in Theorem 8.13, and the proof of Theorem 8.13 continues to hold in the diffusion setting if the Bernstein inequality for martingales in Corollary 8.10 in van de Geer (2000) is replaced with the Bernstein-type deviation inequality (**BI**). □

Proof of Theorem 2.5.9. The basic idea of the proof is the same as in the classical article of Lepski (1990). The main difficulties arising in the current diffusion framework were already mentioned in Section 2.5.5. In order to deal with the specific problem due to the fact that the variance of the estimators considered in the procedure is proportional to the value of the unknown density at the point x_0, we apply the same strategy as in Butucea (2001).

Let $\beta \in [\beta_*, \beta_T]$, $L \in [L_*, L^*]$, $\beta' \in (d/2, \beta]$, and fix $j \in \{1, \ldots, d\}$. Denote by γ_{Ti}, $i \in \mathbb{N}$, functions of T such that $\lim_{T\to\infty}\gamma_{Ti} = 0$. For $\psi_{T,\beta}$ and $C_j(\beta, L; \rho, \sigma)$ introduced in (2.5.8) and (2.5.10), respectively, recall that $\psi_{\beta,L}^j = \psi_{T,\beta}\, C_j(\beta, L; \rho, \sigma)$. Denote by $\widetilde{T}^j(\beta)$ the *effective noise level under adaptation*, defined as

$$\widetilde{T}^j(\beta) := \left(\frac{4d\rho(x_0)a_{jj}\log T}{\beta T}\right)^{1/2}.$$

Recall the definitions of the bandwidth $\widehat{h}_{T,\beta'}^j$ and the thresholding sequence $\widehat{\eta}_{T,\beta'}^j$, namely

$$\begin{aligned} \widehat{h}_{T,\beta'}^j &= \left(\frac{4d\widehat{\rho}_T(x_0)a_{jj}\log T}{\beta' T}\right)^{1/2\beta'}, \\ \widehat{\eta}_{T,\beta'}^j &= \left(\frac{4d\widehat{\rho}_T(x_0)a_{jj}\log T}{\beta' T}\right)^{\frac{\beta'-d/2}{2\beta'}} \|K_{\beta'}\|_{L^2(\mathbb{R}^d)}, \end{aligned} \tag{2.6.40}$$

and consider their deterministic counterparts

$$h_{T,\beta'}^j := \left(\frac{4d\rho(x_0)a_{jj}\log T}{\beta' T}\right)^{1/2\beta'} = \left(\widetilde{T}^j(\beta')\right)^{1/\beta'} \tag{2.6.41}$$

and

$$\eta_{T,\beta'}^j := \left(\frac{4d\rho(x_0)a_{jj}\log T}{\beta' T}\right)^{\frac{\beta'-d/2}{2\beta'}} \|K_{\beta'}\|_{L^2(\mathbb{R}^d)} = (h_{T,\beta'}^j)^{\beta'-d/2}\; \|K_{\beta'}\|_{L^2(\mathbb{R}^d)}.$$

Set

$$\widetilde{\beta} = \widetilde{\beta}(\beta, \beta') := \begin{cases} \beta' + \frac{d}{2}, & \text{if } \frac{d}{2} \leq \beta' \leq \frac{\beta}{2} + \frac{d}{4}, \\ \beta, & \text{if } \frac{\beta}{2} + \frac{d}{4} < \beta' \leq \beta. \end{cases}$$

Define $\overline{\delta}_T := (\log T)^{-1}$, and introduce the random event

$$A^j_{T,\beta'} := \left\{ \left| (\widehat{h}^j_{T,\beta'} / h^j_{T,\beta'})^{\widetilde{\beta}' - d/2} - 1 \right| \leq \overline{\delta}_T \right\} \tag{2.6.42}$$

and the associated deterministic set

$$\mathcal{H}^j_{T,\beta'} := \left\{ h : \left| (h/h^j_{T,\beta'})^{\widetilde{\beta}' - d/2} - 1 \right| \leq \overline{\delta}_T \right\}.$$

Note that there exists a positive constant c_0 such that

$$\mathcal{H}^j_{T,\beta'} \subseteq \left\{ h : \left| (h/h^j_{T,\beta'}) - 1 \right| \leq c_0 \overline{\delta}_T \right\} =: H^j_{T,\beta'}.$$

Denote the kernel estimator of $\operatorname{div}(a_j \rho)(x_0)$ with deterministic bandwidth $h \in \mathcal{H}^j_{T,\beta'}$ by

$$g^j_{T,\beta'}(x_0, h) := \frac{2}{Th^d} \int_0^T K_{\beta'} \left(\frac{X_u - x_0}{h} \right) \mathrm{d}X^j_u, \tag{2.6.43}$$

and set

$$g^j_{T,\beta'}(x_0) := g^j_{T,\beta'}(x_0, h^j_{T,\beta'}). \tag{2.6.44}$$

Define

$$\mathsf{s}^j_T(\beta) := 2(h^j_{T,\beta})^{-d/2} \sqrt{\frac{\rho(x_0) a_{jj}}{T}} \, \|K_\beta\|_{L^2(\mathbb{R}^d)},$$

and let

$$d_T(\beta) := \sqrt{\frac{d \log T}{\beta}} \tag{2.6.45}$$

such that $\mathsf{s}^j_T(\beta) d_T(\beta) = \eta^j_{T,\beta}$. For $\beta' \leq \beta$, introduce the auxiliary sequence

$$\tau^j_T(\beta') := \mathsf{s}^j_T(\beta') \left(\sqrt{d^2_T(\beta') - d^2_T(\beta)} + \left(\frac{\log T}{\beta_T} \right)^{1/4} \right).$$

Auxiliary results. The first auxiliary result concerns the performance of the preliminary density estimator $\widehat{\rho}_T(x_0)$ which is contained in the adaptive bandwidth (see (2.6.40)). The following lemma states that weak regularity conditions on the bandwidth and the kernel used for defining $\widehat{\rho}_T(x_0)$ ensure that the usage of a random bandwidth does not substantially affect the estimation procedure.

Lemma 2.6.3. *Consider the estimator $\widehat{\varrho}_T(x_0)$ defined in (2.5.15), assume that the bandwidth h_T is such that*

$$\lim_{T \to \infty} \frac{T h_T^{2d}}{(\log T)^3} = \infty \quad \text{and} \quad \limsup_{T \to \infty} h_T^d T^{\alpha_0} < \infty,$$

for some fixed $\alpha_0 \in (0, 1/2)$, and that the kernel Q is bounded, positive and such that $\int_{\mathbb{R}^d} \|u\| |Q(u)| \mathrm{d}u < \infty$. Then, for any $\beta' \in (d/2, \beta]$ and some sufficiently small constant $\alpha > 0$ fixed,

$$\mathbf{P}_b \Big((A^j_{T,\beta'})^{\mathrm{c}} \Big) \leq 2 \exp \left(- \frac{T \left((1 - \alpha) h_T^d \overline{\delta}_T \epsilon \right)^2}{2 \|Q\|^2_\infty} \right) = o_T(1).$$

Proof. The proof is analogue to the proof of Lemma 3.2 in Butucea (2001). For the sake of completeness, we include a brief sketch. Recall that $\overline{\delta}_T = (\log T)^{-1}$ such that the above conditions imply that $\lim_{T\to\infty} Th_T^{2d}\overline{\delta}_T^2(\log T)^{-1} = \infty$ and $h_T^{\beta_*+1-d/2}(\overline{\delta}_T)^{-1} \to 0$. For $\alpha := (\widetilde{\beta}' - d/2)/(2\beta') < 1$, it holds, for any $x, y > 0$, $|x^\alpha - 1| \leq |x-1|$ and $|x \vee \epsilon - y \vee \epsilon| \leq |x-y|$. Thus, since $\rho(x_0) \geq \epsilon$,

$$\begin{aligned}
\mathbf{P}_b\Big(\big(A^j_{T,\beta'}\big)^{\mathrm{c}}\Big) &\overset{(2.6.42)}{=} \mathbf{P}_b\left(\Big|\big(\widehat{h}^j_{T,\beta'}/h^j_{T,\beta'}\big)^{\widetilde{\beta}'-d/2} - 1\Big| > \overline{\delta}_T\right)\\
&= \mathbf{P}_b\left(\Big|\Big(\frac{\widehat{\rho}_T(x_0)}{\rho(x_0)}\Big)^{\alpha} - 1\Big| > \overline{\delta}_T\right)\\
&\leq \mathbf{P}_b\left(\Big|\frac{\widehat{\rho}_T(x_0)}{\rho(x_0)} - 1\Big| > \overline{\delta}_T\right)\\
&\leq \mathbf{P}_b\Big(\big|\widehat{\varrho}_T(x_0) \vee \epsilon - \rho(x_0) \vee \epsilon\big| > \overline{\delta}_T\epsilon\Big)\\
&\leq \mathbf{P}_b\Big(\big|\widehat{\varrho}_T(x_0) - \mathbf{E}_b\widehat{\varrho}_T(x_0)\big| + \big|\mathbf{E}_b\widehat{\varrho}_T(x_0) - \rho(x_0)\big| > \overline{\delta}_T\epsilon\Big).
\end{aligned}$$

With arguments similar to those used in the proof of case 2) of Lemma 2.2.4, it can be shown that there exists some positive constant C such that

$$\big|\mathbf{E}_b\widehat{\varrho}_T(x_0) - \rho(x_0)\big| \leq Ch_T^{\beta_*+1-d/2} = o(\overline{\delta}_T).$$

Consequently, for small $\alpha > 0$ fixed,

$$\mathbf{P}_b\Big(\big(A^j_{T,\beta'}\big)^{\mathrm{c}}\Big) \leq \mathbf{P}_b\Big(\big|\widehat{\varrho}_T(x_0) - \mathbf{E}_b\widehat{\varrho}_T(x_0)\big| > (1-\alpha)\overline{\delta}_T\epsilon\Big).$$

The Hoeffding-type specification of the Bernstein-type inequality (**BI**) entails, for any $r > 0$,

$$\mathbf{P}_b\Big(\big|\widehat{\varrho}_T(x_0) - \mathbf{E}_b\widehat{\varrho}_T(x_0)\big| > r\Big) \leq 2\exp\left(-\frac{Th_T^{2d}r^2}{2\|Q\|_\infty^2}\right).$$

For $r = (1-\alpha)\overline{\delta}_T\epsilon$ and since $T\big(h_T^d\overline{\delta}_T\big)^2 \to_{T\to\infty} \infty$, the assertion follows. □

Lemma 2.6.4 (Bound on the bias)**.** *Consider the estimator $g^j_{T,\beta'}(x_0,h)$ defined in (2.6.43), and let*

$$\mathsf{b}_{\beta,\beta'} := (\mathsf{b}')^{\frac{d}{2}-\widetilde{\beta}}\left((2\pi)^{-d}\int_{\mathbb{R}^d}\frac{\|\lambda\|^{4\beta'-2\widetilde{\beta}}}{\big(1+\|\lambda\|^{2\beta'}\big)^2}\mathrm{d}\lambda\right)^{\frac{1}{2}}, \quad \textit{where } \mathsf{b}' := \mathsf{b}(\beta') = \left(\frac{2\beta'-d}{d}\right)^{\frac{1}{2\beta'}}.$$

For any $\mathcal{H}_{T,\beta'} \ni h > 0$,

$$\sup_{b\in\Pi^j(\beta,L)}\Big|\mathbf{E}_b g^j_{T,\beta'}(x_0,h) - \operatorname{div}(a_j\rho)(x_0)\Big| \leq 2Lh^{\widetilde{\beta}-d/2}\mathsf{b}_{\beta,\beta'}. \tag{2.6.46}$$

Furthermore,

$$\sup_{d/2<\beta'\leq\beta<\infty}\mathsf{b}_{\beta,\beta'} < \infty, \quad \limsup_{\delta\to 0}\sup_{\beta,\beta'\in[\beta_*,\infty):|\beta-\beta'|\leq\delta}\frac{\mathsf{b}_{\beta,\beta'}}{\mathsf{b}_{\beta,\beta}} \leq 1. \tag{2.6.47}$$

Proof. It follows from the definition of the estimators that

$$\begin{aligned}
\Big|\mathbf{E}_b g^j_{T,\beta'}(x_0,h) - \operatorname{div}(a_j\rho)(x_0)\Big| &= \Big|2\int_{\mathbb{R}^d}K_{\beta',h}(y-x_0)b^j(y)\rho(y)\mathrm{d}y - \operatorname{div}(a_j\rho)(x_0)\Big|\\
&= \Big|K_{\beta',h} * \operatorname{div}(a_j\rho)(x_0) - \operatorname{div}(a_j\rho)(x_0)\Big|.
\end{aligned}$$

The Fourier inversion formula implies that

$$\operatorname{div}(a_j\rho)(y) = (2\pi)^{-d} \int_{\mathbb{R}^d} \phi_{\operatorname{div}(a_j\rho)}(\lambda) \mathrm{e}^{-\mathrm{i}\lambda^t y} \mathrm{d}\lambda, \quad y \in \mathbb{R}^d. \tag{2.6.48}$$

Thus, applying (2.6.48) and the Fourier inversion formula to the convolution $K_{\beta',h} * \operatorname{div}(a_j\rho)$, one obtains

$$\begin{aligned}
&\left| K_{\beta',h} * \operatorname{div}(a_j\rho)(x_0) - \operatorname{div}(a_j\rho)(x_0) \right| \\
&\quad = (2\pi)^{-d} \left| \int_{\mathbb{R}^d} \phi_{\operatorname{div}(a_j\rho)}(\lambda) (\phi_{K_{\beta',h}}(\lambda) - 1) \mathrm{e}^{-\mathrm{i}\lambda^t x_0} \mathrm{d}\lambda \right| \\
&\quad = (2\pi)^{-d} \left| \int_{\mathbb{R}^d} \phi_{\operatorname{div}(a_j\rho)}(\lambda) \Big(\big(1 + \|h\lambda/\mathsf{b}'\|^{2\beta'}\big)^{-1} - 1 \Big) \mathrm{e}^{-\mathrm{i}\lambda^t x_0} \mathrm{d}\lambda \right| \\
&\quad = (2\pi)^{-d} \left| \int_{\mathbb{R}^d} \phi_{\operatorname{div}(a_j\rho)}(\lambda) \, \|h\lambda/\mathsf{b}'\|^{2\beta'} \left(1 + \|h\lambda/\mathsf{b}'\|^{2\beta'}\right)^{-1} \mathrm{d}\lambda \right| \\
&\quad \leq \left((2\pi)^{-d} \int_{\mathbb{R}^d} \left| \phi_{\operatorname{div}(a_j\rho)}(\lambda) \right|^2 \|\lambda\|^{2\widetilde{\beta}} \mathrm{d}\lambda \right)^{1/2} \left(\frac{h}{\mathsf{b}'} \right)^{\widetilde{\beta}} \qquad (2.6.49) \\
&\qquad\qquad \times \left((2\pi)^{-d} \int_{\mathbb{R}^d} \frac{\|h\lambda/\mathsf{b}'\|^{4\beta' - 2\widetilde{\beta}}}{\left(1 + \|h\lambda/\mathsf{b}'\|^{2\beta'}\right)^2} \mathrm{d}\lambda \right)^{1/2} \\
&\quad \leq 2L \mathsf{b}_{\beta,\beta'} \, h^{\widetilde{\beta} - d/2}. \qquad (2.6.50)
\end{aligned}$$

Here we used the scaling properties of Fourier transforms for evaluating $\phi_{K_{\beta',h}}(\lambda)$ and Cauchy–Schwarz to obtain (2.6.49). The remaining assertions are Lemma 1(ii), (iii) in [KT04]. □

The principal importance of exponential bounds on the stochastic error of estimators considered in the adaptive procedure was already indicated in Section 2.5.5. The Bernstein-type deviation inequality (**BI**) and its implication, the basic uniform exponential inequality stated in Lemma 2.5.13, will be applied now to derive more specific bounds on the stochastic error of the estimators $g^j_{T,\beta}$ defined according to (2.6.43). The following function classes are defined analogously to Butucea (2001),

$$\begin{aligned}
\mathcal{K}_1 &:= \left\{ K_{\beta',h}(\cdot) := h^{-d} K_{\beta'}\left((\cdot - x_0)/h\right) : h \in H^j_{T,\beta'} \right\}, \\
\mathcal{K}_2 &:= \left\{ K_{\beta',h} - K_{\beta',h_{T,\beta'}} : h \in H^j_{T,\beta'} \right\}.
\end{aligned}$$

Note first that there exists some positive constant $K_{\max}$ such that, for any $\beta > d/2$,

$$\|K_\beta\|_\infty = \sup_{x \in \mathbb{R}^d} |K_\beta(x)| \leq (2\pi)^{-d} \left(1 + \int_{\|y\|>1} \frac{\mathrm{d}y}{1 + \|y\|^{2\beta}} \right) \leq K_{\max} < \infty.$$

Consequently,

$$\sup_{h \in H^j_{T,\beta'}} \|K_{\beta',h}\|_\infty \leq (h^j_{T,\beta'})^{-d} \|K_{\beta'}\|_\infty \leq (h^j_{T,\beta'})^{-d} K_{\max}. \tag{2.6.51}$$

Bochner's lemma implies that

$$\sup_{h \in H^j_{T,\beta'}} \int_{\mathbb{R}^d} K^2_{\beta',h}(y) \mathrm{d}\mu(y) = (h^j_{T,\beta'})^{-d} \|K_{\beta'}\|^2_{L^2(\mathbb{R}^d)} \rho(x_0) \; (1 + o_T(1)), \tag{2.6.52}$$

where $o_T(1) \to 0$ uniformly in β, β'. For $h_1 \neq h_2 \in H^j_{T,\beta'}$, it further holds

$$
\begin{aligned}
\left\|K_{\beta',h_1} - K_{\beta',h_2}\right\|_{L^2(\mu)} &\leq \left\|K_{\beta',h_1} - K_{\beta',h_2}\right\|_{L^2(\mathbb{R}^d)} \\
&= (2\pi)^{-d/2} \left\|\phi_{K_{\beta'}}(h_1\ \cdot) - \phi_{K_{\beta'}}(h_2\ \cdot)\right\|_{L^2(\mathbb{R}^d)} \\
&\leq (2\pi)^{-d/2} \left(\int_{\mathbb{R}^d} \frac{|h_2^{2\beta'} - h_1^{2\beta'}|^2 \|y\|^{4\beta'}}{\left(1+\|h_1 y\|^{2\beta'}\right)^2 \left(1+\|h_2 y\|^{2\beta'}\right)^2} \,\mathrm{d}y\right)^{1/2} \\
&\leq h_2^{2\beta'} |1-(h_1/h_2)^{2\beta'}| \left((2\pi)^{-d} \int_{\mathbb{R}^d} \frac{\|y\|^{4\beta'}}{\left(1+\|h_2 y\|^{2\beta'}\right)^2} \,\mathrm{d}y\right)^{1/2} \\
&= h_2^{-d/2} |1-(h_1/h_2)^{2\beta'}| \left((2\pi)^{-d} \int_{\mathbb{R}^d} \frac{\|\lambda\|^{4\beta'}}{\left(1+\|\lambda\|^{2\beta'}\right)^2} \,\mathrm{d}\lambda\right)^{1/2} \\
&\leq O(1)\ (h^j_{T,\beta'})^{-d/2} |1-(h_1/h_2)^{2\beta'}|,
\end{aligned}
$$

where the last line follows from assertion (2.6.47) in Lemma 2.6.4. As concerns $\mathcal{K}_2$, note that

$$
\begin{aligned}
&\sup_{h \in H^j_{T,\beta'}} \left\|K_{\beta',h} - K_{\beta',h_{T,\beta'}}\right\|_\infty \\
&\quad\leq \sup_{h \in H^j_{T,\beta'}} \sup_{\lambda \in \mathbb{R}^d} (2\pi)^{-d/2} \left|\int_{\mathbb{R}^d} \left(\phi_{K_{\beta',h}}(y) - \phi_{K_{\beta',h_{T,\beta'}}}(y)\right) \mathrm{e}^{\mathrm{i}\lambda^t y} \mathrm{d}y\right| \\
&\quad\leq \sup_{h \in H^j_{T,\beta'}} (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{\left|h^{2\beta'} - (h^j_{T,\beta'})^{2\beta'}\right| \|y\|^{2\beta'}}{\left(1+\|hy\|^{2\beta'}\right)^2 \left(1+\|h^j_{T,\beta'} y\|^{2\beta'}\right)^2} \,\mathrm{d}y \\
&\quad\leq O(1) (h^j_{T,\beta'})^{-d} \left|1 - (h/h^j_{T,\beta'})^{2\beta'}\right| \\
&\quad\leq O(1) (h^j_{T,\beta'})^{-d}\ \beta_T \overline{\delta}_T
\end{aligned}
$$

and

$$
\begin{aligned}
\sup_{h \in H^j_{T,\beta'}} \left\|K_{\beta',h} - K_{\beta',h_{T,\beta'}}\right\|_{L^2(\mu)} &\leq O(1) \sup_{h \in H^j_{T,\beta'}} (h^j_{T,\beta'})^{-d/2} \left|1 - (h/h^j_{T,\beta'})^{2\beta'}\right| \\
&\leq O(1) (h^j_{T,\beta'})^{-d/2}\ \beta_T \overline{\delta}_T. \qquad (2.6.53)
\end{aligned}
$$

In particular, for $i \in \{1,2\}$,

$$
\int_0^1 \max\left\{\sqrt{\log N_{[\,]}(\varepsilon, \mathcal{K}_i, L^2(\mu))}, 1\right\} \mathrm{d}\varepsilon \leq \sqrt{\log T}.
$$

For $h \in H^j_{T,\beta}$, let

$$
\begin{aligned}
Z^j_{T,\beta}(h) &:= g^j_{T,\beta}(x_0, h) - \mathbf{E}_b g^j_{T,\beta}(x_0, h) \qquad (2.6.54) \\
&= \frac{2}{T} \int_0^T K_{\beta,h}(X_u) \mathrm{d}X^j_u - \int_{\mathbb{R}^d} K_{\beta,h}(y) \operatorname{div}(a_j \rho)(y) \mathrm{d}y.
\end{aligned}
$$

Lemma 2.6.5. *Grant Assumptions* (**BI**) *and* (**SG+**). *For any $\beta' > d/2$, the stochastic error $Z^j_{T,\beta'}(\cdot)$ defined according to (2.6.54) has the following properties:*

(a) *For any*

$$
u \in \left[\tau^j_T(\beta'),\ R_1\ \mathsf{s}^j_T(\beta') \sqrt{\log T}\right], \qquad R_1 > 0 \text{ an absolute constant,} \qquad (2.6.55)
$$

there exist some sufficiently small γ, independent of β', and some universal constant $c_1 > 0$ such that

$$\mathbf{P}_b\Big(\sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta'}(h)| > u\Big) \le c_1 \exp\left(-\frac{1}{2}\Big(\frac{u(1-\gamma)}{\mathsf{s}^j_T(\beta')}\Big)^2\right) + o(T^{-1}). \tag{2.6.56}$$

(b) *For any*

$$u \in [R_1\ \mathsf{s}^j_T(\beta')\sqrt{\log T},\ R_2], \qquad R_1, R_2 > 0 \text{ absolute constants,}$$

it holds, for some constants $c_2, c_3 > 0$,

$$\mathbf{P}_b\Big(\sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta'}(h)| > u\Big) \le c_2 \exp\left(-c_3\Big(\frac{u}{\mathsf{s}^j_T(\beta')}\Big)^2\right). \tag{2.6.57}$$

(c) *Assume that $\beta' < \beta$. Then, uniformly in $\beta \in \mathcal{B}_T$,*

$$\begin{aligned}\sup_{\beta'\in\mathcal{B},\ \beta'<\beta} m \sup_{b\in\Pi^j(\beta,L)} (\psi^j_{\beta,L})^{-2}\ \mathbf{E}_b\bigg(\Big(\sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta'}(h)|\Big)^2 &\\ \times \mathbb{1}\Big\{\sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta'}(h)| > \tau^j_T(\beta')\Big\}\bigg) &\to 0,\\ \sup_{\beta'\in\mathcal{B},\ \beta'<\beta} m \sup_{b\in\Pi^j(\beta,L)} (\psi^j_{\beta,L})^{-2}\ \mathbf{E}_b\bigg(\Big(\sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta'}(h)|\Big)^2 &\\ \times \mathbb{1}\Big\{\sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta'}(h)| > \sqrt{\mathsf{s}^j_T(\beta')\ \psi^j_{\beta,L}}\Big\}\bigg) &\to 0.\end{aligned}$$

Proof. The assertions are analogue to the statements in Lemma 4.3, Lemma 4.5 and Theorem 4.6 in Butucea (2001). The inequalities in (a) and (b) follow from the Bernstein-type deviation inequality (**BI**) and its implication (**BI**+) stated in Lemma 2.5.13. For deriving the inequality (2.6.56) with the specific factor 1/2 in the exponent, we however have to go into greater detail.

We start with proving (2.6.56). Throughout the proof, $D_1, D_2, \dots$ denote positive constants. For fixed $\delta \in (0,1)$ and arbitrary $u, T > 0$, write

$$\mathbf{P}_b\Big(\Big|Z^j_{T,\beta'}(h^j_{T,\beta'})\Big| > u\Big) \le \mathsf{t}_1 + \mathsf{t}_2,$$

for

$$\mathsf{t}_1 := \mathbf{P}_b\Big(\Big|\frac{1}{T}\int_0^T \Big(2K_{\beta',h^j_{T,\beta'}}(X_u)b^j(X_u) - \int_{\mathbb{R}^d} K_{\beta',h^j_{T,\beta'}}(y)\operatorname{div}(a_j\rho)(y)\mathrm{d}y\Big)\mathrm{d}u\Big| > \delta u\Big),$$

and, denoting $M_t(K) := 2\int_0^t K(X_u)\sum_{r=1}^d \sigma_{jr}\ \mathrm{d}W^r_u$, $t > 0$,

$$\mathsf{t}_2 := \mathbf{P}_b\Big(\big|M_T(K_{\beta',h^j_{T,\beta'}})\big| > (1-\delta)u\Big). \tag{2.6.58}$$

For any $u \le R_1\ \mathsf{s}^j_T(\beta')\sqrt{\log T}$, we have

$$\begin{aligned}u\|K_{\beta,h^j_{T,\beta'}}\|_\infty &\le u(h^j_{T,\beta'})^{-d}\ K_{\max} \le R_1\ \mathsf{s}^j_T(\beta')(h^j_{T,\beta'})^{-d}\sqrt{\log T}\ K_{\max}\\ &\le D_1(\mathsf{s}^j_T(\beta'))^2\ T(h^j_{T,\beta'})^{-d/2+\beta'},\end{aligned}$$

such that, since $\beta' > d/2$,

$$\frac{u\|K_{\beta,h}\|_\infty}{\left(\mathsf{s}_T^j(\beta')\right)^2 T} = o_T(1).$$

Furthermore, the enhanced spectral gap assumption (**SG**+) gives

$$\varsigma(2K_{\beta,h^j_{T,\beta'}}) \le D_2 (h^j_{T,\beta'})^{2-d}.$$

Thus, for T sufficiently large,

$$C_B\Big(\varsigma(2K_{\beta,h^j_{T,\beta'}}) + \delta u \;\|K_{\beta,h^j_{T,\beta'}}\|_\infty\Big) \le \left(\mathsf{s}_T^j(\beta')\right)^2 T.$$

The Bernstein-type deviation inequality (**BI**) therefore implies that

$$\mathsf{t}_1 \le 2\exp\left(-\frac{T\delta^2u^2}{2\Big(\varsigma(2K_{\beta,h^j_{T,\beta'}}) + \delta u \;\|K_{\beta,h^j_{T,\beta'}}\|_\infty\Big)}\right) \le 2\exp\left(-\frac{\delta^2u^2}{2\left(s_T^j(\beta')\right)^2}\right). \tag{2.6.59}$$

In view of (2.6.52), there exists some sufficiently small $\delta_{T0} > 0$ such that

$$a_{jj}\;\|2K_{\beta',h}\|^2_{L^2(\mu)} = \frac{\left(\mathsf{s}_T^j(\beta')\right)^2}{1-\delta_{T0}}.$$

Consequently, Bernstein's inequality for continuous martingales gives, for any $h > 0$,

$$\begin{aligned}
\mathbf{P}_b\Big(|M_T(K_{\beta',h})| > T(1-\delta)u;\; \langle M(K_{\beta',h})\rangle_T \le T a_{jj} c_P^{-1}\;\varsigma(2K_{\beta',h})\Big)&\\
\le 2\exp\left(-\frac{T(1-\delta)^2u^2}{2a_{jj}\;\|2K_{\beta',h}\|^2_{L^2(\mu)}}\right)&\\
\le 2\exp\left(-\frac{(1-\delta)^2u^2\;(1-\delta_{T0})}{2\left(\mathsf{s}_T^j(\beta')\right)^2}\right)&.
\end{aligned} \tag{2.6.60}$$

For any $h > 0$, (**BI**) entails that

$$\begin{aligned}
&\mathbf{P}_b\Big(\langle M(K_{\beta',h})\rangle_T > T a_{jj} c_P^{-1}\;\varsigma(2K_{\beta',h})\Big)\\
&\quad= \mathbf{P}_b\Big(\frac{1}{T}\int_0^T K^2_{\beta',h}(X_u)\mathrm{d}u > c_P^{-1}\;\varsigma(K_{\beta',h})\Big)\\
&\quad\le \exp\left(-\frac{T\;c_P^{-2}\varsigma^2(K_{\beta',h})}{2C_B\Big(\varsigma(K^2_{\beta',h}) + c_P^{-1}\|K^2_{\beta',h}\|_\infty\;\varsigma(K_{\beta',h})\Big)}\right).
\end{aligned}$$

Plugging in the specific choice $h = h^j_{T,\beta'}$, it follows

$$\mathbf{P}_b\Big(\langle M(K_{\beta',h^j_{T,\beta'}})\rangle_T > T a_{jj} c_P^{-1}\;\varsigma(2K_{\beta',h^j_{T,\beta'}})\Big) = o(T^{-1}), \tag{2.6.61}$$

using again that $\beta' > d/2$. Adding the upper bounds (2.6.59), (2.6.60) and (2.6.61), we obtain, for some small $\gamma > 0$,

$$\mathbf{P}_b\Big(\Big|Z^j_{T,\beta'}(h^j_{T,\beta'})\Big| > u\Big) \le 2\exp\left(-\frac{u^2(1-\gamma)}{2\left(\mathsf{s}_T^j(\beta')\right)^2}\right) + o(T^{-1}). \tag{2.6.62}$$

Consider the sequence

$$\delta_{T1} := \beta_T\bar{\delta}_T\sqrt{\log T} = (\log\log T)^{\delta_1}(\log T)^{-1/2} \to 0,$$

and note that, for any $u > 0$,

$$\mathbf{P}_b\Bigg(\sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta'}(h)| > u\Bigg) \le \mathbf{P}_b\Bigg(\sup_{h\in H^j_{T,\beta'}} \Big|Z^j_{T,\beta'}(h) - Z^j_{T,\beta'}(h^j_{T,\beta'})\Big| > u\delta_{T1}\Bigg) + \mathbf{P}_b\Bigg(\Big|Z^j_{T,\beta'}(h^j_{T,\beta'})\Big| > u(1-\delta_{T1})\Bigg).$$

Since $u(1-\delta_{T1}) \le u \le R_1\ \mathsf{s}^j_T(\beta')\sqrt{\log T}$, (2.6.62) gives an upper bound on the latter summand. For T large enough, it further holds

$$\Big[\tau^j_T(\beta')\delta_{T1},\ R_1\delta_{T1}\mathsf{s}^j_T(\beta')\ \sqrt{\log T}\Big] \subseteq \left[\frac{\beta_T\overline{\delta}_T\sqrt{\log T}}{\sqrt{T}\big(h^j_{T,\beta'}\big)^{d/2}},\ \beta_T\overline{\delta}_T\right].$$

Taking into account (2.6.53) and since

$$\int_0^1 \max\Big\{\sqrt{\log N_{[\,]}(\varepsilon,\mathcal{K}_2,L^2(\mu)},1\Big\}\,\mathrm{d}\varepsilon \le D_4\frac{\beta_T\overline{\delta}_T\sqrt{\log T}}{\big(h^j_{T,\beta'}\big)^{d/2}},$$

the uniform exponential inequality (**BI**+) implies that

$$\mathbf{P}_b\Bigg(\sup_{h\in H^j_{T,\beta'}} \Big|Z^j_{T,\beta'}(h) - Z^j_{T,\beta'}(h^j_{T,\beta'})\Big| > u\delta_{T1}\Bigg) \le C_1\exp\left(-\frac{D_5T\big(h^j_{T,\beta'}\big)^d(u\delta_{T1})^2}{(\beta_T\overline{\delta}_T)^2}\right). \tag{2.6.63}$$

Summing the upper bounds due to (2.6.62) and (2.6.63), we obtain (2.6.56).

The inequality stated in (b) follows as an application of (**BI**+) with

$$K := \big(h^j_{T,\beta'}\big)^{-d}K_{\max} \quad\text{and}\quad M^2 := \big(h^j_{T,\beta'}\big)^{-d}\|K_{\beta'}\|^2_{L^2(\mathbb{R}^d)}\rho(x_0).$$

Part (c) is proven similarly to Theorem 4.6 in Butucea (2001) by noting that there exists some positive constant R_3 such that

$$\sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta'}(h)| \le R_3\big(h^j_{T,\beta'}\big)^{-d}$$

and decomposing

$$\mathbf{E}_b\Bigg(\Bigg(\sup_{h\in H^j_{T,\beta'}}|Z^j_{T,\beta'}(h)|\Bigg)^2\mathbb{1}\Bigg\{\sup_{h\in H^j_{T,\beta'}}|Z^j_{T,\beta'}(h)| > \tau^j_T(\beta')\Bigg\}\Bigg) \le \mathcal{I}_1+\mathcal{I}_2+\mathcal{I}_3,$$

for

$$\mathcal{I}_1 := \int_{\tau^j_T(\beta')}^{R_1\mathsf{s}^j_T(\beta')\sqrt{\log T}} \mathbf{P}_b\Bigg(\sup_{h\in H^j_{T,\beta'}}|Z^j_{T,\beta'}(h)| > u\Bigg)\mathrm{d}u^2,$$

$$\mathcal{I}_2 := \int_{R_1\mathsf{s}^j_T(\beta')\sqrt{\log T}}^{R_2} \mathbf{P}_b\Bigg(\sup_{h\in H^j_{T,\beta'}}|Z^j_{T,\beta'}(h)| > u\Bigg)\mathrm{d}u^2,$$

$$\mathcal{I}_3 := \int_{R_2}^{R_3\big(h^j_{T,\beta'}\big)^{-d}} \mathbf{P}_b\Bigg(\sup_{h\in H^j_{T,\beta'}}|Z^j_{T,\beta'}(h)| > u\Bigg)\mathrm{d}u^2.$$

This decomposition provides sufficiently tight upper bounds on the terms $\mathcal{I}_1$ to $\mathcal{I}_3$ by exploiting the interplay of the interval length and uniform exponential bounds on the integrands as they follow from parts (a) and (b) proven above. $\square$

The next lemma contains a decomposition of the normalizing factor $\psi^j_{\beta,L}$ and some relations which are needed later in the proof. For $h^j_{T,\beta}$ defined according to (2.6.41), denote

$$\mathsf{b}^j_{T,\beta'} := 2L\ \mathsf{b}_{\beta,\beta'}\ (h^j_{T,\beta'})^{\widetilde{\beta}-d/2}. \tag{2.6.64}$$

Lemma 2.6.6. *Let $\beta \in [d/2,\infty)$, $L \in [L_*, L^*]$, and denote $\nu = (\beta, L)$. It then holds*

$$\psi^j_\nu = (2L)^{d/(2\beta)} \left(\eta^j_{T,\beta} + (h^j_{T,\beta})^{\beta-d/2}\ \mathsf{b}_{\beta,\beta}\right). \tag{2.6.65}$$

Furthermore, there exist positive constants $D_1, \ldots, D_5$, depending only on β_*, L_*, L^*, d and σ, such that

$$D_1 \le \psi^j_\nu / \eta^j_{T,\beta} \le D_2, \tag{2.6.66}$$

and, for $\beta' \in [d/2,\infty)$, $\beta' < \beta$,

$$\frac{D_3}{\beta_T}\left(\frac{\log T}{T}\right)^{\kappa(\beta')-\kappa(\beta)} \le \frac{\psi^j_{\beta',L}}{\psi^j_\nu} \le D_4 T^{\kappa(\beta)-\kappa(\beta')} \tag{2.6.67}$$

and

$$\frac{\left(\mathsf{b}^j_{T,\beta'}\right)^2 + \left(\tau^j_T(\beta')\right)^2}{\left(\psi^j_\nu\right)^2} \le D_5 \log T\ T^{2\kappa(\beta)-2\kappa(\beta')}. \tag{2.6.68}$$

Proof. Recall that the extremal function $g_{\beta,L,\varepsilon}$ as introduced in Section 2.5.5 satisfies

$$g_{\beta,L,\varepsilon}(0) \overset{(2.5.34)}{=} \sup\left\{f(0) : \int_{\mathbb{R}^d} f^2(x)\mathrm{d}\lambda\!\!\lambda(x) \le \varepsilon,\ \eta^2_\beta(f) \le L\right\} = \omega_{\beta,L}(\varepsilon),$$

for $f := \operatorname{div}(a_j\rho)$. It follows from renormalization arguments that, for any $x \in \mathbb{R}^d$,

$$g_{\beta,2L,\widetilde{T}^j(\beta)}(x) \quad = \quad a_1 g_{\beta,1,1}(b_1 x),$$

where $b_1 = (2L/\widetilde{T}^j(\beta))^{1/\beta}$ and

$$a_1 = 2Lb_1^{d/2-\beta} = (2L)^{d/(2\beta)}\left(\widetilde{T}^j(\beta)\right)^{1-d/(2\beta)} = (2L)^{d/(2\beta)}\left(h^j_{T,\beta}\right)^{\beta-d/2}.$$

Since

$$g_{\beta,1,1}(0) \overset{(2.5.36)}{=} \mathbb{I}_\beta^{-1}\ \mathsf{b}^{-\beta+d/2}\widetilde{K}_\beta(0) \overset{(2.5.37)}{=} \mathbb{I}_\beta\ \mathsf{b}^{-\beta+d/2}\ \frac{2\beta}{d}, \tag{2.6.69}$$

this implies that

$$\begin{aligned} g_{\beta,2L,\widetilde{T}^j(\beta)}(0) &= (2L)^{d/(2\beta)}\left(\widetilde{T}^j(\beta)\right)^{1-d/(2\beta)}\mathbb{I}_\beta\ \mathsf{b}^{-\beta+d/2}\ \frac{2\beta}{d} \\ &= (2L)^{d/(2\beta)}\left(\frac{4d\rho(x_0)a_{jj}\log T}{\beta T}\right)^{\frac{\beta-d/2}{2\beta}}\left(\frac{2\beta-d}{d}\right)^{-\frac{\beta-d/2}{2\beta}}\ \mathbb{I}_\beta\ \frac{2\beta}{d} = \psi^j_\nu. \end{aligned} \tag{2.6.70}$$

The relation $g_{\beta,1,1}(0) = \|K_\beta\|_{L^2(\mathbb{R}^d)} + \mathsf{b}_{\beta,\beta}$ from Lemma 2.5.12 thus yields the decomposition

$$\begin{aligned} \psi^j_\nu &\overset{(2.6.70)}{=} g_{\beta,2L,\widetilde{T}^j(\beta)}(0) \overset{(2.6.69)}{=} a_1 g_{\beta,1,1}(0) \\ &\overset{(2.5.39)}{=} a_1\left(\|K_\beta\|_{L^2(\mathbb{R}^d)} + \mathsf{b}_{\beta,\beta}\right) = (2L)^{d/(2\beta)}\left(\eta^j_{T,\beta} + (h^j_{T,\beta})^{\beta-d/2}\mathsf{b}_{\beta,\beta}\right). \end{aligned}$$

Assertions (2.6.66) and (2.6.67) follow analogously to the proof of the relations (44)–(46) in Lemma 4 in [KT04] (pp. 453–454). Finally, for the proof of (2.6.68), we refer to the proof of Lemma 3.5 in Butucea (2001).

□

Main part of the proof of the upper bound. Define $\beta^- = \beta^-(\beta)$ by

$$\beta^- := \beta - \frac{\beta_T^+}{\log T}, \tag{2.6.71}$$

where $\beta_T^+ := (\log\log T)^{\delta_3}$, for some $\delta_3 \in (\delta_2, 1)$. We follow the standard approach and decompose the risk successively. Let $\nu = (\beta, L)$, and set

$$\mathcal{R}_{T,\nu}^+(j) := \sup_{b \in \Pi^j(\beta,L)} \left(\psi_\nu^j\right)^{-2} \mathbf{E}_b \left(\left|\widetilde{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|^2 \mathbb{1}\{\widehat{\beta}_T^j \geq \beta^-\} \right),$$
$$\mathcal{R}_{T,\nu}^-(j) := \sup_{b \in \Pi^j(\beta,L)} \left(\psi_\nu^j\right)^{-2} \mathbf{E}_b \left(\left|\widetilde{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|^2 \mathbb{1}\{\widehat{\beta}_T^j < \beta^-\} \right).$$

(I) We first consider the case $\widehat{\beta}_T^j \geq \beta^-$, and we show that

$$\limsup_{T\to\infty} \sup_{\nu \in \mathcal{B}_T} \mathcal{R}_{T,\nu}^+(j) \leq 1. \tag{2.6.72}$$

Define $\overline{\beta} = \overline{\beta}(\beta)$ via the equation

$$\left(\frac{\log T}{4L^2 T}\right)^{1/(2\beta)} = \left(\frac{\log T}{T}\right)^{1/(2\overline{\beta})}. \tag{2.6.73}$$

Let $\beta^+ \in \mathcal{G}$ be the largest grid point $\leq \overline{\beta}$, and assume that T is large enough for ensuring

$$\beta^- = \beta - \frac{\beta_T^+}{\log T} = \beta - \frac{(\log\log T)^{\delta_3}}{\log T} < \beta^+. \tag{2.6.74}$$

Denote

$$\mathcal{G}_1 = \mathcal{G}_1(\beta) := \left\{\beta' \in \mathcal{G} : \beta^- \leq \beta' \leq \beta^+\right\}, \tag{2.6.75}$$
$$\mathcal{G}_2 = \mathcal{G}_2(\beta) := \left\{\beta' \in \mathcal{G} : \beta^+ < \beta' \leq \beta_T\right\}, \tag{2.6.76}$$

and rewrite

$$\begin{aligned}
\mathcal{R}_{T,\nu}^+(j) &= \sup_{b \in \Pi^j(\beta,L)} \left(\psi_\nu^j\right)^{-2} \mathbf{E}_b \left(\left|\widetilde{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|^2 \mathbb{1}\{\widehat{\beta}_T^j \geq \beta^-\} \right)\\
&= \sup_{b \in \Pi^j(\beta,L)} \left(\psi_\nu^j\right)^{-2} \mathbf{E}_b \left(\left|\widetilde{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|^2 \mathbb{1}\{\widehat{\beta}_T^j \in \mathcal{G}_1 \cup \mathcal{G}_2\} \right).
\end{aligned}$$

Let $\beta' \in \mathcal{G}_1 = \mathcal{G}_1(\beta)$ and $b \in \Pi^j(\beta, L)$, and assume that T is so large that $\widetilde{\beta}(\beta,\beta') = \beta$. It holds

$$\begin{aligned}
&\sup_{h \in \mathcal{H}_{T,\beta'}^j} \left|\mathbf{E}_b g_{T,\beta'}^j(x_0,h) - \operatorname{div}(a_j\rho)(x_0)\right|\\
&\qquad\leq \sup_{h \in \mathcal{H}_{T,\beta'}^j} \left(2Lh^{\widetilde{\beta}-d/2}\mathsf{b}_{\beta,\beta'}\right)\\
&\qquad\leq 2L(h_{T,\beta'}^j)^{\beta-d/2}\mathsf{b}_{\beta,\beta'}\ (1+\overline{\delta}_T)\\
&\qquad= 2L\left(\frac{4d\rho(x_0)a_{jj}\log T}{\beta' T}\right)^{\frac{\beta-d/2}{2\beta'}} \mathsf{b}_{\beta,\beta'}\ (1+\overline{\delta}_T)\\
&\qquad\leq 2L\left(\frac{4d\rho(x_0)a_{jj}}{\beta'}\right)^{\frac{\beta-d/2}{2\beta'}} \left(\frac{\log T}{T}\right)^{\frac{\beta-d/2}{2\overline{\beta}}} \mathsf{b}_{\beta,\beta'}\ (1+\overline{\delta}_T)\\
&\qquad= 2L\left(\frac{4d\rho(x_0)a_{jj}}{\beta'}\right)^{\frac{\beta-d/2}{2\beta'}} \left(\frac{\log T}{4L^2T}\right)^{\frac{\beta-d/2}{2\beta}} \mathsf{b}_{\beta,\beta'}\ (1+\overline{\delta}_T)\\
&\qquad= \Lambda(\beta,\beta')(2L)^{d/(2\beta)}(h_{T,\beta}^j)^{\beta-d/2}\mathsf{b}_{\beta,\beta'}\ (1+\overline{\delta}_T),
\end{aligned}$$

where

$$\Lambda(\beta,\beta') := (4d\rho(x_0)a_{jj})^{\frac{\beta-d/2}{2\beta'}-\frac{\beta-d/2}{2\beta}}(\beta')^{-\frac{\beta-d/2}{2\beta'}}\beta^{\frac{\beta-d/2}{2\beta}}.$$

We used Lemma 2.6.4, the facts that $\mathcal{H}^j_{T,\beta'} \subset H^j_{T,\beta'}$, that $\beta' \leq \overline{\beta}$ and the definition of $\overline{\beta}$ to arrive at the above upper bound.

We now argue analogously to the proof of the upper bound in [KT04] (see pp. 461–463). For any $\beta' \in \mathcal{G}_1$, there exists some positive constant C such that

$$|\beta-\beta'| \leq C\beta_T^+(\log T)^{-1}. \tag{2.6.77}$$

Since $\Lambda(\beta,\beta')$ is uniformly continuous in $\beta,\beta' \in [\beta_*,\infty)$, this implies that $\Lambda(\beta,\beta') \leq 1+\gamma_{T1}$. Furthermore, for any $\beta' \in \mathcal{G}_1$, $\beta \in [\beta_*,\beta_T]$, it holds $\mathsf{b}_{\beta,\beta'} \leq \mathsf{b}_{\beta,\beta}\ (1+\gamma_{T2})$. Consequently, for any $\beta' \in \mathcal{G}_1$, $b \in \Pi^j(\beta,L)$,

$$\sup_{h\in\mathcal{H}^j_{T,\beta'}} \left|\mathbf{E}_b g^j_{T,\beta'}(x_0,h) - \operatorname{div}(a_j\rho)(x_0)\right| \leq (2L)^{d/(2\beta)}(h^j_{T,\beta})^{\beta-d/2}\mathsf{b}_{\beta,\beta}\ (1+\gamma_{T3}).$$

Similar arguments (also see the derivation of line (54) on p. 1591 in [KT01]) yield

$$\begin{aligned}
\eta^j_{T,\beta^+} &= \left(\widetilde{T}^j(\beta^+)\right)^{\frac{\beta^+-d/2}{\beta^+}}\|K_{\beta^+}\|_{L^2(\mathbb{R}^d)}\\
&\leq \left(\frac{4d\rho(x_0)a_{jj}}{\beta^+}\right)^{\frac{\beta^+-d/2}{2\beta^+}}\left(\frac{\log T}{T}\right)^{\frac{\overline{\beta}-d/2}{2\overline{\beta}}}\|K_{\beta^+}\|_{L^2(\mathbb{R}^d)}\ (1+\gamma_{T4})\\
&= \left(\frac{4d\rho(x_0)a_{jj}}{\beta^+}\right)^{\frac{\beta^+-d/2}{2\beta^+}}\left(\frac{\log T}{T}\right)^{1/2}\left(\frac{\log T}{4L^2T}\right)^{-d/(4\beta)}\|K_{\beta^+}\|_{L^2(\mathbb{R}^d)}\ (1+\gamma_{T4})\\
&\leq \left(\frac{4d\rho(x_0)a_{jj}\log T}{\beta T}\right)^{\frac{\beta-d/2}{2\beta}}\|K_{\beta}\|_{L^2(\mathbb{R}^d)}(2L)^{d/(2\beta)}\ (1+\gamma_{T5})\\
&= (2L)^{d/(2\beta)}\eta^j_{T,\beta}\ (1+\gamma_{T5}).
\end{aligned} \tag{2.6.78}$$

Recall the definition of the stochastic error $Z^j_{T,\beta}(\cdot)$, for any $h \in H^j_{T,\beta}$ given as

$$Z^j_{T,\beta}(h) = g^j_{T,\beta}(x_0,h) - \mathbf{E}_b g^j_{T,\beta}(x_0,h).$$

Whenever $\widehat{\beta}^j_T = \beta' \in \mathcal{G}_1$ and the event $A^j_{T,\beta'}$ holds, the above arguments imply that

$$\begin{aligned}
&\left|\widetilde{g}^j_T(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|\\
&\quad= \left|\widehat{g}^j_{T,\beta'}(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|\\
&\quad\leq \sup_{h\in\mathcal{H}^j_{T,\beta'}}\left|g^j_{T,\beta'}(x_0,h) - \operatorname{div}(a_j\rho)(x_0)\right|\\
&\quad\leq \sup_{h\in\mathcal{H}^j_{T,\beta'}}\left\{\left|g^j_{T,\beta'}(x_0,h) - \mathbf{E}_b g^j_{T,\beta'}(x_0,h)\right| + \left|\mathbf{E}_b g^j_{T,\beta'}(x_0,h) - \operatorname{div}(a_j\rho)(x_0)\right|\right\}\\
&\quad\leq \sup_{h\in\mathcal{H}^j_{T,\beta'}}\left|Z^j_{T,\beta'}(h)\right| + (2L)^{d/(2\beta)}(h^j_{T,\beta})^{\beta-d/2}\mathsf{b}_{\beta,\beta}\ (1+\gamma_{T3}) &(2.6.79)\\
&\quad\leq \sup_{h\in H^j_{T,\beta'}}\left|Z^j_{T,\beta'}(h)\right| + \psi^j_\nu\ (1+\gamma_{T3}). &(2.6.80)
\end{aligned}$$

The last line holds true since $\mathcal{H}_{T,\beta'} \subset H_{T,\beta'}$ and in view of the decomposition of the normalizing factor ψ^j_ν according to (2.6.65). If $\widehat{\beta}^j_T \geq \beta^+$, the definition of the estimator $\widehat{\beta}^j_T$ according to (2.5.23) implies that

$$\left|\widehat{g}^j_{T,\widehat{\beta}^j_T}(x_0) - \widehat{g}^j_{T,\beta^+}(x_0)\right| \leq \widehat{\eta}^j_{T,\beta^+}. \tag{2.6.81}$$

Therefore, if $\widehat{\beta}_T^j = \beta' \in \mathcal{G}_2$, it holds on $A^j_{T,\beta+}$,

$$\begin{aligned}
&|\widetilde{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)| \\
&\quad\le \Big(|\widehat{g}^j_{T,\beta'}(x_0) - \widehat{g}^j_{T,\beta+}(x_0)| + |\widehat{g}^j_{T,\beta+}(x_0) - \operatorname{div}(a_j\rho)(x_0)|\Big) \\
&\quad\le \widehat{\eta}^j_{T,\beta+} + |\widehat{g}^j_{T,\beta+}(x_0) - \operatorname{div}(a_j\rho)(x_0)| \\
&\quad\le \sup_{h\in\mathcal{H}^j_{T,\beta+}} \Big\{\eta^j_{T,\beta+}\left(h/h^j_{T,\beta}\right)^{\beta-d/2} + |g^j_{T,\beta+}(x_0,h) - \mathbf{E}_b g^j_{T,\beta+}(x_0,h)| \\
&\qquad\qquad + |\mathbf{E}_b g^j_{T,\beta+}(x_0,h) - \operatorname{div}(a_j\rho)(x_0)|\Big\} \\
&\quad\overset{(2.6.79)}{\le} \eta^j_{T,\beta+}\ (1+\overline{\delta}_T) + \sup_{h\in H^j_{T,\beta+}} |Z^j_{T,\beta+}(h)| + (2L)^{d/(2\beta)}\big(h^j_{T,\beta}\big)^{\beta-d/2}\mathsf{b}_{\beta,\beta}\ (1+\gamma_{T3}) \\
&\quad\overset{(2.6.78)}{\le} \sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta+}(h)| + \Big((2L)^{d/(2\beta)}\eta^j_{T,\beta}\ (1+\gamma_{T6}) \\
&\qquad\qquad + (2L)^{d/(2\beta)}\big(h^j_{T,\beta}\big)^{\beta-d/2}\mathsf{b}_{\beta,\beta}\ (1+\gamma_{T3})\Big).
\end{aligned}$$

In view of the decomposition (2.6.65), this last line implies that

$$|\widetilde{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)|\mathbb{1}\{\widehat{\beta}_T^j = \beta' \in \mathcal{G}_2\}\ \mathbb{1}\{A^j_{T,\beta+}\} \le \sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta+}(h)| + \psi^j_\nu\ (1+\gamma_{T7}). \quad (2.6.82)$$

Thus, using (2.6.80) and (2.6.82),

$$\begin{aligned}
&\big(\psi^j_\nu\big)^{-2}\ \mathbf{E}_b\bigg(\big|\widetilde{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)\big|^2\ \mathbb{1}\{\widehat{\beta}_T^j \in \mathcal{G}_1\cup\mathcal{G}_2\}\bigg) \\
&\quad\le \sum_{\beta'\in\mathcal{G}_1} \mathbf{E}_b\bigg(\Big(1+\gamma_{T3} + \big(\psi^j_\nu\big)^{-1}\sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta'}(h)|\Big)^2\ \mathbb{1}\{\widehat{\beta}_T^j = \beta'\}\ \mathbb{1}\{A^j_{T,\beta'}\}\bigg) \\
&\qquad + \sum_{\beta'\in\mathcal{G}_1} \mathbf{E}_b\bigg(\big(\psi^j_\nu\big)^{-2}\ |\widetilde{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)|^2\ \mathbb{1}\{\widehat{\beta}_T^j = \beta'\}\ \mathbb{1}\{(A^j_{T,\beta'})^{\mathrm{c}}\}\bigg) \\
&\qquad + \sum_{\beta'\in\mathcal{G}_2} \mathbf{E}_b\bigg(\Big(1+\gamma_{T7} + \big(\psi^j_\nu\big)^{-1}\sup_{h\in H^j_{T,\beta+}} |Z^j_{T,\beta+}(h)|\Big)^2\ \mathbb{1}\{\widehat{\beta}_T^j = \beta'\}\ \mathbb{1}\{A^j_{T,\beta+}\}\bigg) \\
&\qquad + \sum_{\beta'\in\mathcal{G}_2} \mathbf{E}_b\bigg(\big(\psi^j_\nu\big)^{-2}|\widetilde{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)|^2\ \mathbb{1}\{\widehat{\beta}_T^j = \beta'\}\ \mathbb{1}\{(A^j_{T,\beta+})^{\mathrm{c}}\}\bigg) \\
&\quad =: \sum_{\beta'\in\mathcal{G}_1}\Big(\mathsf{p}^j_1(\beta') + \mathsf{p}^j_2(\beta')\Big) + \sum_{\beta'\in\mathcal{G}_2}\Big(\mathsf{p}^j_3(\beta') + \mathsf{p}^j_4(\beta')\Big), \text{ say.}
\end{aligned} \quad (2.6.83)$$

The terms $\mathsf{p}^j_1(\cdot),\dots,\mathsf{p}^j_4(\cdot)$ are now considered separately. Note first that, for any $\beta' \in \mathcal{G}_1$,

$$\begin{aligned}
\mathsf{p}^j_1(\beta') &\le \Big(1+\gamma_{T3} + \sqrt{\mathsf{s}^j_T(\beta')\big(\psi^j_\nu\big)^{-1}}\Big)^2\ \mathbf{P}_b\Big(\widehat{\beta}_T^j = \beta'\Big) \\
&\quad + \mathbf{E}_b\bigg(\Big(1+\gamma_{T3} + \big(\psi^j_\nu\big)^{-1}\sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta'}(h)|\Big)^2 \\
&\qquad\qquad \times \mathbb{1}\Big\{\sup_{h\in H_{T,\beta'}} |Z^j_{T,\beta'}(h)| > \sqrt{\mathsf{s}^j_T(\beta')\psi^j_\nu}\Big\}\bigg).
\end{aligned} \quad (2.6.84)$$

Since (2.6.77) holds for any $\beta' \in \mathcal{G}_1$, (2.6.67) implies that

$$\psi^j_{\beta',L}/\psi^j_\nu \leq D_4\, T^{\frac{\beta-d/2}{2\beta}-\frac{\beta'-d/2}{2\beta'}} \leq D_4 \exp\left(C\beta_T^+\right).$$

Together with (2.6.66), this yields

$$\eta^j_{T,\beta'} \leq D_1^{-1}\ \psi^j_{\beta',L} \leq D_1^{-1} D_4\ \exp\left(C\beta_T^+\right)\psi^j_\nu.$$

Consequently,

$$\begin{aligned}\frac{\mathsf{s}^j_T(\beta')}{\psi^j_\nu} &\leq \frac{D_1^{-1}D_4\ \exp\left(C\beta_T^+\right)\ \mathsf{s}^j_T(\beta')}{\eta^j_{T,\beta'}} \leq \frac{D_1^{-1}D_4\ \exp\left(C\beta_T^+\right)}{d_T(\beta_T)}\\ &\leq D_1^{-1}D_4\ \exp\left(C\beta_T^+\right)\sqrt{\frac{\beta_T}{\log T}} =: A_T.\end{aligned}$$

The summand in (2.6.84) is bounded from above by the sum of the terms

$$\begin{aligned}&2\ (1+\gamma_{T3})^2\ \mathbf{P}_b\Big(\sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta'}(h)| > \sqrt{\mathsf{s}^j_T(\beta')\psi^j_\nu}\Big)\\ &\qquad\qquad\qquad \overset{(2.6.56)}{\leq} 2c_1\ (1+\gamma_{T3})^2 \exp\left(-\frac{\psi^j_\nu(1-\gamma_T)^2}{2\mathsf{s}^j_T(\beta')}\right)\end{aligned}$$

and

$$2\ (\psi^j_\nu)^{-2}\ \mathbf{E}_b\bigg(\Big(\sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta'}(h)|\Big)^2\ \mathbb{1}\Big\{\sup_{h\in H^j_{T,\beta'}} |Z^j_{T,\beta'}(h)| > \sqrt{\mathsf{s}^j_T(\beta')\psi^j_\nu}\Big\}\bigg).$$

Part (c) of Lemma 2.6.5 entails that the latter term tends to zero, uniformly in $\beta' \in \mathcal{G}_1$. Therefore,

$$\mathsf{p}^j_1(\beta') \leq \left(1+\gamma_{T3}+\sqrt{A_T}\right)^2\ \mathbf{P}_b\Big(\widehat{\beta}^j_T = \beta'\Big) + O(1)\exp\left(-\frac{(1-\gamma_T)^2}{2A_T}\right) + o_T(1).$$

Recall that the cardinality m of the grid $\mathcal{G}$ satisfies

$$m \leq k_1^{-1}\beta_T(\log T)^{\delta_1} = k_1^{-1}(\log T)^{\delta_1}(\log\log T)^{\delta_2}.$$

By construction, $\beta^+ \in \mathcal{G}_1$, such that $\mathsf{p}^j_3(\beta')$ is upper-bounded analogously. Consequently,

$$\begin{aligned}\sum_{\beta'\in\mathcal{G}_1}\mathsf{p}^j_1(\beta') + \sum_{\beta'\in\mathcal{G}_2}\mathsf{p}^j_3(\beta') &\leq \Big(1+\max\left\{\gamma_{T3},\gamma_{T7}\right\}+\sqrt{A_T}\Big)^2\ \mathbf{P}_b\Big(\widehat{\beta}^j_T \in \mathcal{G}_1\cup\mathcal{G}_2\Big)\\ &\qquad\qquad + O(1)m\exp\left(-\frac{(1-\gamma_T)^2}{2A_T}\right) + o_T(1)\\ &\leq 1+o_T(1).\end{aligned}$$

For $\mathsf{p}^j_2(\cdot)$ and any $\beta' \in \mathcal{G}_1$, there exists some constant c_0 such that

$$\begin{aligned}\mathsf{p}^j_2(\beta') &\leq (\psi^j_\nu)^{-2}\ \mathbf{E}_b\left(\left|\widehat{g}^j_{T,\beta'}(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|^2\ \mathbb{1}\{(A^j_{T,\beta'})^{\mathrm{c}}\}\right)\\ &\leq (\psi^j_\nu)^{-2}\ \left(\mathbf{E}_b\left(\left|\widehat{g}^j_{T,\beta'}(x_0) - \operatorname{div}(a_j\rho)(x_0)\right|^4\right)\right)^{1/2}\ \left(\mathbf{P}_b\Big((A^j_{T,\beta'})^{\mathrm{c}}\Big)\right)^{1/2}\\ &\leq c_0\left(\mathbf{P}_b\Big((A^j_{T,\beta'})^{\mathrm{c}}\Big)\right)^{1/2}.\end{aligned}$$

Lemma 2.6.3 then implies that $\mathsf{p}_2^j(\beta')$ is exponentially small, for any $\beta' \in \mathcal{G}_1$, such that

$$\sum_{\beta' \in \mathcal{G}_1} \mathsf{p}_2^j(\beta') \to 0.$$

Analogously, it follows that $\sum_{\beta' \in \mathcal{G}_2} \mathsf{p}_4^j(\beta') \to 0$, completing finally the verification of (2.6.72).

(II) It is proven now that

$$\lim_{T\to\infty} \sup_{\nu \in \mathcal{B}_T} \mathcal{R}^-_{T,\nu}(j) = 0. \tag{2.6.85}$$

Recall that the estimators $g^j_{T,\beta}(x_0,h)$ and $g^j_{T,\beta}(x_0)$ are defined as

$$g^j_{T,\beta}(x_0,h) = \frac{2}{Th^d}\int_0^T K_\beta\left(\frac{X_u - x_0}{h}\right)\mathrm{d}X_u^j, \quad g^j_{T,\beta}(x_0) = g^j_{T,\beta}(x_0, h^j_{T,\beta}).$$

Let $\beta' \in \mathcal{G}_T$, and assume that the event $A^j_{T,\beta'}$ holds. In view of the definition of the stochastic error $Z^j_{T,\beta'}$ in (2.6.54) and taking into account Lemma 2.6.4, it holds, whenever $b \in \Pi^j(\beta, L)$,

$$\begin{aligned}
&\big|\widehat{g}^j_{T,\beta'}(x_0) - \operatorname{div}(a_j\rho)(x_0)\big| \\
&\quad = \big|\widehat{g}^j_{T,\beta'}(x_0, \widehat{h}^j_{T,\beta'}) - \operatorname{div}(a_j\rho)(x_0)\big| \\
&\quad \le \sup_{h \in \mathcal{H}^j_{T,\beta'}} \big|g^j_{T,\beta'}(x_0,h) - \operatorname{div}(a_j\rho)(x_0)\big| \\
&\quad \le \sup_{h \in \mathcal{H}^j_{T,\beta'}} \Big\{\big|g^j_{T,\beta'}(x_0,h) - \mathbf{E}_b g^j_{T,\beta'}(x_0,h)\big| + \big|\mathbf{E}_b g^j_{T,\beta'}(x_0,h) - \operatorname{div}(a_j\rho)(x_0)\big|\Big\} \\
&\quad \le \sup_{h \in \mathcal{H}^j_{T,\beta'}} \Big\{\big|Z^j_{T,\beta'}(h)\big| + 2Lh^{\widetilde{\beta}-d/2}\mathsf{b}_{\beta,\beta'}\Big\}.
\end{aligned}$$

Then, using the definition of $\mathsf{b}^j_{T,\beta'}$ in (2.6.64),

$$\begin{aligned}
&\sum_{\beta' \in \mathcal{G}, \beta' < \beta^-} \sup_{b \in \Pi^j(\beta,L)} (\psi^j_\nu)^{-2}\, \mathbf{E}_b\Big(\big|\widetilde{g}^j_T(x_0) - \operatorname{div}(a_j\rho)(x_0)\big|^2\, \mathbb{1}\{\widehat{\beta}^j_T = \beta'\}\, \mathbb{1}\{A^j_{T,\beta'}\}\Big) \\
&\quad \le \sum_{\beta' \in \mathcal{G}, \beta' < \beta^-} \sup_{b \in \Pi^j(\beta,L)} (\psi^j_\nu)^{-2}\, \mathbf{E}_b\bigg(\Big(\mathsf{b}^j_{T,\beta'}(1+\overline{\delta}_T) + \sup_{h \in H^j_{T,\beta'}} \big|Z^j_{T,\beta'}(h)\big|\Big)^2\, \mathbb{1}\{\widehat{\beta}^j_T = \beta'\}\bigg) \\
&\quad \le 2\,(1+\overline{\delta}_T)^2 \sum_{\beta' \in \mathcal{G}, \beta' < \beta^-} \sup_{b \in \Pi^j(\beta,L)} (\psi^j_\nu)^{-2}(\mathsf{b}^j_{T,\beta'})^2\, \mathbf{P}_b\Big(\widehat{\beta}^j_T = \beta'\Big) \\
&\qquad + 2 \sum_{\beta' \in \mathcal{G}, \beta' < \beta^-} \sup_{b \in \Pi^j(\beta,L)} (\psi^j_\nu)^{-2}\, \mathbf{E}_b\bigg(\Big(\sup_{h \in H^j_{T,\beta'}} \big|Z^j_{T,\beta'}(h)\big|\Big)^2\, \mathbb{1}\{\widehat{\beta}^j_T = \beta'\}\bigg) \\
&\quad \le 2 \sum_{\beta' \in \mathcal{G}, \beta' < \beta^-} \sup_{b \in \Pi^j(\beta,L)} (\psi^j_\nu)^{-2}\Big((\mathsf{b}^j_{T,\beta'})^2(1+\overline{\delta}_T)^2 + \big(\tau^j_T(\beta')\big)^2\Big)\, \mathbf{P}_b\Big(\widehat{\beta}^j_T = \beta'\Big) \\
&\qquad + 2 \sum_{\beta' \in \mathcal{G}, \beta' < \beta^-} \sup_{b \in \Pi^j(\beta,L)} (\psi^j_\nu)^{-2} \\
&\qquad\qquad \times \mathbf{E}_b\bigg(\Big(\sup_{h \in H^j_{T,\beta'}} \big|Z^j_{T,\beta'}(h)\big|\Big)^2\, \mathbb{1}\Big\{\sup_{h \in H^j_{T,\beta'}} \big|Z^j_{T,\beta'}(h)\big| > \tau^j_T(\beta')\Big\}\bigg) \\
&\quad =: g^j_1(\nu) + g^j_2(\nu).
\end{aligned} \tag{2.6.86}$$

The term $g^j_1(\nu)$ is bounded from above by exploiting the fact that the probability to underestimate the value of β by $\widehat{\beta}^j_T$ substantially is small, whenever $b \in \Pi^j(\beta, L)$. Recall that m is the cardinality of the grid $\mathcal{G}$.

Lemma 2.6.7 (Probability of undershooting)**.** *Let $\beta \in [\beta_*, \infty)$, $\beta' \in \mathcal{G}$, $\beta' < \beta^-$, $L \in [L_*, L^*]$, and $\nu = (\beta, L)$. Then there exists some constant K such that*

$$\sup_{b \in \Pi^j(\beta,L)} \mathbf{P}_b\Big(\widehat{\beta}_T^j = \beta'\Big) \le K \ m \ (T^{-d/(2\beta')} + o(T^{-1})). \tag{2.6.87}$$

Proof. Throughout the proof, $C, C', C_1, C_2, \ldots$ denote positive constants which can depend only on β_*, L_*, L^*, d and σ. We first proceed analogously to the proof of Lemma 4.8 in Butucea (2001).

Denote by $\overline{\beta}' = \overline{\beta}'(\beta')$ the smallest element of the grid $\mathcal{G}$ which is greater than β'. It follows from the definition of $\widehat{\beta}_T^j$ in (2.5.23) that, whenever $\widehat{\beta}_T^j = \beta'$, there exists at least one β'', $\beta'' < \beta' \le \overline{\beta}'$, for which

$$|\widehat{g}^j_{T,\overline{\beta}'}(x_0) - \widehat{g}^j_{T,\beta''}(x_0)| > \widehat{\eta}^j_{T,\beta''}.$$

Consequently, and since $\mathcal{G}$ has cardinality m,

$$\begin{aligned}
\sup_{b \in \Pi^j(\beta,L)} \mathbf{P}_b\Big(\widehat{\beta}_T^j = \beta'\Big) &\le \sum_{\beta'' \in \mathcal{G}, \beta'' \le \beta'} \sup_{b \in \Pi^j(\beta,L)} \mathbf{P}_b\left(|\widehat{g}^j_{T,\overline{\beta}'}(x_0) - \widehat{g}^j_{T,\beta''}(x_0)| > \widehat{\eta}^j_{T,\beta''}\right) \\
&\le m \max_{\beta'' \in \mathcal{G}, \beta'' \le \beta'} \sup_{b \in \Pi^j(\beta,L)} \mathbf{P}_b\left(|\widehat{g}^j_{T,\overline{\beta}'}(x_0) - \widehat{g}^j_{T,\beta''}(x_0)| > \widehat{\eta}^j_{T,\beta''}\right) \\
&\le m \max_{\beta'' \in \mathcal{G}, \beta'' \le \beta'} \sup_{b \in \Pi^j(\beta,L)} \left(\mathsf{q}_1^j(\beta'') + \mathsf{q}_2^j(\beta'')\right),
\end{aligned}$$

where

$$\begin{aligned}
\mathsf{q}_1^j(\beta'') &:= \mathbf{P}_b\bigg(\Big\{|\widehat{g}^j_{T,\overline{\beta}'}(x_0) - \widehat{g}^j_{T,\beta''}(x_0)| > \widehat{\eta}^j_{T,\beta''}\Big\} \bigcap \Big\{A^j_{T,\overline{\beta}'} \cap A^j_{T,\beta''}\Big\}\bigg), \\
\mathsf{q}_2^j(\beta'') &:= \mathbf{P}_b\bigg(\Big\{|\widehat{g}^j_{T,\overline{\beta}'}(x_0) - \widehat{g}^j_{T,\beta''}(x_0)| > \widehat{\eta}^j_{T,\beta''}\Big\} \bigcap \Big\{(A^j_{T,\overline{\beta}'})^{\mathrm{c}} \cup (A^j_{T,\beta''})^{\mathrm{c}}\Big\}\bigg).
\end{aligned}$$

The definition of the event $A^j_{T,\beta}$ according to (2.6.42) entails that, whenever $A^j_{T,\beta''}$ holds,

$$\widehat{\eta}^j_{T,\beta''} = \eta^j_{T,\beta''} \left(\widehat{h}^j_{T,\beta''}/h^j_{T,\beta''}\right)^{\widetilde{\beta''}-d/2} \ge \eta^j_{T,\beta''}(1 - \overline{\delta}_T)$$

which yields

$$\begin{aligned}
\mathsf{q}_1^j(\beta'') &\le \mathbf{P}_b\bigg(\Big\{|\widehat{g}^j_{T,\overline{\beta}'}(x_0) - \widehat{g}^j_{T,\beta''}(x_0)| > \eta^j_{T,\beta''}(1 - \overline{\delta}_T)\Big\} \bigcap \Big\{A^j_{T,\overline{\beta}'} \cap A^j_{T,\beta''}\Big\}\bigg) \\
&\le \mathbf{P}_b\bigg(\sup_{h \in \mathcal{H}^j_{T,\overline{\beta}'}} |g^j_{T,\overline{\beta}'}(x_0, h) - \operatorname{div}(a_j\rho)(x_0)| \\
&\qquad\qquad + \sup_{h \in \mathcal{H}^j_{T,\beta''}} |g^j_{T,\beta''}(x_0, h) - \operatorname{div}(a_j\rho)(x_0)| > \eta^j_{T,\beta''}(1 - \overline{\delta}_T)\bigg).
\end{aligned}$$

Furthermore, Lemma 2.6.4 gives

$$\begin{aligned}
\sup_{h \in \mathcal{H}^j_{T,\overline{\beta}'}} &|g^j_{T,\overline{\beta}'}(x_0, h) - \operatorname{div}(a_j\rho)(x_0)| + \sup_{h \in \mathcal{H}^j_{T,\beta''}} |g^j_{T,\beta''}(x_0, h) - \operatorname{div}(a_j\rho)(x_0)| \\
&\overset{(2.6.46)}{\le} L\left(\left(h^j_{T,\overline{\beta}'}\right)^{\widetilde{\beta}-d/2} \mathsf{b}_{\beta,\overline{\beta}'} + \left(h^j_{T,\beta''}\right)^{\widetilde{\beta}-d/2} \mathsf{b}_{\beta,\beta''}\right)(1 + \overline{\delta}_T) \qquad (2.6.88) \\
&\qquad + \sup_{h \in \mathcal{H}^j_{T,\overline{\beta}'}} |Z^j_{T,\overline{\beta}'}(h)| + \sup_{h \in \mathcal{H}^j_{T,\beta''}} |Z^j_{T,\beta''}(h)|.
\end{aligned}$$

We continue with deriving an upper bound on the first summands in (2.6.88) in terms of the threshold $\eta^j_{T,\beta''}$, using arguments as in the proof of Lemma 5 in [KT04]. It can be seen from the definition of the respective quantities that

$$\psi^j_{\beta'',L} = \psi_{T,\beta''}\ C_j(\beta, L; \rho, \sigma) \geq C_1\ \big(\widetilde{T}^j(\beta'')\big)^{1-d/(2\beta'')},$$

and therefore

$$\begin{aligned}\big(\psi^j_{\beta'',L}\big)^{-1}\big(h^j_{T,\beta''}\big)^{\widetilde{\beta}-d/2} &\leq C_1^{-1}\big(\widetilde{T}^j(\beta'')\big)^{\frac{\widetilde{\beta}-\beta''}{\beta''}} = C_2 T^{-\frac{\widetilde{\beta}-\beta''}{2\beta''}}\, d_T(\beta'')^{\frac{\widetilde{\beta}-\beta''}{\beta''}}\\ &= C_2 \exp\left(-\frac{\widetilde{\beta}-\beta''}{2\beta''}\,(\log T - C_3 \log\log T)\right). \qquad (2.6.89)\end{aligned}$$

Since, as on p. 457 in [KT04],

$$\widetilde{\beta} - \beta'' \geq \min\{\beta - \beta'', d/2\} \geq \min\{\beta - \beta^-, d/2\} \geq \beta_T^+ (\log T)^{-1},$$

(2.6.89) in particular gives, for some positive constants K_1, K_2,

$$\big(h^j_{T,\beta''}\big)^{\widetilde{\beta}-d/2} \leq \delta_1(T)\psi^j_{\beta'',L}, \qquad \text{where } \delta_1(T) := K_1 \exp\Big(-K_2 \beta_T^{-1}\beta_T^+\Big) \to 0.$$

The same method as in (2.6.89) can be used for showing that

$$\big(\psi^j_{\beta'',L}\big)^{-1}\big(h^j_{T,\overline{\beta}'}\big)^{\widetilde{\beta}-d/2} \leq C_4 \delta_1(T).$$

Consequently,

$$\begin{aligned}L\left(\big(h^j_{T,\overline{\beta}'}\big)^{\widetilde{\beta}-d/2}\mathsf{b}_{\beta,\overline{\beta}'} + \big(h^j_{T,\beta''}\big)^{\widetilde{\beta}-d/2}\mathsf{b}_{\beta,\beta''}\right)(1+\overline{\delta}_T)&\\ \leq C_5\ \psi^j_{\beta'',L}\ \delta_1(T)(1-\overline{\delta}_T)&\\ \overset{(2.6.66)}{\leq} C_6\ \eta^j_{T,\beta''}\ \delta_1(T)(1-\overline{\delta}_T).& \qquad (2.6.90)\end{aligned}$$

Combining (2.6.88) and (2.6.90), we thus obtain

$$\begin{aligned}\mathsf{q}^j_1(\beta'') &\leq \mathbf{P}_b\Big(\sup_{h\in H^j_{T,\beta''}} \big|Z^j_{T,\beta''}(h)\big| > \eta^j_{T,\beta''}(1-2C_6\delta_1(T))(1-\overline{\delta}_T)\Big)\\ &\qquad + \mathbf{P}_b\Big(\sup_{h\in H^j_{T,\overline{\beta}'}} \big|Z^j_{T,\overline{\beta}'}(h)\big| > \eta^j_{T,\beta''}C_6\delta_1(T)(1-\overline{\delta}_T)\Big)\\ &=: \mathsf{q}^j_{1a}(\beta'') + \mathsf{q}^j_{1b}(\beta'').\end{aligned}$$

For some large enough constant R_1, it holds

$$\begin{aligned}\tau^j_T(\beta'') &\leq \mathsf{s}^j_T(\beta'')\left(d_T(\beta'') + \left(\frac{\log T}{\beta_T}\right)^{1/4}\right)\\ &\leq \eta^j_{T,\beta''}(1-2C_6\delta_1(T))(1-\overline{\delta}_T) \leq R_1 \mathsf{s}^j_T(\beta'')\sqrt{\log T}.\end{aligned}$$

(2.6.56) in Lemma 2.6.5(a) then gives

$$\begin{aligned}\mathsf{q}^j_{1a}(\beta'') &\leq c_1 \exp\left(-\frac{\big(\eta^j_{T,\beta''}\big)^2}{2\big(\mathsf{s}^j_T(\beta'')\big)^2}(1-2C_6\delta_1(T))^2(1-\overline{\delta}_T)^2(1-\gamma_T)\right)^2 + o(T^{-1})\\ &\leq c_1 \exp\left(-\frac{d_T^2(\beta'')}{2}(1-4C_6\delta_1(T))\right) + o(T^{-1}) \leq O(1)\ T^{-\frac{d}{2\beta''}} + o(T^{-1}).\end{aligned}$$

For the second term, Lemma 2.6.5(b) entails

$$\mathsf{q}^j_{1b}(\beta'') \le c_2 \exp\Big(-c_3\big(\eta^j_{T,\beta''}/\mathsf{s}^j_T(\overline{\beta}')\big)^2 C_6^2\ \delta_1^2(T)(1-\overline{\delta}_T)^2(1-\gamma_T)^2\Big). \tag{2.6.91}$$

Note that

$$\left(\frac{\eta^j_{T,\beta''}}{\mathsf{s}^j_T(\overline{\beta}')}\right)^2 = \left(\frac{\mathsf{s}^j_T(\beta'')d_T(\beta'')}{\mathsf{s}^j_T(\overline{\beta}')}\right)^2 = \left(\frac{h^j_{T,\beta''}}{h^j_{T,\overline{\beta}'}}\right)^{-d} \frac{\|K_{\beta''}\|^2_{L^2(\mathbb{R}^d)}}{\|K_{\overline{\beta}'}\|^2_{L^2(\mathbb{R}^d)}}\ d_T^2(\beta'').$$

We now argue similarly to the proof of Lemma 5 in [KT04]. Since $\overline{\beta}' > \beta''$, the definition of the grid according to (2.5.14) implies that only two cases are of interest, namely (i) $\overline{\beta}' - \beta'' \ge (\log T)^{-1/2}$ and (ii) $k_1(\log T)^{-\delta_1} < \overline{\beta}' - \beta'' < (\log T)^{-1/2}$. In case (i), one obtains

$$\left(\frac{h^j_{T,\beta''}}{h^j_{T,\overline{\beta}'}}\right)^{d} \frac{\|K_{\overline{\beta}'}\|^2_{L^2(\mathbb{R}^d)}}{\|K_{\beta''}\|^2_{L^2(\mathbb{R}^d)}} \le C_7\ \beta_T^2\ \exp\left(-\frac{C_8}{\beta_T^2}\sqrt{\log T}\right),$$

and in case (ii), it holds $h^j_{T,\beta''}/h^j_{T,\overline{\beta}'} \le O(1)$ and

$$\frac{\|K_{\overline{\beta}'}\|_{L^2(\mathbb{R}^d)}}{\|K_{\beta''}\|_{L^2(\mathbb{R}^d)}} \le \frac{\|K_{\overline{\beta}'} - K_{\beta''}\|_{L^2(\mathbb{R}^d)}}{\|K_{\beta''}\|_{L^2(\mathbb{R}^d)}} + 1 \le C_9(\log T)^{-1/2} + 1.$$

(For details, see [KT04], pp. 458–459.) Summing up,

$$\begin{aligned}\left(\frac{h^j_{T,\beta''}}{h^j_{T,\overline{\beta}'}}\right)^{d} \frac{\|K_{\overline{\beta}'}\|^2_{L^2(\mathbb{R}^d)}}{\|K_{\beta''}\|^2_{L^2(\mathbb{R}^d)}} &\le D_{11}\max\left\{\beta_T^2\exp\left(-\frac{C_4}{\beta_T^2}(\log T)^{1/2}\right),\ C_5(\log T)^{-1/2}+1\right\}\\ &\le \delta_2(T)+1\end{aligned}$$

for $\delta_2(T) := \max\left\{\beta_T^2\exp\left(-C_4\beta_T^{-2}(\log T)^{1/2}\right), C_5(\log T)^{-1/2}\right\}$. Thus,

$$\left(\frac{\eta^j_{T,\beta''}}{\mathsf{s}^j_T(\overline{\beta}')}\right)^2 \ge \frac{d_T^2(\beta'')}{\delta_2(T)+1} \ge d_T^2(\beta')(1-\delta_2(T)),$$

such that (2.6.91) gives

$$\mathsf{q}^j_{1b}(\beta'') \le 2\exp\big(-D_7 d_T^2(\beta')\delta_1^2(T)(1-\delta_2(T))(1-\gamma_T)\big). \tag{2.6.92}$$

In view of the definitions of $\delta_1(T)$ and $\delta_2(T)$, this last expression tends to zero faster than $T^{-d/(2\beta')}$. Since Lemma 2.6.3 entails

$$\mathsf{q}^j_2(\beta'') \le \mathbf{P}_b\Big((A^j_{T,\overline{\beta}'})^{\mathrm{c}}\Big) + \mathbf{P}_b\Big((A^j_{T,\beta''})^{\mathrm{c}}\Big) \le 4\exp\left(-\frac{T((1-\alpha)h_T^d\overline{\delta}_T\rho_T^*)^2}{2\|Q\|_\infty^2}\right), \tag{2.6.93}$$

the same remark applies to $\mathsf{q}^j_2(\beta'')$, completing the verification of (2.6.87). □

Let $\beta \in [\beta_*, \beta_T]$, $\beta' \in \mathcal{G}$, $L \in [L_*, L^*]$, $\nu = (\beta, L)$. By means of Lemma 2.6.7 and using relation (2.6.68) in Lemma 2.6.6, we obtain

$$\begin{aligned}g_1^j(\nu) &= 2\sum_{\beta'\in\mathcal{G},\beta'<\beta^-}\ \sup_{b\in\Pi^j(\beta,L)} (\psi^j_\nu)^{-2}\Big((\mathsf{b}^j_{T,\beta'})^2(1+\overline{\delta}_T)^2 + (\tau^j_T(\beta'))^2\Big)\mathbf{P}_b\Big(\widehat{\beta}^j_T = \beta'\Big)\\ &\le 2D_5\ K\ m \sum_{\beta'\in\mathcal{G},\beta'<\beta^-} \log T\ T^{-d/(2\beta)+d/(2\beta')}\ (T^{-d/(2\beta')} + o(T^{-1})).\end{aligned}$$

The grid $\mathcal{G} = \mathcal{G}_T$ is defined such that

$$\operatorname{card}(\mathcal{G}) = m \le \beta_T k_1^{-1} (\log T)^{\delta_1} = k_1^{-1} (\log T)^{\delta_1} (\log\log T)^{\delta_2}.$$

Consequently,

$$\lim_{T\to\infty} \sup_{\nu\in\mathcal{B}_T} g_1^j(\nu) = 0,$$

and the first assertion in Lemma 2.6.5(c) immediately gives

$$\lim_{T\to\infty} \sup_{\nu\in\mathcal{B}_T} g_2^j(\nu) = 0.$$

Note finally that, for any $\beta' \in \mathcal{G}$, there exists some constant c_0' such that

$$\begin{aligned}
&\sup_{b\in\Pi^j(\beta,L)} (\psi_\nu^j)^{-2}\, \mathbf{E}_b \Big(|\widehat{g}_{T,\beta'}^j(x_0) - \operatorname{div}(a_j\rho)(x_0)|^2\, \mathbb{1}\{\widehat{\beta}_T^j = \beta'\}\, \mathbb{1}\{(A_{T,\beta'}^j)^{\mathrm{c}}\}\Big)\\
&\quad \le \sup_{b\in\Pi^j(\beta,L)} (\psi_\nu^j)^{-2} \Big(\mathbf{E}_b \Big(|\widehat{g}_{T,\beta'}^j(x_0) - \operatorname{div}(a_j\rho)(x_0)|^4\Big)\Big)^{1/2} \Big(\mathbf{P}_b\Big((A_{T,\beta'}^j)^{\mathrm{c}}\Big)\Big)^{1/2}\\
&\quad \le c_0' \Big(\mathbf{P}_b\Big((A_{T,\beta'}^j)^{\mathrm{c}}\Big)\Big)^{1/2}.
\end{aligned}$$

Lemma 2.6.3 shows that the last term is exponentially small and independent both of β and β' such that

$$\sum_{\beta'\in\mathcal{G},\beta'<\beta^-} \sup_{b\in\Pi^j(\beta,L)} (\psi_\nu^j)^{-2}\, \mathbf{E}_b \Big(|\widetilde{g}_T^j(x_0) - \operatorname{div}(a_j\rho)(x_0)|^2\, \mathbb{1}\{\widehat{\beta}_T^j = \beta'\}\, \mathbb{1}\{(A_{T,\beta'}^j)^{\mathrm{c}}\}\Big) = o_T(1),$$

thus completing the proof of (2.6.85). □

Proof of Theorem 2.5.10. It remains to verify (2.5.27). The proof of part (b) actually is along the lines of the proof of Theorem 2 in [KT04]. For the sake of completeness, it is given here at full length. For any $\beta > d/2$ and $L > 0$, denote $\psi_{\beta,L}^{\mathrm{Kol}} := \psi_{T,\beta}\, C_{\mathrm{Kol}}(\beta, L; \rho)$, for $\psi_{T,\beta}$ and $C_{\mathrm{Kol}}(\beta, L; \rho)$ defined in (2.5.8) and (2.5.25), respectively.

By assumption (2.5.26), there exists some $\delta \in (0, 1/2)$ such that, for T large enough,

$$(\psi_{\beta_0,L}^{\mathrm{Kol}})^{-2}\, \mathscr{R}_{T,\beta_0,L}(\breve{g}_T(x_0)) \le (1-2\delta)^3.$$

Fix $\beta_1 \in (\beta_0, \beta_T]$. If the definitions of ρ (cf. (2.6.17)) and $g_{T,1}^j$ (cf. (2.6.22)) are suitably modified, it is shown as in the proof of Theorem 2.5.5 that, for any $\beta_0' \in (\beta_1, \beta_T]$,

$$\begin{aligned}
&\inf_{\widehat{g}_T} \max\Big\{ (\psi_{\beta_1,L}^{\mathrm{Kol}})^{-2}\, \mathscr{R}_{T,\beta_0',L}(\widehat{g}_T(x_0)), (\psi_{\beta_0,L}^{\mathrm{Kol}})^{-2}\, \mathscr{R}_{T,\beta_0,L}(\widehat{g}_T(x_0))\Big\}\\
&\qquad \ge \inf_{\widehat{g}_T} \max\Big\{ \mathbf{E}_0 \Big((\psi_{\beta_1,L}^{\mathrm{Kol}})^{-2}\, |\widehat{g}_T(x_0) - g_{T,0}^j(x_0)|^2\Big),\\
&\qquad\qquad \mathbf{E}_1 \Big((\psi_{\beta_0,L}^{\mathrm{Kol}})^{-2}\, |\widehat{g}_T(x_0) - g_{T,1}^j(x_0)|^2\Big)\Big\}\\
&\qquad \ge (1-\delta)^3.
\end{aligned}$$

For sufficiently large T, it thus holds, for any estimator $\widehat{g}_T$ of $\partial_j\rho$,

$$\begin{aligned}
&\max\Big\{ (\psi_{\beta_1,L}^{\mathrm{Kol}})^{-2}\, \mathscr{R}_{T,\beta_0',L}(\widehat{g}_T(x_0)), (1-2\delta)^3\Big\}\\
&\quad \ge \max\Big\{ (\psi_{\beta_1,L}^{\mathrm{Kol}})^{-2}\, \mathscr{R}_{T,\beta_0',L}(\widehat{g}_T(x_0)), (\psi_{\beta_0,L}^{\mathrm{Kol}})^{-2}\, \mathscr{R}_{T,\beta_0,L}(\widehat{g}_T(x_0))\Big\} \ge (1-\delta)^3.
\end{aligned}$$

Consequently, for T large enough,

$$\mathscr{R}_{T,\beta_0',L}(\widehat{g}_T(x_0)) \geq (1-\delta)^3 \, \left(\psi^{\mathrm{Kol}}_{\beta_1,L}\right)^2.$$

Lemma 2.5.3 (resp. its proof) implies that there exists some constant $c' > 0$ such that

$$\inf_{\widehat{g}_T} \mathscr{R}_{T,\beta_0,L}(\widehat{g}_T(x_0)) \geq c' T^{-\frac{\beta_0 - d/2}{\beta_0}}.$$

Summing up, for any $\beta_0' > \beta_1$ and some positive constant c'',

$$\begin{aligned}
\frac{\mathscr{R}_{T,\beta_0',L}(\check{g}_T(x_0))}{\mathscr{R}_{T,\beta_0',L}(\widetilde{g}_T(x_0))} \frac{\mathscr{R}_{T,\beta_0,L}(\check{g}_T(x_0))}{\mathscr{R}_{T,\beta_0,L}(\widetilde{g}_T(x_0))} &\geq \frac{(1-\delta)^3 \left(\psi^{\mathrm{Kol}}_{\beta_1,L}\right)^2}{\left(\psi^{\mathrm{Kol}}_{\beta_0',L}\right)^2} \left(\frac{c' T^{\frac{\beta_0 - d/2}{\beta_0}}}{\left(\psi^{\mathrm{Kol}}_{\beta_0,L}\right)^2} \right) \\
&\geq c'' \left(\frac{\log T}{T} \right)^{\frac{d(\beta_1 - \beta_0')}{2\beta_0'\beta_1}} (\log T)^{-\frac{\beta_0 - d/2}{\beta_0}} \to_{T\to\infty} \infty.
\end{aligned}$$

□

Proof of Theorem 2.5.11. The proof is similar to the proof of Theorem 2.5.9 as will be sketched in the sequel. Let $\beta \in [\beta_*, \beta_T]$, $L \in [L_*, L^*]$, and $\beta' \in (d/2, \beta]$. Denote again by γ_{Ti}, $i \in \mathbb{N}$, functions of T such that $\lim_{T\to\infty} \gamma_{Ti} = 0$. Let $\psi_{\beta,L} := \psi_{T,\beta} \, C(\beta, L; \rho, \sigma)$, for $\psi_{T,\beta}$ and $C(\beta, L; \rho, \sigma)$ introduced in (2.5.8) and (2.5.31), respectively. Denote

$$\widetilde{T}(\beta) := \left(\frac{4\rho(x_0)\|\sigma\|_{S_2}^2 \log T}{\beta T} \right)^{1/2}, \qquad h_{T,\beta} := \left(\frac{4\rho(x_0)\|\sigma\|_{S_2}^2 \log T}{\beta T} \right)^{1/(2\beta)},$$
$$\eta_{T,\beta} := h_{T,\beta}^{\beta - d/2} \|K_\beta\|_{L^2(\mathbb{R}^d)},$$

and define $\widetilde{\beta}$, $A_{T,\beta}$, $\mathcal{H}_{T,\beta}$ and $H_{T,\beta}$ analogue to the proof of Theorem 2.5.9 as

$$\begin{aligned}
\widetilde{\beta} = \widetilde{\beta}(\beta, \beta') &:= \begin{cases} \beta' + \frac{d}{2}, & \text{if } \frac{d}{2} \leq \beta' \leq \frac{\beta}{2} + \frac{d}{4}, \\ \beta, & \text{if } \frac{\beta}{2} + \frac{d}{4} < \beta' \leq \beta, \end{cases} \\
A_{T,\beta'} &:= \left\{ \left| \left(\widehat{h}_{T,\beta'}/h_{T,\beta'}\right)^{\widetilde{\beta}' - d/2} - 1 \right| \leq \overline{\delta}_T \right\}, \\
\mathcal{H}_{T,\beta'} &:= \left\{ h : \left| \left(h/h_{T,\beta'}\right)^{\widetilde{\beta}' - d/2} - 1 \right| \leq \overline{\delta}_T \right\}, \\
H_{T,\beta'} &:= \left\{ h : \left| \left(h/h_{T,\beta'}\right) - 1 \right| \leq c_0 \overline{\delta}_T \right\},
\end{aligned}$$

for some positive constant c_0 and $\overline{\delta}_T := (\log T)^{-1}$. Given $h \in \mathcal{H}_{T,\beta}$, let

$$g_{T,\beta}(x_0, h) := \frac{2}{Th^d} \int_0^T K_\beta \left(\frac{X_u - x_0}{h} \right) \mathrm{d}X_u,$$

define

$$g_{T,\beta}(x_0) := g_{T,\beta}(x_0, h_{T,\beta}), \qquad \mathsf{s}_T(\beta) := 2h_{T,\beta}^{-d/2} \sqrt{\frac{\rho(x_0)}{T}} \|\sigma\|_{S_2} \, \|K_\beta\|_{L^2(\mathbb{R}^d)},$$

and

$$d_T(\beta) := \frac{\eta_{T,\beta}}{\mathsf{s}_T(\beta)} = \sqrt{\frac{\log T}{\beta}}.$$

Finally, let

$$\tau_T(\beta') := \mathsf{s}_T(\beta') \left(\left(d_T^2(\beta') - d_T^2(\beta) \right)^{1/2} + \left(\frac{\log T}{\beta_T} \right)^{1/4} \right).$$

Most of the auxiliary results which were used for proving Theorem 2.5.9 admit (more or less) straightforward extensions to the present setting. In particular, the following analogue of Lemma 2.6.5 holds true.

Lemma 2.6.8. *Grant Assumptions* (**BI**) *and* (**SG**+)*. Let* $\beta > d/2$*, and denote*

$$Z_{T,\beta}(h) := g_{T,\beta}(x_0, h) - \mathbf{E}_b g_{T,\beta}(x_0, h), \qquad h \in H_{T,\beta}. \tag{2.6.94}$$

Then, for any $\beta' > d/2$*, it holds:*

(a) *For any*

$$u \in \left[\tau_T(\beta'),\ R_1'\ \mathsf{s}_T(\beta')\sqrt{\log T}\right], \qquad R_1' > 0 \text{ an absolute constant,}$$

there exist some sufficiently small γ'*, independent of* β'*, and some universal constant* $c_1' > 0$ *such that*

$$\mathbf{P}_b\left(\sup_{h \in H_{T,\beta'}} \|Z_{T,\beta'}(h)\| > u \right) \leq c_1' \exp\left(-\frac{u^2(1-\gamma')}{2\mathsf{s}_T^2(\beta')} \right) + o(T^{-1}).$$

(b) *For any*

$$u \in \left[R_1'\ \mathsf{s}_T(\beta')\sqrt{\log T},\ R_2'\right], \qquad R_1', R_2' > 0 \text{ absolute constants,}$$

it holds, for some constants $c_2', c_3' > 0$,

$$\mathbf{P}_b\left(\sup_{h \in H_{T,\beta'}} \|Z_{T,\beta'}(h)\| > u \right) \leq c_2' \exp\left(-\frac{c_3' u^2}{\mathsf{s}_T^2(\beta')} \right). \tag{2.6.95}$$

(c) *Assume that* $\beta' < \beta$*. Then, uniformly in* $\beta \in \mathcal{B}_T$*,*

$$\sup_{\beta' \in \mathcal{B},\ \beta' < \beta} m \sup_{b \in \Pi(\beta,L)} \left(\psi_{\beta,L}\right)^{-2} \mathbf{E}_b\Bigg(\Big(\sup_{h \in H_{T,\beta'}} \|Z_{T,\beta'}(h)\| \Big)^2 \times \mathbb{1}\Big\{ \sup_{h \in H_{T,\beta'}} \|Z_{T,\beta'}(h)\| > \tau_T(\beta') \Big\} \Bigg) \to 0,$$

$$\sup_{\beta' \in \mathcal{B},\ \beta' < \beta} m \sup_{b \in \Pi(\beta,L)} \left(\psi_{\beta,L}\right)^{-2} \mathbf{E}_b\Bigg(\Big(\sup_{h \in H_{T,\beta'}} \|Z_{T,\beta'}(h)\| \Big)^2 \times \mathbb{1}\Big\{ \sup_{h \in H_{T,\beta'}} \|Z_{T,\beta'}(h)\| > \sqrt{\mathsf{s}_T(\beta')\ \psi_{\beta,L}} \Big\} \Bigg) \to 0.$$

Proof. The assertions follow from slight modifications of the proof of Lemma 2.6.5, taking into account Remark 2.5.14. □

For $\nu = (\beta, L) \in \mathcal{B}_T$, define β^- and β_T^+ as before as

$$\beta^- := \beta - \frac{\beta_T^+}{\log T}, \qquad \beta_T^+ := (\log\log T)^{\delta_3}, \qquad \text{for } \delta_3 \in (\delta_2, 1).$$

Set

$$\begin{aligned}
\mathcal{R}^+_{T,\nu} &:= \psi_\nu^{-2} \sup_{b\in\Pi(\nu)} \mathbf{E}_b \left(\left\| \widetilde{g}_T(x_0) - \operatorname{div}(a\rho)(x_0) \right\|^2 \mathbb{1}\{\widehat{\beta}_T \geq \beta^-\} \right), \\
\mathcal{R}^-_{T,\nu} &:= \psi_\nu^{-2} \sup_{b\in\Pi(\nu)} \mathbf{E}_b \left(\left\| \widetilde{g}_T(x_0) - \operatorname{div}(a\rho)(x_0) \right\|^2 \mathbb{1}\{\widehat{\beta}_T < \beta^-\} \right).
\end{aligned}$$

(I) Some rather straightforward modifications of the proof of part (**I**) of Theorem 2.5.9 yield

$$\limsup_{T\to\infty} \sup_{\nu\in\mathcal{B}_T} \mathcal{R}^+_{T,\nu} \leq 1. \tag{2.6.96}$$

Indeed, analogously to the proof of (2.6.65) in Lemma 2.6.6, it is shown that

$$\psi_\nu = (2L)^{d/(2\beta)} \sqrt{d} \left(\eta_{T,\beta} + h_{T,\beta}^{\beta-d/2}\ \mathsf{b}_{\beta,\beta} \right). \tag{2.6.97}$$

As in (2.6.73), define $\overline{\beta} = \overline{\beta}(\beta)$ via the equation

$$\left(\frac{\log T}{4L^2 T} \right)^{1/(2\beta)} = \left(\frac{\log T}{T} \right)^{1/(2\overline{\beta})}.$$

Let $\beta^+ \in \mathcal{G}$ be the largest grid point $\leq \overline{\beta}$, and assume that T is large enough to ensure that $\beta^- < \beta^+$ and that $\widetilde{\beta} = \beta$. Consider the sets $\mathcal{G}_1 = \mathcal{G}_1(\beta)$ and $\mathcal{G}_2 = \mathcal{G}_2(\beta)$ as defined in (2.6.75) and (2.6.76), respectively. Since $\widetilde{\eta}_\beta(\operatorname{div}(a\rho)) \leq 2L\sqrt{d}$, the same steps as in the proof of Lemma 2.6.4 imply that

$$\begin{aligned}
&\left\| \mathbf{E}_b g_{T,\beta'}(x_0,h) - \operatorname{div}(a\rho)(x_0) \right\| \\
&\quad = (2\pi)^{-d} \left\| \int_{\mathbb{R}^d} \phi_{\operatorname{div}(a\rho)}(\lambda) (\phi_{K_{\beta',h}}(\lambda) - 1) \mathrm{e}^{-\mathrm{i}\lambda^t x_0} \mathrm{d}\lambda \right\| \\
&\quad \leq \left((2\pi)^{-d} \int_{\mathbb{R}^d} \left\| \phi_{\operatorname{div}(a\rho)}(\lambda) \right\|^2 \|\lambda\|^{2\beta} \mathrm{d}\lambda \right)^{1/2} \left(\frac{h}{\mathsf{b}'} \right)^\beta \\
&\qquad\qquad \times \left((2\pi)^{-d} \int_{\mathbb{R}^d} \frac{\|h\lambda/\mathsf{b}'\|^{4\beta'-2\beta}}{(1+\|h\lambda/\mathsf{b}'\|^{2\beta'})^2} \mathrm{d}\lambda \right)^{1/2} \\
&\quad \leq 2L\sqrt{d}\ \mathsf{b}_{\beta,\beta'}\ h^{\beta-d/2}.
\end{aligned} \tag{2.6.98}$$

Therefore, whenever the event $A_{T,\beta'}$ holds,

$$\begin{aligned}
&\sup_{h\in\mathcal{H}_{T,\beta'}} \left\| \mathbf{E}_b g_{T,\beta'}(x_0,h) - \operatorname{div}(a\rho)(x_0) \right\| \\
&\qquad \leq 2L\sqrt{d}\ \mathsf{b}_{\beta,\beta'} \sup_{h\in\mathcal{H}_{T,\beta'}} h^{\beta-d/2} \\
&\qquad \leq 2L\sqrt{d}\ \mathsf{b}_{\beta,\beta'} \left(\frac{4\rho(x_0)\|\sigma\|_{S_2}^2}{\beta'} \right)^{\frac{\beta-d/2}{2\beta'}} \left(\frac{\log T}{4L^2 T} \right)^{\frac{\beta-d/2}{2\beta}} (1+\overline{\delta}_T).
\end{aligned}$$

Arguing as in the derivation of (2.6.80) and (2.6.78), we obtain, for any $\beta' \in \mathcal{G}_1$,

$$\begin{aligned}
\sup_{h\in\mathcal{H}_{T,\beta'}} \left\| \mathbf{E}_b g_{T,\beta'}(x_0,h) - \operatorname{div}(a\rho)(x_0) \right\| &\leq (2L)^{d/(2\beta)} \sqrt{d}\ h_{T,\beta}^{\beta-d/2} \mathsf{b}_{\beta,\beta}\ (1+\gamma_{T1}) \\
&\overset{(2.6.97)}{\leq} \psi_\nu\ (1+\gamma_{T1})
\end{aligned}$$

and

$$\eta_{T,\beta^+} \leq (2L)^{d/(2\beta)} \eta_{T,\beta}\ (1+\gamma_{T2}).$$

Thus, whenever $\widehat{\beta}_T = \beta' \in \mathcal{G}_1$ and the event $A_{T,\beta'}$ holds, one has

$$\|\widetilde{g}_T(x_0) - \operatorname{div}(a\rho)(x_0)\| \leq \sup_{h \in \mathcal{H}_{T,\beta'}} \|Z_{T,\beta'}(h)\| + \psi_\nu\ (1+\gamma_{T1}).$$

Similarly, if $\widehat{\beta}_T = \beta' \in \mathcal{G}_2$ and the event A_{T,β^+} holds,

$$\begin{aligned}
&\|\widetilde{g}_T(x_0) - \operatorname{div}(a\rho)(x_0)\| \\
&\quad \leq \|\widehat{g}_{T,\widehat{\beta}_T}(x_0) - g_{T,\beta^+}(x_0)\| + \|g_{T,\beta^+}(x_0) - \operatorname{div}(a\rho)(x_0)\| \\
&\quad \leq \sqrt{d}\ \widehat{\eta}_{T,\beta^+} + \sup_{h \in \mathcal{H}_{T,\beta^+}} \|Z_{T,\beta^+}(h)\| + (2L)^{d/(2\beta)}\sqrt{d}\ h_{T,\beta}^{\beta-d/2}\mathsf{b}_{\beta,\beta}\ (1+\gamma_{T1}) \\
&\quad \leq \sup_{h \in \mathcal{H}_{T,\beta^+}} \|Z_{T,\beta^+}(h)\| + (2L)^{d/(2\beta)}\sqrt{d}\ \left(h_{T,\beta}^{\beta-d/2}\mathsf{b}_{\beta,\beta}\ (1+\gamma_{T1}) + \eta_{T,\beta}\ (1+\gamma_{T3})\right) \\
&\quad \leq \sup_{h \in H_{T,\beta^+}} \|Z_{T,\beta^+}(h)\| + \psi_\nu\ (1+\gamma_{T4}).
\end{aligned}$$

Decomposing analogously to (2.6.83), it follows

$$\begin{aligned}
&\psi_\nu^{-2}\ \mathbf{E}_b\Big(\|\widetilde{g}_T(x_0) - \operatorname{div}(a\rho)(x_0)\|^2\ \mathbb{1}\{\widehat{\beta}_T \in \mathcal{G}_1 \cup \mathcal{G}_2\}\Big) \\
&\quad \leq \sum_{\beta' \in \mathcal{G}_1} \underbrace{\mathbf{E}_b\left(\Big(1+\gamma_{T1} + \psi_\nu^{-1} \sup_{h \in H_{T,\beta'}} \|Z_{T,\beta'}(h)\|\Big)^2\ \mathbb{1}\{\widehat{\beta}_T = \beta'\}\ \mathbb{1}\{A_{T,\beta'}\}\right)}_{=:\mathsf{p}(\beta')} \\
&\qquad + \sum_{\beta' \in \mathcal{G}_1} \mathbf{E}_b\left(\psi_\nu^{-2}\ \|\widetilde{g}_T(x_0) - \operatorname{div}(a\rho)(x_0)\|^2\ \mathbb{1}\{\widehat{\beta}_T = \beta'\}\ \mathbb{1}\{A^{\mathrm{c}}_{T,\beta'}\}\right) \\
&\qquad + \sum_{\beta' \in \mathcal{G}_2} \mathbf{E}_b\left(\Big(1+\gamma_{T4} + \psi_\nu^{-1} \sup_{h \in H_{T,\beta^+}} \|Z_{T,\beta^+}(h)\|\Big)^2\ \mathbb{1}\{\widehat{\beta}_T = \beta'\}\ \mathbb{1}\{A_{T,\beta^+}\}\right) \\
&\qquad + \sum_{\beta' \in \mathcal{G}_2} \mathbf{E}_b\left(\psi_\nu^{-2}\ \|\widetilde{g}_T(x_0) - \operatorname{div}(a\rho)(x_0)\|^2\ \mathbb{1}\{\widehat{\beta}_T = \beta'\}\ \mathbb{1}\{A^{\mathrm{c}}_{T,\beta^+}\}\right).
\end{aligned}$$

Exemplarily, we consider the term $\mathsf{p}(\cdot)$. For any $\beta' \in \mathcal{G}_1$, it holds

$$\begin{aligned}
\mathsf{p}(\beta') &\leq \left(1+\gamma_{T1} + \sqrt{\mathsf{s}_T(\beta')\psi_\nu^{-1}}\right)^2 \mathbf{P}_b\Big(\widehat{\beta}_T = \beta'\Big) \\
&\quad + 2\ (1+\gamma_{T1})^2\ \mathbf{P}_b\Big(\sup_{h \in H_{T,\beta'}} \|Z_{T,\beta'}(h)\| > \sqrt{\mathsf{s}_T(\beta')\psi_\nu}\Big) \qquad (2.6.99) \\
&\quad + 2\ \psi_\nu^{-2}\ \mathbf{E}_b\bigg(\Big(\sup_{h \in H_{T,\beta'}} \|Z_{T,\beta'}(h)\|\Big)^2\ \mathbb{1}\Big\{\sup_{h \in H_{T,\beta'}} \|Z_{T,\beta'}(h)\| > \sqrt{\mathsf{s}_T(\beta')\psi_\nu}\Big\}\bigg). \qquad (2.6.100)
\end{aligned}$$

Lemma 2.6.8(b) entails that the term in (2.6.99) is bounded from above by some constant multiple of

$$2\ (1+\gamma_{T1})^2 \exp\left(-c_3' \frac{\psi_\nu(1-\gamma_T)}{\mathsf{s}_T(\beta')}\right).$$

Note that

$$\eta_{T,\beta'} \leq L_2 \exp(C\beta_T^+)\psi_\nu$$

such that

$$\frac{\mathsf{s}_T(\beta')}{\psi_\nu} \leq \frac{L_2 \exp(C\beta_T^+)\mathsf{s}_T(\beta')}{\eta_{T,\beta'}} = L_2 \exp(C\beta_T^+)\sqrt{\frac{\beta}{\log T}} =: A_T'.$$

Lemma 2.6.8(c) implies that the term in (2.6.100) tends to zero, uniformly in $\beta' \in \mathcal{G}_1$. Similar arguments yield upper bounds on the remaining parts of the above decomposition (also cf. the conclusion of the proof of part (**I**) of Theorem 2.5.9). Summing up, we get

$$\mathcal{R}^+_{T,\nu} \leq \Big(1 + \gamma_{T2} + \sqrt{A'_T}\Big)^2 \ \mathbf{P}_b\big(\widehat{\beta}_T \in \mathcal{G}_1 \cup \mathcal{G}_2\big) + Cm \exp\big(- 2A'^{-1}_T\big) + o_T(1).$$

Inserting the definition of A'_T and m, we obtain (2.6.96).

(II) We now show that

$$\lim_{T\to\infty} \sup_{\nu \in \mathcal{B}_T} \mathcal{R}^-_{T,\nu} = 0. \tag{2.6.101}$$

Let $\beta \in [\beta_*, \beta_T]$, $\beta' \in \mathcal{G}$, $\beta' < \beta^- = \beta - \beta^+_T / \log T$, $L \in [L_*, L^*]$. Let $b \in \Pi(\beta, L)$, $\nu = (\beta, L)$. It was already noted that a straightforward modification of Lemma 2.6.4 implies that

$$\sup_{b\in\Pi(\nu)} \|\mathbf{E}_b g_{T,\beta'}(x_0,h) - \operatorname{div}(a\rho)(x_0)\| \leq 2L\sqrt{d}\ \mathsf{b}_{\beta,\beta'}\ h^{\widetilde{\beta}-d/2}$$

such that, for any T large enough,

$$\begin{aligned}
\sup_{h\in\mathcal{H}_{T,\beta'}} &\Big(\|\mathbf{E}_b g_{T,\beta'}(x_0,h) - \operatorname{div}(a\rho)(x_0)\|\ \mathbb{1}\{A_{T,\beta'}\}\Big) \\
&\qquad\leq 2L\sqrt{d}\ \mathsf{b}_{\beta,\beta'} \sup_{h\in\mathcal{H}_{T,\beta'}} \Big(h^{\widetilde{\beta}-d/2}\ \mathbb{1}\{A_{T,\beta'}\}\Big) \\
&\qquad\leq \mathsf{b}_{T,\beta'}\ (1+\overline{\delta}_T),
\end{aligned}$$

where $\mathsf{b}_{T,\beta'} := 2L\sqrt{d}\ \mathsf{b}_{\beta,\beta'}\ h^{\widetilde{\beta}-d/2}_{T,\beta'}$. Consequently,

$$\begin{aligned}
&\|\widehat{g}_{T,\beta'}(x_0) - \operatorname{div}(a\rho)(x_0)\|\ \mathbb{1}\{A_{T,\beta'}\} \\
&\qquad\leq \sup_{h\in\mathcal{H}_{T,\beta'}} \Big(\|\mathbf{E}_b g_{T,\beta'}(x_0,h) - \operatorname{div}(a\rho)(x_0)\|\ \mathbb{1}\{A_{T,\beta'}\} + \|Z_{T,\beta'}(h)\|\Big) \\
&\qquad\leq \mathsf{b}_{T,\beta'}\ (1+\overline{\delta}_T) + \sup_{h\in\mathcal{H}_{T,\beta'}} \|Z_{T,\beta'}(h)\|,
\end{aligned}$$

where the stochastic error $Z_{T,\beta'}(\cdot)$ is defined according to (2.6.94). Notably, we get

$$\begin{aligned}
\sum_{\beta'\in\mathcal{G},\beta'<\beta^-} \sup_{b\in\Pi(\nu)} \psi_\nu^{-2}\ \mathbf{E}_b\Big(&\|\widehat{g}_{T,\beta'}(x_0) - \operatorname{div}(a\rho)(x_0)\|^2 \\
&\times \mathbb{1}\{\widehat{\beta}_T = \beta'\}\ \mathbb{1}\{A_{T,\beta'}\}\Big) \leq g_1(\nu) + g_2(\nu),
\end{aligned}$$

for

$$\begin{aligned}
g_1(\nu) &:= 2 \sum_{\beta'\in\mathcal{G},\beta'<\beta^-} \sup_{b\in\Pi(\nu)} \psi_\nu^{-2}\ \Big(\mathsf{b}^2_{T,\beta'}(1+\overline{\delta}_T)^2 + \tau_T^2(\beta')\Big)\ \mathbf{P}_b\big(\widehat{\beta}_T = \beta'\big), \\
g_2(\nu) &:= 2 \sum_{\beta'\in\mathcal{G},\beta'<\beta^-} \sup_{b\in\Pi(\nu)} \psi_\nu^{-2} \\
&\qquad\times\ \mathbf{E}_b\Big(\Big(\sup_{h\in H_{T,\beta'}} \|Z_{T,\beta'}(h)\|\Big)^2\ \mathbb{1}\Big\{\sup_{h\in H_{T,\beta'}} \|Z_{T,\beta'}(h)\| > \tau_T(\beta')\Big\}\Big).
\end{aligned}$$

As in the proof of the upper bound for component-wise estimation of $\operatorname{div}(a\rho)$, we proceed by exploiting that the "probability of undershooting," that is, the probability of underestimating the value of the true smoothness parameter β by the estimator $\widehat{\beta}_T$, is small, uniformly over $b \in \Pi(\nu)$.

Lemma 2.6.9. *Let $\beta \in [\beta_*, \infty)$, $\beta' \in \mathcal{G}$, $\beta' < \beta^-$, $L \in [L_*, L^*]$, and $\nu = (\beta, L)$. Then, for $m = \operatorname{card}(\mathcal{G})$, there exists some constant K such that*

$$\sup_{b \in \Pi(\nu)} \mathbf{P}_b\Big(\widehat{\beta}_T = \beta'\Big) \leq K \ m \ T^{-d/(2\beta')}.$$

Proof. The proof is similar to the proof of Lemma 2.6.7. Note first that the definition of $\widehat{\beta}_T$ according to (2.5.30) implies that

$$\begin{aligned}
\sup_{b \in \Pi(\nu)} \mathbf{P}_b(\widehat{\beta}_T = \beta') &\leq \sum_{\beta'' \in \mathcal{G}, \beta'' \leq \beta'} \sup_{b \in \Pi(\nu)} \mathbf{P}_b\left(\|\widehat{g}_{T,\overline{\beta}'}(x_0) - \widehat{g}_{T,\beta''}(x_0)\| > \sqrt{d}\ \widehat{\eta}_{T,\beta''}\right) \\
&\leq m \max_{\beta'' \in \mathcal{G}, \beta'' \leq \beta'} \sup_{b \in \Pi(\nu)} \mathbf{P}_b\left(\|\widehat{g}_{T,\overline{\beta}'}(x_0) - \widehat{g}_{T,\beta''}(x_0)\| > \sqrt{d}\ \widehat{\eta}_{T,\beta''}\right) \\
&\leq m \max_{\beta'' \in \mathcal{G}, \beta'' \leq \beta'} \sup_{b \in \Pi(\nu)} (\mathsf{q}_1 + \mathsf{q}_2), \qquad (2.6.102)
\end{aligned}$$

where

$$\begin{aligned}
\mathsf{q}_1 = \mathsf{q}_1(\overline{\beta}', \beta'') &:= \mathbf{P}_b\Big(\{\|\widehat{g}_{T,\overline{\beta}'}(x_0) - \widehat{g}_{T,\beta''}(x_0)\| > \sqrt{d}\ \widehat{\eta}_{T,\beta''}\} \bigcap \{A_{T,\overline{\beta}'} \cap A_{T,\beta''}\}\Big), \\
\mathsf{q}_2 = \mathsf{q}_2(\overline{\beta}', \beta'') &:= \mathbf{P}_b\Big(\{\|\widehat{g}_{T,\overline{\beta}'}(x_0) - \widehat{g}_{T,\beta''}(x_0)\| > \sqrt{d}\ \widehat{\eta}_{T,\beta''}\} \bigcap \Big\{A^{\mathrm{c}}_{T,\overline{\beta}'} \cup A^{\mathrm{c}}_{T,\beta''}\Big\}\Big).
\end{aligned}$$

Consider the sequence

$$\delta_{T3} := \exp\left(-\frac{k_2(\log T)^{1-\delta}}{8\beta_T^2}\right) \to 0.$$

Plugging in the definition of the respective quantities, we obtain

$$\begin{aligned}
\frac{\mathsf{b}_{T,\beta''}}{\delta_{T3}\eta_{T,\beta''}} &= \frac{L\sqrt{d} h_{T,\beta''}^{\widetilde{\beta''}-d/2} \mathsf{b}_{\beta,\beta''}}{\delta_{T3}\mathsf{s}_{T,\beta''}} \sqrt{\frac{\beta''}{\log T}} \\
&\leq O(1)\sqrt{\beta_T} h_{T,\beta''}^{\widetilde{\beta''}-\beta''}\, \delta_{T3}^{-1} \\
&\leq O(1)\sqrt{\beta_T} \exp\left(\frac{k_2(\log T)^{1-\delta}}{8\beta_T^2} - \frac{(\widetilde{\beta''} - \beta'')\log T}{2\beta_T}\right) = o(1).
\end{aligned}$$

Furthermore,

$$\frac{\mathsf{b}_{T,\overline{\beta}'}}{\delta_{T3}\eta_{T,\beta''}} = \frac{L\sqrt{d} h_{T,\beta''}^{\widetilde{\overline{\beta}'}-d/2} \mathsf{b}_{\beta,\overline{\beta}'}}{\delta_{T3}\mathsf{s}_{T,\beta''}} \sqrt{\frac{\beta''}{\log T}} \leq O(1)\sqrt{\beta_T} \sqrt{\frac{h_{T,\beta''}}{h_{T,\overline{\beta}'}}}\delta_{T3}^{-1} = o(1).$$

Consequently, there exists some positive constant L_3 such that, for T large enough,

$$\big(\mathsf{b}_{T,\overline{\beta}'} + \mathsf{b}_{T,\beta''}\big)\ (1 + \overline{\delta}_T) \leq L_3\sqrt{d}\ \delta_{T3}\eta_{T,\beta''}\ (1 - \overline{\delta}_T).$$

Therefore,

$$
\begin{aligned}
\mathsf{q}_1 &\leq \mathbf{P}_b\Big(\sup_{h\in\mathcal{H}_{T,\overline{\beta}'}} \|g_{T,\overline{\beta}'}(x_0,h) - \operatorname{div}(a\rho)(x_0)\| \\
&\qquad\qquad + \sup_{h\in\mathcal{H}_{T,\beta''}} \|g_{T,\beta''}(x_0,h) - \operatorname{div}(a\rho)(x_0)\| > \sqrt{d}\ \eta_{T,\beta''}(1-\overline{\delta}_T)\Big) \\
&\leq \mathbf{P}_b\Big(\big(\mathsf{b}_{T,\overline{\beta}'} + \mathsf{b}_{T,\beta''}\big)\ (1+\overline{\delta}_T) \\
&\qquad\qquad + \sup_{h\in H_{T,\overline{\beta}'}} \|Z_{T,\overline{\beta}'}(h)\| + \sup_{h\in H_{T,\beta''}} \|Z_{T,\beta''}(h)\| > \sqrt{d}\ \eta_{T,\beta''}(1-\overline{\delta}_T)\Big) \\
&\leq \mathbf{P}_b\Big(\sup_{h\in H_{T,\overline{\beta}'}} \|Z_{T,\overline{\beta}'}(h)\| + \sup_{h\in H_{T,\beta''}} \|Z_{T,\beta''}(h)\| > (1-L_3\delta_{T3})\sqrt{d}\ \eta_{T,\beta''}(1-\overline{\delta}_T)\Big) \\
&\leq \mathsf{q}_{1a} + \mathsf{q}_{1b},
\end{aligned}
$$

where

$$
\begin{aligned}
\mathsf{q}_{1a} &= \mathsf{q}_{1a}(\beta'') := \mathbf{P}_b\Big(\sup_{h\in H_{T,\beta''}} \|Z_{T,\beta''}(h)\| > (1-2L_3\delta_{T3})\sqrt{d}\ \eta_{T,\beta''}\ (1-\overline{\delta}_T)\Big), \\
\mathsf{q}_{1b} &= \mathsf{q}_{1b}(\overline{\beta}',\beta'') := \mathbf{P}_b\Big(\sup_{h\in H_{T,\overline{\beta}'}} \|Z_{T,\overline{\beta}'}(h)\| > L_3\delta_{T3}\sqrt{d}\ \eta_{T,\beta''}\ (1-\overline{\delta}_T)\Big).
\end{aligned}
$$

Part (a) of Lemma 2.6.8 implies that

$$
\begin{aligned}
\mathsf{q}_{1a} &\leq c_1' \exp\left(-\frac{(1-2L_3\delta_{T3})^2\ d\ \eta^2_{T,\beta''}\ (1-\overline{\delta}_T)^2\ (1-\gamma')}{2\mathsf{s}_T^2(\beta'')}\right) \\
&\leq c_1' \exp\left(-\frac{(1-L_5\ \delta_{T3})\ d\log T}{2\beta''}\right) \leq L_6\ T^{-d/(2\beta'')}.
\end{aligned}
$$

Lemma 2.6.8(b) entails that there exist some positive constants c_2', c_3' such that

$$
\mathsf{q}_{1b} \leq c_2' \exp\left(-c_3' \frac{\delta_{T3}^2 \eta^2_{T,\beta''}(1-\overline{\delta}_T)^2 T}{\mathsf{s}_T^2(\overline{\beta}')}\right).
$$

Since

$$
\begin{aligned}
\frac{\delta_{T3}\eta_{T,\beta''}}{\mathsf{s}_T(\overline{\beta}')} &= \frac{\delta_{T3}\mathsf{s}_T(\beta'')}{\mathsf{s}_T(\overline{\beta}')}\sqrt{\frac{\log T}{\beta''}} \geq \frac{\delta_{T3} h_{T,\beta''}^{-d/2}\|K_{\beta''}\|_{L^2(\mathbb{R}^d)}}{h_{T,\overline{\beta}'}^{-d/2}\|K_{\overline{\beta}'}\|_{L^2(\mathbb{R}^d)}}\sqrt{\frac{\log T}{\beta''}} \\
&\geq O(\delta_{T3})\ T^{d/(4\beta'')-d/(4\overline{\beta}')}\sqrt{\frac{\log T}{\beta''}} \to +\infty,
\end{aligned}
$$

the upper bound on q_{1b} is asymptotically negligible as compared to $T^{-d/(2\beta')}$. Summing up, $\mathsf{q}_1 \leq O(1)T^{-d/(2\beta')}$. Lemma 2.6.3 entails that q_2 is also asymptotically negligible, and in view of (2.6.102), we obtain the assertion. □

Let $\beta \in [\beta_*, \beta_T]$, $\beta' \in \mathcal{G}$, $L \in [L_*, L^*]$, $\nu = (\beta, L)$. The very same arguments as in the conclusion of the proof of part (**II**) of Theorem 2.5.9 entail that

$$
\lim_{T\to\infty}\sup_{\nu\in\mathcal{B}_T} g_1(\nu) = 0 \quad \text{and} \quad \lim_{T\to\infty}\sup_{\nu\in\mathcal{B}_T} g_2(\nu) = 0.
$$

In combination with another application of Lemma 2.6.3, we obtain (2.6.101). □

3. Donsker theorems for multidimensional ergodic diffusions

We study weak convergence of empirical processes of multidimensional ergodic diffusions under the basic assumption that the underlying invariant measure satisfies Poincaré's inequality. Results from classical empirical process theory such as Ossiander's bracketing CLT and the Giné–Zinn CLT are revisited. We further prove increased regularity for diffusions with finite invariant measure by investigating smoothed versions of the empirical diffusion process.

3.1. Introduction

Let $(X_t)_{t\geq 0}$ be an ergodic diffusion process on $E \subset \mathbb{R}^d$ with unique invariant probability measure μ. Denote its infinitesimal generator on $L^2(E,\mu) =: L^2(\mu)$ by A, assume that $X_0 \sim \mu$, and consider $0 \not\equiv f \in L^2_0(\mu)$. Recall that the additive functional

$$\left(\int_0^t f(X_u)\mathrm{d}u\right)_{t\geq 0} \tag{3.1.1}$$

of the diffusion $(X_t)_{t\geq 0}$ is said to satisfy a central limit theorem (CLT) if

$$\frac{1}{\sqrt{t}}\int_0^t f(X_u)\mathrm{d}u \Rightarrow_{t\to\infty} Z \sim \mathcal{N}(0,1). \tag{3.1.2}$$

Functional central limit theorems provide an extension of the CLT. For instance, the functional CLT due to Bhattacharya (1982) states that, for any fixed $t \geq 0$ and any fixed function f of the form $f = AF$,

$$\frac{1}{\sqrt{n}}\left(\int_0^{nt} f(X_u)\mathrm{d}u\right)_{t\geq 0} \Rightarrow_{n\to\infty} \sigma(f)W, \tag{3.1.3}$$

where $\sigma^2(f) := -2\int_E F(x)f(x)\mu(\mathrm{d}x)$ and $W = (W_t)_{t\geq 0}$ is a standard Brownian motion. The passage to a continuous-time result is obvious and allows to recover a CLT, namely

$$\frac{1}{\sqrt{t}}\int_0^t f(X_u)\mathrm{d}u \Rightarrow_{t\to\infty} Z \sim \mathcal{N}\left(0,\sigma^2(f)\right). \tag{3.1.4}$$

The Cramér–Wold device immediately gives a multidimensional extension of (3.1.4). Given any finite set of functions $f_1,\ldots,f_m$ of the form $f_i = AF_i$, $i = 1,\ldots,m$, the law of the m-dimensional process with components

$$\frac{1}{\sqrt{t}}\int_0^t f_i(X_u)\mathrm{d}u, \qquad i = 1,\ldots,m,$$

converges weakly to an m-dimensional centered Gaussian distribution with asymptotic covariances given by

$$-\int_E f_i(x)Af_j(x)\mu(\mathrm{d}x) - \int_E Af_i(x)f_j(x)\mu(\mathrm{d}x), \qquad i,j = 1,\ldots,m.$$

The aim in the sequel is to provide an extension of the CLT mentioned above in a different direction and to describe classes of functions on which limit results in the spirit of (3.1.4) hold uniformly. This question is the central issue of empirical process theory. Given i.i.d. random variables $X_1, \dots, X_n \sim \mathbf{P}$ and a class of functions $\mathcal{F} \subset L^2(\mathbf{P})$, the empirical process indexed by $\mathcal{F}$ is given as

$$(\mathbb{G}_n(f))_{f\in\mathcal{F}} := \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n (f(X_i) - \mathbf{E}f(X_1))\right)_{f\in\mathcal{F}}. \tag{3.1.5}$$

Lindeberg's CLT gives convergence of the finite-dimensional marginals whenever the variance of $f(X_1)$ is finite. In order to extend this result to a CLT in the space $\ell^\infty(\mathcal{F})$ of uniformly bounded functions $z : \mathcal{F} \to \mathbb{R}$, the existence of a tight version of the limiting Gaussian process does not suffice. Rather one has to verify additional side conditions related to the random geometry of $\mathcal{F}$ in order to get a *uniform* CLT or so-called *Donsker theorem*.

For ergodic diffusion processes, one standard strategy for proving a (finite-dimensional) CLT similar to (3.1.4) is to use a martingale representation and to decompose the additive functional (3.1.1) into the sum of an L^2 martingale and a remainder term which vanishes in the limit. The proof of the CLT is then reduced to a CLT for martingales. The same strategy has been applied by van der Vaart and van Zanten (2005) for proving their Donsker theorems for scalar ergodic diffusions with finite speed measure. In particular, it is shown in their article that the empirical process of a regular scalar diffusion on an interval $I \subseteq \mathbb{R}$ with finite speed measure behaves substantially different as compared to the classical empirical process (3.1.5). In fact, weak convergence of the empirical process takes place in $\ell^\infty(\mathcal{F})$ if and only if the limit exists as a tight, Borel measurable map. Contrary to the situation for i.i.d. random elements, no additional (entropy) conditions restricting the size of the class $\mathcal{F}$ are required. The proof of this result is heavily based on an analysis of diffusion local time and relies on the fact that the empirical measure of a univariate regular diffusion is continuous with respect to Lebesgue measure. For dimension $d \geq 2$, diffusion local time does not exist. In particular, the empirical measure of a multivariate diffusion is no longer Lebesgue-continuous. While some relation between the empirical process of the diffusion to a family of local martingales can be established in the multidimensional case, too, it is however not clear whether necessary and sufficient conditions can be obtained in this way. There exist different possible origins for the formulation of Donsker theorems:

(a) CLT via Poisson equation in L^2;

(b) CLT from the existence of asymptotic variance in the reversible case.

The strategy in (a) starts by solving the Poisson equation for some fixed $0 \not\equiv f \in L_0^2(\mu)$. If $AF = f$ admits some regular enough solution F, Itô's formula implies that, for every $t \geq 0$ and $\varepsilon > 0$,

$$\int_0^{\varepsilon^{-1}t} f(X_u)\mathrm{d}u = F(X_{\varepsilon^{-1}t}) - F(X_0) - M_t^\varepsilon,$$

where $(M_t^\varepsilon)_{t\geq 0}$ is a local martingale. The functional CLT for L^2 local martingales due to Rebolledo (1980) may then be applied to obtain convergence of the finite-dimensional marginal laws. In the reversible setting, a similar approach allows to prove finite-dimensional weak convergence whenever the Kipnis–Varadhan condition is satisfied (see Section 3.2.2 below). We first show that Poincaré's inequality can be used to verify an increment condition on the empirical process of X and allows to state sufficient criteria for Donsker theorems. Furthermore, (**PI**) implies a Bernstein-type deviation inequality which holds for any μ-integrable bounded function h with $\mu(h) = 0$, namely, for any $t, r > 0$,

$$\mathbf{P}_\mu\left(\frac{1}{t}\int_0^t h(X_u)\,\mathrm{d}u > r\right) \leq \exp\left(-t\min\left\{\frac{r^2}{8c_P\ \mathrm{Var}_\mu(h)}, \frac{r}{2c_P\ \|h\|_\infty}\right\}\right) \tag{3.1.6}$$

(cf. Proposition 1.2 in Cattiaux and Guillin (2008); also see Theorem 3.1 in Guillin et al. (2009)). Bernstein-type inequalities in this spirit are perfectly tailored to the application of techniques from empirical process theory. In particular, (3.1.6) allows to state a sufficient condition for Donsker theorems in terms of bracketing metric entropy. The sharp Bernstein-type deviation inequality due to Lezaud (2001) applies to symmetric Markov processes, still under the assumption that Poincaré's inequality is satisfied. It replaces the variance term on the right-hand side of (3.1.6) with the asymptotic variance of h appearing in the CLT, given as

$$\lim_{t\to\infty} \frac{1}{t} \operatorname{Var}_{\mathbf{P}_\mu}\left(\int_0^t h\left(X_u\right) \mathrm{d}u\right),$$

and allows for a more refined analysis. In general, Bernstein-type deviation inequalities provide some control of the Orlicz norm, and standard chaining techniques can be applied to derive conditions for asymptotic equicontinuity. One obstacle is that both (3.1.6) and the inequality due to Lezaud (2001) are restricted to bounded functions, and the question arises how *unbounded* function classes might be treated in the highest possible generality. Bernstein's inequality for continuous martingales is one resort. The approach is conceptually different and relies on the idea of martingale approximation as it is also used for proving finite-dimensional weak convergence. Given any continuous local martingale M vanishing at 0, Bernstein's inequality for continuous martingales (see, e.g., Revuz and Yor (1999), p. 153) states that

$$\forall\ t, r > 0, \qquad \mathbf{P}_\mu\left(\sup_{s\le t} M_s \ge rt; \langle M, M\rangle_t \le ct\right) \le \exp\left(-\frac{tr^2}{2c}\right), \tag{3.1.7}$$

thus giving a subgaussian bound whenever some control of the form

$$\frac{1}{t}\langle M, M\rangle_t \le c \tag{3.1.8}$$

for all t is available. Here, $\langle M, M\rangle_t$ denotes the martingale quadratic variation of M at time t. While (3.1.7) is not restricted to bounded functions, it however requires verifying a condition in the spirit of (3.1.8). We shall investigate the different approaches in more detail in the sequel.

3.2. Preliminaries

3.2.1. Notation and definitions

Let $(X_t)_{t\ge0}$ be an ergodic Markov process on a Borel measurable space $(E, \mathscr{B}(E))$, $E \subset \mathbb{R}^d$, with invariant measure μ as introduced in Section 1.1.2. Define the associated Markov semigroup $(P_t)_{t\ge0}$ by $(P_t f)(x) := \mathbf{E}_\mu\left(f\left(X_t\right) | X_0 = x\right)$, and denote by A the infinitesimal generator of (P_t) with domain $\mathbb{D}_A$ in $L^2(\mu)$. Usually, only a dense subset of the domain $\mathbb{D}_A$ is known, and so it is convenient to assume that there exists an algebra $\mathcal{A}$ of bounded functions on E which is dense in $\mathbb{D}_A$ and in all $L^p(\mu)$ spaces and stable by A (cf. Ledoux (2000) and Bakry (2008)). The *carré du champ* operator Γ is defined by

$$\Gamma(f, g) := A(fg) - fAg - gAf, \qquad f, g \in \mathcal{A}.$$

In the current diffusion framework, where A is a second order differential operator (cf. (1.1.1)), it holds

$$\Gamma(f)(x) := \Gamma(f, f)(x) = \langle a(x)\nabla f(x), \nabla f(x)\rangle, \qquad f \in \mathcal{A}.$$

Let $A = -\int_0^{+\infty} \lambda \mathrm{d}E_\lambda$ be the spectral decomposition of A on $L^2(\mu)$. The Dirichlet form $\mathcal{E}(f, g)$ is defined by

$$\mathbb{D}_{\mathcal{E}} = \mathbb{D}_{\sqrt{-A}} = \left\{h \in L^2(\mu) : \int_0^{+\infty} \lambda \mathrm{d}\left\langle E_\lambda h, h\right\rangle_\mu < +\infty\right\},$$

$$\mathcal{E}(f, g) = \left\langle\sqrt{-A}f, \sqrt{-A}g\right\rangle_\mu = \int_0^{+\infty} \lambda \mathrm{d}\left\langle E_\lambda f, g\right\rangle_\mu, \qquad f, g \in \mathbb{D}_{\mathcal{E}}.$$

Inequalities which connect the $L^p(\mu)$ norms of some function f to L^q norms of the gradient $\sqrt{\Gamma(f)}$ will prove useful in the sequel. Recall that μ satisfies a Poincaré inequality if there exists some constant c_P such that

$$\forall\, f \in \mathcal{A}, \qquad \operatorname{Var}_\mu(f) := \int f^2 \mathrm{d}\mu - \left(\int f \mathrm{d}\mu\right)^2 \leq c_P \mathcal{E}(f,f) = c_P \int \Gamma(f,f) \mathrm{d}\mu. \qquad (\mathbf{PI})$$

3.2.2. Finite-dimensional weak convergence

Let $\mathcal{F} \subset L_0^2(\mu)$. It was already mentioned that convergence of the marginals of $(\mathbb{G}_t(f))_{f\in\mathcal{F}}$ can be established in different situations and under different assumptions on $\mathcal{F}$.

Case 1. *Uniform CLT via Poisson equation in L^2.* Assume that $\mathcal{F} \subset \mathbb{D}_{A^{-1}}$ such that, for any $f \in \mathcal{F}$, there exists $F \in \mathbb{D}_A$ with $AF = f$. By Dynkin's formula, $(M_t^f)_{t\geq 0}$, defined by

$$M_t^f := F(X_t) - F(X_0) - \int_0^t (AF)(X_u)\,\mathrm{d}u,$$

is a locally square-integrable martingale relative to $\mathbf{P}_\mu$, and if $F^2 \in \mathbb{D}_A$, it holds

$$\langle M^f \rangle_t = \int_0^t \Gamma(F,F)(X_u)\,\mathrm{d}u.$$

Letting $M_{n,t}^f := M_{nt}^f/\sqrt{n}$, it follows from the law of large numbers that

$$\langle M_{n,\cdot}^f \rangle_t \to_{n\to\infty} \int \Gamma(F,F) \mathrm{d}\mu.$$

Since, for any fixed $t \geq 0$, $F(X_{nt})/\sqrt{n} \to_{n\to\infty} 0$ in probability, it follows from the martingale CLT that (3.1.3) and (3.1.4) hold. In the case of an infinite-dimensional extension of a CLT which is obtained via the Poisson equation in L^2, the limiting process $\mathbb{G}$ in a Donsker theorem must be centered Gaussian with covariance function

$$\begin{aligned}\mathbf{E}_\mu \mathbb{G}(f)\mathbb{G}(g) &= \int \Gamma(A^{-1}f, A^{-1}g)\mathrm{d}\mu \\ &= -\int \left(\left[A^{-1}f\right] g + \left[A^{-1}g\right] f\right) \mathrm{d}\mu, \qquad f,g \in \mathcal{F}. \end{aligned} \tag{3.2.1}$$

The pseudo-metric induced by $\mathbb{G}$ on $\mathcal{F}$ is given by

$$\begin{aligned} d_{\mathbb{G}}^2(f,g) &= \int \Gamma(A^{-1}(f-g))\mathrm{d}\mu \\ &= -2\int (f-g)\left[A^{-1}(f-g)\right] \mathrm{d}\mu, \qquad f,g\in\mathcal{F}. \end{aligned} \tag{3.2.2}$$

Remark 3.2.1. The CLT for Markov processes via solution of the Poisson equation related to the infinitesimal generator has been studied amply in the literature; cf. Section VIII.3f in Jacod and Shiryaev (2002). One classical reference for continuous Markov processes is Bhattacharya (1982). More recently, the approach has been extended to obtain CLTs for the Euler scheme with decreasing step of Brownian diffusions (see Lamberton and Pagès (2002)) and for Lévy-driven SDEs (Panloup (2008)).

Remark 3.2.2. It follows from results on the Poisson equation in a bounded domain that the equation $AF = f$ on $\mathbb{R}^d$ has a solution when f is compactly supported and the diffusion X is uniformly elliptic. In the general case, Pardoux and Veretennikov (2001) solve the problem under ellipticity conditions in specific Sobolev spaces and give upper bounds for F and its first derivatives. Further explicit criteria for solving the Poisson equation in L^2 are given in Cattiaux et al. (2012).

Case 2. *Reversible setting: Uniform CLT from the existence of asymptotic variance.* For reversible μ, a CLT holds even if the Poisson equation cannot be solved. Given $f \in L_0^2(\mu)$, the *Kipnis–Varadhan condition* is satisfied if

$$V := \int_0^\infty \left(\int (P_s f)^2 \mathrm{d}\mu\right) \mathrm{d}s < \infty. \tag{3.2.3}$$

In this case, the CLT holds under $\mathbf{P}_\mu$ with $\mathrm{Var}_\mu(\mathbb{G}_t) \sim_{t\to\infty} 4V$. Let $\mathcal{F} \subset L_0^2(\mu)$ be such that any $f \in \mathcal{F}$ satisfies (3.2.3). It follows from the above that the pseudo-metric induced by the limiting Gaussian process $\mathbb{G}$ on $\mathcal{F}$ reads

$$d_{\mathbb{G}}^2(f,g) = \int_0^\infty \left(\int (P_s(f-g))^2 \mathrm{d}\mu\right) \mathrm{d}s, \qquad f,g \in \mathcal{F}. \tag{3.2.4}$$

For functions $f,g \in \mathbb{D}_{A^{-1}}$, the expressions in (3.2.2) and (3.2.4) coincide.

3.2.3. Weak convergence of the empirical diffusion process

The classes of functions on E to be considered in the sequel will be countable families $\mathcal{F}$ of (real-valued) measurable functions f on $(E, \mathcal{B}(E))$ such that $\|f(x)\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}} |f(x)| < \infty$ for all $x \in E$. By restricting attention to countable classes, we circumvent the various measurability questions that the study of empirical processes raises. A convenient condition however is admissibility or, more precisely, Dudley's notion of Suslin admissible classes; see Dudley (1999), Chapter 5.

Definition 3.2.3. Let $\mathcal{F} \subset L_0^1(\mu)$ be a countable class of measurable functions, and define for any $t > 0$ the random map

$$\mathbb{G}_t(f) := \frac{1}{\sqrt{t}} \int_0^t f(X_u)\mathrm{d}u, \qquad f \in \mathcal{F}.$$

We say that $\mathcal{F}$ is a *Donsker class* (for μ) if $\mathbb{G}_t(f) \rightsquigarrow_{\ell^\infty(\mathcal{F})} \mathbb{G}$, where $\mathbb{G}$ is a tight, Borel measurable element of $\ell^\infty(\mathcal{F})$.

Weak convergence in $\ell^\infty(\mathcal{F})$ to a tight Borel measurable limit is equivalent to finite-dimensional weak convergence and (i) asymptotic tightness, or, equivalently, (ii) equicontinuity with respect to a semimetric d such that $(\mathcal{F}, d)$ is totally bounded.

By definition, $\mathcal{F}$ can be Donsker only if there exists a version of the Gaussian process $\mathbb{G}$ which is a tight Borel measurable map into $\ell^\infty(\mathcal{F})$. Tightness of the Gaussian process $\mathbb{G}$ indexed by $\mathcal{F}$ is equivalent to saying that $\mathcal{F}$ is pregaussian when we consider this term in the diffusion context in accord with classical empirical process usage.

Definition 3.2.4. (i) We call $\mathcal{F} \subset L^2(\mu)$ *pregaussian* (for μ) if the centered Gaussian process $\mathbb{G}$ appearing in the CLT admits a version with almost all sample paths bounded and continuous on $\mathcal{F}$ with respect to the metric $d_{\mathbb{G}}^2(f,g) := \varsigma^2(f-g)$, where

$$\varsigma^2(f) := \lim_{t\to\infty} \mathrm{Var}_\mu\left(\mathbb{G}_t(f)\right) = \lim_{t\to\infty} \mathrm{Var}_\mu\left(\frac{1}{\sqrt{t}}\int_0^t f(X_u)\mathrm{d}u\right) \tag{3.2.5}$$

is the asymptotic variance appearing in the CLT.

(ii) Given a separable, infinite-dimensional Hilbert space H, a set $C \subset H$ is called a *GC-set* if the restriction of the isonormal Gaussian process $\mathbb{L}$ on H can be chosen such that its sample functions are uniformly continuous on C.

A set
$$\mathcal{G} \subset \mathcal{L}_0^2(\mathbf{P}) := \left\{ f \in \mathcal{L}^2(\mathbf{P}) : \int f \mathrm{d}\mathbf{P} = 0 \right\}$$
is pregaussian if and only if the corresponding set in $L_0^2(\mathbf{P})$ is a GC-set; here, as usual, $L_0^2(\mathbf{P})$ denotes the set of all equivalence classes of elements of $\mathcal{L}_0^2(\mathbf{P})$ for $\mathbf{P}$-a.s. equality. In order to make this definition useful in the current framework, we consider Case 1. Assume that the finite-dimensional distributions (fidis) of the random maps $\mathbb{G}_t$ converge weakly to those of a centered Gaussian random map $\mathbb{G}$ with covariance structure $\operatorname{cov}(\mathbb{G}(f), \mathbb{G}(g)) = \int \Gamma(A^{-1}f, A^{-1}g)\mathrm{d}\mu$. Define the norm $\|\cdot\|_{\mathbb{D}_A}$ on the domain of A as in (1.1.4), and assume that the carré du champ associated with A satisfies Poincaré's inequality. Then the couple $(\mathbb{D}_A, \|\cdot\|_{\mathbb{D}_A})$ defines a separable Hilbert space (see Lemma A.2.1 in the Appendix). This observation enables one to use classical results on the continuity of Gaussian processes in order to verify pregaussianness as defined above.

3.3. Approaches to uniform CLTs

Weak convergence in $\ell^\infty(\mathcal{F})$ is equivalent to finite-dimensional weak convergence and asymptotic uniform equicontinuity. The question of convergence of the fidis was briefly discussed in Section 3.2.2. In particular, it was pointed out that CLTs can be obtained via the Poisson equation and, for reversible μ, from the existence of the asymptotic variance. There exists a wealth of extensions and alternative approaches to proving CLTs for ergodic Markov processes. We shall not further deepen this question but refer to Cattiaux et al. (2012) for a comprehensive recent review of situations where finite-dimensional convergence can be established. For proving uniform CLTs, asymptotic equicontinuity remains to be verified, and this is where the empirical process machinery comes into play. In this section, we shall focus on the approach via solving the Poisson equation in L^2 such that the latter task reduces to showing that, for every $\varepsilon, \eta > 0$, there exists some $\delta > 0$ such that
$$\limsup_{t\to\infty} \mathbf{P}_\mu \left(\sup_{d_{\mathbb{G}}(f,g)<\delta} |\mathbb{G}_t(f) - \mathbb{G}_t(g)| > \varepsilon \right) < \eta, \qquad f, g \in \mathcal{F} \subset \mathbb{D}_{A^{-1}},$$
where $d_{\mathbb{G}}$ is given by (3.2.2).

3.3.1. Poincaré's inequality and an Ossiander-type bracketing CLT

The first result is convenient for statistical applications as it permits to use known results on bracketing numbers. Recall that $N_{[\,]}(\varepsilon, \mathcal{F}, L^2(\mu))$ denotes the ε-entropy with bracketing.

Proposition 3.3.1 (Rohde and Strauch (2010), Theorem 3.2)**.** *Let X be an ergodic stationary diffusion process satisfying condition* (D2)*, and let $\mathcal{F} \subset \mathbb{D}_{A^{-1}}$ be a countable class of bounded functions. If $\mathcal{F}$ satisfies*
$$\int_0^\infty \sqrt{\log N_{[\,]}(\varepsilon, \mathcal{F}, L^2(\mu))}\mathrm{d}\varepsilon < \infty, \tag{3.3.1}$$
then $\mathcal{F}$ is Donsker.

Proof. It remains to prove asymptotic equicontinuity. Ossiander's result is about L^2-bracketing. The proof of this classical bracketing CLT as it is given in Dudley (1999), pp. 239–244, is based on chaining arguments which are also valid when the pseudo-metric $d_{\mathbb{G}}$ is used. The only ingredient of the proof which does not apply in the diffusion context is the classical Bernstein-inequality which can be replaced with the Bernstein-type inequality (3.1.6). Furthermore, by

Cauchy–Schwarz and Poincaré's inequality,

$$\operatorname{Var}_\mu(\mathbb{G}_t(Ag)) = \langle g, Ag\rangle_\mu \le \|g\|_{L^2(\mu)}\|Ag\|_{L^2(\mu)} \lesssim \|Ag\|^2_{L^2(\mu)},$$

hence $d_{\mathbb{G}}(A^{-1}f, A^{-1}h) \lesssim \|f-h\|_{L^2(\mu)}$, so the bracketing entropy numbers with respect to $d_{\mathbb{G}}$ can be upper-bounded by $L^2(\mu)$-bracketing. □

3.3.2. Applications of the sharp Bernstein-type inequality for symmetric Markov processes

In this section, we require a sharper version of (3.1.6) due to Lezaud (2001). Using Kato's perturbation theory, he proved that for any $h \in L^1_0(\mu)$, it holds

$$\forall t, r > 0, \qquad \mathbf{P}_\mu\left(\left|\frac{1}{t}\int_0^t h(X_u)\mathrm{d}u\right| > r\right) \le 2\exp\left(-\frac{tr^2}{2\left(\varsigma^2(h) + c_P\, r\|h\|_\infty\right)}\right), \tag{3.3.2}$$

where $\varsigma^2(h)$ is the asymptotic variance appearing in the CLT (cf. (3.2.5)). If $\varsigma^2(h) \lesssim \|h\|^2_\infty$ for all h as above, then the Bernstein-type concentration inequality (3.3.2) implies the Poincaré inequality (**PI**) (cf. the comment below Theorem 1.2 in Gao et al. (2010) and the subsequent remark), and (**PI**) can be seen as some kind of minimal assumption for (3.3.2) to hold for all bounded functions h.

A sufficient condition for asymptotic tightness

We first consider the more general framework described in Theorem 1.2.7 in Talagrand (2005). Definitions and relevant properties of the γ_α functionals are given in Appendix A.3. Let T be a set equipped with two distances ϖ_1, ϖ_2, and consider a process $(Z_t)_{t\in T}$ with $\mathbf{E}Z_t = 0$ which satisfies the following increment condition,

$$\Pr\left(|Z_s - Z_t| \ge u\right) \le 2\exp\left(-\min\left\{\frac{u}{\varpi_1(s,t)}, \frac{u^2}{\varpi_2(s,t)^2}\right\}\right), \qquad s,t \in T.$$

Then, for all $u, \delta > 0$, $k \in \mathbb{N}$, the following inequality holds,

$$\mathbf{E}\sup_{\varpi_2(s,t)<\delta}|Z_s - Z_t| \le L\left(\inf\sup_{t\in T}\sum_{n>k} 2^n\Delta_1(A_n(t)) + \inf\sup_{t\in T}\sum_{n>k} 2^{n/2}\Delta_2(A_n(t)) + \delta 2^{2^k}\right), \tag{3.3.3}$$

where the infimum is always taken over all admissible sequences $(\mathcal{A}_n)$ of partitions of T. For the proof, define the functionals

$$\zeta_1(k) := \inf\sup_{t\in T}\sum_{n>k} 2^n\Delta_1(A_n(t)), \qquad \zeta_2(k) := \inf\sup_{t\in T}\sum_{n>k} 2^{n/2}\Delta_2(A_n(t)), \quad k \in \mathbb{N}.$$

Fix $k > 0$. Theorem 1.3.6 in Talagrand (2005) states that $\gamma_\alpha(T,d) \le L(\alpha)\sup\gamma_\alpha(F,d)$, where the supremum is taken over all finite $F \subset T$. Thus, it is no loss of generality to assume that T is finite. Inspection of the proof of Theorem 1.2.7 in Talagrand (2005) yields

$$\begin{aligned}\mathbf{E}\sup_{t\in T}\left|Z_t - Z_{\pi_k(t)}\right| &\le L\sup_{t\in T}\sum_{n>k}\left(2^n\varpi_1(\pi_n(t),\pi_{n-1}(t)) + 2^{n/2}\varpi_2(\pi_n(t),\pi_{n-1}(t))\right)\\ &\le L\left(\zeta_1(k) + \zeta_2(k)\right),\end{aligned}$$

where $\pi_n(t)$ denotes the point in $A_n(t)$ which is closest to t (in terms of the respective metric). The remainder of the proof is along the lines of the proof of Theorem 11.14 in Ledoux and Talagrand (1991). For the sake of completeness, we sketch the remaining steps here. Define

$$U := \{(x, y) \in \mathcal{A}_k \times \mathcal{A}_k : \exists u, v \in T \text{ such that } \varpi_2(u, v) < \delta, \pi_k(u) = x, \pi_k(v) = y\}.$$

For $(x, y) \in U$, fix $u_{x,y}, v_{x,y}$ such that $\pi_k(u_{x,y}) = x$, $\pi_k(v_{x,y}) = y$, and $\varpi_2(u_{x,y}, v_{x,y}) < \delta$. By Lemma 11.3 in Ledoux and Talagrand (1991), it holds

$$\mathbf{E} \sup_{x,y\in U} |Z_{u_{x,y}} - Z_{v_{x,y}}| \le \delta\sqrt{\operatorname{card} U} \le 2^{2^k}\delta.$$

Consider arbitrary $s, t \in T$ with $\varpi_2(s, t) < \delta$, and let $x = \pi_k(s)$, $y = \pi_k(t)$. It then follows that $(x, y) \in U$ and, since $\pi_k(u_{x,y}) = \pi_k(s) = x$, $\pi_k(v_{x,y}) = \pi_k(t) = y$,

$$|Z_s - Z_t| \le \sup_{x,y\in U} |Z_{u_{x,y}} - Z_{v_{x,y}}| + 4 \sup_{t\in T} \left|Z_t - Z_{\pi_k(t)}\right|.$$

This proves (3.3.3). Returning to the case of a symmetric Markov process $(X_t)_{t\ge0}$ whose carré du champ satisfies (3.3.2), let $\mathcal{F} \subseteq L^1_0(\mu)$ be a class of bounded functions. Consider the collection $(\mathbb{G}_t(f))_{f\in\mathcal{F}}$ of functionals $\mathbb{G}_t(f) := t^{-1/2}\int_0^t f(X_u)\mathrm{d}u$, $t > 0$, indexed by $\mathcal{F}$. Then the above reasoning implies that, for all $u, t, \delta > 0$, $k \in \mathbb{N}$,

$$\begin{aligned} \mathbf{P}_\mu\left(\sup_{f,g\in\mathcal{F}: d_{\mathbb{G}}(f,g)<\delta} |\mathbb{G}_t(f - g)| > u\right) \quad &\le \quad \frac{L}{u}\Big(\inf \sup_{f\in\mathcal{F}} \sum_{n>k} 2^{n/2}\Delta_2(A_n(f)) \\ &\quad + \frac{1}{\sqrt{t}} \inf \sup_{f\in\mathcal{F}} \sum_{n>k} 2^n\Delta_\infty(A_n(f)) + \delta 2^{2^k}\Big), \end{aligned} \tag{3.3.4}$$

where the infimum again is taken over all admissible sequences of partitions of $\mathcal{F}$.

Proposition 3.3.2. *Assume that μ satisfies (3.3.2). If $\mathcal{F}$ is pregaussian and if there exists an admissible sequence $(\mathcal{A}_n)_{n\ge0}$ of partitions of $\mathcal{F}$ such that*

$$\lim_{k\to\infty} \sup_{f\in\mathcal{F}} \sum_{n>k} 2^n\Delta_\infty(A_n(f)) = o_{\mathbf{P}}(1), \tag{3.3.5}$$

then $(\mathbb{G}_t(f))_{f\in\mathcal{F}}$ is asymptotically $d_{\mathbb{G}}$-equicontinuous.

Proof. Since $\mathcal{F}$ is pregaussian, Theorem A.3.3 implies that there exists an admissible sequence of partitions $(\mathcal{A}_n)_{n\ge0}$ of $\mathcal{F}$ such that

$$\lim_{k\to\infty} \sup_{f\in\mathcal{F}} \sum_{n>k} 2^{n/2}\Delta_2(A_n(f)) = 0. \tag{3.3.6}$$

Given $\eta > 0$, it follows from (3.3.4), the above condition (3.3.5) and (3.3.6) that $\delta > 0$ can be chosen such that

$$\lim_{t\to\infty} \mathbf{P}_\mu\left(\sup_{d_{\mathbb{G}}(f,g)<\delta} \frac{1}{\sqrt{t}}\left|\int_0^t (f - g)(X_s)\mathrm{d}s\right| > \varepsilon\right) < \eta.$$

□

Remark 3.3.3. Assume that $h_1, \dots, h_m \in L^1_0(\mu)$ satisfy the increment condition (3.3.2), and that $\max_{i=1,\dots,m} \operatorname{Var}_\mu(\mathbb{G}_t(h_i)) \le \varsigma^2$ and $\max_{i=1,\dots,m} \|h_i\|_\infty \le M$. Analogous to the proof of Lemma 2.2.10 in van der Vaart and Wellner (1996), it can be shown that, for any $p \ge 1$,

$$\left\|\max_{i=1,\dots,m} \frac{1}{\sqrt{t}} \int_0^t h_i(X_u)\,\mathrm{d}u\right\|_{L^p(\mu)} \lesssim \sqrt{\log(1+m)}\left(\varsigma + \frac{c_P M}{\sqrt{t}}\sqrt{\log(1+m)}\right). \tag{3.3.7}$$

The chaining method provides extensions of bounds on finite suprema as in (3.3.7) to the infinite-dimensional case. For a class $\mathcal{F} \subset L_0^1(\mu)$ of functions satisfying (3.3.2), the generic chaining technique immediately gives an upper bound of the form

$$\mathbf{E}_\mu \sup_{f\in\mathcal{F}} \left| \frac{1}{\sqrt{t}} \int_0^t f(X_u)\mathrm{d}u \right| \lesssim \gamma_2(\mathcal{F}, d_{\mathbb{G}}) + \frac{\gamma_1(\mathcal{F}, d_\infty)}{\sqrt{t}}.$$

Excursion: The effect of pregaussianness on the asymptotic equicontinuity criterion

We now state a result similar in spirit to Theorem 3.2 in Giné and Zinn (1984) which describes how pregaussianness bears on the equicontinuity condition. Their theorem states that a class $\mathcal{F}$ of uniformly bounded functions on $(S, \mathscr{B}(S), \mathbb{P})$ is $\mathbb{P}$-Donsker if and only if it is $\mathbb{P}$-pregaussian and it holds for any (or all) $\eta > 0$

$$\sup\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i (f-g)(X_i) : f, g \in \mathcal{F}, \|f-g\|_{L^2(\mathbb{P})} < \frac{\eta}{\sqrt{n}} \right\} \to_{n\to\infty} 0$$

in probability, where (ε_i) denotes a Rademacher (randomizing) sequence. Symmetrization plays a central role in the proof of this result (cf. also Ledoux and Talagrand (1991), pp. 405–407) which essentially consists of applying Sudakov's minoration principle and real exponential bounds which follow from comparison of Rademacher averages to Gaussian averages. Since the symmetrization device is not directly applicable in the diffusion framework, we use a different idea and decompose the function class $\mathcal{F}$. The Bernstein-type inequality (3.3.2) shows that the empirical process $\mathbb{G}_t$ exhibits a tail behaviour which is a mixture of a subexponential and a sub-Gaussian part. The sub-Gaussian part is controlled by the pregaussian hypothesis while the subexponential part can be dealt with by means of Sudakov's minoration. Given a function class $\mathcal{F}$ and any $\delta > 0$, let

$$\mathcal{F}_\delta := \{f - g : f, g \in \mathcal{F}, d_{\mathbb{G}}(f,g) < \delta\}. \tag{3.3.8}$$

Note that, while the original result due to Giné and Zinn (1984) is formulated in terms of the $L^2(\mathbb{P})$ norm, $\mathcal{F}_\delta$ in (3.3.9) is defined via $d_{\mathbb{G}}$ rather than $L^2(\mu)$.

Proposition 3.3.4 (Rohde and Strauch (2010), Theorem 3.1)**.** *Let $\mathcal{F} \subset L_0^1(\mu)$ be a countable class of uniformly bounded functions. The empirical process $(\mathbb{G}_t(f))_{f\in\mathcal{F}}$ is asymptotically equicontinuous if and only if $\mathcal{F}$ is pregaussian and it holds for any $\eta > 0$,*

$$\mathbf{E}_\mu \left\| \frac{1}{\sqrt{t}} \int_0^t f(X_u)\,\mathrm{d}s \right\|_{\mathcal{F}_{(\eta/\sqrt{t})^{1/2}}} \to_{t\to\infty} 0. \tag{3.3.9}$$

Proof. It is clear that the conditions on $\mathcal{F}$ are necessary for asymptotic equicontinuity of $\mathcal{F}$. The first part of the proof of sufficiency is along the lines of the proof of Theorem 3.2 in Giné and Zinn (1984). Let $\eta > 0$ be fixed, and set $\delta_t := (\eta/\sqrt{t})^{1/2}$, $m_t := N(\delta_t/2, d_{\mathbb{G}}, \mathcal{F})$. Pregaussianness of $\mathcal{F}$ implies by Sudakov's minoration (cf., e.g., Corollary 3.18 in Ledoux and Talagrand (1991)) that

$$\lim_{t\to\infty} \delta_t \sqrt{\log m_t} = 0. \tag{3.3.10}$$

Furthermore, there exists a finite set $\mathcal{G} := \mathcal{G}(\eta, t) = \{g_1, \ldots, g_{m_t}\}$ maximal in $\mathcal{F}$ such that $d_{\mathbb{G}}(g_i, g_j) > \delta_t$ for all $1 \le i < j \le m_t$. Since $\mathcal{G}$ forms a δ_t-covering of $\mathcal{F}$ with respect to $d_{\mathbb{G}}$, the triangle inequality yields $\|\mathbb{G}_t\|_{\mathcal{F}_\eta} \le \|\mathbb{G}_t\|_{\mathcal{F}_{\delta_t}} + \|\mathbb{G}_t\|_{\mathcal{G}_\eta}$. In view of (3.3.9), it remains to bound

$\mathbf{E}_\mu \|\mathbb{G}_t\|_{\mathcal{G}_\eta}$ from above. This is done by means of a slight modification of Lemma A.1 in van der Vaart (1996). For any $f \in (\mathcal{F} - \mathcal{F})$, let $\sigma_f^2 := \mathbf{E}_\mu \Gamma(f)$, $c_f := \sup_{x\in E} |f(x)|$, and consider the decomposition

$$|\mathbb{G}_t(f)| = |\mathbb{G}_t(f)| \cdot \mathbb{1}\left\{|\mathbb{G}_t(f)| \leq \frac{\sigma_f^2}{c_f}\sqrt{t}\right\} + |\mathbb{G}_t(f)| \cdot \mathbb{1}\left\{|\mathbb{G}_t(f)| > \frac{\sigma_f^2}{c_f}\sqrt{t}\right\}. \tag{3.3.11}$$

The sharp Bernstein-type inequality (3.3.2) then gives

$$\begin{aligned}
\mathbf{P}_\mu\left(|\mathbb{G}_t(f)| \cdot \mathbb{1}\left\{|\mathbb{G}_t(f)| \leq \frac{\sigma_f^2}{c_f}\sqrt{t}\right\} > x\right) &\leq 2\exp\left(-\frac{x^2}{4\sigma_f^2}\right), \\
\mathbf{P}_\mu\left(|\mathbb{G}_t(f)| \cdot \mathbb{1}\left\{|\mathbb{G}_t(f)| > \frac{\sigma_f^2}{c_f}\sqrt{t}\right\} > x\right) &\leq 2\exp\left(-\frac{x}{4c_f/\sqrt{t}}\right).
\end{aligned}$$

This implies (cf. the proof of Lemma A.1 in van der Vaart (1996)) that

$$\mathbf{E}_\mu \psi_1\left(\frac{|\mathbb{G}_t(f)|\mathbb{1}\left\{|\mathbb{G}_t(f)| > \frac{\sigma_f^2}{c_f}\sqrt{t}\right\}}{Kc_f/\sqrt{t}}\right) \leq 1, \qquad \mathbf{E}_\mu \psi_2\left(\frac{|\mathbb{G}_t(f)|\mathbb{1}\left\{|\mathbb{G}_t(f)| \leq \frac{\sigma_f^2}{c_f}\sqrt{t}\right\}}{K\sigma_f}\right) \leq 1,$$

where K is a sufficiently large (universal) constant, not depending on c_f and σ_f, and ψ_α are the Young functions $\psi_\alpha(x) := \exp(x^\alpha) - 1$, $\alpha = 1, 2$. Inequality (2.10) in Arcones and Giné (1993) then gives for $c' := \sup_{f\in\mathcal{F}} c_f$,

$$\mathbf{E}_\mu \max_{1\leq i<j\leq m_t} |\mathbb{G}_t(f_i - f_j)|\, \mathbb{1}\left\{|\mathbb{G}_t(f_i - f_j)| > \frac{\sigma_{f_i-f_j}^2}{c_{f_i-f_j}}\sqrt{t}\right\} \lesssim \log(m_t)\frac{c'}{\sqrt{t}}.$$

For estimating the first term in (3.3.11), we use the generic chaining bound in Theorem 1.2.6 in Talagrand (2005) to obtain

$$\mathbf{E}_\mu \max_{1\leq i<j\leq m_t} |\mathbb{G}_t(f_i - f_j)|\, \mathbb{1}\left\{|\mathbb{G}_t(f_i - f_j)| \leq \frac{\sigma_{f_i-f_j}^2}{c_{f_i-f_j}}\sqrt{t}\right\} \lesssim \gamma_2\left(\mathcal{G}_\eta, d_{\mathbb{G}}\right).$$

Pregaussianness implies that $\lim_{\eta\to 0} \gamma_2\left(\mathcal{F}_\eta, d_{\mathbb{G}}\right) = 0$, and combined with (3.3.10) this completes the verification of the asymptotic equicontinuity condition. □

Remark 3.3.5. An alternative result on the influence of pregaussianness on the Donsker property can be derived from Theorem 2.6.2 in Talagrand (2005). Loosely speaking, it asserts that a countable class $\mathcal{F} \subset L^2(\mathbb{P})$ is $\mathbb{P}$-Donsker if it can be decomposed into the sum of one class $\mathcal{F}_1$ for which the random distances $\left(\sum_{i=1}^n n^{-1}|f(X_i)|^2\right)^{1/2}$ are controlled by the pregaussian hypothesis and another class $\mathcal{F}_2$ for which small $L^1(\mathbb{P}_n)$ perturbations have their $L^2(\mathbb{P}_n)$ distances bounded by a fixed Gaussian distance.

The importance of the above results on the effect of pregaussianness is that they can be used to formulate simplified conditions for asymptotic equicontinuity. In particular, they provide a starting point for deriving (modified versions of) uniform CLTs under minimal conditions (in the best case, reduced to pregaussianness). Versions of the Giné–Zinn result have been applied in Giné and Nickl (2008) and Radulović and Wegkamp (2009) for proving Donsker-type theorems for smoothed empirical processes under only pregaussian conditions; a detailed discussion of their results will be given in Section 3.6. Mendelson and Zinn (2006) use the alternative characterization sketched above (also cf. Theorem 1 in Talagrand (1987a), Theorem 3.1.3 in Giné and Zinn (1986), Corollary 14.9 in Ledoux and Talagrand (1991)) to construct a modified empirical process which converges under only pregaussian conditions. Their approach is based on the generic chaining technique which is applied to build a modified empirical process for which only the existence of the limiting Gaussian process is required to obtain both tail estimates and the CLT for the modified process. The variant of the empirical process they consider is based on almost admissible sequences and therefore only of limited use in practice. Note however that, replacing Bernstein's inequality for independent random variables with the sharp Bernstein-type deviation inequality (3.3.2), the proof of Theorem 2.7 in Mendelson and Zinn (2006) can be modified to construct a respective variant of the empirical diffusion process.

Beyond Poincaré's inequality

Empirical process theory has matured over the last decades, and the theory now provides tools for proving results known from the classical setting also in more general situations without too much effort. Giné (2007) points out strategies to prove CLTs for classical empirical processes in the following order,

(1) arguments from probability of Banach spaces,

(2) VC-type arguments,

(3) bracketing results, and

(4) conditions in terms of random entropies.

A bracketing result has already been stated in Section 3.3.1, and (1) and (4) will be investigated in this section. As concerns (2), we merely note that it is possible to discretize the empirical process and to work with the discretized version, exploiting some mixing properties. In particular, the symmetrization device can be applied after suitable decoupling to the discretized version, such that sufficient conditions ensuring asymptotic equicontinuity (such as Vapnik–Chervonenkis type conditions) can be derived. We do not pursue this strategy here but refer the reader to Rio (2000) for further results in this spirit.

Random entropy criteria

The proof of the result of van der Vaart and van Zanten (2005) on scalar empirical diffusion processes comprises the following steps:

(1) Reduction to the natural scale case such that it suffices to prove the theorem for diffusions X for which the identity is a scale function;

(2) asymptotic equivalence with uniform weak convergence of continuous local martingales;

(3) proof of convergence of the finite-dimensional marginals of the approximating martingales by means of the martingale CLT;

(4) verification of asymptotic equicontinuity by means of a limit theorem for diffusion local time.

While (2) and (3) have their counterparts in our previous analysis of weak convergence properties of the empirical processes of multidimensional diffusions, item (4) remains open. The first part of the proof of asymptotic equicontinuity in the scalar diffusion case consists in applying Bernstein's inequality for continuous martingales, and this can be done in the multidimensional setting, too. The approach gives a criterion for asymptotic tightness which does not rely on Poincaré's inequality.

Proposition 3.3.6. *Assume that μ is reversible, and define the pseudo-metric $d_{\mathbb{G}}$ as in (3.2.4). If $\mathcal{F}$ is pregaussian, if any $f \in \mathcal{F}$ satisfies the Kipnis–Varadhan condition (3.2.3), and if*

$$\sup_{\substack{f,g\in\mathcal{F}:\\ d_{\mathbb{G}}(f,g)>0}} \frac{\frac{1}{t}\int_0^\infty \int_0^t (P_s(f-g))^2(X_u)\mathrm{d}u\mathrm{d}s}{4\int_0^\infty \left(\int (P_s(f-g))^2\mathrm{d}\mu\right)\mathrm{d}s} = O_{\mathbf{P}}(1), \tag{3.3.12}$$

then

$$(\mathbb{G}_t(f))_{f\in\mathcal{F}} \rightsquigarrow_{\ell^\infty(\mathcal{F})} (\mathbb{G}(f))_{f\in\mathcal{F}},$$

where $\mathbb{G}$ is a centered Gaussian process with covariance structure

$$\mathbf{E}_\mu \mathbb{G}(f)\mathbb{G}(g) = 4\int_0^\infty \int P_s f P_s g \mathrm{d}\mu\mathrm{d}s.$$

Proof. Pregaussianness of $\mathcal{F}$ implies that $(\mathcal{F}, d_{\mathbb{G}})$ is totally bounded. For fixed $\varepsilon > 0$ and $f \in \mathcal{F}$, define the resolvent

$$g_\varepsilon := R_\varepsilon f := \int_0^\infty \mathrm{e}^{-\varepsilon s} P_s f \mathrm{d}s. \tag{3.3.13}$$

Thus, $g_\varepsilon \in \mathbb{D}_A$, $\varepsilon g_\varepsilon - A g_\varepsilon = f$, and

$$\int_0^t f(X_u)\mathrm{d}u = M_t(g_\varepsilon) + \varepsilon\int_0^t g_\varepsilon(X_u)\mathrm{d}u + g_\varepsilon(X_0) - g_\varepsilon(X_t),$$

where

$$M_t(g_\varepsilon) := g_\varepsilon(X_t) - g_\varepsilon(X_0) - \int_0^t (Ag_\varepsilon)(X_u)\mathrm{d}u$$

is a martingale. Since $\varepsilon\|g_\varepsilon\|^2 \to_{\varepsilon\to 0} 0$ by (3.2.3), the finite-dimensional marginals of $(\mathbb{G}_t(f))_{f\in\mathcal{F}}$ converge by means of the martingale CLT. Given $f, f' \in \mathcal{F}$ and any fixed $\varepsilon > 0$, define $g_\varepsilon, g'_\varepsilon$ according to (3.3.13). By the preceding, it suffices to prove asymptotic equicontinuity of the approximating martingales $M_t(\cdot)$. Bernstein's inequality for continuous martingales (cf. (3.1.7)) implies that, for any $\eta > 0$ and all $t > 0$,

$$\mathbf{P}_\mu\left(\frac{1}{\sqrt{t}}\left|M_t(g_\varepsilon - g'_\varepsilon)\right| > \eta; \frac{1}{t}\langle M(g_\varepsilon - g'_\varepsilon)\rangle_t \leq K^2 d^2_{\mathbb{G}}(f,f')\right) \leq 2\exp\left(-\frac{\eta^2}{2K^2 d^2_{\mathbb{G}}(f,f')}\right),$$

that is, $f \mapsto t^{-1/2}M_t(g_\varepsilon)\cdot \mathbb{1}\{t^{-1}\langle M(g_\varepsilon)\rangle_t \leq K^2\int (P_s f)^2\mathrm{d}\mu\}$ is subgaussian with respect to $Kd_{\mathbb{G}}$. Since $\mathcal{F}$ is pregaussian, there exists an admissible sequence $(A_n)_{n\geq 0}$ of partitions of $\mathcal{F}$ such that

$$\lim_{k\to\infty}\sup_{f\in\mathcal{F}}\sum_{n>k} 2^{n/2}\Delta_2(A_n(f)) = 0.$$

Now, for any $\zeta > 0$, there exists $\delta > 0$ such that, uniformly in t,

$$\mathbf{E}_\mu \sup_{d_{\mathbb{G}}(f,f')<\delta} \left|M_t(g_\varepsilon - g'_\varepsilon)\right| \cdot \mathbb{1}\left\{t^{-1}\langle M(g_\varepsilon - g'_\varepsilon)\rangle_t \leq K^2 d^2_{\mathbb{G}}(f,f')\right\} < \zeta. \tag{3.3.14}$$

Combining (3.3.12) and (3.3.14) then gives the assertion. □

3.4. Smoothed empirical diffusion processes: Introduction and motivation

The aim is to obtain Donsker-type theorems under necessary and sufficient conditions, and the strategy in the sequel is to study some modified version of the empirical diffusion process, the *smoothed empirical diffusion process.* All of the above considerations are based on the preprint Rohde and Strauch (2010) which arose in joint work with Angelika Rohde.

When estimating a linear functional or a collection of linear functionals of the form $\int_E f \mathrm{d}\mu$, given a continuous record $(X_u)_{0\le u\le t}$ of observations of X up to time $t>0$, a natural estimator is based on the empirical measure $\mu_t := t^{-1}\int_0^t \delta_{X_u}\mathrm{d}u$, namely

$$\int_E f\mathrm{d}\mu_t = \frac{1}{t}\int_0^t f\left(X_u\right)\mathrm{d}u. \tag{3.4.1}$$

If the invariant measure μ admits a Lebesgue density ρ, a competing estimator is found by constructing a kernel estimator $\widehat{\rho}_{t,h}$ of ρ and applying the plug-in rule. This approach yields the estimator

$$\int_E f(x)\widehat{\rho}_{t,h}(x)\mathrm{d}x = \int_E f(x)\left(\frac{1}{th^d}\int_0^t K\left(\frac{x-X_u}{h}\right)\mathrm{d}u\right)\mathrm{d}x, \tag{3.4.2}$$

where $K:\mathbb{R}^d\to\mathbb{R}$ is some sufficiently smooth kernel and $h=h_t$ denotes the bandwidth. It is a priori not clear whether one estimator dominates the other in some aspects or whether they show different behavior when a large class of linear functionals is considered simultaneously.

In the sequel, we study the smoothed empirical diffusion process related to (3.4.2), for any $f\in L^1(\mu)$ defined as

$$\begin{aligned}\mathbb{S}_{t,h}(f) &:= \sqrt{t}\int_E f(x)\left(\widehat{\rho}_{t,h}(x)-\rho(x)\right)\mathrm{d}x\\ &= \sqrt{t}\int_E f(x)\left(\frac{1}{t}\int_0^t K_h\left(x-X_u\right)\mathrm{d}u-\rho(x)\right)\boldsymbol{\lambda}(\mathrm{d}x)\\ &= \frac{1}{\sqrt{t}}\int_0^t\left(f*K_h\left(X_u\right)-\int_E f\mathrm{d}\mu\right)\mathrm{d}u,\end{aligned} \tag{3.4.3}$$

where $K_h(x)=h^{-d}K\left(h^{-1}x\right)$ and $\boldsymbol{\lambda}$ denotes Lebesgue measure. The mean-squared error for the estimation of a d-dimensional density at a fixed point in a Hölder ball with parameter $\beta>0$ is known to be of order $n^{-\beta/(2\beta+d)}$, n being the number of i.i.d. observations. In contrast, Corollary 1 in Dalalyan and Reiß (2007) states that the convergence rate for the mean-squared pointwise risk of an estimator $\widehat{\rho}_t$ for the invariant density ρ of a diffusion X, based on the observation $(X_u)_{0\le u\le t}$, can be upper-bounded by $t^{-1/2}(\log t)^2$ for $d=2$ and $t^{-(\beta+1)/(2\beta+d)}$ for $d\ge 3$, respectively. As will be shown subsequently, this effect gets visible when studying the smoothed version of the empirical diffusion process related to (3.4.2), based on some kernel estimator $\widehat{\rho}_{t,h}$ of the invariant density ρ.

The form of the smoothed empirical diffusion process in (3.4.3) is similar to a representation of empirical processes of scalar diffusions which is obtained by means of the occupation times formula. Indeed, let X be a one-dimensional diffusion on an open interval $I\subseteq\mathbb{R}$ with speed measure m which is assumed to be finite, $m(I)<\infty$. The diffusion thus defined is ergodic, and its invariant measure is given by $\mu:=m/m(I)$. The occupation times formula implies that one may write

$$\begin{aligned}\frac{1}{\sqrt{t}}\int_0^t\left(f(X_u)-\int_I f\mathrm{d}\mu\right)\mathrm{d}u &= \frac{1}{\sqrt{t}}\int_I l_t(x)f(x)m(\mathrm{d}x)-\sqrt{t}\int_I\frac{f(x)}{m(I)}m(\mathrm{d}x)\\ &= \sqrt{t}\int_I f(x)\left(\frac{l_t(x)}{t}-\frac{1}{m(I)}\right)m(\mathrm{d}x),\end{aligned} \tag{3.4.4}$$

where $l_t(x)$ denotes diffusion local time of X at time t in the point $x\in I$. In view of the obvious similarity of (3.4.3) and (3.4.4) and the results of Dalalyan and Reiß (2007), the hope is that one can also establish increased regularity for the smoothed empirical diffusion process in (3.4.3).

3.5. Modified empirical CLTs for Donsker classes and beyond

Assume that $\mathcal{F} \subset \mathbb{D}_{A^{-1}}$ such that any $f \in \mathcal{F}$ can be written as $f = AF$ for some $F \in \mathbb{D}_A$, and let $\mathcal{G} := A^{-1}\mathcal{F}$. Without further mentioning, $\mathcal{F}$ and $\mathcal{G}$ will be assumed to be countable classes of functions. The decomposition of the smoothed empirical process $\mathbb{S}_{t,h}$ into the centered smoothed empirical process and the (inevitable) bias term reads

$$\mathbb{S}_{t,h}(f) = \sqrt{t}\left(\frac{1}{t}\int_0^t (f * K_h)(X_u)\,\mathrm{d}u - \int (f * K_h)\,\mathrm{d}\mu\right) + \sqrt{t}\int (Ag * K_h)\mathrm{d}\mu, \tag{3.5.1}$$

where $K_h(x) := h^{-d}K(x/h)$ for some compactly supported kernel K on $\mathbb{R}^d$ with $\int K\mathrm{d}\lambda\!\!\lambda = 1$. Alternatively, one might decompose as follows,

$$\mathbb{S}_{t,h}(f) = \frac{1}{\sqrt{t}}\int_0^t (Ag * K_h - A(g * K_h))(X_u)\,\mathrm{d}u + \frac{1}{\sqrt{t}}\int_0^t A(g * K_h)(X_u)\mathrm{d}u. \tag{3.5.2}$$

The idea based on (3.5.2) is to prove asymptotic equicontinuity of the "intermediary process"

$$\mathbb{H}_{t,h}(g) := \mathbb{G}_t\left(A\left(g * K_h\right)\right) = \frac{1}{\sqrt{t}}\int_0^t A\left(g * K_h\right)(X_u)\mathrm{d}u, \tag{3.5.3}$$

and to show that, uniformly over $g \in \mathcal{G}$,

$$\mathbb{H}_{t,h}(g) = \mathbb{S}_{t,h}(Ag) + o_{\mathbf{P}}(1).$$

This approach will be investigated in detail in the next section. The process $\mathbb{H}_{t,h}$ is well-defined on any subspace of the domain $\mathbb{D}_A$ which is locally translation-invariant; cf. Proposition A.2.2 in Appendix A.2.

The "standard" decomposition in (3.5.1) allows to state sufficient conditions for weak convergence. In particular, an analogue of the theorem in van der Vaart (1994) on weak convergence of smoothed empirical processes holds whenever the carré du champ of the diffusion X satisfies Poincaré's inequality.

Proposition 3.5.1. *Let $\mathcal{F} \subset \mathbb{D}_{A^{-1}}$ be a translation-invariant countable Donsker class of measurable functions, and assume that μ satisfies* (**PI**). *Further assume that, for every t, $\mathcal{F} \subset L^1(|m_t|)$ and $\int_{\mathbb{R}^d} \|f(\cdot - y)\|_{2,\mu}\mathrm{d}|m_t|(y) < \infty$ for all $f \in \mathcal{F}$. If the following hypotheses hold,*

$$\sup_{f\in\mathcal{F}} \mathbf{E}_\mu\left(\int_E (f(\cdot + y) - f)K_{h_t}(y)\mathrm{d}y\right)^2 = \sup_{f\in\mathcal{F}} \mathbf{E}_\mu(f * K_{h_t} - f)^2 \to_{t\to\infty} 0; \tag{\textbf{H1}}$$

$$\sup_{f\in\mathcal{F}} \sqrt{t}\left|\mathbf{E}_\mu \int_E (f(\cdot + y) - f)K_{h_t}(y)\mathrm{d}y\right| = \sup_{f\in\mathcal{F}} \sqrt{t}\,|\mathbf{E}_\mu(f * K_{h_t} - f)| \to_{t\to\infty} 0, \tag{\textbf{H2}}$$

then $\mathbb{S}_{t,h}(f) \rightsquigarrow_{\ell^\infty(\mathcal{F})} \mathbb{G}$, where $\mathbb{G}$ is the centered Gaussian random map with covariance structure given by (3.2.1).

Proof. The proof is along the lines of the proof of the Theorem in van der Vaart (1994), taking into account the slight correction due to Giné and Nickl (2008) and incorporating the modifications necessary due to the divergence from the i.i.d. setting. Precisely, let $\mathcal{M}$ be a collection of signed measures of finite variation such that $\sup_{m\in\mathcal{M}} \|m\|_{\mathrm{TV}} < \infty$. By assumption, the functions $y \mapsto f(x - y)$ and $y \mapsto \|f(\cdot - y)\|_{2,\mathbf{P}}$ are in $L^1(|m|)$ for all $x \in \mathbb{R}^d$. Since $(\mathbb{D}_A, \|\cdot\|_{\mathbb{D}_A})$ is a separable Hilbert space (cf. Lemma A.2.1), one may extend Theorem 5.3 in Dudley and show that the closure in $\|\cdot\|_{\mathbb{D}_A}$ for the coarsest topology of the symmetric convex hull of any Donsker class is again Donsker; cf. the proof of Lemma 2 in Giné and Nickl (2008). It then follows from Proposition A.2.2 in the Appendix that the class $\{f * m : f \in \mathcal{F}, m \in \mathcal{M}\}$ is Donsker. The remainder parts of the proof can be taken verbatim from the proof of the Theorem in van der Vaart (1994). $\square$

Entropy conditions and conditions on the order of the kernel K may be used for verifying (**H1**) and (**H2**).

Lemma 3.5.2 (cf. Theorem 6.2 in Rohde and Strauch (2010)). *Let $\beta > d/2$, and assume that the invariant density $\rho \in C^{\beta+1}(\mathbb{R}^d)$. If $th^{2\beta} \to 0$ and, in addition,*

(a) *K is a kernel of order $\lfloor \beta \rfloor$, then*

$$\lim_{t\to\infty} \sqrt{t}\,|\mathbf{E}_\mu\,(f * K_h)| = 0;$$

(b) *the diffusion coefficient $a(x) \equiv a$ is constant, $\mathcal{G} \subset C^1(\mathbb{R}^d)$ and*

$$\left\{\delta^\alpha g : g \in \mathcal{G}, |\alpha| \le 1, \alpha \in \{0,1\}^d\right\} \subset C^{\beta+1}(\mathbb{R}^d),$$

then, for any kernel K of order $2\lfloor \beta \rfloor - 1$,

$$\lim_{t\to\infty} \mathbf{E}_\mu \sup_{g\in\mathcal{G}} \frac{1}{\sqrt{t}} \left| \int_0^t \left(A\,(g * K_h) - (Ag) * K_h\right)(X_u)\,\mathrm{d}u \right| = 0.$$

Proof. (a) Taylor expansion of $\rho(y)$ and $\rho(x)$ up to the order $\lfloor\beta\rfloor$ gives the upper bound

$$\sqrt{t}\,|\mathbf{E}_\mu(f * K_h)| = \sqrt{t}\left| \int f(y) \left(\int (\rho(x) - \rho(y)) K_h(x-y)\mathrm{d}x \right) \mathrm{d}y \right| \lesssim \sqrt{t}h^\beta = o(1).$$

(b) Denote by $P_x^{(b^i)}$ and $Q_x^{(\nabla g)^i}$ the Taylor expansion of b^i of order $\lfloor\beta\rfloor$ at $x \in \mathbb{R}^d$ and the Taylor expansion of $(\nabla g)^i$ of order $\lfloor\beta\rfloor - 1$ at $x \in \mathbb{R}^d$, respectively. Since K is of order $2\lfloor\beta\rfloor - 1$, it holds for the remainder terms $R_{x,y}^{(b^i)} := b^i(y) - P_x^{(b^i)}(y)$ and $R_{x,y}^{(\nabla g)^i} := (\nabla g)^i(y) - Q_{x,y}^{(\nabla g)^i}(y)$,

$$\begin{aligned}
&\int \left(b^i(y) - b^i(x)\right) K_h(x-y)(\nabla g)^i(y)\mathrm{d}y \\
&\quad = \int \left(P_x^{(b^i)}(y) - b^i(x) + R_{x,y}^{(b^i)}\right) K_h(x-y) \left(Q_x^{(\nabla g)^i}(y) + R_{x,y}^{(\nabla g)^i}\right) \mathrm{d}y \\
&\quad = \int R_{x,y}^{(b^i)} K_h(x-y) \left(Q_x^{(\nabla g)^i}(y) + R_{x,y}^{(\nabla g)^i}\right) \mathrm{d}y \\
&\qquad + \int \left(P_x^{(b^i)}(y) - b^i(x)\right) K_h(x-y) R_{x,y}^{(\nabla g)^i} \mathrm{d}y.
\end{aligned}$$

Hölder continuity of b and ∇g implies that

$$\sup_{x,y\in C_\varepsilon : \|x-y\|\le h} \left|R_{x,y}^{(b^i)}\right| \lesssim h^\beta, \qquad \sup_{x,y\in C_\varepsilon : \|x-y\|\le h} \left|R_{x,y}^{(\nabla g)^i}\right| \lesssim h^{\beta-1}.$$

The Taylor approximation $Q_x^{(\nabla g)^i}(y)$ depends continuously on x, y and all partial derivatives up to order $\lfloor\beta\rfloor - 1$ which are uniformly bounded in absolute value. Thus,

$$\sup_{g\in\mathcal{G}} \sup_{x,y\in\overline{C_\varepsilon}} \left|Q_x^{(\nabla g)^i}(y)\right| < \infty.$$

It remains to note that $\sup_{x,y\in C_\varepsilon:\|x-y\|\le h} \left|P_x^{(b^i)}(y) - b^i(x)\right| \lesssim h$. Summarizing, we obtain

$$\begin{aligned}
&\mathbf{E}_\mu \sup_{g\in\mathcal{G}} \frac{1}{\sqrt{t}} \left| \int_0^t \left((Ag) * K_h - A\,(g * K_h)\right)(X_u)\,\mathrm{d}u \right| \\
&\quad \le \sqrt{t} \int_E \sup_{g\in\mathcal{G}} \left| \int (b(y) - b(x))^t\, \nabla_w g(y) K_h(x-y)\mathrm{d}y \right| \mathrm{d}\mu(x) \lesssim \sqrt{t}h^\beta.
\end{aligned}$$

□

Assumption (**H1**) may be verified similarly to the i.i.d. setting. We will not deepen this question but refer to Giné and Nickl (2008) for further results on this issue.

3.6. Analyzing the smoothed empirical diffusion process

Our interest is in deriving uniform CLTs under minimal assumptions. The snag of Proposition 3.5.1 is that it only gives Donsker-type theorems for Donsker classes. It was however pointed out by Radulović and Wegkamp (2000) in the i.i.d. setting that some smoothed version of the empirical process may – in some cases – converge in law in $\ell^\infty(\mathcal{F})$ to a tight limit if $\mathcal{F}$ itself is not Donsker, under only the pregaussian hypothesis on $\mathcal{F}$.

In the classical i.i.d. framework, uniform CLTs for smoothed empirical processes indexed by pregaussian function classes were proven recently by Giné and Nickl (2008) (subsequently, [GN08], for short) and Radulović and Wegkamp (2009) [RW09]. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbf{P}$ for some law $\mathbf{P}$ on $\mathbb{R}^d$ satisfying $\mathrm{d}\mathbf{P}(x) = p_0(x)\mathrm{d}\lambda(x)$. The smoothed empirical process $\mathbb{H}_n$ and the centered smoothed empirical process $\mathbb{H}_n^0$ are defined as

$$\mathbb{H}_n(f) := \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n (f * K_{h_n})(X_i) - \int f\mathrm{d}\mathbf{P}\right),$$

and

$$\mathbb{H}_n^0(f) := \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n (f * K_{h_n})(X_i) - \int (f * K_{h_n})\mathrm{d}\mathbf{P}\right),$$

respectively, where $K_h(x) := h^{-d}K((x-\cdot)/h)$ for some kernel $K : \mathbb{R}^d \to \mathbb{R}$. Both [GN08] and [RW09] adapt Theorem 3.2 in Giné and Zinn (1984) which characterizes the impact of pregaussianness on the asymptotic equicontinuity criterion in order to prove uniform CLTs under only pregaussian conditions. Theorem 3 in [GN08] finally reduces the problem of proving asymptotic equicontinuity of the smoothed empirical process to verifying that, in outer probability,

$$\lim_{n\to\infty}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(X_i)\right\|_{\mathcal{H}'_{n^{-1/4}}} = 0,$$

where

$$\mathcal{H}_\delta := \left\{\int_{\mathbb{R}^d} (f-g)(y) K_h(y-x)\mathrm{d}y : f, g \in \mathcal{F},\ \|f-g\|_{L^2(\mathbf{P})} \le \delta\right\}$$

and $(\varepsilon_i)_{i=1}^\infty$ denotes a Rademacher sequence. This condition is still difficult to verify in general, and [GN08] suggest to use either uniform entropy bounds or bracketing entropy bounds for $\mathcal{H}_\delta$. They prove that smoothed empirical processes indexed by some specific subset of the Besov space $\mathcal{B}^s_{11}(\mathbb{R})$, $s < 1$ – which is known to be $\mathbf{P}$-pregaussian but not $\mathbf{P}$-Donsker – satisfies a uniform CLT. The proof uses the suggested bracketing entropy bound which results in the condition that $h_n n^{(2s-1)/(8s(s-1))} \to \infty$. [RW09] take a slightly different approach. Their general condition implying asymptotic equicontinuity of the centered smoothed empirical process $\mathbb{H}_n^0$ reads

$$\mathbf{E}\sup_{\mathbf{P}f^2 \le \eta/\sqrt{n}} \left|\mathbb{H}_n^0(f)\right| \to 0.$$

In the sequel, [RW09] use the decomposition by the Cauchy–Schwarz inequality,

$$\begin{aligned}
&\mathbf{E}^*\sup_{\mathbf{P}f^2 \le \eta/\sqrt{n}} \left|\sqrt{n}\int f(x)\left(\widehat{p}_n(x) - \mathbf{E}\widehat{p}_n(x)\right)\mathrm{d}x\right| \\
&\quad \le \mathbf{E}^*\sup_{\mathbf{P}f^2 \le \eta/\sqrt{n}} \left(\sqrt{n\mathbf{P}f^2}\right)\left(\int \frac{\left(\widehat{p}_n(x) - \mathbf{E}\widehat{p}_n(x)\right)^2}{p(x)}\mathrm{d}x\right)^{1/2} \\
&\quad \lesssim \left(\sqrt{n}\int \frac{\operatorname{Var}\widehat{p}_n(x)}{p(x)}\mathrm{d}x\right)^{1/2}.
\end{aligned}$$

This approach is not suitable for investigating the smoothed empirical diffusion process. Proposition 3.3.4 is analogue to Giné and Zinn (1984)'s Theorem 3.2 in that it also characterizes the impact of pregaussianness on the asymptotic equicontinuity criterion, but the constrained set $\mathcal{F}_\delta$ is defined in terms of $d_{\mathbb{G}}$, and Poincaré's inequality only gives an upper bound of the form $\|A\cdot\|_{\mu,2}$ on this metric. Moreover, the above approach needs that the variance decreases to zero slightly faster than $1/\sqrt{n}$. Our proof for the smoothed empirical diffusion process is conceptually different.

3.6.1. The empirical diffusion process indexed by smoothed functions

Throughout this section, $((X_t),\mathbf{P}_\mu)$ is assumed to be a stationary ergodic diffusion, taking values in some set $E\subset\mathbb{R}^d$ with non-void interior $E\setminus\partial E$, which satisfies the basic conditions (D1)-(D5) introduced in Section 1.1.3. Assume that $\mathcal{F}\subset\mathbb{D}_{A^{-1}}$ is pregaussian and such that $\mathcal{G}:=A^{-1}\mathcal{F}\subset W^{2,2}(\mu)$. Furthermore, suppose that $\mathcal{G}$ possesses an envelope $G\in L^1(\mu)$ with compact support $C\subset E\setminus\partial E$. If not explicitly stated otherwise, $K\in C_c^2(\mathbb{R}^d)$ denotes some kernel which admits the representation $K_{h,x}(z)=h^{-d}\widetilde{K}\left(\|x-z\|/h\right)$. Without loss of generality, we assume that the support of K is the closed d-dimensional unit ball $B_0(1)$. Subsequently, $(\mathbb{G}(f))_{f\in\mathcal{F}}$ denotes a centered Gaussian process with covariance structure as defined in (3.2.1).

Convergence of the marginals. Let $\mathcal{G}:=A^{-1}\mathcal{F}\subset\mathbb{D}_A$. Since (P_t) is a strongly continuous semigroup, we have $C_c^2\subset\mathbb{D}_A$ such that, by Proposition A.2.2, $g*K_h\in\mathbb{D}_A$ for any $g\in\mathcal{G}$ and h sufficiently small. In particular, $\mathbf{E}_\mu A\left(g*K_h\right)=\mathbf{E}_\mu Ag=0$. Dynkin's formula implies that $(M_t^g)_{t\geq 0}$ with $M_t^g:=g\left(X_t\right)-g\left(X_0\right)-\int_0^t(Ag)\left(X_u\right)\mathrm{d}u$ is a martingale, and letting

$$M_{t,h}^g:=g*K_h\left(X_t\right)-g*K_h\left(X_0\right)-\int_0^t A\left(g*K_h\right)\left(X_u\right)\mathrm{d}u,$$

the collection $\left(t^{-1/2}M_{t,h_t}^g(s)\right)_{0\leq s\leq t}$ is a triangular array of martingales.

Let $g_1,\dots,g_m$ be any finite set of functions from $\mathcal{G}$, and let $\varepsilon>0$ be fixed. Note that

$$\begin{aligned}
&\mathbf{P}_\mu\left(\max_{i=1,\dots,m}\left|\frac{1}{\sqrt{t}}M_{t,h_t}^{g_i}-\frac{1}{\sqrt{t}}M_t^{g_i}\right|>\varepsilon\right)\\
&\quad\leq m\max_{i=1,\dots,m}\mathbf{P}_\mu\left(\left|\frac{1}{\sqrt{t}}M_{t,h_t}^{g_i}-\frac{1}{\sqrt{t}}M_t^{g_i}\right|>\varepsilon\right)\\
&\quad\leq\frac{m}{\varepsilon^2}\max_{i=1,\dots,m}\mathbf{E}_\mu\left|\frac{1}{\sqrt{t}}\left(M_{t,h_t}^{g_i}-M_t^{g_i}\right)\right|^2\\
&\quad=-\frac{2m}{\varepsilon^2}\max_{i=1,\dots,m}\int\left(g_i-g_i*K_{h_t}\right)^t A\left(g_i-g_i*K_{h_t}\right)\mathrm{d}\mu \qquad (3.6.1)\\
&\quad=\frac{m}{\varepsilon^2}\max_{i=1,\dots,m}\int\nabla_w\left(g_i-g_i*K_{h_t}\right)^t(y)a(y)\nabla_w\left(g_i-g_i*K_{h_t}\right)(y)\mathrm{d}\mu(y)\\
&\quad=\frac{m}{\varepsilon^2}\max_{i=1,\dots,m}\int\left(\nabla_w g_i-(\nabla_w g_i)*K_{h_t}\right)^t(y)a(y)\left(\nabla_w g_i-(\nabla_w g_i)*K_{h_t}\right)(y)\rho(y)\mathrm{d}y\\
&\quad=o(1)
\end{aligned}$$

as $t\to\infty$, since (K_{h_t}) is a Dirac sequence and $\sup_{y\in C}\|a(y)\rho(y)\|_{S_2}$ is bounded. Here, $\|\cdot\|_{S_2}$ denotes the Frobenius norm. The expression for the variance in (3.6.1) is deduced from Bhattacharya (1982), and we then used subsequently Lemmata A.2.3 and A.2.4 from the Appendix.

Furthermore,

$$\begin{aligned}
\mathbf{E}_\mu \sup_{g\in\mathcal{G}} \left|\mathbb{H}_{t,h_t}(g) - \frac{1}{\sqrt{t}} M^g_{t,h_t}\right| &= \frac{1}{\sqrt{t}} \mathbf{E}_\mu \sup_{g\in\mathcal{G}} \left|g * K_{h_t}(X_t) - g * K_{h_t}(X_0)\right| \\
&\leq \frac{2}{\sqrt{t}} \mathbf{E}_\mu \sup_{g\in\mathcal{G}} \left(g * |K_{h_t}|(X_0)\right) \\
&\leq \frac{2}{\sqrt{t}} \int_E \int |K_{h_t}(y)|\, \mathrm{d}y \mathrm{d}\mu(z) \to_{t\to\infty} 0,
\end{aligned}$$

and therefore, as $t \to \infty$ and for $h = h_t \searrow 0$,

$$\frac{1}{\sqrt{t}} \int_0^t A(g_i * K_{h_t})(X_u)\mathrm{d}u = \frac{1}{\sqrt{t}} M^{g_i}_t + o_{\mathbf{P}}(1), \qquad i = 1, \ldots, m. \tag{3.6.2}$$

Convergence of the finite-dimensional distributions now follows from the martingale CLT for triangular arrays.

Asymptotic equicontinuity. It follows from (3.6.2) that it suffices to prove asymptotic equicontinuity of the triangular array of approximating martingales $\left(t^{-1/2} M^g_{t,h_t}\right)_{g\in\mathcal{G}}$. Define

$$\|M_{t,h_t}\|_{d_{\mathbb{G}}} := \sup_{g,g'\in\mathcal{G}: d_{\mathbb{G}}(g,g')>0} \frac{\sqrt{\langle M^g_{\cdot,h_\cdot} - M^{g'}_{\cdot,h_\cdot}\rangle_t}}{d_{\mathbb{G}}(g,g')}.$$

For any fixed $K > 0$, $\varepsilon > 0$,

$$\begin{aligned}
&\limsup_{t\to\infty} \mathbf{P}_\mu \left(\sup_{d_{\mathbb{G}}(g,g')\leq\delta} \frac{1}{\sqrt{t}} \left|M^g_{t,h_t} - M^{g'}_{t,h_t}\right| > \varepsilon \right) \\
&\quad\leq \limsup_{t\to\infty} \mathbf{P}_\mu \left(\sup_{d_{\mathbb{G}}(g,g')\leq\delta} \frac{1}{\sqrt{t}} \left|M^g_{t,h_t} - M^{g'}_{t,h_t}\right| > \varepsilon; \|M_{t,h_t}\|_{d_{\mathbb{G}}} \leq K \right) \\
&\qquad + \limsup_{t\to\infty} \mathbf{P}_\mu \left(\|M_{t,h_t}\|_{d_{\mathbb{G}}} > K \right).
\end{aligned} \tag{3.6.3}$$

For the first term, it follows from Bernstein's inequality for continuous local martingales (3.1.7) that

$$\mathbf{P}_\mu \left(\frac{1}{\sqrt{t}} \left|M^g_{t,h_t} - M^{g'}_{t,h_t}\right| > \varepsilon; \|M_{t,h_t}\|_{d_{\mathbb{G}}} \leq K \right) \leq 2\exp\left(-\frac{\varepsilon^2}{2K^2 d_{\mathbb{G}}(g,g')} \right),$$

that is, the random map

$$f \mapsto \frac{1}{\sqrt{t}} M^{A^{-1}f}_{t,h_t} \mathbb{1}\left\{ \|M_{t,h_t}\|_{d_{\mathbb{G}}} \leq K \right\}$$

is sub-Gaussian with respect to $K d_{\mathbb{G}}$. Thus, by pregaussianness, the first term on the right-hand side of (3.6.3) vanishes as $\delta \searrow 0$. It remains to be shown that the second term in (3.6.3) vanishes asymptotically. Itô's formula yields the following expression for the quadratic variation,

$$\begin{aligned}
\frac{1}{t} \left\langle M^g_{\cdot,h_\cdot} \right\rangle_t &= \frac{1}{t} \int_0^t \left((\nabla_w g) * K_{h_t}\right)^t (X_u)\, a(X_u)\, (\nabla_w g) * K_{h_t}(X_u)\, \mathrm{d}u \\
&= \frac{1}{t} \int_C \int_0^t \left((\nabla_w g)(y) K_{h_t}(y - X_u)\right)^t a(X_u)\, (\nabla_w g)(y) K_{h_t}(y - X_u)\, \mathrm{d}u\mathrm{d}y.
\end{aligned}$$

By Hölder's inequality, the last expression can be bounded from above by some constant multiple of

$$\int_C (\nabla_w g)^t(y) \left(\frac{1}{t} \int_0^t a(X_u) \left|K_{h_t}\right|(y - X_u)\mathrm{d}u \right) (\nabla_w g)(y)\mathrm{d}y.$$

Hence, using the definition of $d_{\mathbb{G}}$, in order to prove asymptotic equicontinuity, we are left with the task of showing that

$$\sup_{g\in\mathcal{G}} \frac{\int_C (\nabla_w g)^t (y) \left(\frac{1}{t}\int_0^t a(X_u) |K_{h_t}| (y - X_u)\,\mathrm{d}u\right) (\nabla_w g)(y)\mathrm{d}y}{\int_E (\nabla_w g)^t a \nabla_w g \mathrm{d}\mu} = O_{\mathbf{P}}(1).$$

The last equation is true whenever

$$\begin{aligned}
&\sup_{y\in C}\left\|\frac{1}{t}\int_0^t a(X_u)|K_{h_t}|(y-X_u)\,\mathrm{d}u\right\|_{S_2} \qquad (3.6.4)\\
&\quad\le \sup_{y\in C}\left\|\int a(z)|K_{h_t}|(y-z)\rho(z)\mathrm{d}z\right\|_{S_2}\\
&\qquad + \sup_{y\in C}\left\|\frac{1}{t}\int_0^t a(X_u)|K_{h_t}|(y-X_u)\,\mathrm{d}u - \int a(z)|K_{h_t}|(y-z)\rho(z)\mathrm{d}z\right\|_{S_2}\\
&\quad = O_{\mathbf{P}}(1).
\end{aligned}$$

Let $x_1,\dots,x_{N_{h_t}}$ be an h_t-net of C with respect to the Euclidean distance. Since C is compact, it holds $N_{h_t}\sim h_t^{-d}$. The first term on the right-hand side of (3.6.4) can be bounded from above by

$$\sup_{y\in C}\left\|\int a(z)h_t^{-d}\mathbb{1}_{B_y(2h_t)}(y-z)\rho(z)\mathrm{d}z\right\|_{S_2} \lesssim \sup_{y\in C}\|a(y)\|_{S_2}h_t^{-d},$$

and it only remains to treat the second term. This will be done by applying the bound on the variance of integral functionals of diffusion processes due to Dalalyan and Reiß (2007) which was already employed in Chapter 2. It will be used now in the following version.

> (Dalalyan and Reiß (2007), Proposition 1) Let $C\subset E\subset\mathbb{R}^d$ be bounded and assume that $\mu\le\kappa\lambda\!\!\lambda$ on C for some positive constant κ. Then, for some constant ϱ depending only on c_P, C_0, d and κ,
>
> $$\mathrm{Var}_\mu\left(\frac{1}{\sqrt{t}}\int_0^t \delta^{-d}\mathbb{1}_{B_y(\delta)}(X_u)\,\mathrm{d}u\right) \le \varrho\delta^{-2d}\lambda\!\!\lambda(B_y(\delta))^2\zeta_d^2(\lambda\!\!\lambda(B_1(\delta))),$$
>
> with
>
> $$\zeta_d(x) := \begin{cases}\max\{1,\log(1/x))^2\}, & \text{if } d=2,\\ x^{1/d-1/2}, & \text{if } d\ge 3.\end{cases}$$

We first note that

$$\zeta_d(\lambda\!\!\lambda(B_x(h_t))) \lesssim \zeta_d\left(h_t^d\right). \tag{3.6.5}$$

For fixed $i,j\in\{1,\dots,d\}$, define for $x\in C$ and $t,h>0$,

$$Z_{x,t}^{h_t} := \sqrt{t}\left(\frac{1}{t}\int_0^t a_{ij}(X_u)h_t^{-d}\mathbb{1}_{B_x(2h_t)}(X_u)\,\mathrm{d}u - \mathbf{E}_\mu a_{ij}(X_0)h_t^{-d}\mathbb{1}_{B_x(2h_t)}(X_0)\right).$$

Plugging (3.6.5) into the above variance inequality gives for any $x\in C$ and $t>0$

$$\mathrm{Var}_\mu\left(Z_{x,t}^{h_t}\right) \lesssim \sup_{z\in C}|a_{ij}(z)|^2\zeta_d^2\left(h_t^d\right).$$

With

$$\sigma_{2,h}^2(x) := \lim_{t\to\infty}\mathrm{Var}_{\mathbf{P}_\mu}\left(Z_{x,t}^h\right) \lesssim \sup_{z\in C}|a_{ij}(z)|^2\zeta_d^2\left(h^d\right)$$

and

$$c_{\infty,h} := \sup_{z\in C}\left|a_{ij}(z)h^{-d}\mathbb{1}_{B_x(2h)}(z)\right| \lesssim \sup_{z\in C}|a_{ij}(z)|\,h^{-d},$$

Lezaud's Bernstein-type inequality (3.3.2) yields the exponential tail bounds

$$\mathbf{P}_\mu\left(|Z_{x,t}^{h_t}| > u\right) \leq 2\exp\left(-\frac{u^2/2}{\sigma_{2,h_t}^2(x) + c_P c_{\infty,h_t} u/\sqrt{t}}\right) \qquad \forall u > 0.$$

By means of Pisier's maximal inequality (cf. (3.3.7)), it follows

$$\mathbf{E}_\mu \sup_{k=1,\dots,N_{h_t}} \frac{1}{\sqrt{t}}|Z_{x_k,t}^{h_t}| \lesssim \sup_{z\in C}|a_{ij}(z)|\frac{1}{\sqrt{t}}\left(\zeta_d\left(h_t^d\right) + \frac{\sqrt{\log\left(\mathrm{e}h_t^{-1}\right)}}{h_t^d\sqrt{t}}\right)\sqrt{\log\left(\mathrm{e}h_t^{-1}\right)}.$$

We have proved

Theorem 3.6.1 (Rohde and Strauch (2010), Theorem 4.3)**.** *Let $((X_t),\mathbf{P}_\mu)$ be a stationary, ergodic, strongly continuous diffusion in $E\subseteq\mathbb{R}^d$ with non-void interior $E\setminus\partial E$, satisfying the conditions* (D1)-(D5) *in Section 1.1.3. Assume that $A^{-1}\mathcal{F}$ is a countable subset of measurable functions in $W^{2,2}(\mu)$, with compact support C in the interior of E, and let $\widetilde{h}_t^{(d)} := t^{-1/d}\log(\mathrm{e}t)$. If $\mathcal{F}$ is pregaussian, then*

$$\left(\frac{1}{\sqrt{t}}\int_0^t A\left(\left(A^{-1}f\right)*K_{h_t}\right)(X_u)\mathrm{d}u\right)_{f\in\mathcal{F}} \rightsquigarrow (\mathbb{G}(f))_{f\in\mathcal{F}} \ \textit{in } \ell^\infty(\mathcal{F}),$$

provided that $h_t = h_t^{(d)} \searrow 0$ and $\widetilde{h}_t^{(d)} = O(h_t^{(d)})$.

3.6.2. Example

Theorem 3.6.1 gives uniform CLTs for Kolmogorov processes as introduced in Section 1.1.3. For any convex set $I\subset\mathbb{R}^d$, let $\mathcal{H}(\beta,L;I)$ denote the isotropic Hölder smoothness class, defined in analogy to the isotropic Hölder balls $\mathcal{H}(\beta,L)\equiv\mathcal{H}(\beta,L;\mathbb{R}^d)$ introduced in (2.2.1), that is,

$$\mathcal{H}(\beta,L;I) := \begin{cases}\left\{f: I\to\mathbb{R}: |f(x)-f(y)|\leq L\|x-y\|^\beta\right\}, & \beta\leq 1,\\ \left\{f\in C^{\lfloor\beta\rfloor}(I): \left|f(x)-P_y^{(f)}(x)\right|\leq L\|x-y\|^\beta\right\}, & \beta>1,\end{cases}$$

Theorem 3.6.2 (Rohde and Strauch (2010), Corollary 6.1)**.** *Let $V\in\mathcal{H}(\beta+1,L)$ for some $L>0$, $\beta>d/2$. Suppose $b=-\nabla V$ satisfies the at-most-linear-growth condition,* $\exp\circ(-2V)\in L^1(\mathbb{R}^d)$ *and*

$$\max_{i=1,\dots,d}\max_{\alpha:|\alpha|\leq\lfloor\beta\rfloor}\left|\partial^\alpha b_i(0)\right| \ \leq\ \gamma \tag{3.6.6}$$

for some $\gamma>0$. Assume that $\mathcal{G}=A^{-1}_{|N_A^\perp\cap\mathbb{D}_A}\mathcal{F}$ satisfies the requirement of Theorem 3.6.1, where $C\subset E$ is convex, $\mathcal{G}\subset C^1(\mathbb{R}^d)$ and

$$\left\{\partial^\alpha g:\ g\in\mathcal{G}, |\alpha|=1, \alpha\in\{0,1\}^d\right\}\subset\mathcal{H}(\beta-1,L;C)$$

for $\beta>d/2$. Let $h_t^{(d)} := t^{-1/d}\log(\mathrm{e}t)$. Then

$$(\mathbb{S}_{t,h_t}(f))_{f\in\mathcal{F}} \ \rightsquigarrow\ (\mathbb{G}(f))_{f\in\mathcal{F}} \quad \textit{in } \ell^\infty(\mathcal{F}),$$

provided that the involved kernel K is of order $2\lfloor\beta\rfloor-1$.

Proof. It follows from the arguments in Section 1.1.3 that conditions (D1)-(D5) are satisfied. In view of Theorem 3.6.1, it suffices to prove that

$$(\mathbb{H}_{t,h_t}(g))_{g\in\mathcal{G}} = (\mathbb{S}_{t,h_t}(Ag))_{g\in\mathcal{G}} + o_{\mathbf{P}}(1).$$

The assumption that $V \in \mathcal{H}(\beta+1, L)$ and (3.6.6) together imply that ρ is Hölder continuous of order $\beta+1$ on every bounded set $D \subset \mathbb{R}^d$. Lemma 3.5.2 then gives $\sqrt{t}\,|\mathbf{E}_\mu\,(f * K_{h_t})| \to_{t\to\infty} 0$ and

$$\lim_{t\to\infty} \mathbf{E}_\mu \sup_{g\in\mathcal{G}} \frac{1}{\sqrt{t}} \left| \int_0^t \left(A\,(g * K_h) - (Ag) * K_h\right)(X_u)\,\mathrm{d}u \right| = 0.$$

□

4. Extensions, concluding remarks and outlook

Two classical questions concerning nonparametric estimation theory and weak convergence properties of diffusion processes in $\mathbb{R}^d$ were investigated in the present work. We identified classes of ergodic diffusion processes for which results similar profound as in the classical i.i.d. framework or for scalar diffusions can be obtained. To conclude, we give a selective overview of results obtained in the previous two chapters. Furthermore, we sketch possible extensions.

4.1. Sharp adaptive estimation for ergodic diffusion processes

The problem of estimating the drift of a multidimensional diffusion process is somewhat complex. A common approach to estimate the drift of *scalar* ergodic diffusions is to exploit the representation of the drift via the invariant density of the diffusion and its derivative. We opted for a similar strategy which in particular led us to restricting attention to ergodic diffusion process solutions of the SDE $\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma\ \mathrm{d}W_t$ whose invariant density ρ satisfies the relation

$$2b^j\rho = \operatorname{div}(a_j\rho) = \sum_{k=1}^d \partial_k\left(a_{jk}\rho\right), \qquad j \in \{1,\ldots,d\}.$$

Under reasonable assumptions, the invariant density ρ can be estimated faster than the divergences $\operatorname{div}(a_j\rho)$. The main part of our investigation of the drift estimation problem thus focuses on analyzing the performance of kernel estimators of the divergences $\operatorname{div}(a_j\rho)$. Sharp results for estimating the components of the vector $\operatorname{div}(a\rho)$ are obtained under general assumptions such as ergodicity, stationarity, and the condition that the carré du champs of the diffusion satisfies Poincaré's inequality. At a later stage, conditions on the growth of the drift coefficient and on its radial behavior (which in particular imply exponential bounds on the invariant density of the associated diffusion process) allow to transfer the results on estimating the components of $\operatorname{div}(a\rho)$ to the original drift estimation problem.

From a practical point of view, the results on exact asymptotics allow to distinguish between competing estimators. The drift estimators studied in the literature typically attain the minimax optimal rate but, at least for moderate sample size T, the influence of the constant may be substantial. The results also deliver insight into the matter of the problem of drift estimation. Recall that $\mathbb{S}_d = 2\pi^{d/2}/\Gamma(d/2)$ denotes the surface of the unit sphere in $\mathbb{R}^d$. Theorem 2.3.3 and Theorem 2.3.10 state that the asymptotically exact constant for the (adequately weighted) global L^2 risk for estimating the j-th drift component over (weighted) L^2 Sobolev classes with integer smoothness index $\beta > 1$ and radius $L > 0$ is given by

$$\frac{(L(2\beta+d))^{\frac{d}{2\beta+d}}}{d}\left(\frac{a_{jj}\beta\ \mathbb{S}_d}{(2\pi)^d(\beta+d)}\right)^{\frac{2\beta}{2\beta+d}}.$$

As compared to the optimal constant for density estimation from i.i.d. observations which is developed in Rigollet and Tsybakov (2007), the constant also depends on the j-th diagonal entry $a_{jj} = \sum_{k=1}^d \sigma_{jk}^2$ of the diffusion matrix $a = \sigma\sigma^t$.

For pointwise adaptive estimation of the j-th drift component of ergodic diffusions with nondegenerate diffusion coefficient over L^2 Sobolev classes of smoothness $\beta > d/2$ with radius $L > 0$, the asymptotically exact constant for the squared pointwise risk is identified as

$$\frac{L^{\frac{d}{2\beta}}}{\rho(x_0)}\,\frac{2\beta}{d}\left(\frac{d^2 a_{jj}\rho(x_0)}{\beta(2\beta-d)}\right)^{\frac{\beta-d/2}{2\beta}}\quad\mathbb{I}_\beta,$$

where

$$\mathbb{I}_\beta := \sqrt{\frac{1}{2\beta}B\left(1+\frac{d}{2\beta},1-\frac{d}{2\beta}\right)\,\frac{\mathbb{S}_d}{(2\pi)^d}}.$$

As in the white noise setting, the constant involves the value of an optimal recovery problem. Additionally, the value of the invariant density $\rho(\cdot)$ at the point $x_0 \in \mathbb{R}^d$ and the j-th diagonal entry of the diffusion matrix appear. Taking into account the shift in the smoothness assumption, the dependence on the diffusion matrix appears to be of the same form as for estimation with respect to the integrated L^2 risk.

The results on exact drift estimation hold for integer Sobolev smoothness $\beta > 1$ (global setting) and (possibly noninteger) Sobolev smoothness $\beta > d/2$ (pointwise case). This finding admits different interpretations. On the one hand, it suggests that asymptotic equivalence of the drift estimation problem for ergodic Kolmogorov diffusions may hold for smaller regularity than considered in Dalalyan and Reiß (2007). The authors do not provide a counterexample to show that asymptotic equivalence fails below the critical regularity they identified (which is of order $(1/2 + 1/\sqrt{2})d$ for $d \geq 2$). Rather they describe the question whether for (Hölder) smoothness classes of smaller regularity asymptotic equivalence fails as a "challenging open problem." The accordance of the exact asymptotic behavior for substantially smaller smoothness indices as obtained in Dalalyan and Reiß (2007) can however also be seen as an indication of the limits of the asymptotic equivalence approach. This last interpretation was suggested in Korostelev and Nussbaum (1999) who study the exact constant of the risk asymptotics in the uniform norm for density estimation from i.i.d. observations. They consider estimation over Hölder classes of arbitrary smoothness index $\beta > 0$ on the unit interval, and they explicitly prove upper and lower asymptotic risk bounds. It is noted that a more concise argument based on asymptotic equivalence of experiments in the Le Cam sense works only in the case $\beta > 1/2$, and under the additional assumption that the densities are uniformly bounded away from 0. Asymptotic equivalence between density estimation and continuous regression is known to fail for $\beta \leq 1/2$, but the minimax risk asymptotics hold for any $\beta > 0$ in the density model and for continuous regression. Korostelev and Nussbaum (1999) establish from this gap the need to deepen the study of *reduced* equivalence.

The precise form of the *risk function* for global drift estimation differs from those commonly considered in the framework of density or regression function estimation from independent observations. Instead of investigating the classical mean-integrated squared error of an estimator $\widetilde{f}$ of some density f, say, and considering

$$\mathbf{E}_f \int_{\mathbb{R}^d} \left|\widetilde{f}(z) - f(z)\right|^2 \mathrm{d}z,$$

we analyze for global estimation of the drift $b : \mathbb{R}^d \to \mathbb{R}^d$ the *weighted* L^2 risk

$$\mathbf{E}_b \int_{\mathbb{R}^d} \left\|\widetilde{b}_T(z) - b(z)\right\|^2 \rho^2(z)\mathrm{d}z.$$

The weighting by the squared invariant density ρ^2 ensures integrability of the drift vector b. This allows to describe the exact asymptotics for drift estimation over $\mathbb{R}^d$, imposing merely growth conditions on the drift function. In particular, we do not restrict attention to estimation on compact sets or under boundedness assumptions on the drift.

The problem of drift estimation for ergodic non-degenerate diffusions, assuming isotropic Sobolev smoothness of the drift function, is classical in some sense. Results on exact adaptive drift estimation with respect to local minimax weighted L^2 risk criteria and on sharp adaptive pointwise estimation mark the beginning of a deeper nonparametric estimation theory for multidimensional ergodic diffusion processes. Of natural interest are the questions of

- extending the multidimensional estimation procedures to anisotropic situations;
- adaptation to unknown smoothness/ unknown dimension under structural assumptions;
- investigating exact asymptotics for drift estimation of discretely observed diffusions;
- identifying exact asymptotics for more general stochastic processes (degenerate diffusions, processes with jump),

to mention just a few. In the sequel, we shall briefly comment on some of these issues.

Alternative smoothness assumptions and complementary lower bound results. It was explicitly noted that the approach of component-wise estimation of $\operatorname{div}(a\rho)$ is to be preferred to estimating the complete vector $\operatorname{div}(a\rho)$ in one step (which is more convenient from the practical point of view). Notably, the component-wise access also allows to consider estimation under inhomogeneous regularity assumptions on the components of the drift vector. Indeed, the upper bounds in Theorem 2.3.8(B) and Theorem 2.5.11 assume isotropic homogeneous Sobolev smoothness of $\operatorname{div}(a\rho)$, namely

$$\sum_{k=1}^d \sum_{|\alpha|\leq\beta} \int_{\mathbb{R}^d} \left|D^\alpha \left(\operatorname{div}(a_k\rho) - \operatorname{div}(a_k\rho_0)\right)\right|^2 \leq 4L\ d, \qquad \beta\in\mathbb{N},\ L>0,$$

and

$$\sum_{k=1}^d (2\pi)^{-d} \int_{\mathbb{R}^d} \|\lambda\|^{2\beta} \left|\phi_{\operatorname{div}(a_k\rho)}(\lambda)\right|^2 \mathrm{d}\lambda \leq 4L^2\ d, \qquad \beta>d/2,\ L>0. \tag{4.1.1}$$

The above smoothness conditions may prove arbitrarily rough, and thus it appears more appropriate to estimate the vector component by component, replacing, e.g., the Sobolev constraint (4.1.1) with the assumption that, for some values $\beta_i > d/2$ and $L_i > 0$,

$$\eta_{\beta_i}(\operatorname{div}(a_i\rho)) = \left((2\pi)^{-d}\int_{\mathbb{R}^d}\|\lambda\|^{2\beta_i}\left|\phi_{\operatorname{div}(a_i\rho)}(\lambda)\right|^2\mathrm{d}\lambda\right)^{1/2} \leq 2L_i, \qquad i=1,\ldots,d. \tag{4.1.2}$$

An interesting open question is to establish lower bounds for estimating the vector $\operatorname{div}(a\rho)$ under some component-wise smoothness constraint in the spirit of (4.1.2), assuming that the drift vector b and the invariant density ρ satisfy **for any** $j\in\{1,\ldots,d\}$ the relation

$$2b^j\rho = \operatorname{div}(a_j\rho) = \sum_{k=1}^d \partial_k\left(a_{jk}\rho\right). \tag{4.1.3}$$

Both the minimax lower bound in Theorem 2.3.2 (global case) and the lower bound in Theorem 2.5.5 (pointwise setting) consider the case where the relation (4.1.3) holds for **one fixed** $j\in\{1,\ldots,d\}$. Inspection of the respective proofs shows that the hypotheses constructed therein indeed satisfy (4.1.3) for precisely one $j\in\{1,\ldots,d\}$. We conjecture however that suitable modifications of the hypotheses defined in the proofs of Theorem 2.3.2 and Theorem 2.5.5 allow to extend the results to this situation. Precisely, for the case of pointwise estimation and any $i\in\{1,\ldots,d\}$, define the sets $\Pi^i(\cdot)$ and $\mathcal{B}_T$ according to (2.5.5) and (2.5.9), respectively. Given any values $\beta_i>d/2$, $L_i>0$, $i=1,\ldots,d$, denote

$$C_i(\beta_i,L_i;\rho,\sigma) := 2L_i^{d/(2\beta)}\frac{2\beta}{d}\left(\frac{d^2\ a_{ii}\ \rho(x_0)}{\beta_i(2\beta_i-d)}\right)^{\frac{\beta_i-d/2}{2\beta_i}}\ \mathbb{I}_{\beta_i},$$

and

$$\widetilde{\Pi}_i(\beta_i, L_i) = \widetilde{\Pi}_i(\beta_i, L_i; \rho, \sigma) := \Big\{ b \in \bigcap_{k=1}^{d} \Pi^k\big(\min_{j=1,\dots,d} \beta_j \big) : \eta_{\beta_i}(\operatorname{div}(a_i\rho)) \le 2L_i \Big\}.$$

We conjecture that the vector of data-driven estimators $\widetilde{g}_T^1, \dots, \widetilde{g}_T^d$ defined according to (2.5.22) is sharp adaptive and satisfies, for any $x_0 \in \mathbb{R}^d$,

$$\limsup_{T\to\infty} \sum_{i=1}^{d} \sup_{(\beta_i,L_i)\in\mathcal{B}_T} \sup_{b\in\widetilde{\Pi}^i(\beta_i,L_i)} \big(\psi_{T,\beta}\ C_i(\beta_i, L_i; \rho, \sigma)\big)^{-2}\ \mathbf{E}_b \big|\widetilde{g}_T^i(x_0) - \operatorname{div}(a_i\rho)(x_0)\big|^2 \le 1$$

and

$$\liminf_{T\to\infty} \sum_{i=1}^{d} \inf_{\widehat{g}_T^i} \sup_{(\beta_i,L_i)\in\mathcal{B}_T} \sup_{b\in\widetilde{\Pi}^i(\beta_i,L_i)} \big(\psi_{T,\beta}\ C_i(\beta_i, L_i; \rho, \sigma)\big)^{-2}\ \mathbf{E}_b \big|\widetilde{g}_T^i(x_0) - \operatorname{div}(a_i\rho)(x_0)\big|^2 \ge 1,$$

where $\inf_{\widehat{g}_T^i}$ denotes the infimum over all estimators of $\operatorname{div}(a_i\rho)$.

One further step would be to assume *anisotropic smoothness* of the components of $\operatorname{div}(a\rho)$ and to work under the condition that, for some values $\beta_{i1}, \dots, \beta_{id} > d/2$ and $L_i > 0$,

$$\left((2\pi)^{-d} \int_{\mathbb{R}^d} \sum_{k=1}^{d} \big|\lambda^k\big|^{2\beta_{ik}} \big|\phi_{\operatorname{div}(a_i\rho)}(\lambda)\big|^2 \mathrm{d}\lambda \right)^{1/2} \le L_i, \qquad i = 1, \dots, d.$$

The problem of nonparametric estimation becomes rather delicate when functions with anisotropic regularities are considered. Even results on minimax rates of convergence in the classical statistical models are relatively scarce; cf. the references on pp. 137–138 in Kerkyacharian et al. (2001). For the moment, we merely note that the generalization of the Lepski-type procedure as used in Section 2.5 to an anisotropic setting is a non-trivial task due to the lack of natural ordering of the bandwidths.

The degenerate case. It was assumed throughout that the diffusion matrix $a = \sigma\sigma^t$ is uniformly elliptic. The sharp results on adaptive estimation over isotropic Sobolev classes in Chapter 2 reveal that the geometry of non-degenerate diffusion matrices does not change the rate of convergence but is only reflected in the constant. (Of course, the geometry and smoothness properties of σ affect the rate of convergence in an indirect manner by determining the smoothness properties of the invariant density.) No attempts have been made to derive lower bounds under Sobolev smoothness assumptions for the degenerate case. The reason for this is that we are interested in formulating results on exact drift estimation under smoothness conditions on the drift, and in the degenerate case it seems appropriate - instead of assuming that the invariant density belongs to some finite smoothness class such as those of Hölder, Sobolev, or Besov type - to focus on estimating the drift of diffusion processes under the basic assumption that the parabolic Hörmander condition is satisfied. For such processes the invariant density ρ belongs to $C^\infty(\mathbb{R}^d)$.

The exact adaptive results on pointwise estimation of functions $f \in C^\infty(\mathbb{R})$, observed in Gaussian white noise of small intensity ε, in Lepski and Levit (1998) may serve as a motivation for developing exact estimation procedures in this framework. To quantify the difficulty of estimating $C^\infty(\mathbb{R})$ functions, the authors introduce a scale of functional classes $\mathcal{A}_K$ which represents a large portion of $C^\infty(\mathbb{R})$, in terms of the rate of decrease of the corresponding Fourier transforms. An essential feature of their model is that the rate of convergence in estimating $f(x)$, for $\varepsilon \to 0$, may vary, over the whole scale $\mathcal{A}_K$, from extremely fast to extremely slow, up to situations where no consistent estimation is possible. Their follow-up paper Lepski and Levit (1999) provides a multivariate generalization and studies adaptive pointwise estimation of smooth functions $f : \mathbb{R}^d \to \mathbb{R}$, corrupted by white Gaussian noise. The functional classes considered in the multivariate setting are narrower than in Lepski and Levit (1998) in that the assumption of a rapidly vanishing Fourier transform ϕ_f is replaced by an exponential decline of ϕ_f. The L^p norm oracle inequalities obtained in this work allow to derive minimax results on (anisotropic versions of) such functional spaces but do not lead to sharp adaptive findings.

Estimation in structural models. Little surprisingly, the result on nonparametric estimation of the drift under Sobolev smoothness assumptions is subject to the curse of dimensionality. This is reflected in the fact that the accuracy of estimators, even for moderate dimension d, is very poor unless for unreasonably large sample size T. If the d-dimensional drift function b however has a simple structure such that it is effectively m-dimensional with $m < d$, the dimensionality reduction principle due to Stone asserts that the quality of estimation should correspond to the effective dimensionality m.

One important feature for nonparametric estimation of Itô processes is the fact that regularity assumptions enter the scene just in order to ensure existence of a (strong) solution of the associated SDE. This observation may affect the estimation philosophy. For example, there exist cases where it is more appropriate to *adapt to the unknown effective dimension* instead of adapting to the unknown smoothness. Samarov and Tsybakov (2007) consider the problem of nonparametric density estimation from i.i.d. observations in a modification of the projection pursuit density estimation model. Precisely, they assume that the unknown density p has the form

$$p(x) = \varphi_d(x)\ g(B^t x), \qquad x \in \mathbb{R}^d, \tag{4.1.4}$$

where B in an unknown $d \times m$ matrix with orthonormal columns, $g : \mathbb{R}^m \to [0, \infty)$ is an unknown function such that, for all $z \in \mathbb{R}^m$, $\max\{g(z), |\nabla g(z)|, \|\nabla^2 g(z)\|\} \le L$, and $\varphi_d(\cdot)$ is the standard d-variate normal density. The authors propose an aggregate estimator which adapts to the unknown index space of unknown dimension m. For the suggested aggregate estimator, they show that the MISE is of order $n^{-4/(m+4)}$. In the context of drift estimation, it is reasonable to assume minimal Hölder smoothness of order $\beta = 2$. Using an adaptation scheme which adapts to the unknown smoothness, the MISE rate $T^{-2\beta/(2\beta+d)}$ can be attained. In comparison to the rate $T^{-4/(4+m)}$ which is attainable for $\beta = 2$ and adaptation to the effective dimension m, it thus appears more appropriate to adapt to m whenever the drift vector has intrinsic dimension $m < 2d/\beta$. It is therefore of interest to develop alternative adaptive procedures *under reasonable structural assumptions* in the spirit of (4.1.4) which adapt to the unknown effective dimension of the drift vector.

A central question in this connection is to identify "reasonable structural assumptions" on the drift coefficient. In search of a general structural model which does not lead to inadequate modeling but still allows to weaken the curse of dimensionality, Goldenshluger and Lepski (2009) advocate the use of an *additive multi-index model* which includes as special cases the more classical models of single-index, additive, projection pursuit and multi-index structure. They study the problem of minimax adaptive estimation of an unknown function $f : \mathbb{R}^d \to \mathbb{R}$ in the multidimensional Gaussian white noise model

$$\mathrm{d}Y(t) = f(t)\mathrm{d}t + \varepsilon\ \mathrm{d}W(t), \qquad t = (t_1, \ldots, t_d) \in \mathcal{D},$$

for an open interval $\mathcal{D} \supset [-1/2, 1/2]^d$, the standard Brownian sheet W in $\mathbb{R}^d$ and noise level $\varepsilon \in (0, 1)$. L^p-norm oracle inequalities which allow to derive minimax adaptive results are established. The proposed estimation procedure is based on the same rule for selection of estimators from a large parametrized collection as in Goldenshluger and Lepski (2008). In the latter article, a pointwise adaptive estimation procedure in the single index model is proposed.

The selection rule presents one possible blueprint for an adaptive estimation scheme of the drift function in structural models.

Density estimation and study of plug-in properties. The question of estimating the invariant density of a diffusion was only addressed on the side. Proposition 1 in Dalalyan and Reiß (2007) shows that the invariant density of ergodic diffusions in certain cases can be estimated considerably faster with respect to pointwise squared error criteria than in the classical statistical models based on independent observations. Their result is based on Assumption (**SG**+) and involves in particular the spectral gap assumption. To the best of our knowledge, *exact results* on invariant density estimation for diffusions satisfying specific functional inequalities do not exist.

One possibility to *extend the L^2 framework* in Dalalyan and Reiß (2007) is to quantify accuracy of an invariant density estimator $\widehat{\rho}_T$ in terms of the L^s-risk,

$$\mathcal{R}_s(\widehat{\rho}_T,\rho) := \left[\mathbf{E}_\mu\Big(\int_{\mathbb{R}^d}|\widehat{\rho}_T(x)-\rho(x)|^s\Big)\right]^{1/s}, \qquad s\in[1,\infty).$$

Generalizations of Poincaré's inequality are available and might serve as tools for bounding the stochastic error. In particular, Theorem 11 in Roberto and Zegarlinski (2007) states conditions which ensure that the Orlicz–Sobolev inequalities introduced there imply an exponential decay to the associated semigroup. This is a generalization of the well-known equivalence of Poincaré's inequality and spectral gap inequality.

In connection with invariant density estimation, it is of interest to know whether there exist cases where an invariant density estimator $\widehat{\rho}_T$ achieves the optimal MISE and L^1 error, and, at the same time, satisfies

$$\sqrt{T}\sup_{f\in\mathcal{F}}\left|\int f(x)\widehat{\rho}_T(x)\mathrm{d}x - \int f(x)\rho(x)\mathrm{d}x\right| = O_{\mathbf{P}}(1)$$

for certain classes of functions $\mathcal{F}\subset L^1(\mu)$. This question is related to the work of Bickel and Ritov (2003) who introduce the notion of the "plug-in property" (PIP) and consider different perspectives thereof. Given i.i.d. observations $X_1,X_2,\ldots,X_n,\ X_i\sim\mathbb{P}_\theta,\ \theta\in\Theta$, a subset of a linear space of functions, assume that r_n^{-1} is the stochastic minimax rate for estimating $\theta\in\Theta$. An estimator $\widehat{\theta}_n$ of θ is called a *weak plug-in estimator* (PIE) for a set $\mathcal{T}$ of functionals if

$$\sup_{\theta\in\Theta}\sup_{\mathcal{T}}\mathbb{E}_\theta\Big(r_n^{-2}\ \|\widehat{\theta}_n-\theta\|^2 + n\big(\tau(\widehat{\theta})-\tau(\theta)\big)^2\Big) < \infty.$$

It is shown in Bickel and Ritov (2003) for the Gaussian white noise model that the concept of weak PIE and minimaxity in this sense are not compatible for very large $\mathcal{T}$ and nonparametric Θ. For other sets such as indicators of all quadrants (that is, distribution functions), appropriate constructions of efficient estimators exist. The study of the PIP of various types of density estimators has been continued by Nickl (2007) (for the nonparametric MLE), Giné and Nickl (2008) (kernel density estimators) or Giné and Nickl (2009) (wavelet density estimators), to mention just a few. All of these results are obtained in the classical i.i.d. framework. It is indicated to accomplish a systematic study of the invariant density estimation problem for ergodic diffusions with an investigation of the PIP. We shall come back to this question in Section 4.2 below.

4.2. Uniform CLTs for multidimensional diffusions

Infinite-dimensional extensions of the classical (functional) CLT for ergodic continuous Markov processes due to Bhattacharya (1982) and under the Kipnis–Varadhan condition were studied in the second part of the work. We found different explicit criteria for establishing uniform CLTs for multidimensional ergodic diffusions satisfying functional inequalities such as the Poincaré hypothesis. In principle, there exist different approaches to deal with empirical diffusion processes.

With a view towards statistical applications, it is often convenient to exploit mixing properties of the diffusion such that results from the well-investigated setting of empirical processes based on independent observations are available. Various *parallels* between classical empirical processes and empirical processes of ergodic diffusions which satisfy Poincaré's inequality were already depicted. Results in this spirit can also be deduced from the mixing properties of the diffusion. Proposition 3.4 in Cattiaux and Guillin (2008) states that, whenever the semigroup P_t satisfies

$$\mathrm{Var}_\mu\left(P_t f\right) \leq \eta(t)\left\|f - \int f \mathrm{d}\mu\right\|_\infty^2,$$

the stationary process is strongly mixing with mixing coefficient $\alpha(t) \leq \sqrt{\eta(t)}$. We found it convenient to use Poincaré's inequality rather than working with mixing coefficients as this allows to use fine results from Markovian semigroup theory. One such example is the sharp Bernstein-type deviation inequality (3.3.2) due to Lezaud (2001). For recent results on Bernstein-type deviation inequalities under mixing conditions, we refer to Merlevède et al. (2009), Merlevède et al. (2011) and references therein.

From the theoretical point of view, it is of interest to scrape out *differences* in the behavior of empirical processes based on diffusions and independent observations, respectively. One motivation for our investigation indeed was the Gaussian characterization of the Donsker property of scalar regular diffusions with finite speed measure due to van der Vaart and van Zanten (2005). Their result reflects regularity properties specific to scalar diffusions, and their proof is heavily based on local time techniques. We took an alternative approach. In order to formulate uniform CLTs under minimal assumptions, we studied smoothed versions of the empirical diffusion process. The idea was to establish exceptional regularity of the modified empirical diffusion process by stating Donsker theorems under pregaussian conditions and minimal assumptions on the bandwidth involved in the smoothed empirical diffusion process

$$\mathbb{S}_{t,h}(f) = \sqrt{t}\left(\int f(x)\left(\widehat{\rho}_{t,h}(x) - \rho(x)\right)\mathrm{d}x\right). \tag{4.2.1}$$

One such result is given in Theorem 3.6.2, and the proof of the preceding Theorem 3.6.1 differs from the classical i.i.d. case. In particular, we do not use explicitly closeness of $\widehat{\rho}_{t,h}$ to the Lebesgue density ρ of the invariant measure in a mean squared sense. The arguments instead are based on martingale approximation and theory of Markovian semigroups, carried out for some auxiliary intermediary process. The analysis requires the combination of different types of tail estimates for additive functionals under the Poincaré hypothesis as well as sharp variance bounds of Dalalyan and Reiß (2007). Notably, the bandwidths h involved are too small for ensuring consistency of the kernel estimator $\widehat{\rho}_{t,h}$ for the invariant density ρ.

We turn to sketching some potential further developments.

Characterization of uniform Donsker classes. In statistical applications, the underlying probability measure ν usually is unknown, and therefore it is of interest to know which classes are Glivenko–Cantelli or Donsker for every ν. A function class $\mathcal{F}$ which is Donsker for any probability measure ν is called *universal Donsker*, and for many universal Donsker classes, the CLT holds even uniformly in ν. There exist different formal definitions of so-called *uniform Donsker classes.* We follow Giné and Zinn (1991). Let

$$\mathrm{BL}_1^{\mathcal{F}} = \mathrm{BL}_1(\ell^\infty(\mathcal{F})) := \left\{g : \ell^\infty(\mathcal{F}) \to \mathbb{R},\ \|g\|_\infty \leq 1,\ \sup_{x,y\in\ell^\infty(\mathcal{F})} \frac{|g(x)-g(y)|}{\|x-y\|_{\mathcal{F}}} \leq 1\right\}.$$

For measures $\mathbb{P}, \mathbb{Q}$, defined on sub-sigma algebras of the Borel sets of $\ell^\infty(\mathcal{F})$, let

$$d_{\mathrm{BL}_1}(\mathbb{P}, \mathbb{Q}) := \sup_{g \in \mathrm{BL}_1^{\mathcal{F}}} \left| \int g \ \mathrm{d}(\mathbb{P} - \mathbb{Q}) \right|.$$

It is well-known that d_{BL_1} metrizes weak convergence in $\mathbb{R}^d$, and it can be shown as in the classical setting that $\mathcal{F}$ is Donsker (in the sense of Definition 3.2.3) if and only if both $\mathcal{F}$ is pregaussian (in the sense of Definition 3.2.4) and

$$\lim_{t \to \infty} d_{\mathrm{BL}_1} \left(\mathscr{L}_{\mu, \mathcal{F}}(\mathbb{G}_t), \mathscr{L}_{\mathcal{F}}(\mathbb{G}_\mu) \right) = 0,$$

where $\mathscr{L}_{\mu,\mathcal{F}}(\mathbb{G}_t)$ denotes the law of the empirical process $\mathbb{G}_t$ indexed by $\mathcal{F}$ under $\mathbf{P}_\mu$. Similarly, $\mathscr{L}_{\mathcal{F}}$ is the law of the limiting Gaussian process $\mathbb{G}_\mu$ indexed by $\mathcal{F}$. As argued by Giné and Zinn (1991), uniformity in all probability measures is useful only in combination with uniformity of $\mathbb{G}_\mu$, and therefore one considers *uniformly pregaussian* classes of functions. We adopt the terminology from the classical framework.

Definition 4.2.1. Let $(E, \mathscr{B}(E))$ be a measurable space, let $\mathscr{P}(E)$ be the set of all probability measures on $(E, \mathscr{B}(E))$, and let $\mathcal{F}$ be a collection of real-valued measurable functions on E.

- $\mathcal{F}$ is called *uniformly pregaussian* if both $\sup_{\mu \in \mathscr{P}(E)} \mathbf{E} \|\mathbb{G}_\mu\|_{\mathcal{F}} < \infty$ and

$$\lim_{\delta \to 0} \sup_{\mu \in \mathscr{P}(E)} \mathbf{E} \sup_{f,g \in \mathcal{F}:\ d_{\mathbb{G}_\mu}(f,g) < \delta} |\mathbb{G}_\mu(f - g)| = 0, \tag{4.2.2}$$

  where $\mathbb{G}_\mu$ is a centered Gaussian process indexed by $\mathcal{F}$ with covariance structure determined by the finite-dimensional CLT (cf. Section 3.2.2), and $d_{\mathbb{G}_\mu}$ is the intrinsic metric on $\mathcal{F}$ induced by $\mathbb{G}_\mu$.

- $\mathcal{F}$ is called a *uniform Donsker class* if both $\mathcal{F}$ is uniformly pregaussian and

$$\lim_{t \to \infty} \sup_{\mu \in \mathscr{P}(E)} d_{\mathrm{BL}_1}(\mathscr{L}_{\mu,\mathcal{F}}(\mathbb{G}_t), \mathscr{L}_{\mathcal{F}}(\mathbb{G}_\mu)) = 0.$$

Theorem 2.3 in Giné and Zinn (1991) gives a sharp characterization of the uniform Donsker property for classical empirical processes based on independent observations and states that $\mathcal{F}$ is uniform Donsker if and only if $\mathcal{F}$ is uniformly pregaussian. The proof of their Theorem comprises several steps and uses symmetrization at crucial points, but we conjecture that the connection between uniform pregaussianness and the uniform Donsker property still holds in the diffusion framework. It remains to work out minimal assumptions on the diffusion processes which are required to establish this relation.

Verifying pregaussianness. Seen individually, results on Donsker properties under pregaussian conditions in the spirit of Theorem 3.6.1 are of certain interest from a theoretical point of view, but they still leave one with the task of verifying that the function class $\mathcal{F}$ is pregaussian. Let us consider the approach to uniform CLTs via solution of the Poisson equation. In this case, the pseudo-metric induced by the limiting Gaussian process $\mathbb{G}$ appearing in the CLT is given by

$$d_{\mathbb{G}}^2(f,g) = -2 \int (f-g) \left[A^{-1}(f-g) \right] \mathrm{d}\mu, \qquad f, g \in \mathcal{F} \subset \mathbb{D}_{A^{-1}}.$$

Cauchy–Schwarz and Poincaré's inequality imply that

$$-2 \left\langle f, A^{-1} f \right\rangle_\mu \leq 2 \|f\|_{\mu,2} \|A^{-1} f\|_{\mu,2} \lesssim \|f\|_{\mu,2}^2. \tag{4.2.3}$$

The above straightforward estimate was already used for establishing the bracketing result in Section 3.3.1. The idea of bracketing is easily proved and understood, with the obstacle that the condition is nowhere close to being necessary. By contrast, the entropy estimates considered first by Vapnik and Chervonenkis are known to be essential in determining whether a class of functions is a (uniform) Glivenko–Cantelli or Donsker class (in the classical sense). Under the Poincaré hypothesis, it is straightforward to state a condition for (uniform) pregaussianness in terms of the combinatorial dimension. This approach again makes (implicit) use of the mixing properties of the diffusion. More interesting results are to be expected when the form of the limiting Gaussian process $\mathbb{G}$ appearing in the CLT can be exploited more directly, but this issue is open for further investigation.

Study of plug-in properties (ctd.). Our focus in the study of the smoothed empirical diffusion process $\mathbb{S}_{t,h}$ (cf. (4.2.1)) was on stating minimal assumptions on the bandwidth h in order to scrape out remarkable regularity properties of (the modified version of) the empirical diffusion process. When investigating the problem of invariant density estimation for multidimensional ergodic diffusions (assuming, e.g., the availability of certain functional inequalities) and with a view towards statistical applications, it is certainly also of interest to address the PIP, a question which is different in spirit. The statistically most relevant version of the PIP is the so-called efficient PIP. Adapted to our framework and using the usual quadratic loss function notions, a formal definition might be stated as follows.

Definition 4.2.2 (cf. Definition 4.1 in Bickel and Ritov (2003))**.** Let $\left\|\widetilde{\rho}_T-\rho\right\|^2 = O_{\mathbf{P}}(r_T^2)$, where r_T is the (stochastic) minimax rate of convergence for estimating $\rho \in \mathcal{P}$. For each $\tau \in \mathcal{T}$, $\mathcal{T}$ a set of functionals, let $\widetilde{\tau}_T$ be an efficient estimator of τ (in the sense of Definition 5.2.7 in Bickel et al. (1998)). An estimator $\widehat{\rho}_T$ is called an *efficient PIE* if

$$\left\|\widehat{\rho}_T-\rho\right\|^2 = O_{\mathbf{P}}(r_T^2)$$

and

$$\sqrt{T}\sup_{\tau\in\mathcal{T}}\left|\tau(\widehat{\rho}_T)-\widetilde{\tau}\right| = o_{\mathbf{P}}(1).$$

As pointed out by Giné and Nickl (2008), the proof of the PIP consists of two parts, namely the use of the right uniform CLT for the variance term, and, more severely, the treatment of the bias term. While Theorem 3.6.1 allows for very fast bandwidths $h_t \searrow 0$ with $O(h_t) = t^{-1/d}\log(\mathrm{e}t)$, Lemma 3.5.2 which is used for proving that, uniformly in $g \in \mathbb{G}$,

$$\left|\mathbb{S}_{t,h}(Ag)-\mathbb{H}_{t,h}(g)\right| = \frac{1}{\sqrt{t}}\left|\int_0^t \left(A\left(g*K_h\right)-(Ag)*K_h\right)(X_u)\,\mathrm{d}u\right| = o_{\mathbf{P}}(1),$$

requires that $th^{2\beta}\to 0$. In particular, this excludes the use of the choice $h\sim t^{-1/(2\beta+d)}$ which gives the optimal rate of the MSE (cf. Corollary 1 in Dalalyan and Reiß (2007)). In the classical i.i.d. framework, Giné and Nickl (2008) suggest to combine the information on the underlying density and on the function class. Similar ideas may be suitable for obtaining sharper bounds on $\mathbb{S}_{t,h}(Ag)-\mathbb{H}_{t,h}(g)$, but this still has to be investigated in more detail. Let us finally remark that explicit regularity conditions on the local characteristics of the SDE associated to the diffusion might be formulated for bounding the approximation error.

Uniform CLTs for ergodic diffusions observed along Euler schemes. In order to contribute to the theory of empirical processes of multivariate ergodic diffusions, we studied the empirical diffusion process $\mathbb{G}_t$ and the smoothed empirical diffusion process $\mathbb{S}_{t,h}$, based on the empirical measure and some weighted empirical measure related to some kernel density estimator, respectively. The investigation can be extended in a different direction by considering (weighted) empirical measures based on approximation schemes for SDEs.

This problem is interesting both from the theoretical and the practical point of view. Consider the SDE

$$\mathrm{d}X_t = b\left(X_t\right)\mathrm{d}t + \sigma\left(X_t\right)\ \mathrm{d}W_t, \tag{4.2.4}$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$, $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times q}$ are continuous mappings, W is a q-dimensional standard Brownian motion, and assume that the diffusion process solution X of (4.2.4) is stationary and ergodic with invariant measure μ. In typical applications, the coefficients b are σ are given by some physical model, and the interest is, e.g., in computing $\int f \mathrm{d}\mu$ for some given $f \in L^1(\mu)$. To do so, one may use numerical approximation schemes or, alternatively, PDE methods for solving the stationary Fokker–Planck equation. It is however known (cf. Pagès and Panloup (2009)) that probabilistic methods are superior in terms of efficiency in higher-dimensional settings (dimension $d \geq 3$) or in degenerate cases.

The problem of approximation of the invariant distribution for Brownian diffusions by means of an adapted Euler scheme with decreasing step $\left(\gamma_k\right)_k$ has been addressed, e.g., by Lamberton and Pagès (2002) and Pagès and Panloup (2009). Lamberton and Pagès (2002) suggest to compute the Euler discretization of (4.2.4), given as

$$X_{n+1} = X_n + \gamma_{n+1} b\left(X_n\right) + \sqrt{\gamma_{n+1}}\sigma\left(X_n\right)U_{n+1}, \qquad n \geq 0,$$

where $\left(U_n\right)_{n\geq 1}$ is an $\mathbb{R}^q$-valued white noise defined on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$, independent of $X_0 \in L^0(\Omega, \mathscr{A}, \mathbb{P})$, and $\gamma := \left(\gamma_n\right)_{n\geq 1}$ is a step sequence vanishing to 0. In a second step, a weighted empirical measure ν_n^η is formed, using a weight sequence $\eta := \left(\eta_n\right)_{n\geq 0}$,

$$\nu_n^\eta(\omega, \mathrm{d}x) := \frac{1}{\sum_{k=1}^n \eta_k} \sum_{k=1}^n \eta_k \delta_{X_{k-1}}(\omega), \qquad \omega \in \Omega,\ n \geq 1.$$

Under certain Lyapunov-type stability assumptions and under conditions on the step sequence γ, it is shown in Theorem 9 in Lamberton and Pagès (2002) that, for every C^2 function F with D^2F bounded and Lipschitz, the following CLT holds

$$\sqrt{\sum_{k=1}^n \gamma_n}\ \nu_n^\gamma(AF) \Rightarrow Z \sim \mathcal{N}(0, \sigma_F^2), \tag{4.2.5}$$

where $\sigma_F^2 := \int_{\mathbb{R}^d} \left|\sigma^t \nabla F(x)\right|^2 \mu(\mathrm{d}x)$. The result in (4.2.5) is similar to the CLT in (3.1.4) which was a starting point for our uniform CLTs in Chapter 3. More recently, Pagès and Panloup (2012) obtained results in the spirit of (3.1.4) for the case of functionals of the path process and its associated Euler scheme with decreasing step. With a view toward applications, it is certainly of interest to establish uniform CLTs for ergodic diffusions observed along Euler schemes and for functionals of the entire path process. Finite-dimensional results as in (4.2.5) may provide an origin for the investigation.

A. Supplementary results

A.1. Auxiliary results from nonparametric statistics and statistics of random processes

This section summarizes different auxiliary results which were used in the proof of the upper and lower bounds presented in Chapter 2. We start with stating results on the question of absolute continuity of probability measures induced by different diffusion processes. Recall that we had to deal with this question for proving lower bounds in the diffusion model (see, e.g., Lemma 2.6.2 in Section 2.6.1). For the proof of the following result and many more related material, we refer to Chapter 7 of Liptser and Shiryaev (2001).

Theorem A.1.1 (Liptser and Shiryaev (2001), Section 7.6.4, pp. 296–297)**.** *Let $b_1, b_2 : \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times r}$, $d, r \in \mathbb{N}$, be such that the SDEs*

$$\mathrm{d}X_t = b_0(X_t)\mathrm{d}t + \sigma(X_t)\ \mathrm{d}W_t, \qquad X_0 = x, \tag{A.1.1}$$

$$\mathrm{d}Y_t = b_1(Y_t)\mathrm{d}t\ + \sigma(Y_t)\ \mathrm{d}W_t, \qquad Y_0 = y, \tag{A.1.2}$$

admit a unique strong solution. Assume further that the system of equations

$$\sigma(x)\alpha(x) = b_1(x) - b_2(x)$$

has a solution for any $x \in \mathbb{R}^d$, and denote by a^- the pseudo-inverse of the matrix $a = \sigma\sigma^t$. Consider the laws induced by the solutions of (A.1.1) and (A.1.2), and denote their restrictions to $(C[0,T], \mathscr{B}(C[0,T]))$ by $\mathbb{P}^{(T)}_{x,b_0}$ and $\mathbb{P}^{(T)}_{x,b_1}$, respectively. If $x = y$ and

$$\mathbf{P}\bigg(\int_0^T \Big(b_1^t(X_u)a^-(X_u)b_1(X_u) + b_2^t(X_u)a^-(X_u)b_2(X_u)\Big)\,\mathrm{d}u < +\infty\bigg) = 1,$$

then $\mathbb{P}^{(T)}_{x,b_1} \ll \mathbb{P}^{(T)}_{x,b_2}$, and the Radon–Nikodym density of $\mathbb{P}^{(T)}_{x,b_1}$ with respect to $\mathbb{P}^{(T)}_{x,b_2}$ is given by

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}^{(T)}_{x,b_1}}{\mathrm{d}\mathbb{P}^{(T)}_{x,b_2}}(Y^T) = \exp\bigg(&\int_0^T (b_1 - b_2)^t(Y_u)a^-(Y_u)\mathrm{d}Y_u \\ &- \frac{1}{2}\int_0^T (b_1 - b_2)^t(Y_u)a^-(Y_u)(b_1 + b_2)(Y_u)\mathrm{d}u\bigg).\end{aligned}$$

If, in addition,

$$\mathbf{P}\bigg(\int_0^T \Big(b_1^t(Y_u)a^-(Y_u)b_1(Y_u) + b_2^t(Y_u)a^-(Y_u)b_2(Y_u)\Big)\,\mathrm{d}u < +\infty\bigg) = 1,$$

then $\mathbb{P}^{(T)}_{x,b_1} \sim \mathbb{P}^{(T)}_{x,b_2}$ and

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}^{(T)}_{x,b_2}}{\mathrm{d}\mathbb{P}^{(T)}_{x,b_1}}(X^T) = \exp\bigg(&\int_0^T (b_1 - b_2)^t(X_u)a^-(X_u)\mathrm{d}X_u \\ &+ \frac{1}{2}\int_0^T (b_1 - b_2)^t(X_u)a^-(X_u)(b_1 + b_2)(X_u)\mathrm{d}u\bigg).\end{aligned}$$

Remark A.1.2. Assume now that the diffusion process solutions X and Y of (A.1.1) and (A.1.2), respectively, are ergodic and admit invariant measures with Lebesgue density $\rho_1 = \rho_{\sigma,b_1}$ and $\rho_2 = \rho_{\sigma,b_2}$, respectively. Suppose that the initial values x and y follow the respective invariant measure such that the processes X and Y are strictly stationary. Consider the laws induced by the stationary solutions of (A.1.1) and (A.1.2), and denote their restrictions to $(C[0,T], \mathscr{B}(C[0,T]))$ by $\mathbb{P}_{b_1}^{(T)}$ and $\mathbb{P}_{b_2}^{(T)}$, respectively. It then follows from Theorem A.1.1 that

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}_{b_1}^{(T)}}{\mathrm{d}\mathbb{P}_{b_2}^{(T)}}(Y^T) &= \frac{\rho_1}{\rho_2}(Y_0)\ \exp\Bigg(\int_0^T (b_1-b_2)^t(Y_u)a^-(Y_u)\mathrm{d}Y_u \\ &\qquad\qquad -\frac{1}{2}\int_0^T (b_1-b_2)^t(Y_u)a^-(Y_u)(b_1+b_2)(Y_u)\mathrm{d}u\Bigg) \\ &= \frac{\rho_1}{\rho_2}(Y_0)\ \exp\Bigg(\int_0^T (b_1-b_2)^t(Y_u)\sigma^-(Y_u)\mathrm{d}W_u \\ &\qquad\qquad -\frac{1}{2}\int_0^T (b_1-b_2)^t(Y_u)a^-(Y_u)(b_1+b_2)(Y_u)\mathrm{d}u\Bigg).\end{aligned} \tag{A.1.3}$$

For proving the lower bound for pointwise estimation of the components of the divergence vector (precisely, for verifying the conditions of Lemma 2.6.2 in Section 2.6.1), we also require the following result.

Lemma A.1.3 (martingale CLT; Theorem VIII.3.11 in Jacod and Shiryaev (2002))**.** *Assume that X is a continuous Gaussian martingale with characteristics $(0, C, 0)$, and that each X^n is a local continuous martingale. If D is a dense subset of $\mathbb{R}_+$, then it holds $X_n \Rightarrow X$ if and only if*

$$\langle X^{n,i}, X^{n,j}\rangle_t \xrightarrow{\mathbf{P}} C_{ij}(t) \qquad \textit{for all } t \in D.$$

The next lemma proves useful in deriving upper bounds on the L^2 risk of pointwise kernel estimators. It is a modification of Theorem 1A in Parzen (1962) who contributes the result to Bochner.

Lemma A.1.4 (Bochner; Parzen (1962))**.** *Let $K_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \ldots, m$, $m \in \mathbb{N}$, be kernel functions which satisfy the following conditions,*

$$\sup_{x\in\mathbb{R}^d} |K_i(x)| < \infty, \qquad \int_{\mathbb{R}^d} K_i^2(x)\mathrm{d}x < \infty, \qquad \lim_{\|x\|\to\infty} \left\{\|x\|K_i^2(x)\right\} = 0,$$

and assume that $g : \mathbb{R}^d \to \mathbb{R}$ satisfies $\int_{\mathbb{R}^d} |g(x)|\mathrm{d}x < \infty$. For $h_T \searrow 0$ and

$$g_T(x) := \frac{1}{h_T^d}\int_{\mathbb{R}^d}\sum_{i=1}^m K_i^2\left(\frac{y-x}{h_T}\right) g(y)\mathrm{d}y,$$

it holds at every continuity point x_0 of $g(\cdot)$,

$$\lim_{T\to\infty} g_T(x_0) = g(x_0)\int_{\mathbb{R}^d}\sum_{i=1}^m K_i^2(y)\mathrm{d}y. \tag{A.1.4}$$

Proof. The proof is brief and along the lines of the proof of Theorem 1A in Parzen (1962). By definition, it holds

$$g_T(x) - g(x)\int_{\mathbb{R}^d}\sum_{i=1}^m K_i^2(y)\mathrm{d}y = \int_{\mathbb{R}^d}\sum_{i=1}^m h_T^{-d}\ K_i^2\left(\frac{u}{h_T}\right)\ (g(u+x)-g(x))\mathrm{d}u.$$

Consequently,

$$
\begin{aligned}
&|g_T(x_0) - g(x_0) \int_{\mathbb{R}^d} \sum_{i=1}^m K_i^2(y)\mathrm{d}y| \\
&\quad \le \max_{\|u\|\le\delta} |g(u+x_0) - g(x_0)| \int_{\|y\|\le \delta h_T^d} \sum_{i=1}^m K_i^2(y)\mathrm{d}y \\
&\qquad + \int_{\|u\|>\delta} \frac{|g(u+x_0)|}{\|u\|} \frac{\|u\|}{h_T^d} \sum_{i=1}^m K_i^2\left(\frac{u}{h_T}\right)\mathrm{d}u + |g(x_0)| \int_{\|u\|>\delta} h_T^{-d} \sum_{i=1}^m K_i^2\left(\frac{u}{h_T}\right)\mathrm{d}u \\
&\quad \le \max_{\|u\|\le\delta} |g(u+x_0) - g(x_0)| \int_{\mathbb{R}^d} \sum_{i=1}^m K_i^2(y)\mathrm{d}y \\
&\qquad + \frac{1}{\delta} \sup_{\|u\|>\delta h_T^{-d}} \left\{ \|u\| \sum_{i=1}^m K_i^2(u) \right\} \int_{\mathbb{R}^d} |g(u)|\mathrm{d}u + |g(x_0)| \int_{\|y\|>\delta h_T^{-d}} \sum_{i=1}^m K_i^2(y)\mathrm{d}y.
\end{aligned}
$$

Letting $T \to \infty$ and then $\delta \to 0$, the assertion follows. □

A.2. Auxiliary results for Chapter 3

All of the following results are taken from the preprint Rohde and Strauch (2010).

Lemma A.2.1 (Rohde and Strauch (2010), Lemma 4.2)**.** *Let A satisfy the Poincaré inequality. Then $(\mathbb{D}_A, \|\cdot\|_{\mathbb{D}_A})$ is a separable Hilbert space.*

Proof. It is clear that $(\mathbb{D}_A, \|\cdot\|_{\mathbb{D}_A})$ is pre-Hilbert. In order to prove completeness, let $(g_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $(\mathbb{D}_A, \|\cdot\|_{\mathbb{D}_A})$. Then $(g_n)_{n\in\mathbb{N}}$ and $(Ag_n)_{n\in\mathbb{N}}$ are Cauchy sequences with respect to $\|\cdot\|_{L^2(\mu)}$. Completeness of $L^2(\mu)$ implies that there exist some g such that $\|g - g_n\|_{L^2(\mu)} \to 0$ and some G such that $\|G - Ag_n\|_{L^2(\mu)} \to 0$. Since A is closed, it follows $G = Ag$, and, in particular, $g \in \mathbb{D}_A$. It remains to prove separability. Note that $\mathscr{R}_A \subset L^2(\mu)$ is separable as a subset of a separable metric space. Let $(f_n)_{n\in\mathbb{N}}$ be a dense subset of $\mathscr{R}_A$, and let $(g_n)_{n\in\mathbb{N}}$ be a dense subset in $\mathbb{D}_A \cap N_A$, where N_A denotes the null-space of A which is a closed subset of $L^2(\mu)$, since A is closed. For any set $S \subset \mathbb{D}_A$, let $A_{|S}$ denote the restriction of A to S. Then the set

$$
\left(A^{-1}_{|N_A^\perp \cap \mathbb{D}_A}(f_n)\right)_{n\in\mathbb{N}} \bigcup (g_n)_{n\in\mathbb{N}}
$$

is countably dense in $(\mathbb{D}_A, \|\cdot\|_{\mathbb{D}_A})$. For the proof, let $g \in \mathbb{D}_A$ be arbitrary. Such g can be written as $g = g^\perp + g_0$ for some $g^\perp \in \mathbb{D}_A \cap N_A^\perp$ and some $g_0 \in N_A$. It holds $\|Ag^\perp\|_{L^2(\mu)} = \|Ag\|_{L^2(\mu)}$. Now let

$$
g_k^\perp \subset \left\{A^{-1}_{|N_A^\perp \cap \mathbb{D}_A}(f_n) : n \in \mathbb{N}\right\}
$$

with $\|A(g_k^\perp - g^\perp)\|_{L^2(\mu)} \to 0$. Poincaré's inequality then gives $\|g_k^\perp - g^\perp\|_{L^2(\mu)} \to 0$. Furthermore, let $g_k^0 \subset \{g_n : n \in \mathbb{N}\}$ such that $\|g_k^0 - g_0\|_{L^2(\mu)} \to 0$. Thus, $\left(g_k^\perp + g_k^0\right)$ is the desired approximation. □

Proposition A.2.2 (Rohde and Strauch (2010), Proposition 4.1)**.** *Assume that $g(\cdot + uh) \in \mathbb{D}_A$ for $h \in [0, h_0]$, $u \in \mathcal{S}^{d-1}$, such that $\|g(\cdot + uh)\|_{\mathbb{D}_A}$ is uniformly bounded in $h \le h_0$, $u \in \mathcal{S}^{d-1}$. Then $g * K_h \in \mathbb{D}_A$ for $h \in [0, h_0]$, and the convolution is contained in the $\|\cdot\|_{\mathbb{D}_A}$-closure of $\|K\|_{\mathrm{TV}}$ times the symmetric convex hull of $\{g(\cdot - y) : \|y\|_2 \le h_0\}$.*

Proof. By Lemma A.2.1, $(\mathbb{D}_A, \|\cdot\|_{\mathbb{D}_A})$ is a separable Hilbert space. This implies in particular that $\mathbb{D}_A$ is closed under finite convex combinations. The arguments given in the proof of Lemma 1 in Giné and Nickl (2008) show that $g * K_h$ is contained in the closure in $\|\cdot\|_{\mathbb{D}_A}$ of $\|K\|_{\mathrm{TV}}$ times the symmetric convex hull of $\left\{g(\cdot - y) : y \in \mathbb{R}^d\right\}$. Since $(\mathbb{D}_A, \|\cdot\|_{\mathbb{D}_A})$ is complete, the above closure in $\|\cdot\|_{\mathbb{D}_A}$ is again contained in $\mathbb{D}_A$. □

Lemma A.2.3 (cf. the proof of Theorem 4.3 in Rohde and Strauch (2010))**.** *For all $g \in W^{2,2}(\mu)$ of compact support in $E \setminus \partial E$, it holds*

$$-2\int gAg\mathrm{d}\mu = \int (\nabla_w g)^t\, a(\cdot)\nabla_w g\mathrm{d}\mu. \tag{A.2.1}$$

Proof. Let $g \in W^{2,2}(\mu)$ of compact support in the interior of E. We regularize g by convolution in a standard way. Denote by $\phi_h(\cdot) := h^{-d}\phi(\cdot/h)$ a Dirac sequence, where $\phi \in C_c^\infty(\mathbb{R}^d)$. Thus, for sufficiently small h, $g*\phi_h \in C_c^2(\mathbb{R}^d)$, and (A.2.1) holds for $g*\phi_h$. Since the diffusion coefficient a is locally bounded, the coordinate of a are uniformly bounded on compacts. Thus,

$$\begin{aligned}
&\int \left(\nabla_w \left(g - g * \phi_h\right)\right)^t a \left(\nabla_w \left(g - g * \phi_h\right)\right) \mathrm{d}x \\
&\quad = \int \left(\nabla_w g - (\nabla_w g) * \phi_h\right)^t (x) a(x) \left(\nabla_w g - (\nabla_w g) * \phi_h\right)(x)\mathrm{d}x \\
&\quad = o(1) \text{ as } h \searrow 0,
\end{aligned}$$

since ϕ_h is a Dirac sequence. □

Lemma A.2.4 (Rohde and Strauch (2010), Lemma A.1)**.** *Assume that $g \in W^{1,2}(\mu)$ and K_h is symmetric. Then it holds μ-a.s.*

$$\partial_w^\alpha \left(g * K_h\right) = \left(\partial_w^\alpha g\right) * K_h \quad \textit{for all } \alpha \in \{0,1\}^d \textit{ with } |\alpha| = 1.$$

Proof. By definition of the weak derivative, it holds for all $\phi \in \mathcal{C}_K^\infty$,

$$\begin{aligned}
\int \partial^\alpha (g * K_h)(x)\phi(x)\mathrm{d}\lambda\!\!\lambda(x) &= -\int \left(g * K_h\right)(x)\partial^\alpha\phi(x)\mathrm{d}\lambda\!\!\lambda(x) \\
&= -\iint g(y)K_h(x-y)\mathrm{d}y\ \partial^\alpha\phi(x)\mathrm{d}\lambda\!\!\lambda(x) \\
&= -\iint \partial^\alpha\phi(x)K_h(x-y)\mathrm{d}x\ g(y)\mathrm{d}\lambda\!\!\lambda(y) \\
&= -\int \left(\partial^\alpha\phi * K_h\right)(y)g(y)\mathrm{d}\lambda\!\!\lambda(y) \\
&= -\int \partial^\alpha \left(\phi * K_h\right)(y)g(y)\mathrm{d}\lambda\!\!\lambda(y) \qquad \text{(A.2.2)} \\
&= \int \partial_w^\alpha g(y) \left(\phi * K_h\right)(y)\mathrm{d}\lambda\!\!\lambda(y) \\
&= \int (\partial_w^\alpha g) * K_h(x)\phi(x)\mathrm{d}\lambda\!\!\lambda(x),
\end{aligned}$$

where the identity $\partial^\alpha(\phi * K_h) = (\partial^\alpha\phi) * K_h$ in (A.2.2) is proved, for instance, in Lemma 5 a) in Giné and Nickl (2008). □

Lemma A.2.5 (Rohde and Strauch (2010), Lemma A.2)**.** *Let A be the generator of an Itô–Feller diffusion in $E \subseteq \mathbb{R}^d$ with continuous drift and diffusion coefficient, b and σ, respectively. Then any $g \in W^{2,2}(\mu)$ of compact support in $E \setminus \partial E$ belongs to $\mathbb{D}_A$, and*

$$Ag(\cdot) \;=\; \frac{1}{2}\sum_{i,j=1}^d a_{ij}(\cdot)\frac{\partial_w^2 g}{\partial_w x_i \partial_w x_j}(\cdot) + \sum_{i=1}^d b_i(\cdot)\frac{\partial_w g}{\partial_w x_i}(\cdot).$$

Proof. Let $g \in W^{2,2}(\mu)$ be of compact support in $E \setminus \partial E$. Let $\phi_h(\cdot) = h^{-d}\phi(\cdot/h)$ be a Dirac sequence with some twice continuously differentiable, symmetric kernel ϕ. Then $g * \phi_h$ is twice continuously differentiable and of compact support, hence $g * \phi_h \in \mathbb{D}_A$ for sufficiently small h, and it holds by Lemma A.2.4 that

$$\partial_w^\alpha (g * \phi_h) \;=\; (\partial_w^\alpha g) * \phi_h \quad \text{for all multi-indices } \alpha \in \{0,1,2\}^d \text{ with } |\alpha| \leq 2.$$

Therefore,

$$A(g * \phi_h)(x) \;=\; \frac{1}{2}\sum_{i,j=1}^d a_{ij}(x)\left(\frac{\partial_w^2 g}{\partial_w x_i \partial_w x_j} * \phi_h\right)(x) + \sum_{i=1}^d b_i(x)\left(\frac{\partial_w g}{\partial_w x_i} * \phi_h\right)(x).$$

Since ϕ_h defines a Dirac sequence, $\|g - g * \phi_h\|_{L^2(\mu)} \to 0$ as $h \searrow 0$ (cf. Theorem 8.14 in Folland (1999)). Let

$$G(x) \;:=\; \frac{1}{2}\sum_{i,j=1}^d a_{ij}(x)\frac{\partial_w^2 g}{\partial_w x_i \partial_w x_j}(x) + \sum_{i=1}^d b_i(x)\frac{\partial_w g}{\partial_w x_i}(x).$$

Denote the union of the supports of g and $g * \phi_h$ by C_h. Then

$$\begin{aligned}
\|A(g * \phi_h) - G\|_{L^2(\mu)} \;&\leq\; \frac{1}{2}\sum_{i,j=1}^d \|a_{ij}\mathbb{1}_{C_h}\|_{\sup}\left\|\frac{\partial_w^2 g}{\partial_w x_i \partial_w x_j} * \phi_h \;-\; \frac{\partial_w^2 g}{\partial_w x_i \partial_w x_j}\right\|_{L^2(\mu)} \\
&\quad + \sum_{i=1}^d \|b_i \mathbb{1}_{C_h}\|_{\sup}\left\|\frac{\partial_w g}{\partial_w x_i} * \phi_h \;-\; \frac{\partial_w g}{\partial_w x_i}\right\|_{L^2(\mu)} \\
&\longrightarrow\; 0 \text{ as } h \searrow 0,
\end{aligned}$$

because (ϕ_h) is a Dirac sequence, $g \in W^{2,2}(\mu)$ and $a(\cdot)$ and $b(\cdot)$ are continuous, hence uniformly bounded on a decreasing sequence of compacts. But this implies $G = Ag$, since A is closed. $\square$

A.3. Continuity and boundedness of Gaussian processes

This section contains a brief summary of results on continuity and boundedness of Gaussian processes. Further references may be found, e.g., in Chapter 2 of Dudley (1999) or Chapter 6 of Marcus and Rosen (2006) The importance of this question is based on the fact that the existence of a bounded and $d_{\mathbb{G}}$-uniformly continuous version of the Gaussian limit $\mathbb{G}$ is necessary for $\mathcal{F}$ to be Donsker. Throughout the sequel, denote by $(\nu_t)_{t\in T}$ a centered Gaussian process, and endow T with the L^2 metric given by the covariance structure of the process, denoted by d. Two stochastic processes X_t and Y_t, defined on the same probability space with the same index set T, are said to be *modifications* of each other if, for each t, $X_t = Y_t$ with probability 1. The well-known *metric entropy criterion* due to Dudley states that there exists a modification ν' of ν with bounded, uniformly continuous sample paths on (T, d) if

$$\int_0^{\operatorname{diam}(T)} \sqrt{\log N(\varepsilon, T, d)}\mathrm{d}\varepsilon < \infty. \tag{A.3.1}$$

The entropy condition in (A.3.1) is only necessary for boundedness and continuity of Gaussian processes with stationary increments; it is not necessary in general. Sharp characterizations of centered Gaussian processes having sample-bounded modifications may be expressed via majorizing measures and the γ_α functionals.

Definition A.3.1 (Dudley (1999), p. 59)**.** Let (T, d) be a metric space and $\mathscr{P}(T)$ the set of all Borel probability measures on T. For $m \in \mathscr{P}(T)$, let

$$\gamma_m(T) := \sup_{x \in T} \int_0^\infty \sqrt{\log \frac{1}{m\left(B_x(r)\right)}} \mathrm{d}r,$$

where $B_x(r) := \{y : d(x, y) < r\}$ is the open ball of center x and radius r. If $\gamma_m(T) < \infty$, then m is called a *majorizing measure* for (T, d). Let $\gamma(T) := \gamma(T, d) := \inf\{\gamma_m(T) : m \in \mathscr{P}(T)\}$.

It then holds $\gamma(T, d) < \infty$ if and only if there exists a majorizing measure on (T, d). Corollary 2.7.30 in Dudley (1999) gives a sufficient condition for the GC property of sets $C \subset H$, H a Hilbert space. Namely, if there exists a probability measure μ on C such that

$$\lim_{\varepsilon \searrow 0} \sup_{x \in C} \int_0^\varepsilon \sqrt{\log \frac{1}{\mu\left(B_x(r)\right)}} \mathrm{d}r = 0,$$

then C is a GC-set in H.

The generic chaining method due to Talagrand gives an alternative characterization which is conceptually less difficult than the majorizing measures approach. In addition, it is closer related to the facts based on metric entropy numbers.

Definition A.3.2 (Definitions 1.2.3 and 1.2.5 in Talagrand (2005))**.** For a metric space (T, d), an *admissible sequence* of T is an increasing sequence $(\mathcal{A}_n)$ of partitions of T such that, for every $n \geq 1$, $\operatorname{card}\mathcal{A}_n \leq 2^{2^n}$ and $\operatorname{card}\mathcal{A}_0 = 1$. Denote by $A_n(t)$ the unique element of $\mathcal{A}_n$ that contains t. For $\alpha > 0$, the γ_α *functional* is defined as

$$\gamma_\alpha(T, d) := \inf \sup_{t \in T} \sum_{n \geq 0} 2^{n/\alpha} d\left(t, A_n\right),$$

where the infimum is taken with respect to all admissible sequences of T and $d\left(t, A_n\right) := \inf_{u \in A_n} d(t, u)$.

A very concise summary of relevant properties and their importance is given in Section 2 in Mendelson (2011). The following result is crucial.

Theorem A.3.3 (Theorem 1.3 in Talagrand (2005))**.** *Under measurability conditions, the following are equivalent:*

1. *The map $t \mapsto \nu_t(\omega)$ is uniformly continuous on T with probability* 1.
2. $\lim_{\delta \to 0} \mathbf{E} \sup_{d(u,t) \leq \delta} |\nu_u - \nu_t| = 0$.
3. *There exists an admissible sequence of T such that*

$$\lim_{s_0 \to \infty} \sup_{t \in T} \sum_{s = s_0}^{\infty} 2^{s/2} d(t, \pi_s(t)) = 0.$$

The above is an analogue of Theorem 1.3 in Talagrand (2005) which is only formulated but not proven there. As Mendelson (2011) notes, the continuity theorem from Talagrand (1987b) can be converted to obtain Theorem 1.3. This proof is carried out on pp. 31–35 in Kuelbs et al. (2013).

Bibliography

Arcones, M. A. and Giné, E. (1993). Limit Theorems for U-Processes. *Ann. Probab.*, 21(3):1494–1542.

Bakry, D. (2008). Functional inequalities for Markov semigroups. `http://hal.archives-ouvertes.fr/docs/00/35/37/24/PDF/tata.pdf`.

Bakry, D., Cattiaux, P., and Guillin, A. (2008). Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré. *J. Funct. Anal.*, 245(3):727–759.

Banon, G. (1978). Nonparametric identification for diffusion processes. *SIAM J. Control Optim.*, 16(3):380–395.

Belitser, E. N. and Levit, B. Y. (1995). On minimax filtering over ellipsoids. *Math. Methods Statist.*, 4(3):259–273.

Bhattacharya, R. N. (1982). On the functional central limit theorem and the law of the iterated logarithm for Markov processes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 60:185–201.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer Series in Statistics. Springer, New York.

Bickel, P. J. and Ritov, Y. (2003). Nonparametric estimators which can be "plugged-in". *Ann. Statist.*, 31(4):1033–1053.

Bovier, A. (2012). Markov processes. Lecture Notes, Universität Bonn. `http://wt.iam.uni-bonn.de/fileadmin/WT/Inhalt/people/Anton_Bovier/lecture-notes/markov.pdf`.

Brown, L. D. and Low, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.*, 24(6):2524–2535.

Brown, L. D. and Zhang, C.-H. (1998). Asymptotic nonequivalence of nonparametric experiments when the smoothness index is $1/2$. *Ann. Statist.*, 26(1):279–287.

Butucea, C. (2001). Exact adaptive pointwise estimation on Sobolev classes of densities. *ESAIM: Prob. Statist.*, 5:1–31.

Cattiaux, P., Chafaï, D., and Guillin, A. (2012). Central limit theorems for additive functionals of ergodic Markov diffusions processes. *ALEA, Lat. Am. J. Probab. Math. Stat.*, 9(2):337–382.

Cattiaux, P. and Guillin, A. (2008). Deviation bounds for additive functionals of Markov processes. *ESAIM: Prob. Statist.*, 12:12–29.

Cavalier, L., Golubev, G. K., Picard, D., and Tsybakov, A. B. (2002). Oracle inequalities for inverse problems. *Ann. Statist.*, 30(3):843–874.

Dalalyan, A. (2005). Sharp adaptive estimation of the drift function for ergodic diffusions. *Ann. Statist.*, 33(6):2507–2528.

Dalalyan, A. and Kutoyants, Y. (2002). Asymptotically efficient trend coefficient estimation for ergodic diffusion. *Math. Methods Statist.*, 11:402–427.

Dalalyan, A. and Kutoyants, Y. (2003). Asymptotically efficient estimation of the derivative of the invariant density. *Statist. Inference Stochast. Process.*, 6(1):89–107.

Dalalyan, A. and Reiß, M. (2007). Asymptotic statistical equivalence for ergodic diffusions: the multidimensional case. *Probab. Theory Relat. Fields*, 137(1):25–47.

Donoho, D. (1992). Asymptotic minimax risk for sup-norm loss: Solution via optimal recovery. *Probab. Theory Relat. Fields*, 99(2):145–170.

Donoho, D. L. and Low, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.*, 20(2):944–970.

Doob, J. L. (2001). *Classical Potential Theory and Its Probabilistic Counterpart.* Classics in Mathematics. Springer-Verlag, Berlin and Heidelberg. Reprint of the 1984 ed.

Dudley, R. M. (1999). *Uniform Central Limit Theorems.* Number 63 in Cambridge Studies in advanced mathematics. Cambridge University Press, Cambridge, UK.

Efroimovich, S. Y. and Pinsker, M. S. (1984). A self-training algorithm for nonparametric filtering. *Avtomat. i Telemekh.*, 11:58–65.

Evans, L. C. (2010). *Partial Differential Equations.* American Math. Soc., Providence, RI, 2. edition.

Folland, G. B. (1999). *Real Analysis.* John Wiley & Sons, Inc., New York, 2nd edition.

Gao, F., Guillin, A., and Wu, L. (2010). Bernstein type's concentration inequalities for symmetric Markov processes. Preprint. `arxiv:1002.2163v2`.

Gill, R. D. and Levit, B. Y. (1995). Applications of the van Trees inequality: a Bayesian Cramér–Rao bound. *Bernoulli*, 1(1–2):59–79.

Giné, E. (2007). Empirical processes and some of their applications. `http://www.math.uconn.edu/~gine/empproc-rev.pdf`.

Giné, E. and Nickl, R. (2008). Uniform central limit theorems for kernel density estimators. *Probab. Theory Relat. Fields*, 141:333–387.

Giné, E. and Nickl, R. (2009). Uniform limit theorems for wavelet density estimators. *Ann. Probab.*, 37(4):1605–1646.

Giné, E. and Zinn, J. (1984). Some limit theorems for empirical processes. *Ann. Probab.*, 14:929–989.

Giné, E. and Zinn, J. (1986). Lectures on the central limit theorem for empirical processes.

Giné, E. and Zinn, J. (1991). Gaussian characterization of uniform Donsker classes of functions. *Ann. Probab.*, 19(2):758–782.

Gnedenko, B. V. (1968). *The Theory of Probability.* Chelsea Publ., New York, NY.

Goldenshluger, A. and Lepski, O. (2008). Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190.

Goldenshluger, A. and Lepski, O. (2009). Structural adaptation via $\mathbb{L}_p$-norm inequalities. *Probab. Theory Relat. Fields*, 143(1):41–71.

Golubev, G. K. and Levit, B. Y. (1996). On the second order minimax estimation of distribution functions. *Math. Methods Statist.*, 5(1):1–31.

Guillin, A., Léonard, C., Wu, L., and Yao, N. (2009). Transportation-information inequalities for Markov processes. *Probab. Theory Relat. Fields*, 144(3–4):669–695.

Itō, K. and McKean, H. P. (1965). *Diffusion Processes and their Sample Paths*, volume 125 of *Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen.* Springer-Verlag, Berlin.

Jacod, J. and Shiryaev, A. N. (2002). *Limit Theorems for Stochastic Processes.* Grundlehren der mathematischen Wissenschaften. Springer, Berlin, second edition.

Kallenberg, O. (2002). *Foundations of Modern Probability.* Probability and its applications. Springer, Berlin.

Karatzas, I. and Shreve, S. E. (1988). *Brownian Motion and Stochastic Calculus.* Graduate texts in mathematics. Springer, Berlin.

Katkovnik, V. Y. (1985). *Nonparametric identification and data smoothing: The method of local approximation.* Theoretical Foundations of Engineering Cybernetics. Nauka, Moskow.

Kerkyacharian, G., Lepski, O., and Picard, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Relat. Fields*, 121:131–170.

Klemelä, J. and Tsybakov, A. B. (2001). Sharp adaptive estimation of linear functionals. *Ann. Statist.*, 29(6):1567–1600.

Klemelä, J. and Tsybakov, A. B. (2004). Exact constants for pointwise adaptive estimation under the Riesz transform. *Probab. Theory Relat. Fields*, 129:441–467.

Korostelev, A. and Nussbaum, M. (1999). The asymptotic minimax constant for sup-norm loss in nonparametric density estimation. *Bernoulli*, 5(6):1099–1118.

Korostelev, A. P. (1993). An asymptotically minimax regression estimator in the uniform norm up to exact constant. *Theory Probab. Appl.*, 38(2):737–743.

Korostelev, A. P. and Nussbaum, M. (1995). Density estimation in the uniform norm and white noise approximation. Preprint No. 154, Weierstrass Institute, Berlin.

Kuelbs, J., Kurtz, T., and Zinn, J. (2013). A CLT for empirical processes involving time dependent data. *Ann. Probab.*, 41(2):785–816.

Kutoyants, Y. (1999). Efficient density estimation for ergodic diffusion processes. *Statist. Inference Stochast. Process.*, 1:131–155.

Kutoyants, Y. A. (2004). *Statistical Inference for Ergodic Diffusion Processes.* Springer Series in Statistics. Springer, New York.

Lamberton, D. and Pagès, G. (2002). Recursive computation of the invariant distribution of a diffusion. *Bernoulli*, 8(3):367–405.

Ledoux, M. (2000). The geometry of Markov diffusion generators. *Annales de la faculté des sciences de Toulouse* 6^{e} *série*, 9(2):305–366.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete.* Springer, Berlin.

Lepski, O. V. (1990). One problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.*, 35:459–470.

Lepski, O. V. (1991). Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.*, 36:682–697.

Lepski, O. V. (1992). On problems of adaptive estimation in white Gaussian noise. In *Topics in nonparametric estimation*, volume 12 of *Adv. Soviet Math.*, pages 87–106. Amer. Math. Soc., Providence, RI.

Lepski, O. V. and Levit, B. Y. (1998). Adaptive minimax estimation of infinitely differentiable functions. *Math. Methods Statist.*, 7(2):123–156.

Lepski, O. V. and Levit, B. Y. (1999). Adaptive nonparametric estimation of smooth multivariate functions. *Math. Methods Statist.*, 8(3):344–370.

Lepski, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.*, 25(6):2512–2546.

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

Liptser, R. S. and Shiryaev, A. N. (2001). *Statistics of Random Processes*, volume 1 General Theory of *Applications of Mathematics: Stochastic Modelling and Applied Probability*. Springer, Berlin, second edition.

Lorenzi, L. and Bertoldi, M. (2007). *Analytical Methods for Markov Semigroups*. Pure and applied mathematics. Chapman & Hall, New York, first edition.

Mandl, P. (1968). *Analytical treatment of one-dimensional Markov processes*. Springer, Berlin.

Marcus, M. B. and Rosen, J. (2006). *Markov Processes, Gaussian Processes, and Local Times*. Number 100 in Cambridge Studies in advanced mathematics. Cambridge University Press, Cambridge, UK.

Mendelson, S. (2011). Discrepancy, chaining and subgaussian processes. *Ann. Probab.*, 39(3):985–1026.

Mendelson, S. and Zinn, J. (2006). Modified empirical CLT's under only pre-Gaussian conditions. In *IMS Lecture Notes - Monograph Series. High Dimensional Probability IV*, pages 173–184. Institute of Mathematical Statistics.

Merlevède, F., Peligrad, M., and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *IMS Lecture Notes - Monograph Series. High Dimensional Probability V: The Luminy Volume*, pages 273–292. Institute of Mathematical Statistics.

Merlevède, F., Peligrad, M., and Rio, E. (2011). A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probab. Theory Relat. Fields*, 151(3–4):435–474.

Metafune, G., Pallara, D., and Rhandi, A. (2005). Global properties of invariant measures. *J. Funct. Anal.*, 223(2):396–424.

Nickl, R. (2007). Donsker-type theorems for nonparametric maximum likelihood estimators. *Probab. Theory Relat. Fields*, 138:411–449.

Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.*, 24(6):2399–2430.

Nussbaum, M. (2006). Minimax risk, Pinsker bound for. `http://onlinelibrary.wiley.com/doi/10.1002/0471667196.ess1098.pub2/full`.

Pagès, G. and Panloup, F. (2009). Approximation of the distribution of a stationary Markov process with application to option pricing. *Bernoulli*, 15(1):146–177.

Pagès, G. and Panloup, F. (2012). Ergodic approximation of the distribution of a stationary diffusion: Rate of convergence. *Ann. Appl. Probab.*, 22(3):1059–1100.

Panloup, F. (2008). Computation of the invariant measure of a Lévy driven SDE: Rate of convergence. *Stochastic Process. Appl.*, 118(8):1351–1384.

Pardoux, E. and Veretennikov, A. Y. (2001). On the Poisson equation and diffusion approximation I. *Ann. Probab.*, 29(3):1061–1085.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33(3):1065–1076.

Pinsker, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transmission*, 16:120–133.

Prakasa Rao, B. (1999). *Statistical Inference for Diffusion Type Processes.* Arnold, Oxford University Press, London, New York.

Qian, Z. and Zheng, W. (2004). A representation formula for transition probability densities of diffusions and applications. *Stochastic Process. Appl.*, 111(1):57–76.

Radulović, D. and Wegkamp, M. (2000). Weak convergence of smoothed empirical processes: Beyond Donsker classes. In *High Dimensional Probability II*, volume 47, pages 89–105. Birkhäuser, Boston.

Radulović, D. and Wegkamp, M. (2009). Uniform Central Limit Theorems for pregaussian classes of functions. In *High Dimensional Probability V: The Luminy Volume*, volume 5, pages 84–102. Institute of Mathematical Statistics.

Rebolledo, R. (1980). Central limit theorems for local martingales. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 51(3):269–286.

Revuz, D. and Yor, M. (1999). *Continuous Martingales and Brownian Motion*, volume 293 of *Grundlehren der mathematischen Wissenschaften.* Springer, Berlin, 3rd edition.

Rigollet, P. and Tsybakov, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.*, 16(3):260–280.

Rio, E. (2000). *Théorie asymptotique des processus aléatoires faiblement dépendants.* Mathématiques & Applications 31. Springer, Berlin.

Roberto, C. and Zegarlinski, B. (2007). Orlicz–Sobolev inequalities for sub-Gaussian measures and ergodicity of Markov semi-groups. *J. Funct. Anal.*, 243(1):28–66.

Rohde, A. and Strauch, C. (2010). Uniform central limit theorems for multidimensional diffusions. Preprint. `arXiv:1010.3604`.

Royer, G. (2007). *An initiation to logarithmic Sobolev inequalities.* Collection SMF: Cours spécialisés. American Mathematical Society.

Samarov, A. and Tsybakov, A. B. (2007). Aggregation of density estimators and dimension reduction. In *Advances in Statistical Modeling and Inference. Essays in Honor of Kjell A. Doksum*, pages 233–251. World Scientific.

Spokoiny, V. G. (2000). Adaptive drift estimation for nonparametric diffusion model. *Ann. Statist.*, 33(3):815–836.

Strassen, V. and Dudley, R. M. (1969). *The central limit theorem and ε-entropy*, volume 89, chapter Lecture Notes in Mathematics, pages 224–231. Springer.

Stroock, D. W. and Varadhan, S. R. S. (1979). *Multidimensional Diffusion Processes*, volume 233 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin and New York.

Taikov, L. V. (1968). Kolmogorov-type inequalities and the best formulas for numerical differentiation. *Math. Notes*, 4(2):233–238.

Talagrand, M. (1987a). Donsker classes and random geometry. *Ann. Probab.*, 15(4):1327–1338.

Talagrand, M. (1987b). Regularity of Gaussian processes. *Acta Math.*, 159:99–149.

Talagrand, M. (2005). *The Generic Chaining*. Springer Monographs in Mathematics. Springer, Berlin.

Tsybakov, A. B. (1986). Robust reconstruction of functions by the local-approximation method. *Problems Inform. Transmission*, 22:133–146.

Tsybakov, A. B. (1998). Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.*, 26(6):2420–2469.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, Berlin.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

van der Vaart, A. (1994). Weak convergence of smoothed empirical processes. *Scand. J. Statist.*, 21:501–504.

van der Vaart, A. W. (1996). New Donsker classes. *Ann. Probab.*, 24(4):2128–2140.

van der Vaart, A. W. and van Zanten, H. (2005). Donsker theorems for diffusions: Necessary and sufficient conditions. *Ann. Probab.*, 33(4):1422–1451.

van der Vaart, A. W. and Wellner, J. W. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer, New York.

van Zanten, H. (2003). On uniform laws of large numbers for ergodic diffusions and consistency of estimators. *Statist. Inference Stochast. Processes*, 6:199–213.

Zusammenfassung

In dieser Dissertation werden zwei Problemkreise für multivariate ergodische Diffusionsprozesse betrachtet.

Zunächst untersuchen wir die Frage der scharf adaptiven Schätzung der Driftfunktion einer Diffusion, die als starke Lösung einer stochastischen Differentialgleichung gegeben ist. Auf der Grundlage einer stetigen Aufzeichnung von Beobachtungen werden exakte, datenbasierte Verfahren für die globale und für die punktweise Schätzung vorgeschlagen, die die optimale Grenzkonstante unter natürlichen Glattheitsannahmen an den Driftkoeffizienten erreichen. Es wird sowohl eine Verbindung zum klassischen Resultat von Pinsker zur Schätzung über Sobolev-Klassen bezüglich des L^2-Risikos als auch zum "optimal recovery"-Problem hergestellt. Die exakten Ergebnisse ermöglichen insbesondere eine Abschätzung des Einflusses der Diffusionsmatrix auf das Problem der Driftschätzung.

Im zweiten Teil der Promotionsschrift studieren wir unendlichdimensionale Erweiterungen des zentralen Grenzwertsatzes für additive Funktionale ergodischer Diffusionen. Zunächst werden die Begriffe der Donsker-Klasse und der prä-Gauß'schen Eigenschaft im Diffusionskontext eingeführt, um dann klassische Resultate aus der Theorie empirischer Prozesse aufzugreifen. Verschiedene Parallelen zum klassischen, auf unabhängigen und identisch verteilten Beobachtungen beruhenden empirischen Prozess werden für Diffusionen, die einer Poincaré-Ungleichung genügen, aufgezeigt. Ferner wird eine gesteigerte Regularität für multivariate ergodische Diffusionen mit endlichem invarianten Maß nachgewiesen. Dieser Effekt wird durch die Untersuchung geglätteter Versionen des empirischen Diffusionsprozesses aufgedeckt.

Lebenslauf

Entfällt aus datenschutzrechtlichen Gründen.