ESAIM: PS 20 (2016) 432–462
DOI: 10.1051/ps/2016017

ESAIM: Probability and Statistics
www.esaim-ps.org

ADAPTIVE CONFIDENCE BANDS FOR MARKOV CHAINS AND DIFFUSIONS: ESTIMATING THE INVARIANT MEASURE AND THE DRIFT *

Jakob Söhl[1] and Mathias Trabs[2]

Abstract. As a starting point we prove a functional central limit theorem for estimators of the invariant measure of a geometrically ergodic Harris-recurrent Markov chain in a multi-scale space. This allows to construct confidence bands for the invariant density with optimal (up to undersmoothing) L^∞-diameter by using wavelet projection estimators. In addition our setting applies to the drift estimation of diffusions observed discretely with fixed observation distance. We prove a functional central limit theorem for estimators of the drift function and finally construct adaptive confidence bands for the drift by using a completely data-driven estimator.

Mathematics Subject Classification. 62G15, 60F05, 60J05, 60J60, 62M05.

Received August 25, 2015. Revised May 24, 2016. Accepted July 1st, 2016.

1. Introduction

Diffusion processes are prototypical examples of the theory of stochastic differential equations as well as of continuous time Markov processes. At the same time diffusions are widely used in applications, for instance, to model molecular movements, climate data or in econometrics. Focusing on Langevin diffusions, we will consider the solution of the stochastic differential equation

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \quad t \geqslant 0,$$

with unknown drift function $b\colon \mathbb{R} \to \mathbb{R}$, a volatility parameter $\sigma > 0$ and with a Brownian motion $W = \{W_t : t \geqslant 0\}$. The problem of statistical estimation based on discrete observations from this model is embedded into the framework of geometrically ergodic Harris-recurrent Markov chains. We study the estimation of the

Keywords and phrases. Adaptive confidence bands, diffusion, drift estimation, ergodic Markov chain, stationary density, Lepski's method, functional central limit theorem.

* *The authors acknowledge intensive and very helpful discussions with Richard Nickl. J.S. thanks the European Research Council (ERC) for support under Grant No. 647812. M.T. is grateful to the Statistical Laboratory of the University of Cambridge for its hospitality during a visit from February to March 2014, where this research was initiated, and to the Deutsche Forschungsgemeinschaft (DFG) for the research fellowship TR 1349/1-1. Part of the paper was carried out while M.T. was employed at the Humboldt-Universität zu Berlin.*

[1] Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, CB3 0WB Cambridge, UK. j.soehl@statslab.cam.ac.uk

[2] Department of Mathematics, University of Hamburg, Bundesstraße 55, 20146 Hamburg, Germany. mathias.trabs@uni-hamburg.de

invariant density of such Markov chains. The drift function b depends nonlinearly on the invariant density μ so that the two estimation problems of b and μ are closely related. We prove functional central limit theorems for estimators of both b and μ in multi-scale spaces. This allows the construction of confidence bands for μ. Owing to the nonlinear dependence the construction of confidence bands for b is more involved. In this more difficult situation and by using a self-similarity assumption we make the additional step of constructing confidence bands for b that shrink at a rate adapting to the unknown smoothness.

Estimating the invariant density of a Markov process has been of interest for a long time. An early treatment is given by Roussas [39], who considered kernel estimators and showed consistency and asymptotic normality of the estimators under the strong Doeblin condition. Rosenblatt [38] analysed kernel estimators under the weaker condition G_2 on the Markov chain. More general δ-sequences were used for the estimation by Castellana and Leadbetter [6], who prove pointwise consistency and under strong mixing assumptions asymptotic normality. Yakowitz [47] shows asymptotic normality of kernel density estimators for the invariant density of Markov chains without using assumptions on the rates of mixing parameter sequences. Adaptive estimation was considered by Lacour [29], who estimates the invariant density and the transition density of Markov chains by model selection and proves that the estimators attain the minimax convergence rate under L^2-loss. For stationary processes, Schmisser [40] estimates the derivatives of the invariant density by model selection, derives the convergence rates of the estimators and pays special attention to the case of discretely observed diffusion processes. We see that asymptotic normality has been widely considered in the nonparametric estimation of invariant densities and thus implicitly also confidence intervals. However, we are not aware of any extensions of the pointwise results to uniform confidence bands for invariant densities, which are, for instance, necessary to construct goodness-of-fit tests of the Kolmogorov–Smirnov type.

The statistical properties of the diffusion model depend crucially on the observation scheme. If the whole path $(X_t)_{0\leqslant t\leqslant T}$ is observed for some time horizon $T > 0$, we speak of *continuous* observations. The case of discrete observations $(X_{k\Delta})_{k=0,\dots,n-1}$ with observation distance $\Delta > 0$ is distinguished into *high-frequency* observations, *i.e.* $\Delta \downarrow 0$, and *low–frequency* observations, where $\Delta > 0$ is fixed. While in the first two settings path properties of the process can be used, statistical inference for low−frequency observations has to rely on the Markovian structure of the observations. A review on parametric estimation in diffusion models is given by Kutoyants [28] and Aït−Sahalia [2]. Nonparametric results are summarized in [20], where also estimators based on low−frequency observations are introduced and analysed. These low−frequency estimators rely on a spectral identification of diffusion coefficients which have been introduced by Hansen and Scheinkman [22] and Hansen *et al.* [23]. On the same observation scheme, Kristensen [27] studies a pseudo-maximum likelihood approach in a semiparametric model. Nonparametric estimation based on random sampling times of the diffusion has been studied in [12]. While we pursue a frequentist approach, the Bayesian approach is also very attractive. Based on low−frequency observations van der Meulen and van Zanten [44] have proved consistency of the Bayesian method and Nickl and Söhl [36] showed posterior contraction rates.

As usual, nonparametric estimators depend on some tuning parameters, such as the bandwidth for classical kernel estimators. Choosing these parameters in a data-driven way, Spokoiny [41] initiated adaptive drift estimation in the diffusion model based on continuous observations. This was further developed by Dalalyan [14] and Löcherbach *et al.* [31]. Based on high-frequency observations, adaptive estimation was studied by Hoffmann [25] as well as Comte *et al.* [13]. In the low−frequency case the question of adaptive estimation has been studied by Chorowski and Trabs [12]. In this work we go one step further not only constructing a (rate optimal) adaptive estimator for the drift, but constructing adaptive confidence bands.

Statistical applications require tests and confidence statements. Negri and Nishiyama [35] as well as Masuda *et al.* [33] have constructed goodness-of-fit tests for diffusions based on high-frequency observations. Low [32] has shown that even in a simple density estimation problem no confidence bands exist which are honest and adaptive at the same time. Circumventing this negative result by a "self-similarity" condition, Giné and Nickl [18] have constructed honest and adaptive confidence bands for density estimation. Hoffmann and Nickl [26] have further studied necessary and sufficient conditions for the existence of adaptive confidence bands and the "self-similarity" condition has led to several recent papers on adaptive confidence

bands, notably Chernozhukov *et al.* [11] and Szabó *et al.* [42]. The present paper extends the theory of adaptive confidence bands beyond the classical nonparametric models of density estimation, white noise regression and the Gaussian sequence model which have been treated in the above papers.

In order to derive confidence bands, we first have to establish a uniform central limit theorem. The empirical measure of the observations $X_0, \dots, X_{(n-1)\Delta}$ is the canonical estimator for the invariant measure of a Markov chain or diffusion. Considering a wavelet projection estimator, we obtain a smoothed version of the empirical measure, which is subsequently used to estimate the drift function in the case of diffusions. Thus a natural starting point is a functional central limit theorem for the invariant measure. Since our observations are not independent, the standard empirical process theory does not apply. Instead we have to use the Markov structure of the chain $(X_{k\Delta})_k$. In the continuous time analogue the Donsker theorem for diffusion processes has been studied by van der Vaart and van Zanten [45]. In the case of low−frequency observations, the estimation problem is ill-posed and we have nonparametric convergence rates under the uniform loss. For the asymptotic behaviour of the estimation error in the uniform norm we would expect a Gumbel distribution as shown by Giné and Nickl [18] in the density estimation case using extreme value theory. Recent papers by Castillo and Nickl [7, 8] show that we can hope for parametric rates and an asymptotic normal distribution if we consider instead a weaker norm for the loss. More precisely, the estimation error can be measured in a multi-scale space where the wavelet coefficients are down-weighted appropriately. The resulting norm corresponds to a negative Hölder norm.

Following this approach and relying on a concentration inequality by Adamczak and Bednorz [1], our first result is a functional central limit theorem for rather general geometrically ergodic, Harris-recurrent Markov chains. This could also be of interest for the theory on Markov chain Monte Carlo (MCMC) methods considering that the central limit theorem measures the distance between a target integral and its approximation,

$$\int_{\mathbb{R}} f(z)\mu(\mathrm{d}z) \quad \text{and} \quad \frac{1}{n}\sum_{k=0}^{n-1} f(Z_k),$$

respectively, where $(Z_k)_k$ is a Markov chain with invariant measure μ (*cf.* [16]). Nevertheless, our focus is on the statistical point of view. The functional central limit theorem immediately yields non-adaptive confidence bands and as in [8] these have an L^∞-diameter shrinking with (almost) the optimal nonparametric rate. This small deviation from the optimal rate corresponds to the usual undersmoothing in the construction of nonparametric confidence sets.

Applying the results for general Markov chains to diffusion processes observed at low frequency, we obtain a functional central limit theorem for estimators of the drift function. Inspired by Giné and Nickl [18], in a last demanding step the smoothness of b and the corresponding size of the confidence band is estimated to find adaptive confidence bands. The adaptive procedure relies on Lepski's method. In order to make the construction of adaptive confidence bands feasible, we impose a self-similarity assumption on the drift function.

This work is organized as follows: in Section 2 we study general Markov chains and prove the functional central limit theorem and confidence bands under appropriate conditions on the chain. These results are applied to diffusion processes in Section 3. The adaptive confidence bands for the drift estimator are constructed in Section 4. Some proofs are postponed to the last two sections.

2. Confidence bands for the invariant probability density of Markov processes

2.1. Preliminaries on Markov chains

We start with recalling some facts from the theory of Markov chains. For all basic definition and results we refer to [34]. Let $Z = (Z_k)$, $k = 0, 1, \dots$, be a time-homogeneous Markov chain with state space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. To fix the notation, let P_x and P_ν denote the probability measure of the chain with initial conditions $Z_0 = x \in \mathbb{R}$

and $Z_0 \sim \nu$, respectively. The corresponding expectations will be denoted by $\mathbb{E}_x$ and $\mathbb{E}_\nu$, the Markov chain transition kernel by $P(x, A)$, $x \in \mathbb{R}$, $A \in \mathcal{B}(\mathbb{R})$. The transition operator is defined by $(Pf)(x) = \mathbb{E}_x[f(Z_1)]$.

From the general theory of Markov chains we know that for a Harris-recurrent Markov chain Z the existence of a unique invariant probability measure μ is equivalent to the *drift condition*

$$PV(x) - V(x) \leqslant -1 + c\mathbb{1}_C(x)$$

for some petite set C, some $c < \infty$ and some non-negative function V, which is finite at some $x_0 \in \mathbb{R}$. If Z is additionally aperiodic, then this drift condition is already equivalent to Z being *ergodic*

$$\|P^n(x, \cdot) - \pi\|_{TV} \to 0, \quad \text{as} \quad n \to \infty, \quad \text{for all } x \in \mathbb{R},$$

denoting the total variation norm of a measure by $\|\cdot\|_{TV}$. If we impose a stronger drift condition, namely the geometric drift towards C, we obtain even *geometric ergodicity*: for a ψ-irreducible and aperiodic Markov chain Z satisfying

$$(PV)(x) - V(x) \leqslant -\lambda V(x) + c\mathbb{1}_C(x), \quad \text{for all } x \in \mathbb{R}, \tag{2.1}$$

for a petite set C, some $\lambda > 0, c < \infty$ and a function $V\colon \mathbb{R} \to [1, \infty)$, it holds for some $r > 1$, $R < \infty$,

$$\sum_{n \geqslant 0} r^n \|P^n(x, \cdot) - \mu\|_{TV} \leqslant RV(x), \quad \text{for all } x \in \mathbb{R}.$$

Note that ψ-irreducibility together with the geometric drift condition (2.1) implies already that Z is positive Harris with invariant probability measure μ.

The geometric ergodicity yields the following central limit theorem (see [10], Thm. II.4.1). The weakest form of ergodicity so that the central limit theorem holds is ergodicity of degree 2 which is slightly weaker than the geometric ergodicity that we have assumed here.

Proposition 2.1. *Let $(Z_k)_{k \geqslant 0}$ be a geometrically ergodic Markov chain with arbitrary initial condition and invariant probability measure μ, then there exists for every bounded function $f = (f_1, \dots, f_d) : \mathbb{R} \to \mathbb{R}^d$ a symmetric, positive semidefinite matrix $\Sigma_f = (\Sigma_{f_i, f_j})_{i,j=1,\dots,d}$ such that*

$$n^{-1/2} \left(\sum_{k=0}^{n-1} f(Z_k) - n\mathbb{E}_\mu[f(Z_0)] \right) \xrightarrow{d} N(0, \Sigma_f), \quad \text{as } n \to \infty.$$

For $i, j \in \{1, \dots, d\}$ the asymptotic covariances are given by

$$\begin{aligned} \Sigma_{f_i, f_j} &:= \lim_{n \to \infty} n^{-1} \operatorname{Cov}_\mu \left(\sum_{k=0}^{n-1} f_i(Z_k), \sum_{k=0}^{n-1} f_j(Z_k) \right) \\ &= \mathbb{E}_\mu\big[(f_i(Z_0) - \mathbb{E}_\mu[f_i])(f_j(Z_0) - \mathbb{E}_\mu[f_j])\big] + \sum_{k=1}^{\infty} \mathbb{E}_\mu\big[(f_i(Z_0) - \mathbb{E}_\mu[f_i])(f_j(Z_k) - \mathbb{E}_\mu[f_j])\big] \\ &\quad + \sum_{k=1}^{\infty} \mathbb{E}_\mu\big[(f_i(Z_k) - \mathbb{E}_\mu[f_i])(f_j(Z_0) - \mathbb{E}_\mu[f_j])\big]. \end{aligned} \tag{2.2}$$

In order to lift this "pointwise" result to a functional central limit theorem, we will in addition need a concentration inequality for a preciser control on how the sum $n^{-1} \sum_{k=0}^{n-1} f(Z_k)$ deviates from the integral $\int f(z)\mu(\mathrm{d}z)$ for finite sample sizes. To this end, we strengthen the aperiodicity assumption to *strong aperiodicity* (see [34], Prop. 5.4.5), that is there exists a set $C \in \mathcal{B}(\mathbb{R})$, a probability measure ν with $\nu(C) > 0$ and a constant $\delta > 0$ such that

$$P(x, B) \geqslant \delta\nu(B), \quad \text{for all } x \in C, B \in \mathcal{B}(\mathbb{R}). \tag{2.3}$$

Any set C satisfying this condition is called *small set*. Recall that any small set is a petite set.

Proposition 2.2 (Thm. 9 by Adamczak and Bednorz [1]). *Let $Z = (Z_k)_{k\geqslant 0}$ be a Harris recurrent, strongly aperiodic Markov chain on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with unique invariant measure μ. For some set $C \in \mathcal{B}(\mathbb{R})$ with $\mu(C) > 0$ let Z satisfy the drift condition (2.1) and the small set condition (2.3).*

Let $f \in L^2(\mu)$ be bounded. For any $0 < \tau < 1$ there are constants K, c_2 depending only on δ, V, λ, c and τ and a constant c_1 depending additionally on the initial value $x \in \mathbb{R}$ such that for any $t > 0$

$$P_x\left(\left|\sum_{k=0}^{n-1} f(Z_k) - n\mathbb{E}_\mu[f(Z_0)]\right| > t\right) \leqslant K\exp\left(-c_1\left(\frac{t}{\|f\|_\infty}\right)^\tau\right) + K\exp\left(-\frac{c_2 t^2}{n\Sigma_f + t\max(\|f\|_\infty(\log n)^{1/\tau}, \Sigma_f^{1/2})}\right),$$

where Σ_f is given by (2.2) with $d = 1$.

As a last ingredient we need to bound the asymptotic variance Σ_f in Propositions 2.1 and 2.2 in terms of $\|\overline{f}\|^2_{L^2(\mu)}$ for the centred function $\overline{f} := f - \int f\mathrm{d}\mu$. The geometric ergodicity only yields a bound $\mathcal{O}(\|\overline{f}\|^2_\infty)$. Therefore, we require that the transition operator is a contraction in the sense that there exists some $\rho \in (0,1)$ satisfying

$$\|Pg\|_{L^2(\mu)} \leqslant \rho\|g\|_{L^2(\mu)} \quad \text{for all } g \in L^2(\mu) \text{ with } \int g\,\mathrm{d}\mu = 0. \tag{2.4}$$

This property is also known as ρ-mixing. It corresponds to a Poincaré inequality (*cf.* [3], Thm. 1.3) and its relation to drift conditions is analysed by Bakry *et al.* [3]. If (2.4) is fulfilled, the Cauchy–Schwarz inequality yields

$$\Sigma_f = \Sigma_{\overline{f}} \leqslant \|\overline{f}\|^2_{L^2(\mu)} + 2\sum_{k=1}^\infty \|\overline{f}\|_{L^2(\mu)}\|P^k\overline{f}\|_{L^2(\mu)} \leqslant \left(1 + 2\sum_{k=1}^\infty \rho^k\right)\|\overline{f}\|^2_{L^2(\mu)} = \tfrac{1+\rho}{1-\rho}\|\overline{f}\|^2_{L^2(\mu)}. \tag{2.5}$$

2.2. A functional central limit theorem

The basic idea is to prove a functional central limit theorem for the invariant probability measure μ by choosing an orthonormal basis, applying the pointwise central limit theorem to the basis functions (Prop. 2.1) and extending this result to finite linear combinations with the help of the concentration inequality (Prop. 2.2). Provided μ has some regularity, the approximation error due to considering only a finite basis expansion of μ will be negligible. Noting that it is straightforward to extend the results to any compact subset of $\mathbb{R}$, we focus on a central limit theorem on a bounded interval $[a,b]$ with $-\infty < a < b < \infty$.

Let $(\varphi_{j_0,l}, \psi_{j,k} : j \geqslant j_0, l, k \in \mathbb{Z})$, for some $j_0 \geqslant 0$, a scaling function φ and a wavelet function ψ, be a regular compactly supported L^2-orthonormal wavelet basis of $L^2(\mathbb{R})$. For the sake of clarity we throughout use Daubechies' wavelets of order $N \in \mathbb{N}$, but any other compactly supported regular wavelet basis can be applied as well. As a standing assumption we suppose that N is chosen large enough such that the Hölder regularity of φ and ψ is larger than the regularity required for the invariant measure. The approximation spaces for resolution levels $J > j_0$ are defined as

$$V_J := \overline{\operatorname{span}}\{\varphi_{j_0,l}, \psi_{j,k} : j = j_0, \dots, J,\ l,k \in \mathbb{Z}\},$$

The projection onto V_J is denoted by π_J. Since j_0 is fixed and to simplify the notation, we write $\psi_{-1,l} := \varphi_{j_0,l}$.

Using the first $n \in \mathbb{N}$ steps $Z_0, Z_1, \dots, Z_{n-1}$ of a realisation of the chain, we define the empirical measure

$$\mu_n := \frac{1}{n}\sum_{k=0}^{n-1} \delta_{Z_k},$$

where δ_x denotes the Dirac measure at the point $x \in \mathbb{R}$. The canonical projection wavelet estimator of μ is given by

$$\widehat{\mu}_J := \pi_J(\mu_n) = \sum_{l\in\mathbb{Z}} \widehat{\mu}_{-1,l}\psi_{-1,l} + \sum_{j=j_0}^J \sum_{k\in\mathbb{Z}} \widehat{\mu}_{j,k}\psi_{j,k}, \qquad \widehat{\mu}_{j,k} := \langle \psi_{j,k}, \mu_n\rangle := \int \psi_{j,k}\,\mathrm{d}\mu_n. \tag{2.6}$$

For any $\psi_{j,k}$ Proposition 2.1 yields that $\sqrt{n}(\mu_n - \mu)(\psi_{j,k})$ converges in distribution for $n \to \infty$ to a Gaussian random variable

$$\mathbb{G}_\mu(j,k) \sim \mathcal{N}(0, \Sigma_{\psi_{j,k}}) \quad \text{with covariances} \quad \mathbb{E}_\mu[\mathbb{G}_\mu(j,k)\mathbb{G}_\mu(l,m)] = \Sigma_{\psi_{j,k},\psi_{l,m}}. \tag{2.7}$$

Using the techniques from Castillo and Nickl [8], this pointwise convergence of μ_n can be extended to a uniform central limit theorem on $[a,b]$ for the projection estimator $\widehat{\mu}_J$ in the *multi-scale sequence spaces* which are defined as follows: noting that the Daubechies wavelets fulfil $\operatorname{supp}\varphi \subseteq [0, 2N-1]$ and $\operatorname{supp}\psi \subseteq [-N+1, N]$ (*cf.* [24], Chap. 7), the sets $L := K_{-1} := \{k \in \mathbb{Z} : 2^{j_0}a - 2N + 1 \leqslant k \leqslant 2^{j_0}b\}$ and $K_j := \{k \in \mathbb{Z} : 2^j a - N \leqslant k \leqslant 2^j b + N - 1\}$ contain all indices of $\varphi_{j_0,\cdot}$ and $\psi_{j,\cdot}$, respectively, whose support intersects with the interval $[a,b]$. For a monotonously increasing weighting sequence $w = (w_j)_{j=-1,j_0,j_0+1,j_0+2,\ldots}$ with $w_j \geqslant 1$ and $w_{-1} := 1$ we define the multi-scale sequence spaces as

$$\mathcal{M} := \mathcal{M}(w) := \left\{ x = (x_{jk}) : \|x\|_{\mathcal{M}(w)} := \sup_{j \in \{-1, j_0, j_0+1, \ldots\}} \max_{k \in K_j} \frac{|x_{jk}|}{w_j} < \infty \right\},$$

Since the Banach space $\mathcal{M}(w)$ is non-separable, we define the separable, closed subspace

$$\mathcal{M}_0 := \mathcal{M}_0(w) := \left\{ x = (x_{jk}) : \lim_{j \to \infty} \max_{k \in K_j} \frac{|x_{jk}|}{w_j} = 0 \right\}.$$

Let us assume that μ is absolutely continuous with respect to the Lebesgue measure and denote the density likewise by μ. If the density is bounded on $D = [a - 2^{-j_0}(2N-1), b + 2^{-j_0}(2N-1)]$, the orthonormality and the support of $(\psi_{j,k})$ and (2.5) yield $\Sigma_{\psi_{j,k}} = \mathcal{O}(\|\mu\|_\infty)$. Standard estimates of the supremum of normal random variables yield that the maximum over the 2^j variables $\mathbb{G}_\mu(j,\cdot)$ of a resolution level j are of the order $\max_k |\mathbb{G}_\mu(j,k)| = \mathcal{O}_P(\sqrt{j})$, see (2.14) below. Since the cardinality of K_j is of the order 2^j, a weighting $w_j = \sqrt{j}$ seems to be appropriate and indeed we conclude as ([8], Prop. 3):

Lemma 2.3. *Let μ admit a Lebesgue density which is bounded on D. Then $\mathbb{G}_\mu$ from (2.7) satisfies $\mathbb{E}[\|\mathbb{G}_\mu\|_{\mathcal{M}(w)}] < \infty$ for the weights $w_j = \sqrt{j}$. Moreover, $\mathcal{L}(\mathbb{G}_\mu)$ is a tight Gaussian Borel probability measure in $\mathcal{M}_0(w)$ if $\sqrt{j}/w_j \to 0$.*

Let us now summarise the assumptions on the Markov chain, which are needed to prove the functional central limit theorem and for the construction of confidence bands. For any regularity $s > 0$, denoting the integer part of s by $[s]$, the *Hölder space* on a domain D is defined by

$$C^s(D) := \left\{ f : D \to \mathbb{R} \,\middle|\, \|f\|_{C^s} := \sum_{k=0}^{[s]} \|f^{(k)}\|_\infty + \sup_{x \neq y} \frac{|f^{[s]}(x) - f^{[s]}(y)|}{|x-y|^{s-[s]}} \right\}.$$

Assumption 2.4. Let $(Z_k)_{k \geqslant 0}$ be a Harris recurrent, strongly aperiodic Markov chain on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with initial condition $Z_0 = x$. Let the invariant probability measure have a density μ in $C^s(D)$ for some $s > 0$ and some sufficiently large set $D \subseteq \mathbb{R}$ containing $[a,b]$. Let the drift condition (2.1) and small set condition (2.3) be satisfied for some $C \in \mathcal{B}(\mathbb{R})$ with $\mu(C) > 0$. Further suppose that the transition operator is an $L^2(\mu)$-contraction fulfilling (2.4) with $\rho \in (0,1)$.

Remark 2.5. As we have discussed above it suffices to verify that the chain $(Z_k)_{k \geqslant 0}$ is ψ-irreducible and satisfies (2.1) and (2.3) in order to conclude that the $(Z_k)_{k \geqslant 0}$ is Harris recurrent, strongly aperiodic and has a unique invariant probability measure.

Now we can show the functional central limit theorem for $\widehat{\mu}_J$ in the space $\mathcal{M}_0(w)$. Note that the natural nonparametric choice J_n given by $2^{J_n} \sim n^{1/(2s+1)}$ satisfies the conditions of the following theorem. Recall that weak convergence of laws $\mathcal{L}(X)$ of random variables X on a metric space (S,d) can be metrised by the *bounded-Lipschitz metric*

$$\beta_S(\mu,\nu) := \sup_{F:\|F\|_{BL}\leqslant 1} \left| \int_S F(x)(\mu(\mathrm{d}x) - \nu(\mathrm{d}x)) \right| \quad \text{with}$$
$$\|F\|_{BL} := \sup_{x\in S} |F(x)| + \sup_{x,y\in S: x\neq y} \frac{|F(x)-F(y)|}{d(x,y)}.$$

Theorem 2.6. *Grant Assumption 2.4 and let $w = (w_j)$ be increasing and satisfy $\sqrt{j}/w_j \to 0$ as $j \to \infty$. Let $J_n \in \mathbb{N}$ fulfil, for some $\tau \in (0,1)$,*

$$\sqrt{n}2^{-J_n(2s+1)/2} w_{J_n}^{-1} = o(1), \qquad (\log n)^{2/\tau} n^{-1} 2^{J_n} J_n = \mathcal{O}(1).$$

Then $\widehat{\mu}_{J_n}$ from (2.6) satisfies, for $n \to \infty$,

$$\sqrt{n}(\widehat{\mu}_{J_n} - \mu) \xrightarrow{d} \mathbb{G}_\mu \quad \text{in } \mathcal{M}_0(w).$$

Proof. We follow the strategy of ([8], Thm. 1). First we deal with the bias term. By the s-Hölder regularity of μ we have ([19], Def. (5.90) and Prop. 5.3.13)

$$\sup_{j,k} 2^{j(2s+1)/2} |\langle \psi_{j,k}, \mu\rangle| < \infty$$

and thus by the assumption on J_n

$$\|\mu - \pi_{J_n}(\mu)\|_{\mathcal{M}} = \sup_{j>J_n} \max_{k\in K_j} w_j^{-1} |\langle \psi_{jk}, \mu\rangle| \lesssim \sup_{j>J_n} w_j^{-1} 2^{-j(2s+1)/2} = o(n^{-1/2}).$$

Defining $\nu_n := \sqrt{n}(\widehat{\mu}_{J_n} - \pi_{J_n}(\mu))$, we decompose the stochastic error, for $J < J_n$ to be specified later,

$$\begin{aligned}\beta_{\mathcal{M}_0}(\mathcal{L}(\nu_n), \mathcal{L}(\mathbb{G}_\mu)) \leqslant{}& \beta_{\mathcal{M}_0}(\mathcal{L}(\nu_n), \mathcal{L}(\nu_n)\circ\pi_J^{-1}) + \beta_{\mathcal{M}_0}(\mathcal{L}(\nu_n)\circ\pi_J^{-1}, \mathcal{L}(\mathbb{G}_\mu)\circ\pi_J^{-1}) \\ &+ \beta_{\mathcal{M}_0}(\mathcal{L}(\mathbb{G}_\mu)\circ\pi_J^{-1}, \mathcal{L}(\mathbb{G}_\mu)). \end{aligned} \tag{2.8}$$

In the sequel we will separately show that all three terms converge to zero. Let $\varepsilon > 0$. By definition of the $\beta_{\mathcal{M}_0}$-norm we estimate the first term by

$$\begin{aligned}\beta_{\mathcal{M}_0}(\mathcal{L}(\nu_n), \mathcal{L}(\nu_n)\circ\pi_J^{-1}) &= \sup_{F:\|F\|_{BL}\leqslant 1} |\mathbb{E}\left[F(\nu_n) - F(\pi_J(\nu_n))\right]| \\ &\leqslant \mathbb{E}\left[\|\sqrt{n}(\pi_{J_n} - \pi_J)(\mu_n - \mu)\|_{\mathcal{M}}\right] \\ &\leqslant \max_{J<j\leqslant J_n} (w_j^{-1} j^{1/2}) \mathbb{E}\left[\max_{J<j\leqslant J_n} \max_{k\in K_j} j^{-1/2} |\langle \sqrt{n}(\mu_n - \mu), \psi_{j,k}\rangle|\right]. \end{aligned} \tag{2.9}$$

By the assumptions on w and due to the factor in front of the expectation, the above display can be bounded by $\varepsilon/3$ if J is chosen large enough and provided that the expectation can be bounded by a constant independent of J and n. To apply the concentration inequality in Proposition 2.2, note that $\Sigma_{\psi_{j,k}} = \mathcal{O}(\|\mu\|_\infty)$ by (2.5) and $\sqrt{j}\|\psi_{j,k}\|_\infty = \sqrt{j}2^{j/2} = \mathcal{O}(\sqrt{n}(\log n)^{-1/\tau})$ for $j \leqslant J_n$. Hence, for any $M > 0$ large enough we obtain for constants $c_i > 0, i = 1,2,\ldots,$

$$
\begin{aligned}
&\mathbb{E}\left[\max_{J<j\leqslant J_n}\max_{k\in K_j} j^{-1/2}|\langle\sqrt{n}(\mu_n-\mu),\psi_{j,k}\rangle|\right]\\
&\leqslant M+\int_M^\infty P\left(\max_{J<j\leqslant J_n}\max_{k\in K_j} j^{-1/2}|\langle\sqrt{n}(\mu_n-\mu),\psi_{j,k}\rangle|>u\right)\mathrm{d}u\\
&\leqslant M+\sum_{J<j\leqslant J_n,k\in K_j}\int_M^\infty P\left(|\langle\sqrt{n}(\mu_n-\mu),\psi_{j,k}\rangle|>\sqrt{j}u\right)\mathrm{d}u\\
&\lesssim M+\sum_{J<j\leqslant J_n}2^j\int_M^\infty\left(\exp\left(-c_1(\log n)j^\tau u^\tau\right)+\exp\left(-c_2ju^2/(1+u)\right)\right)\mathrm{d}u\\
&\lesssim M+\sum_{J<j\leqslant J_n}2^j\left(\frac{\mathrm{e}^{-c_3(jM)^\tau\log n}}{j^\tau\log n}+\frac{\mathrm{e}^{-c_4jM}}{j}\right)\\
&\lesssim M+\mathrm{e}^{-c_5JM^\tau}\lesssim M+1,
\end{aligned}
\tag{2.10}
$$

where we have used in the next to last estimate that $J_n\lesssim\log n$ and thus $j^\tau\log n\gtrsim j$ for all $j\leqslant J_n$.

To bound the second term in (2.8), we use Proposition 2.1 and the Cramér–Wold device to see that it is smaller than $\varepsilon/3$ for fixed J and n sufficiently large. It remains to consider the third term in (2.8) which can be estimated similarly to (2.9), using that $\mathbb{E}[\sup_j\max_k j^{-1/2}|\mathbb{G}_\mu(j,k)|]<\infty$ by Lemma 2.3. □

2.3. The construction of confidence bands

Using the multi-scale central limit theorem, we now construct confidence bands for the density of the invariant probability measure. For some confidence level $\alpha\in(0,1)$ the natural idea is to take

$$
\mathcal{C}_n(\zeta_\alpha):=\left\{f:\|f-\widehat{\mu}_{J_n}\|_{\mathcal{M}}<\frac{\zeta_\alpha}{\sqrt{n}}\right\}=\left\{f:\sup_{j,k}w_j^{-1}|\langle f-\widehat{\mu}_{J_n},\psi_{j,k}\rangle|<\frac{\zeta_\alpha}{\sqrt{n}}\right\},
$$

where ζ_α is chosen such that $P(\|\mathbb{G}_\mu\|_{\mathcal{M}}<\zeta_\alpha)\geqslant 1-\alpha$. For this set the asymptotic coverage follows immediately from Theorem 2.6. However, $\mathcal{C}_n(\zeta_\alpha)$ is too large in terms of the $L^\infty([a,b])$-diameter

$$
|\mathcal{C}_n(\zeta_\alpha)|_\infty:=\sup\left\{\sup_{x\in[a,b]}|f(x)-g(x)|:f,g\in\mathcal{C}_n(\zeta_\alpha)\right\}.
$$

To obtain the (nearly) optimal L^∞-diameter, we need to control the large resolution levels. As suggested by Castillo and Nickl [8], we use *a priori* knowledge of the regularity s to define

$$
\overline{\mathcal{C}}_n:=\overline{\mathcal{C}}_n(\zeta_\alpha,s,u_n):=\mathcal{C}_n(\zeta_\alpha)\cap\left\{f:\|f\|_{C^s}\leqslant u_n\right\}
\tag{2.11}
$$

for a sequence $u_n\to\infty$.

Proposition 2.7. *Grant Assumption 2.4 with $s>0$ and let $w=(w_j)$ be increasing and satisfy $\sqrt{j}/w_j\to 0$. For $\alpha\in(0,1)$ let $\zeta_\alpha>0$ be such that $P(\|\mathbb{G}_\mu\|_{\mathcal{M}}\geqslant\zeta_\alpha)\leqslant\alpha$ and choose $J_n:=J_n(s)$ such that*

$$
2^{J_n}=\left(\frac{n}{\log n}\right)^{1/(2s+1)}.
$$

Then the confidence set $\overline{\mathcal{C}}_n=\overline{\mathcal{C}}_n(\zeta_\alpha,s,u_n)$ from (2.11) with $u_n:=w_{J_n}/\sqrt{J_n}$ satisfies

$$
\liminf_{n\to\infty}P(\mu\in\overline{\mathcal{C}}_n)\geqslant 1-\alpha\quad\text{and}\quad|\overline{\mathcal{C}}_n|_\infty=\mathcal{O}_P\left(\left(\frac{n}{\log n}\right)^{-s/(2s+1)}u_n\right).
$$

Proof. Let us first verify $\liminf_{n\to\infty} P(\mu \in \overline{\mathcal{C}}_n) \geqslant 1-\alpha$. Since $\mu \in C^s(u_n)$ for large enough n, Theorem 2.6 yields

$$\liminf_{n\to\infty} P(\mu \in \overline{\mathcal{C}}_n) = \liminf_{n\to\infty} P(\sqrt{n}\|\widehat{\mu}_{J_n} - \mu\|_{\mathcal{M}} < \zeta_\alpha) \geqslant P(\|\mathbb{G}_\mu\|_{\mathcal{M}} < \zeta_\alpha) \geqslant 1-\alpha.$$

To bound the diameter let $f, g \in \overline{\mathcal{C}}_n$. Using $\|f - \widehat{\mu}_{J_n}\|_{\mathcal{M}} = \mathcal{O}_P(n^{-1/2})$ and $f - g \in C^s(2u_n)$, we obtain

$$\begin{aligned}
\|f-g\|_{L^\infty([a,b])} &\lesssim \sum_{j\leqslant J_n} 2^{j/2} \max_{k\in K_j} |\langle f-g, \psi_{j,k}\rangle| + \sum_{j>J_n} 2^{j/2} \max_{k\in K_j} |\langle f-g, \psi_{j,k}\rangle| \\
&\leqslant \sum_{j\leqslant J_n} 2^{j/2} \left(\max_{k\in K_j} |\langle f-\widehat{\mu}_{J_n}, \psi_{j,k}\rangle| + \max_{k\in K_j} |\langle g-\widehat{\mu}_{J_n}, \psi_{j,k}\rangle| \right) \\
&\quad + \sum_{j>J_n} 2^{-js} 2^{j(s+1/2)} \max_{k\in K_j} |\langle f-g, \psi_{j,k}\rangle| \\
&\leqslant \left(\|f-\widehat{\mu}_{J_n}\|_{\mathcal{M}} + \|g-\widehat{\mu}_{J_n}\|_{\mathcal{M}}\right) \sum_{j\leqslant J_n} 2^{j/2} w_j + \|f-g\|_{C^s} \sum_{j>J_n} 2^{-js} \\
&= \mathcal{O}_P\left(n^{-1/2} 2^{J_n/2} w_{J_n}\right) + \mathcal{O}_P\left(2^{-J_n s} u_n\right) \\
&= \mathcal{O}_P\left(n^{-1/2} 2^{J_n/2} J_n^{1/2} u_n\right) + \mathcal{O}_P\left(2^{-J_n s} u_n\right). \qquad (2.12)
\end{aligned}$$

Plugging in the choice of J_n, we finally have $n^{-1/2}\sqrt{2^{J_n} J_n} \lesssim (n/\log n)^{-s/(2s+1)} = 2^{-J_n s}$. □

A multi-scale confidence band as in (2.11) allows for the construction of a classical L^∞-band on $[a,b]$ around $\widehat{\mu}_{J_n}$ as follows: let us denote the almost optimal diameter by $\rho_n := \left(\frac{n}{\log n}\right)^{-s/(2s+1)} u_n$. As we can deduce from (2.12), there is a constant $D > 0$ such that we have $\|f - \widehat{\mu}_{J_n}\|_{L^\infty([a,b])} \leqslant D\rho_n$ for any $f \in \overline{\mathcal{C}_n}$. Hence, the band

$$\widetilde{\mathcal{C}}_n := \left\{ f\colon [a,b] \to \mathbb{R} \middle| \|f - \widehat{\mu}_{J_n}\|_{L^\infty([a,b])} \leqslant D\rho_n \right\}$$

contains $\overline{\mathcal{C}}_n$ which only improves the coverage. Consequently, $\widetilde{\mathcal{C}}_n$ is an L^∞-confidence band with level α which shrinks with almost optimal rate ρ_n. In addition, a multi-scale confidence band as in (2.11) allows for simultaneous confidence intervals in all wavelet coefficients. This is especially useful for goodness-of-fit tests where the optimal L^∞-diameter of $\overline{\mathcal{C}}_n$ is a measure of the power of the test.

In order to apply the confidence band (2.11) we need the regularity s of the invariant density and a critical value ζ_α such that $P(\|\mathbb{G}_\mu\|_{\mathcal{M}} < \zeta_\alpha) \geqslant 1-\alpha$ for $\alpha \in (0,1)$. Adaptive confidence bands will be presented later in the context of diffusions. So let us suppose for a moment that the regularity s is known. Then the problem reduces to the construction of the critical value to which the remainder of this section is devoted.

A first observation is that if several independent copies of the diffusion are observed then one could calculate for each copy an estimator $\widehat{\mu}_{J_n}$ and obtain estimators for the values ζ_α from the distribution of the estimators $\widehat{\mu}_{J_n}$ around their joint mean. Since the assumption of many independent copies is not realistic we will not pursue this further. Instead of the consistent estimation of the lowest possible ζ_α we restrict ourselves to estimating an upper bound, which yields possibly more conservative confidence sets. By the concentration of Gaussian measures we know for any $\kappa > 0$ that

$$P\left(\|\mathbb{G}_\mu\|_{\mathcal{M}} \geqslant \mathbb{E}[\|\mathbb{G}_\mu\|_{\mathcal{M}}] + \kappa\right) \leqslant \mathrm{e}^{-\kappa^2/(2\Sigma)},$$

where $\Sigma := \sup_{j,k} \mathbb{E}[|\mathbb{G}_\mu(j,k)|^2] = \sup_{j,k} \Sigma_{\psi_{j,k}}$ (see for example [30], Thm. 7.1). Hence, an upper bound for ζ_α is given by

$$\sqrt{2\Sigma \log \alpha^{-1}} + \mathbb{E}[\|\mathbb{G}_\mu\|_{\mathcal{M}}].$$

The expected value $\mathbb{E}[\|\mathbb{G}_\mu\|_{\mathcal{M}}]$ can be bounded as in Proposition 2 by Castillo and Nickl [8], depending on Σ again. We obtain the following upper bound for ζ_α:

Lemma 2.8. *Let $j_0 \geqslant 1$ and $w = (w_j)$ satisfy $w_{-1} = \sqrt{j_0}$ and $\inf_j w_j/\sqrt{j} \geqslant 1, j \geqslant j_0$, and define $\Sigma := \sup_{j,k} \Sigma_{\psi_{j,k}}$. Then $P(\|\mathbb{G}_\mu\|_{\mathcal{M}} \geqslant \overline{\zeta}_\alpha) \leqslant \alpha$ holds for*

$$\overline{\zeta}_\alpha(\Sigma) := \left(\sqrt{2\log \alpha^{-1}} + 2C + \tfrac{32}{3C}2^{-2j_0}\right)\sqrt{\Sigma} \tag{2.13}$$

with $C := (\sup_{j\geqslant j_0}(4\log|K_j| + 2\log 2)/j)^{1/2}$.

Proof. The cardinality of K_j is denoted by $|K_j|$. Recall that a standard normal random variable Z satisfies

$$\mathbb{E}\left[\mathrm{e}^{Z^2/4}\right] = \sqrt{2} \quad \text{and} \quad P(Z > \kappa) \leqslant \frac{1}{\kappa\sqrt{2\pi}}\mathrm{e}^{-\kappa^2/2}, \kappa > 0.$$

For each $j \geqslant j_0$ and $\kappa = 2\sup_k \Sigma_{\psi_{j,k}}^{1/2}$ Jensen's inequality thus yields

$$\begin{aligned}\mathbb{E}\left[\max_k |\mathbb{G}_\mu(j,k)|\right] &\leqslant \kappa\left(\log \mathbb{E}\left[\mathrm{e}^{\max_k |\mathbb{G}_\mu(j,k)|^2/\kappa^2}\right]\right)^{1/2} \\ &\leqslant 2\sup_k \Sigma_{\psi_{j,k}}^{1/2}\left(\log|K_j| + \tfrac{1}{2}\log 2\right)^{1/2} \leqslant C\sqrt{\Sigma j},\end{aligned} \tag{2.14}$$

for the constant $C := (\sup_j(4\log|K_j|+2\log 2)/j)^{1/2}$. Theorem 7.1 in [30] yields for all t, T such that $t \geqslant \Sigma^{1/2}CT$ and $T > 1$

$$\begin{aligned}P\big(\|\mathbb{G}_\mu\|_{\mathcal{M}} > t\big) &\leqslant \sum_j P\left(|\max_k \mathbb{G}_\mu(j,k) - \mathbb{E}[\max_k \mathbb{G}_\mu(j,k)]| > tw_j - \mathbb{E}[\max_k |\mathbb{G}_\mu(j,k)|]\right) \\ &\leqslant \sum_j P\left(|\max_k \mathbb{G}_\mu(j,k) - \mathbb{E}[\max_k \mathbb{G}_\mu(j,k)]| > (t-C)\sqrt{\Sigma j}\right) \\ &\leqslant 2\sum_j \mathrm{e}^{-jt^2(T-1)^2/(2\Sigma T^2)}.\end{aligned}$$

Recall that $j = j_0, j_0, j_0+1, j_0+2, \ldots$ in the above sum. Using Fubini's theorem and the Gaussian tail bound, we conclude

$$\begin{aligned}\mathbb{E}[\|\mathbb{G}_\mu\|_{\mathcal{M}}] &\leqslant \Sigma^{1/2}CT + \int_{\Sigma^{1/2}CT}^\infty P(\|\mathbb{G}_\mu\|_{\mathcal{M}} > t)\mathrm{d}t \leqslant \Sigma^{1/2}CT + 2\Sigma^{1/2}\sum_j \int_{CT}^\infty \mathrm{e}^{-jt^2(1-1/T)^2/2}\mathrm{d}t \\ &\leqslant \Sigma^{1/2}CT + \frac{2\Sigma^{1/2}T}{C(T-1)^2}\sum_j \mathrm{e}^{-(2\log 2)j(T-1)^2} \leqslant \Sigma^{1/2}CT + \frac{4\Sigma^{1/2}T2^{-2(T-1)^2 j_0}}{C(T-1)^2(1-2^{-2(T-1)^2})}.\end{aligned}$$

Choosing the $T = 2$, we obtain $\mathbb{E}[\|\mathbb{G}_\mu\|_{\mathcal{M}}] \leqslant (2C + \frac{32}{3C}2^{-2j_0})\Sigma^{1/2}$. □

From the above lemma we see that Σ is the key quantity for the construction of the critical values ζ_α. A natural estimator for Σ is $\widehat{\Sigma}_n := (\max_{j\leqslant J_n,k} \widehat{\Sigma}_{\psi_{j,k}})$, where $\widehat{\Sigma}_{\psi_{j,k}}$ are estimators of $\Sigma_{\psi_{j,k}}$ based on n observations. Since J_n tends to infinity, the maximum over all $j \leqslant J_n$ converges to the supremum over all j so that we are asymptotically estimating the right quantity. For the estimators $\widehat{\Sigma}_{\psi_{j,k}}$ we propose the initial monotone sequence estimators based on autocovariations by Geyer [16], which are consistent over-estimates, and this yields almost surely

$$\liminf_{n\to\infty} \widehat{\Sigma}_n \geqslant \Sigma,$$

which suffices for our purposes.

The estimation of $\Sigma_{\psi_{j,k}}$ amounts to the estimation of the asymptotic variance Σ_f in (2.2) for a known function f and this problem is studied in the MCMC-literature. In addition to the sequence estimators, Geyer [16] discusses two other constructions together with their advantages and disadvantages. Robert [37] constructs another estimator applying renewal theory, which is however difficult to calculate. A more recent estimator using i.i.d. copies of the process X is given by Chauveau and Diebolt [9].

As an alternative to the above estimation of Σ in (2.13) an upper bound could be estimated as follows: Using (2.5) we can bound Σ from above,

$$\Sigma \leqslant \sup_{j,k} \tfrac{1+\rho}{1-\rho}\|\overline{\psi}_{j,k}\|_{L^2(\mu)} \leqslant \sup_{j,k} \tfrac{1+\rho}{1-\rho}\|\psi_{j,k}\|_{L^2}\|\mu\|_\infty = \tfrac{1+\rho}{1-\rho}\|\mu\|_\infty,$$

where we can plug in estimators for $\|\mu\|_\infty$ and ρ. Considering a wavelet $\psi_{j,k}$ localised around the maximum of μ we see that the second inequality should provide a good bound. To estimate $\|\mu\|_\infty$ a calculation along the lines of the bound (2.12) shows that for $\mu \in C^s(D)$ with J_n as in Proposition 2.7

$$\|\widehat{\mu}_{J_n} - \mu\|_{L^\infty([a,b])} = \mathcal{O}_P\left(\left(\frac{\log n}{n}\right)^{s/(2s+1)} u_n\right), \tag{2.15}$$

where $u_n = w_{J_n}/\sqrt{J_n}$. Provided the supremum of μ is attained in $[a,b]$ or μ admits some positive global Hölder regularity, we conclude that $\|\mu\|_\infty$ can be estimated by $\|\widehat{\mu}_{J_n}\|_\infty$ with the above rate and is in particular a consistent estimator, which is all that is needed. For the estimation of ρ we observe that it is the second largest eigenvalue of the transition operator P_Δ. Gobet *et al.* [20] estimate this eigenvalue in a reflected diffusion model by constructing first an empirical transition matrix for the transition operator restricted to a finite dimensional space and then taking the second largest eigenvalue of the empirical transition matrix as an estimator for ρ, there denoted by κ_1. They give a rate for their estimator, in particular the estimator is consistent.

Let us finally note that the estimation of ζ_α can be circumvented by a Bayesian approach as studied by Castillo and Nickl [8] as well as Szabó *et al.* [42] in simpler statistical problems. The papers analyse Bayesian credible sets in the density estimation model and in the white noise regression model as well as in the Gaussian sequence model and show that they are frequentist confidence sets. Estimating the drift of a diffusion from low−frequency observations is a more complicated statistical model. Consistency of the Bayesian approach in this setting has been established by van der Meulen and van Zanten [44] and has been extended to the multi-dimensional case by Gugushvili and Spreij [21]. Recently Nickl and Söhl [36] have shown Bayesian posterior contraction rates for scalar diffusions with unknown drift and unknown diffusion coefficient observed at low frequency.

3. Application to diffusion processes

3.1. Estimation of the invariant density and its consequences

We now apply the results from the previous section to diffusion processes. At the same time we extend the results from inference on the invariant probability measure to confidence bands for the drift function. Let us consider the diffusion

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \qquad t \geqslant 0, \quad X_0 = x, \tag{3.1}$$

with a Brownian motion W_t, an unknown drift function $b\colon \mathbb{R} \to \mathbb{R}$, a volatility parameter $\sigma > 0$ and starting point $x \in \mathbb{R}$. We observe X at equidistant time points $0, \Delta, 2\Delta, \dots, (n-1)\Delta$ for some fixed observation distance $\Delta > 0$ and sample size $n \to \infty$. Our aim is inference on the drift b.

Underlying the sequence of observations $(X_{\Delta k})_{k \geqslant 0}$ is a Markov structure described by the transition operator

$$P_\Delta f(x) := \mathbb{E}[f(X_\Delta)|X_0 = x].$$

The semi-group $(P_t : t \geqslant 0)$ has the infinitesimal generator L on the space of twice continuously differentiable functions given by

$$Lf(x) = L_b f(x) := b(x)f'(x) + \tfrac{\sigma^2}{2}f''(x). \tag{3.2}$$

If there is an invariant density $\mu = \mu_b$, the operator L is symmetric with respect to the scalar product of $L^2(\mu) = \{f : \int |f|^2 \mathrm{d}\mu < \infty\}$. We impose the following assumptions on the diffusion:

Assumption 3.1. In model (3.1) let b be continuously differentiable and satisfy $b \in C^s(D)$ for $s \geqslant 1$ and a sufficiently large set $D \subseteq \mathbb{R}$ containing the interval $[a, b]$ for $a < b$. Let σ be in a fixed bounded interval away from the origin. Suppose that b' is bounded and that there are $M, r > 0$ such that

$$\operatorname{sign}(x) b(x) \leqslant -r, \quad \text{for all } |x| \geqslant M.$$

More precisely, we will need $D = [a - 2^{1-j_0}(2N-1), b + 2^{1-j_0}(2N-1)]$. Due to the global Lipschitz continuity and the assumptions on the drift, equation (3.1) has a unique strong solution. Moreover, X_t is a Markov process with invariant probability density given by

$$\mu(x) = C_0 \sigma^{-2} \exp\left(2\sigma^{-2} \int_0^x b(y)\mathrm{d}y\right), \quad x \in \mathbb{R}, \tag{3.3}$$

with normalization constant $C_0 > 0$ (*cf.* [4], Chaps. 1,4). The corresponding Markov chain Z with $Z_k = X_{k\Delta}$ satisfies Assumption 2.4 from the previous section.

Proposition 3.2. *If the diffusion process* (3.1) *satisfies Assumption* 3.1*, then the Markov chain* $(X_{k\Delta})_{k \geqslant 0}$ *satisfies Assumption* 2.4 *where* $\mu \in C^{s+1}(D)$.

Proof. By a time-change argument we can set $\sigma = 1$ without loss of generality. Gihman and Skorohod [17], Thm. 13.2) have given an explicit formula for the transition density $p_\Delta(x, y)$ with respect to the Lebesgue measure (*i.e.*, $P_\Delta(x, B) = \int_B p_\Delta(x, y)\mathrm{d}y$ for all $B \in \mathcal{B}(\mathbb{R})$. In particular, $p_\Delta(x, y)$ is strictly positive and thus Z is ψ-irreducible, where ψ is given by the Lebesgue measure on $\mathbb{R}$.

Moreover, $(x, y) \mapsto p_\Delta(x, y)$ is continuous such that for any compact interval $C \subseteq \mathbb{R}$ we have $\delta := \delta(C) := \inf_{x,y \in C} p_\Delta(x, y) > 0$ and the small set condition (2.3) is satisfied:

$$P_\Delta(x, B) = \int_B p_\Delta(x, y)\mathrm{d}y \geqslant \delta \int_{B \cap C} \mathrm{d}y = \delta |C| \nu(B),$$

where $|C|$ denotes the Lebesgue measure of C and ν is the uniform distribution on C. It also follows that the Markov chain is strongly aperiodic.

To show the drift condition (2.1), we first construct a Lyapunov function for the infinitesimal generator (which is the continuous time analogue of the drift operator $P - \mathrm{Id}$), that is we find a function $V \geqslant 1$ such that

$$LV(x) \leqslant -\lambda V(x) + c\mathbb{1}_C(x), \quad x \in \mathbb{R}. \tag{3.4}$$

Let V be a smooth function with $V(x) = \mathrm{e}^{a|x|}$ for $|x| > R$ for some $R > 0$. Due to the assumptions on b, we then obtain for these x and R large enough

$$LV(x) = \frac{1}{2} V''(x) + b(x) V'(x) = \left(\frac{a^2}{2} + a \operatorname{sign}(x) b(x)\right) V(x) \leqslant -\lambda V(x)$$

for sufficiently small a, λ and thus the previous inequality is satisfied with $C = [-R, R]$. To carry this result over to the drift condition (2.1), we adopt the approach by Galtchouk and Pergamenshchikov ([15], Prop. 6.4): itô's formula yields for all $0 \leqslant t \leqslant \Delta$

$$V(X_t) = V(x) + \int_0^t L(V)(X_s)\mathrm{d}s + \int_0^t V'(X_s)\mathrm{d}W_s.$$

We note that Fubini's theorem yields $\mathbb{E}_\mu[\int_0^\Delta V'(X_s)^2 \mathrm{d}s] = \int_0^\Delta \mathbb{E}_\mu[V'(X_0)^2]\mathrm{d}s < \infty$ for constants a small enough by (3.3) and by the assumptions on b. Consequently we have $\mathbb{E}_x[\int_0^\Delta V'(X_s)^2 \mathrm{d}s] < \infty$ for almost all $x \in \mathbb{R}$.

By the explicit formula of $p_\Delta(x,y)$ we conclude that $\mathbb{E}_x[\int_0^\Delta V'(X_s)^2\mathrm{d}s] < \infty$ for all $x \in \mathbb{R}$. Hence, the stochastic integral is a martingale (under P_x) and $Z(t) := P_t V(x)$ satisfies

$$Z'(t) = \mathbb{E}_x[L(V)(X_t)] = -\lambda Z(t) + \psi(t), \quad \psi(t) := \mathbb{E}_x[L(V)(X_t) + \lambda V(X_t)],$$

where we have $\psi(t) \leqslant cP_x(X_t \in C) \leqslant c$ by (3.4). Solving this differential equation, we obtain for all $t \in [0,\Delta]$

$$Z(t) = Z(0)\mathrm{e}^{-\lambda t} + \int_0^t \mathrm{e}^{-\lambda(t-s)}\psi(s)\mathrm{d}s \leqslant V(x)\mathrm{e}^{-\lambda t} + c\frac{1-\mathrm{e}^{-\lambda\Delta}}{\lambda}.$$

Therefore, the drift condition follows:

$$P_\Delta V(x) - V(x) \leqslant (\mathrm{e}^{-\lambda\Delta} - 1)V(x) + \frac{c}{\lambda} \leqslant -\widetilde{\lambda}V(x) + \frac{c}{\lambda}\mathbb{1}_{\{|x|\leqslant R\}}(x),$$

where $R > 0$ and $\widetilde{\lambda} > 0$ are chosen such that $(1 - \mathrm{e}^{-\lambda\Delta} - \widetilde{\lambda})V(x) > c/\lambda$ for $|x| > R$. In combination with the ψ-irreducibility the drift condition shows that the Markov chain is positive Harris recurrent.

Since our diffusion is symmetric, in the sense that the transition operator is symmetric with respect to $L^2(\mu)$, we argue as Bakry *et al.* ([3] Sect. 4.3), using that the Poincaré inequality is implied by a Lyapunov–Poincaré inequality and we thus have the contraction property (2.4) ([3], Thm. 1.3). Finally, the smoothness of b in combination with the formula for the invariant probability density (3.3) imply that μ is in $C^{s+1}(D)$. □

Theorem 2.6 and Proposition 2.7 yield immediately

Corollary 3.3. *Grant Assumption 3.1 and let $w = (w_j)$ be increasing and satisfy $\sqrt{j}/w_j \to 0$. Then the wavelet projection estimator $\widehat{\mu}_{J_n}$ from (2.6) with $2^{J_n} = (n/\log n)^{1/(2s+3)}$ satisfies*

$$\sqrt{n}(\widehat{\mu}_{J_n} - \mu) \xrightarrow{d} \mathbb{G}_\mu \quad \text{in } \mathcal{M}_0(w).$$

Moreover, the confidence band $\overline{\mathcal{C}}_n = \overline{\mathcal{C}}_n(\zeta_\alpha, s+1, u_n)$ from (2.11) with critical value ζ_α such that $P(\|\mathbb{G}_\mu\|_{\mathcal{M}} \geqslant \zeta_\alpha) \leqslant \alpha$ and $u_n = w_{J_n}/\sqrt{J_n}$ satisfies

$$\liminf_{n\to\infty} P(\mu \in \overline{\mathcal{C}}_n) \geqslant 1-\alpha \quad \text{and} \quad |\overline{\mathcal{C}}_n|_\infty = \mathcal{O}_P\left(\left(\frac{n}{\log n}\right)^{-(s+1)/(2s+3)} u_n\right).$$

3.2. Drift estimation *via* plug-in

Supposing from now on that $\sigma = 1$ and rewriting the formula of the invariant measure (3.3), we see that

$$b(x) = \frac{1}{2}\big(\log\mu(x)\big)'. \tag{3.5}$$

Obviously, b depends on μ in a nonlinear way and the estimation problem is ill-posed because b is a function of the derivative μ'. In general, the same calculation leads to a formula for the function $b(x)/\sigma^2$. Note that all shape properties of the drift function, like monotonicity, extrema, *etc.* are already determined by b/σ^2. As demonstrated by Gobet *et al.* [20], the information on σ is encoded in the transition operator of the underlying Markov chain. However, the estimation procedure in this latter article is quite involved and the construction of adaptive confidence bands in the general setting is beyond the scope of the present article. In the following we always set $\sigma = 1$. Note that if we have an estimator for σ at hand, for instance from a short high-frequency time series of the diffusion, the results easily carry over to an unknown volatility $\sigma > 0$.

Denoting the set of continuous functions on the real line by $C(\mathbb{R})$, we introduce the map

$$\xi\colon \big\{f \in C^1(\mathbb{R}) : f > 0, \|f\|_{L^1} = 1\big\} \to C(\mathbb{R}), \quad f \mapsto \frac{f'}{2f},$$

which is one-to-one with inverse function $\xi^{-1}(g) = \exp(2\int_0^{\cdot} g(y)\mathrm{d}y - c_g)$ with normalization constant $c_g \in \mathbb{R}$ and for any function g in the range of ξ. We can thus estimate the drift function of the diffusion by the plug-in estimator $\xi(\widehat{\mu}_{J_n})$.

Using the confidence set $\overline{\mathcal{C}}_n(\zeta_\alpha, s+1, u_n)$ for the invariant density μ from (2.11), a confidence band for the drift can be constructed *via*

$$\mathcal{D}_n := \mathcal{D}_n(\zeta_\alpha, s, u_n) := \big\{\xi(f) : f \in \overline{\mathcal{C}}_n(\zeta_\alpha, s+1, u_n)\big\}. \tag{3.6}$$

Since ξ is one-to-one, an immediate consequence of Corollary 3.3 is that we have for the coverage probability $\liminf_{n\to\infty} P(b \in \mathcal{D}_n) = \liminf_{n\to\infty} P(\mu \in \overline{\mathcal{C}}_n) \geqslant 1-\alpha$. To bound the diameter of $\mathcal{D}_n$, we first note that ξ is locally Lipschitz continuous: for $f, g \in C^1(\mathbb{R})$ both bounded away from zero on $[a,b]$ we have in $L^\infty([a,b])$

$$\begin{aligned}\|\xi(f) - \xi(g)\|_\infty &= \frac{1}{2}\left\|\frac{f'}{f} - \frac{g'}{g}\right\|_\infty \leqslant \frac{1}{2}\left\|\frac{f'-g'}{f}\right\|_\infty + \frac{1}{2}\left\|\frac{g'}{fg}(g-f)\right\|_\infty \\ &\leqslant \frac{1}{2}\|f^{-1}\|_\infty\|f'-g'\|_\infty + \|\xi(g)\|_\infty\|f^{-1}\|_\infty\|f-g\|_\infty \\ &\leqslant \|f^{-1}\|_\infty\left(\tfrac{1}{2} + \|\xi(g)\|_\infty\right)\|f-g\|_{C^1([a,b])}. \end{aligned} \tag{3.7}$$

For $f, g \in \overline{\mathcal{C}}_n$ we conclude in $L^\infty([a,b])$

$$\begin{aligned}\|\xi(f) - \xi(g)\|_\infty &\leqslant \|\xi(f) - \xi(\mu)\|_\infty + \|\xi(g) - \xi(\mu)\|_\infty \\ &\leqslant \left(\frac{1}{2} + \|b\|_\infty\right)\left(\|f^{-1}\|_\infty\|f-\mu\|_{C^1([a,b])} + \|g^{-1}\|_\infty\|g-\mu\|_{C^1([a,b])}\right).\end{aligned}$$

Analogously to (2.12) the choice $2^{J_n} = (n/\log n)^{1/(2s+3)}$ yields

$$\|f-\mu\|_{C^1([a,b])} = \mathcal{O}_P\left(\left(\frac{n}{\log n}\right)^{-s/(2s+3)} u_n\right) \quad \text{for all } f \in \overline{C}_n(\zeta_\alpha, s+1, u_n).$$

We conclude that f^{-1} is uniformly bounded in $L^\infty([a,b])$ for all $f \in \overline{\mathcal{C}}_n$. Hence, we have proved

Proposition 3.4. *Grant Assumption 3.1 with $\sigma = 1, s > 0$ and let $w = (w_j)$ satisfy $\sqrt{j}/w_j \to 0$. Then the confidence set $\mathcal{D}_n = \mathcal{D}_n(\zeta_\alpha, s, u_n)$ from (3.6) with critical value ζ_α satisfying $P(\|\mathbb{G}_\mu\|_{\mathcal{M}} \geqslant \zeta_\alpha) \leqslant \alpha$, $u_n = w_{J_n}/\sqrt{J_n}$ and J_n chosen such that $2^{J_n} = (n/\log n)^{1/(2s+3)}$ fulfils*

$$\liminf_{n\to\infty} P(b \in \mathcal{D}_n) \geqslant 1-\alpha \quad \text{and} \quad \big|\mathcal{D}_n\big|_\infty = \mathcal{O}_P\left(\left(\frac{n}{\log n}\right)^{-s/(2s+3)} u_n\right).$$

Let us comment on the rate appearing in the previous proposition. Since the identification (3.5) incorporates the derivative of the invariant measure, drift estimation is an inverse problem, which is ill-posed of degree one. Therefore, the minimax rate for the pointwise or L^2-loss is $n^{-s/(2s+3)}$. Considering the uniform loss, we obtain the rate $(n/\log n)^{-s/(2s+3)}$. Finally, $u_n \to \infty$ is the payment for undersmoothing (by using a weighting sequence slightly larger than $\sqrt{j}$). Note that we obtain a faster rate than Gobet *et al.* [20] who have proved that the minimax rate for drift estimation for the mean integrated squared error is $n^{-s/(2s+5)}$ if there is additionally an unknown volatility function in front of the Brownian motion in (3.1).

In fact the map ξ is not only Lipschitz continuous, but even Hadamard differentiable (on appropriate function spaces) with derivative at μ

$$\xi'_\mu(h) = \frac{1}{2}\left(\frac{h}{\mu}\right)', \quad h \in \mathcal{M}_0(w). \tag{3.8}$$

Using the delta method ([46], Thm. 20.8), we obtain a functional central limit theorem for the plug-in estimator $\xi(\widehat{\mu}_{J_n})$.

Theorem 3.5. *Grant Assumption 3.1 with $\sigma = 1$ and let $w = (w_j)$ be increasing and satisfy $\sqrt{j}/w_j \to 0$, $w_j \leqslant 2^{j\delta}$ for some $\delta \in (0, 1/2)$. Let $J_n \in \mathbb{N}$ fulfil, for some $\tau \in (0,1)$,*

$$2^{(9/4+\delta/2)J_n} w_{J_n} n^{-1/2} = o(1), \qquad \sqrt{n} 2^{-J_n(2s+3)/2} w_{J_n}^{-1} = o(1), \qquad (\log n)^{2/\tau} n^{-1} 2^{J_n} J_n = \mathcal{O}(1).$$

For $\widetilde{w}_j \geqslant 2^{j(1+\delta)}$ we have as $n \to \infty$

$$\sqrt{n}(\xi(\widehat{\mu}_{J_n}) - b) \xrightarrow{d} \xi'_\mu(\mathbb{G}_\mu) \quad \text{in } \mathcal{M}_0(\widetilde{w}).$$

The proof of this theorem is postponed to Section A.1. Similarly as in (2.11) confidence bands for the drift function can alternatively be constructed by

$$\overline{\mathcal{D}}_n(\overline{\zeta}_\alpha, s, u_n) := \left\{ f : \|f - \xi(\widehat{\mu}_{J_n})\|_{\mathcal{M}(\widetilde{w})} < \frac{\overline{\zeta}_\alpha}{\sqrt{n}}, \|f\|_{C^s} \leqslant u_n \right\},$$

for $\alpha \in (0, 1/2)$, a quantile $\overline{\zeta}_\alpha$ such that $P(\|\xi'_\mu(\mathbb{G}_\mu)\|_{\mathcal{M}(\widetilde{w})} < \overline{\zeta}_\alpha) \geqslant 1 - \alpha$ and a sequence $u_n \to \infty$. With $2^{J_n} = (n/\log n)^{1/(2s+3)}$ and $u_n = \widetilde{w}_{J_n} 2^{-J_n}/\sqrt{J_n}$ this leads to asymptotic coverage of at least $1 - \alpha$ and a diameter decaying at rate $(n/\log n)^{-s/(2s+3)} u_n$. Note that in contrast to $\mathcal{D}_n$ the diameter of $\overline{\mathcal{D}}_n$ is slightly suboptimal due to the $\delta > 0$ that appeared in Theorem 3.5 and which presumably could be removed by a more technical proof. Based on a direct estimator of the drift we will construct a similar confidence band with the optimal diameter (up to undersmoothing) in the next section.

Comparing both constructions of confidence sets, we see that $\mathcal{D}_n$ can be understood as the variance stabilised version of $\overline{\mathcal{D}}_n$: the critical value of $\mathcal{D}_n$ depends on the unknown μ only through the covariance structure of the limit processes $\mathbb{G}_\mu$ which seems to be unavoidable due to the underlying Markov chain structure. In contrast $\overline{\zeta}_\alpha$ depends additionally on μ through the derivative ξ'_μ. As a consequence the confidence band $\overline{\mathcal{D}}_n$ has the same diameter everywhere while the diameter of $\mathcal{D}_n$ changes.

3.3. A direct approach to estimate the drift

Instead of relying only on the estimator $\widehat{\mu}_{J_n}$ of the invariant density and the plug-in approach, we can use a direct approach to estimate the drift and to obtain its confidence bands. Although there is a one-to-one correspondence between the drift function b and the invariant measure μ, the drift is both the canonical parameter of our model and the main parameter of interest in the context of diffusions. Since we aim for adapting to the regularity of b, the direct estimation approach is natural and, additionally, the resulting confidence bands will have a constant diameter.

Motivated by formula (3.5), we define our drift estimator for integers $J, U > 0$ as

$$\widehat{b}_{J,U} = \frac{1}{2}\pi_J\big(\log \widehat{\mu}_{J+U}\big)' = \frac{1}{2}\sum_{j \leqslant J}\sum_{k \in K_j} \Big\langle \big(\log \widehat{\mu}_{J+U}\big)', \psi_{j,k} \Big\rangle \psi_{j,k}, \tag{3.9}$$

using the wavelet projection estimator $\widehat{\mu}_{J+U}$ from (2.6). In contrast to the plug-in estimator in the previous section the underlying bias-variance trade-off is now driven by the estimation problem of b and the outer projection π_J onto level J. However, in order to linearise the estimation error, we need a stable prior estimator of μ such that we cannot simply use the empirical measure μ_n but instead use its projection onto some resolution level $J + U$ which is strictly larger than J. As a rule of thumb, $U = U_n$ can be chosen such that $2^{U_n} = \log n$ implying that an additional bias term from estimating μ is negligible. Linearising the estimation error, we obtain

$$\Big\langle \widehat{b}_{J,U} - b, \psi_{j,k} \Big\rangle = -\Big\langle \widehat{\mu}_{J+U} - \mu, \frac{\psi'_{j,k}}{2\mu} \Big\rangle + \langle R_{J+U}, \psi_{j,k} \rangle, \quad j \leqslant J, k \in K_j, \tag{3.10}$$

where the remainder is of order $o_P(n^{-1/2})$ for appropriate choices of $J = J_n$, *cf.* Lemma 5.1 below. In view of the linear error term and our findings in Section 2, the limit process $\mathbb{G}_b$ in the multi-scale space $\mathcal{M}_0$ will be given by

$$\mathbb{G}_b(j,k) \sim \mathcal{N}(0, \Sigma_{j,k}) \quad \text{where } \Sigma_{j,k} := \Sigma_{f,f} \text{ is given by (2.2) with } f = \psi'_{j,k}/(2\mu) \tag{3.11}$$

with covariances $\mathbb{E}[\mathbb{G}_b(j,k)\mathbb{G}_b(l,m)] = \Sigma_{f_1,f_2}$ from (2.2) with $f_1 = \psi'_{j,k}/(2\mu)$ and $f_2 = \psi'_{l,m}/(2\mu)$. The ill-posedness of the problem is reflected by $\psi'_{j,k}$ being a factor 2^j larger than $\psi_{j,k}$. We thus need larger weights for high resolution levels to ensure that $\mathbb{G}_b$ takes values in $\mathcal{M}_0(w)$.

Definition 3.6. A weighting sequence $w = (w_j)$ is called *admissible*, if it is monotonously increasing, satisfies $\sqrt{j}2^j/w_j \to 0$ as $j \to \infty$ and if there is some $\delta \in (1,2]$ such that $j \mapsto 2^{j\delta}/w_j$ is monotonously increasing for large j.

The last condition in the definition is a mild technical assumption that we will need in the multi-scale central limit theorem below. For instance, any weighting sequence $w_j = u_j\sqrt{j}2^j$ with $u_j = j^p$ for some polynomial rate $p > 0$ is admissible of degree one for any $\delta \in (1,2]$. Note that admissibility of w implies in particular that $w_j \lesssim 2^{j\delta}$ which allows to compare the $\|\cdot\|_\infty$-norm with the $\|\cdot\|_{\mathcal{M}}$-norm. We find an analogous result to Lemma 2.3 (*cf.* [8], Prop. 3).

Lemma 3.7. *$\mathbb{G}_b$ from (3.11) satisfies $\mathbb{E}[\|\mathbb{G}_b\|_{\mathcal{M}(w)}] < \infty$ for the weights w given by $w_j = \sqrt{j}2^j$. Moreover, $\mathcal{L}(\mathbb{G}_b)$ is a tight Gaussian Borel probability measure in $\mathcal{M}_0(w)$ for any admissible sequence w.*

For the following result suppose that the wavelet basis $(\varphi_{j_0,l}, \psi_{j,k} : j \geqslant j_0, l \in L, k \in K_j)$ of $L^2(\mathbb{R})$ is sufficiently regular (*i.e.*, satisfies (5.2) with $\gamma \geqslant 3/2 + \delta$), for instance, Daubechies' wavelets of order $N \geqslant 20$.

Theorem 3.8. *Grant Assumption 3.1 with $\sigma = 1$ and let $w = (w_j)$ be admissible. Let $J_n \to \infty, U_n \to \infty$ fulfil*

$$\sqrt{n}2^{-J_n(s+1/2)}w_{J_n}^{-1} = o(1), \quad \sqrt{n}2^{-(J_n+U_n)(2s+1)} = o(1), \quad n^{-1/2}2^{2(J_n+U_n)}(J_n+U_n) = o(1).$$

Then $\widehat{b}_{J,U}$ from (3.9) satisfies, as $n \to \infty$,

$$\sqrt{n}(\widehat{b}_{J_n,U_n} - b) \xrightarrow{d} \mathbb{G}_b \quad \text{in } \mathcal{M}_0(w)$$

for the tight Gaussian random variable $\mathbb{G}_b$ in $\mathcal{M}_0(w)$ given by (3.11).

The proof of this theorem is postponed to Section 5. The first condition on J_n is the bias condition for b in $\mathcal{M}_0$. The latter two conditions on $J_n + U_n$ are determined by a bias and a variance condition for μ which we will need to bound the remainder $R_{J_n+U_n}$ from (3.10) in L^∞. If $\delta < 1/2 + s$ in Definition 3.6, then the second condition is strictly weaker than the first one.

Similarly to the confidence band for μ in Proposition 2.7 we can now construct a confidence band for the drift function b. For some $\alpha \in (0,1)$ we consider

$$\mathcal{E}_n := \mathcal{E}_n(\zeta_\alpha, s, u_n) := \left\{ f : \|f - \widehat{b}_{J_n,U_n}\|_{\mathcal{M}} < \frac{\zeta_\alpha}{\sqrt{n}}, \|f\|_{C^s} \leqslant u_n \right\}, \tag{3.12}$$

where ζ_α is chosen such that $P(\|\mathbb{G}_b\|_{\mathcal{M}} < \zeta_\alpha) \geqslant 1 - \alpha$ and $(u_n)_n$ is a diverging sequence.

Proposition 3.9. *Grant Assumption 3.1 with $\sigma = 1, s \geqslant 1$ and let $w = (w_j)$ be admissible. For $\alpha \in (0,1)$ let $\zeta_\alpha > 0$ satisfy $P(\|\mathbb{G}_b\|_{\mathcal{M}} \geqslant \zeta_\alpha) \leqslant \alpha$ and choose $J_n := J_n(s)$ and $U_n \to \infty$ such that*

$$2^{J_n} = \left(\frac{n}{\log n}\right)^{1/(2s+3)} \quad \text{and} \quad 2^{U_n} = \mathcal{O}(\log n).$$

Then the confidence set $\mathcal{E}_n = \mathcal{E}_n(\zeta_\alpha, s, u_n)$ from (3.12) with $u_n := w_{J_n}2^{-J_n}/\sqrt{J_n}$ satisfies

$$\liminf_{n\to\infty} P(b \in \mathcal{E}_n) \geqslant 1 - \alpha \quad \text{and} \quad |\mathcal{E}_n|_\infty = \mathcal{O}_P\left(\left(\frac{n}{\log n}\right)^{-s/(2s+3)} u_n\right).$$

Proof. The proof is essentially the same as for the confidence band of the invariant probability density. We show that the asymptotic coverage probability is at least $1-\alpha$ and obtain for $f, g \in \mathcal{E}_n$ as in (2.12) the bound

$$\|f-g\|_\infty = \mathcal{O}_P\Big(n^{-1/2}2^{J_n/2}w_{J_n}\Big) + \mathcal{O}_P\Big(2^{-J_n s}u_n\Big)$$

Using $u_n = w_{J_n}2^{-J_n}/\sqrt{J_n}$ we thus have

$$\|f-g\|_\infty = \mathcal{O}_P\Big(n^{-1/2}2^{3J_n/2}J_n^{1/2}u_n\Big) + \mathcal{O}_P\Big(2^{-J_n s}u_n\Big).$$

The choice of J_n yields $n^{-1/2}\sqrt{2^{3J_n}J_n} \lesssim (n/\log n)^{-s/(2s+3)} = 2^{-J_n s}$. □

4. Adaptive confidence bands for drift estimation

Inspired by Giné and Nickl [18], we will now construct an adaptive version of the confidence set $\mathcal{E}_n$ from (3.12). To this end we estimate the regularity s of the drift with a Lepski-type method. For some maximal regularity $r > 1$, let the integers $0 < J_{\min} < J_{\max}$ be given by

$$2^{J_{\min}} \sim \left(\frac{n}{\log n}\right)^{1/(2r+3)}, \quad 2^{J_{\max}} \sim \frac{n^{1/4}}{(\log n)^2}.$$

Note that $J_{\min}, J_{\max}$ depend on the sample size n, which is suppressed in the notation. If we knew in advance that b has regularity r, then we would choose the resolution level $J_{\min}$. The upper bound $J_{\max}$ is chosen such that $J_{\max}+U_n$ satisfies the third condition in Theorem 3.8. The set in which we will adaptively choose the optimal resolution level for regularities $s \in [1, r]$ is defined by

$$\mathcal{J}_n := [J_{\min}, J_{\max}] \cap \mathbb{N}.$$

Similar to Giné and Nickl ([18], Lem. 2), we show under the following assumption on b that the optimal truncation level can be consistently estimated up to a fixed integer.

Assumption 4.1. Let $b \in C^s(D), s \geqslant 1$, satisfy for constants $0 < d_1 < d_2 < \infty$ and an integer $J_0 > 0$ that

$$d_1 2^{-Js} \leqslant \|\pi_J(b) - b\|_{L^\infty([a,b])} \leqslant d_2 2^{-Js}, \quad \forall J \geqslant J_0. \tag{4.1}$$

The second inequality in (4.1) is the well known Jackson inequality which is satisfied for all usual choices of wavelet basis. The first inequality is the main condition here, called *self-similarity assumption*. It excludes the cases where the bias would be smaller than the usual order 2^{-Js}. Although the estimator $\widehat{b}_{J,U}$ would profit from a smaller bias, we cannot hope for a consistent estimation of the optimal projection level and the resulting regularity index s if (4.1) (or a slight generalisation by Bull [5]) is violated. Indeed, Hoffmann and Nickl [26] have shown that this kind of condition is necessary to construct adaptive and honest confidence bands. On the other hand, it has been proved by Giné and Nickl [18] that the set of functions that do not satisfy the self-similarity assumption is nowhere dense in the Hölder norm topologies. In that sense, the self-similarity assumption is satisfied by "typical" functions. We will give an illustrative example next. Probabilistic examples for self-similar functions are those Gaussian processes which can be represented as stochastic series expansions like the Karhunen–Loève expansion for Brownian motion or typical examples of Bayesian priors. Naturally, more regular functions $b \in C^r(D)$ for some $r > s$ cannot satisfy Assumption 4.1. For a further discussion and examples we refer to (Giné and Nickl [18], Sect. 3.5) as well as Bull [5].

Example 4.2. Let b be a smooth function on $\mathbb{R}$ except for some point $x_0 \in (a,b)$ where the sth order derivative $b^{(s)}$ has a jump for some integer $s \geqslant 1$. Locally around x_0 the function b can be approximated by a Taylor polynomial $b(x) = \beta_s^{\pm}(x-x_0)^s + \sum_{l=0}^{s-1}\beta_l(x-x_0)^l + \mathcal{O}(|x-x_0|^{s+1})$ with coefficients $\beta_0, \dots, \beta_{s-1} \in \mathbb{R}$ and where

the coefficient of order s is some $\beta_s^+ \in \mathbb{R}$ or $\beta_s^- \in \mathbb{R}$ depending on $x \geqslant x_0$ and $x < x_0$, respectively. Due to the jump of $b^{(s)}$ we have $\beta_s^+ \neq \beta_s^-$.

Choose k_j as the nearest integer of $2^j x_0$, implying that x_0 is in the middle of the support of ψ_{j,k_j}. Using the elementary estimate $|\langle f, \psi_{j,k}\rangle| \leqslant \|f\|_{L^\infty(\operatorname{supp}\psi_{j,k})} 2^{-j/2}\|\psi\|_{L^1}$, we obtain

$$\|\pi_J(b) - b\|_{L^\infty([a,b])} = \left\| \sum_{j>J,k} \langle b, \psi_{j,k}\rangle \psi_{j,k} \right\|_{L^\infty([a,b])} \gtrsim \sup_{j>J} 2^{j/2} |\langle b, \psi_{j,k_j}\rangle|.$$

For sufficiently large j the regularity of the wavelet basis, being thus orthogonal to polynomials, and their compact support yield

$$\begin{aligned} 2^{j/2}|\langle b, \psi_{j,k_j}\rangle| &\geqslant 2^j \left| \int \big(b(x) - \beta_s^-(x-x_0)^s + \sum_{l=0}^{s-1} \beta_l (x-x_0)^l\big)\psi(2^j x - k_j)\mathrm{d}x \right| + \mathcal{O}(2^{-j(s+1)}) \\ &\geqslant 2^j |\beta_s^+ - \beta_s^-| \left| \int_{x>x_0} (x-x_0)^s \psi(2^j x - k_j)\mathrm{d}x \right| + \mathcal{O}(2^{-j(s+1)}) \\ &\geqslant 2^{-js} |\beta_s^+ - \beta_s^-| \min_{\varepsilon \in [-1/2, 1/2]} \left| \int_{y>0} y^s \psi(y+\varepsilon)\mathrm{d}y \right| + \mathcal{O}(2^{-j(s+1)}). \end{aligned}$$

We conclude for J sufficiently large that $\|\pi_J(b) - b\|_{L^\infty([a,b])} \gtrsim 2^{-Js}$.

The oracle choice J_n^* which balances the bias $\|\pi_J(b) - b\|_{L^\infty([a,b])}$ and the main stochastic error is given by

$$J_n^* := J_n^*(s) = \min\left\{ J \in \mathcal{J}_n : (d_2+1)2^{-Js} \leqslant \frac{K}{4}\sqrt{\frac{2^{3J}J}{n}} \right\}$$

for some suitable constant $K > 0$ depending only on ψ, $\inf\{\mu(x) : x \in \cup_{l\in L} \operatorname{supp}\varphi_{j_0,l}\}$ and the maximal asymptotic variance $\widetilde{\Sigma} = \sup_{j,k} \Sigma_{j,k}$ where the latter two quantities can be replaced by the consistent estimators which we have discussed in Section 2.3. We see easily that $2^{J_n^*} \sim (\frac{n}{\log n})^{1/(2s+3)}$. Following Lepski's approach, we define the estimator for J_n^* by

$$\widehat{J}_n = \min\left\{ J \in \mathcal{J}_n : \|\widehat{b}_J - \widehat{b}_j\|_{L^\infty([a,b])} \leqslant K\sqrt{\frac{2^{3j}j}{n}} \ \forall j > J, j \in \mathcal{J}_n \right\}. \tag{4.2}$$

Lemma 4.3. *Grant Assumptions 3.1 and 4.1 for $s \in [1, r]$ with some $r > 1$ and $\sigma = 1$. Let w be admissible. Then there are a constant $K > 0$ depending only on ψ, $\inf\{\mu(x) : x \in \cup_{l\in L} \operatorname{supp}\varphi_{j_0,l}\}$ and the maximal asymptotic variance $\widetilde{\Sigma} = \sup_{j,k} \Sigma_{j,k}$, an integer $M > 0$ depending only on d_1, d_2, K and for any $\tau \in (0,1)$ there are constants $C, c > 0$ depending on τ, K, ψ such that*

$$P\big(\widehat{J}_n \notin [J_n^* - M, J_n^*]\big) \leqslant C\big(n^{-cJ_{\min}^\tau} + \mathrm{e}^{-cJ_{\min}}\big) \to 0.$$

The proof of this lemma relies on the concentration result in Proposition 2.2 and is postponed to Section A.2. Applying that $\widehat{J}_n$ is a reasonable estimator of J_n^*, we obtain a completely data-driven estimator

$$\widehat{b} := \widehat{b}_{\widehat{J}_n, U_n}, \quad \text{with } \widehat{J}_n \text{ from (4.2) and } 2^{U_n} = \log n. \tag{4.3}$$

Corollary 4.4. *In the situation of Lemma 4.3 the adaptive estimator $\widehat{b}$ defined by (4.3) satisfies $\|\widehat{b} - b\|_{\mathcal{M}} = \mathcal{O}_P(n^{-1/2})$ and $\|\widehat{b} - b\|_{L^\infty([a,b])} = \mathcal{O}_P\big((n/\log n)^{-s/(2s+3)} u_n\big)$ with $u_n := w_{J_n^*} 2^{-J_n^*}/\sqrt{J_n^*}$. Further for every $m \in \{0, 1, \ldots, M\}$ we have*

$$\sqrt{n}\left(\widehat{b}_{J_n^* - m, U_n} - b\right) \xrightarrow{d} \mathbb{G}_b \quad \text{in } \mathcal{M}_0(w)$$

as $n \to \infty$ for the tight Gaussian random variable $\mathbb{G}_b$ in $\mathcal{M}_0(w)$ given by (3.11).

Proof. Combining Lemma 4.3 and Theorem 3.8, there is for any $\delta > 0$ a constant $C > 0$ such that for n large enough

$$P\big(\sqrt{n}\|\widehat{b} - b\|_{\mathcal{M}} > C\big) \leqslant \sum_{J=J_n^*-M}^{J_n^*} P\big(\sqrt{n}\|\widehat{b}_J - b\|_{\mathcal{M}} > C\big) + o(1) \leqslant (M+1)\delta.$$

Since M is a finite constant, we have $\|\widehat{b}-b\|_{\mathcal{M}} = \mathcal{O}_P(n^{-1/2})$. Using that $2^{\widehat{J}_n} \sim 2^{J_n^*}$, a calculation similar to (2.12) yields the bound for the uniform norm. For the second claim notice that the estimators $\widehat{b}_{J_n^*-m,U_n}$ satisfy the conditions of Theorem 3.8. □

The bound for the uniform risk is slightly suboptimal because u_n diverges arbitrary slowly (depending on the choice of w) to infinity. Using direct estimates of the $\|\cdot\|_\infty$-norm in the proofs in Section 5, this additional factor could be circumvented. However, it can be interpreted as an additional factor that corresponds to a slight undersmoothing which is often used to have a negligible bias in the construction of confidence bands.

Another consequence of Lemma 4.3 is that we can consistently estimate the regularity s of b. For a sequence of random variables (v_n) with $v_n^{-1} = o_P(1)$ we define the estimator

$$\widehat{s}_n := \max\left(1, \frac{\log n - \log\log n}{2(\log 2)(\widehat{J}_n + v_n)} - \frac{3}{2}\left(1 + \frac{v_n}{\widehat{J}_n}\right)\right). \tag{4.4}$$

Using that $2^{J_n^*} \sim (n/\log n)^{1/(2s+3)}$, we derive from Lemma 4.3 the following corollary. The proof can be found in Section A.3.

Corollary 4.5. *In the situation of Lemma 4.3 the estimator $\widehat{s}_n$ given by (4.4) satisfies for any sequence of random variables (v_n) with $v_n^{-1} = o_P(1)$*

$$P(\widehat{s}_n \leqslant s) \to 1 \quad \text{and} \quad s - \widehat{s}_n = \mathcal{O}_P\left(\frac{v_n}{J_n^*}\right).$$

With the estimator $\widehat{b}$ from (4.3) we can now construct our adaptive confidence bands as follows. By a Bonferroni correction we take care of the possible dependence between the estimators $\widehat{b}_{J_n^*-m,U_n}$ and the adaptive choice $\widehat{J}_n$ of the resolution level. In this way sample splitting can be avoided, which was also used by Bull [5]. For any level $\alpha \in (0,1)$ let $\beta = \alpha/(M+1)$ and define

$$\widetilde{\mathcal{E}}_n := \widetilde{\mathcal{E}}_n(\widehat{\zeta}_{\beta,n}, \widehat{s}_n, t_n) := \left\{ f : \|f - \widehat{b}\|_{\mathcal{M}} < \frac{\widehat{\zeta}_{\beta,n}}{\sqrt{n}}, \|f\|_{C^{\widehat{s}_n}} \leqslant t_n \right\}, \tag{4.5}$$

where (t_n) is a sequence of random variables with $t_n^{-1} = o_P(1)$ and $\widehat{\zeta}_{\beta,n}$ is an (over-)estimator of the critical value ζ_β given by $P(\|\mathbb{G}_b\|_{\mathcal{M}} < \zeta_\beta) \geqslant 1-\beta$ similarly to the construction in Section 2.3. Now we can state our final theorem:

Theorem 4.6. *Grant Assumptions 3.1 and 4.1 for $\sigma = 1, s \in [1, r]$ with some $r > 1$. Let $w = (w_j)$ be admissible and define $u_n := w_{J_n^*} 2^{-J_n^*}/\sqrt{J_n^*}$ as well as $\widehat{u}_n := w_{\widehat{J}_n} 2^{-\widehat{J}_n} \widehat{J}_n^{-1/2}$. For $\alpha \in (0,1)$ set $\beta := \alpha/(M+1)$ with M as in Lemma 4.3. Let $\zeta_\beta > 0$ be given by $P(\|\mathbb{G}_b\|_{\mathcal{M}} \geqslant \zeta_\beta) \leqslant \beta$ and let $\widehat{\zeta}_{\beta,n}$ be an (over-)estimator satisfying $P(\widehat{\zeta}_{\beta,n} \geqslant \zeta_\beta - \varepsilon) \to 1$ for all $\varepsilon > 0$. Then the confidence set $\widetilde{\mathcal{E}}_n = \widetilde{\mathcal{E}}_n(\widehat{\zeta}_{\beta,n}, \widehat{s}_n, t_n)$ given by (4.5), where we choose $t_n := \sqrt{\widehat{u}_n}$ and $\widehat{s}_n$ according to (4.4) with $v_n = o_P(\log \widehat{u}_n)$ and $v_n^{-1} = o_P(1)$, satisfies*

$$\liminf_{n\to\infty} P(b \in \widetilde{\mathcal{E}}_n) \geqslant 1 - \alpha \quad \text{and} \quad |\widetilde{\mathcal{E}}_n|_\infty = \mathcal{O}_P\left(\left(\frac{n}{\log n}\right)^{-s/(2s+3)} u_n\right).$$

Proof. We will adapt the proof of Proposition 3.9 to the estimated quantities $\widehat{J}_n, \widehat{s}_n$ and $\widehat{\zeta}_\beta$. By Corollary 4.5 the probability of the event $\{\widehat{s}_n \leqslant s\}$ converges to one. Due to $v_n^{-1} = o_P(1)$, we thus have $b \in C^{\widehat{s}_n}(v_n)$ with probability tending to one. Using Lemma 4.3 and Corollary 4.4, we infer

$$\limsup_{n\to\infty} P(b \notin \widetilde{\mathcal{E}}_n) = \limsup_{n\to\infty} \sum_{m=0}^{M} P\Big(\sqrt{n}\|\widehat{b}_{J_n^*-m,U_n} - b\|_{\mathcal{M}} \geqslant \widehat{\zeta}_{\beta,n}\Big) \leqslant (M+1)P(\|\mathbb{G}_b\|_{\mathcal{M}} \geqslant \zeta_\beta) \leqslant \alpha.$$

We conclude that $\liminf_{n\to\infty} P(b \in \widetilde{\mathcal{E}}_n) \geqslant 1-\alpha$.

To estimate the diameter, we proceed as in (2.12). Applying additionally Corollary 4.5, we obtain for any $f, g \in \widetilde{\mathcal{E}}_n$

$$\begin{aligned}
\|f-g\|_\infty &\lesssim \sum_{j\leqslant J_n^*} 2^{j/2} \max_k |\langle f-g, \psi_{j,k}\rangle| + \sum_{j>J_n^*} 2^{j/2} \max_k |\langle f-g, \psi_{j,k}\rangle| \\
&\leqslant \big(\|f-\widehat{b}\|_{\mathcal{M}} + \|g-\widehat{b}\|_{\mathcal{M}}\big) \sum_{j\leqslant J_n^*} 2^{j/2} w_j + \|f-g\|_{C^{\widehat{s}_n}} \sum_{j>J_n^*} 2^{-j\widehat{s}_n} \\
&= \mathcal{O}_P\big(n^{-1/2} 2^{J_n^*/2} w_{J_n^*}\big) + \mathcal{O}_P(t_n 2^{-J_n^* s + \mathcal{O}_P(v_n)}) \\
&= \mathcal{O}_P\Big(n^{-1/2} 2^{3J_n^*/2} (J_n^*)^{1/2} u_n + 2^{-J_n^* s} \widehat{u}_n\Big),
\end{aligned}$$

where we have plugged in the choices of t_n, u_n and v_n and $\widehat{u}_n \lesssim u_n$ with probability converging to one. Since $2^{J_n^*} \sim (n/\log n)^{1/(2s+3)}$, the assertion follows. □

The confidence bands are constructed explicitly and this helps to verify that the confidence bands are *honest*, *i.e.* the coverage is achieved uniformly over some set of the unknown parameter. The general philosophy being that uniformity in the assumptions leads to uniformity in the statements, the detailed derivation of honesty is tedious so that we only sketch it here. The main ingredients of the proof are the central limit theorem and the concentration inequality for Markov chains. In the original version of the concentration inequality, Theorem 9 by Adamczak and Bednorz [1], the constants are given explicitly in terms of the assumptions and thus the concentration inequality is uniform in the underlying Markov chain Z. It is also to be expected that the central limit theorem holds uniformly in the bounded-Lipschitz metric with respect to Z although this is not explicitly contained in the statement. With these uniform ingredients a uniform version of Theorem 2.6 can be proved, where the convergence in distribution is again metrised in the bounded-Lipschitz metric. In combination with a uniform bound on the Lebesgue densities of $\|\mathbb{G}_\mu\|_{\mathcal{M}}$ this leads to honest confidence bands in Proposition 2.7. Thanks to the explicit derivation of Assumption 2.4 from Assumption 3.1, uniformity in the diffusion model carries over to uniformity in the Markov chain and we see that the confidence bands in Proposition 3.4 are honest. Likewise a uniform version of Theorem 3.8 can be proved. Provided the random variables $\|\mathbb{G}_b\|_{\mathcal{M}}$ have uniformly bounded Lebesgue densities this uniform version entails honest and adaptive confidence bands for the drift in Theorem 4.6.

5. Proof of Theorem 3.8

In the sequel we use the notation

$$J^+ = J + U$$

for the projection level of $\widehat{\mu}_{J^+}$ and we define

$$S := \bigcup_{l\in L} \operatorname{supp} \varphi_{j_0,l} \subseteq [a - 2^{-j_0}(2N-1), b + 2^{-j_0}(2N-1)].$$

To analyse the estimation error of the wavelet coefficients $\langle \widehat{b}_{J,U}, \psi_{j,k}\rangle$, we apply the following linearisation lemma:

Lemma 5.1. *Grant Assumption 3.1 with $\sigma = 1$. For $j \in \{-1, j_0, \dots, J\}$ and $k \in K_j$ we have*

$$\langle \widehat{b}_{J,U} - b, \psi_{j,k}\rangle = -\big\langle \widehat{\mu}_{J^+} - \mu, \frac{\psi'_{j,k}}{2\mu}\big\rangle + \langle R_{J^+}, \psi_{j,k}\rangle,$$

where the remainder is given by $R_{J^+} = -\frac{\widehat{\mu}_{J^+}-\mu}{2\widehat{\mu}_{J^+}}\Big(\frac{\widehat{\mu}_{J^+}-\mu}{\mu}\Big)'$. If $J^+ = J_n^+$ satisfies for some $\tau \in (0,1)$

$$2^{-J_n^+(s+1)} = o(1), \qquad (\log n)^{2/\tau} n^{-1} 2^{J_n^+} J_n^+ = \mathcal{O}(1),$$

then

$$\|R_{J_n^+}\|_{L^\infty(S)} = \mathcal{O}_P\big(n^{-1} J_n^+ 2^{2J_n^+} + 2^{-J_n^+(2s+1)}\big).$$

Proof. Writing $\eta := (\widehat{\mu}_{J^+} - \mu)/\mu$, the chain rule yields

$$\frac{1}{2}(\log \widehat{\mu}_{J^+})' - b = \frac{1}{2}\big(\log(1+\eta)\big)' = \frac{\eta'}{2(1+\eta)} = \frac{1}{2}\Big(\frac{\widehat{\mu}_{J^+}-\mu}{\mu}\Big)' + R_{J^+},$$

where the remainder is given by

$$R_{J^+} := -\frac{\eta\eta'}{2(1+\eta)} = -\frac{\widehat{\mu}_{J^+}-\mu}{2\widehat{\mu}_{J^+}}\left(\frac{\widehat{\mu}_{J^+}-\mu}{\mu}\right)'.$$

Using integration by parts with vanishing boundary terms, the wavelet coefficients corresponding to the linear term can be written as

$$\frac{1}{2}\langle((\widehat{\mu}_{J^+}-\mu)\mu^{-1})', \psi_{j,k}\rangle = -\frac{1}{2}\langle \widehat{\mu}_{J^+} - \mu, \psi'_{j,k}\mu^{-1}\rangle.$$

Let us bound the remainder, starting with $\|\widehat{\mu}'_{J_n^+} - \mu'\|_{L^\infty(S)}$. Decomposing the uniform error into a bias and a stochastic error term, we obtain

$$\begin{aligned}\|\widehat{\mu}'_{J_n^+} - \mu'\|_{L^\infty(S)} &\leqslant \Big\|\sum_{j\leqslant J_n^+,k}\langle \mu_n - \mu, \psi_{j,k}\rangle\psi'_{j,k}\Big\|_{L^\infty(S)} + \Big\|\sum_{j>J_n^+,k}\langle \mu, \psi_{j,k}\rangle\psi'_{j,k}\Big\|_{L^\infty(S)}\\ &=: V_n + B_n.\end{aligned}$$

Using the localisation property of the wavelet function $\|\sum_k |\psi(\bullet - k)|\|_\infty \lesssim 1$ (which holds for ψ' as well) and the regularity of $\mu \in C^{s+1}(D)$, implying $\sup_{j,k:\operatorname{supp}\psi_{j,k}\cap S\neq\emptyset} 2^{j(s+3/2)}|\langle \mu, \psi_{j,k}\rangle| < \infty$, the bias can be estimated by

$$B_n \lesssim \sum_{j>J_n^+}\max_{k:\operatorname{supp}\psi_{j,k}\cap S\neq\emptyset}|\langle \mu, \psi_{j,k}\rangle|\,\Big\|\sum_k|\psi'_{j,k}|\Big\|_{L^\infty(S)} \lesssim \sum_{j>J_n^+} 2^{-js} \lesssim 2^{-J_n^+ s}.$$

For the stochastic error term we obtain similarly

$$V_n \lesssim \sum_{j\leqslant J_n^+} 2^{3j/2}\max_{k:\operatorname{supp}\psi_{j,k}\cap S\neq\emptyset}|\langle \mu_n - \mu, \psi_{j,k}\rangle|.$$

The maximum of 2^j subgaussian random variables is of order $\mathcal{O}_P(\sqrt{j})$. More precisely, Proposition 2.2 and the assumptions on J_n^+ yield for any $j_0 \leqslant j \leqslant J_n^+$ and $\tau \in (0,1)$, similarly to (2.10),

$$\begin{aligned}P\left(\max_{k:\operatorname{supp}\psi_{j,k}\cap S\neq\emptyset}\sqrt{n}|\langle \mu_n - \mu, \psi_{j,k}\rangle| \geqslant \sqrt{j}t\right) &\lesssim 2^j\exp\big(-c_1(\log n)j^\tau t^\tau\big) + 2^j\exp\big(-c_2 j(t\wedge t^2)\big)\\ &\leqslant \exp\big(j(\log 2 - c_1 j^{\tau-1}(\log n)t^\tau)\big) + \exp\big(j(\log 2 - c_2(t\wedge t^2))\big).\end{aligned}$$

Using $J_n^+ \lesssim \log n$, the right-hand side of the previous display is arbitrarily small for large enough t. An analogous estimate holds for the scaling functions $\psi_{-1,\cdot}$. Therefore,

$$\|\widehat{\mu}'_{J_n^+} - \mu'\|_\infty = \mathcal{O}_P\left(\sum_{j_0 \leqslant j \leqslant J_n^+} 2^{3j/2}\sqrt{j/n}\right) + \mathcal{O}\big(2^{-J_n^+ s}\big) = \mathcal{O}_P\left(2^{3J_n^+/2}\sqrt{J_n^+/n}\right) + \mathcal{O}\big(2^{-J_n^+ s}\big).$$

Analogously, we have

$$\|\widehat{\mu}_{J_n^+} - \mu\|_\infty = \mathcal{O}_P\left(2^{J_n^+/2}\sqrt{J_n^+/n}\right) + \mathcal{O}\big(2^{-J_n^+(s+1)}\big).$$

Since μ is bounded away from zero on S, the choice of J_n^+ yields in particular that we have $\lim_{n\to\infty} P(\inf_{x\in S}\widehat{\mu}_{J_n^+}(x) > \frac{1}{2}\inf_{x\in S}\mu(x)) = 1$. We conclude

$$\begin{aligned}\|R_{J_n^+}\|_{L^\infty(S)} &= \mathcal{O}_P\left(\|\widehat{\mu}'_{J_n^+} - \mu'\|_{L^\infty(S)}\|\widehat{\mu}_{J_n^+} - \mu\|_{L^\infty(S)} + \|\widehat{\mu}_{J_n^+} - \mu\|^2_{L^\infty(S)}\right)\\ &= \mathcal{O}_P\left(\left(2^{3J_n^+/2}\sqrt{J_n^+/n} + 2^{-J_n^+ s}\right)\left(2^{J_n^+/2}\sqrt{J_n^+/n} + 2^{-J_n^+(s+1)}\right)\right),\end{aligned}$$

which shows the asserted bound for $\|R_{J_n^+}\|_{L^\infty(S)}$. □

The linearised stochastic error term can be decomposed into

$$\begin{aligned}-\frac{1}{2}\langle\widehat{\mu}_{J^+} - \mu, \psi'_{j,k}\mu^{-1}\rangle &= -\sum_{l\leqslant J^+,m}\langle\mu_n - \mu, \psi_{l,m}\rangle\left\langle\psi_{l,m}, \frac{\psi'_{j,k}}{2\mu}\right\rangle + \left\langle(\mathrm{Id} - \pi_{J^+})\mu, \frac{\psi'_{j,k}}{2\mu}\right\rangle\\ &= -\left\langle\mu_n - \mu, \frac{\psi'_{j,k}}{2\mu}\right\rangle + \left\langle\mu_n - \mu, (\mathrm{Id} - \pi_{J^+})\left(\frac{\psi'_{j,k}}{2\mu}\right)\right\rangle + \left\langle(\mathrm{Id} - \pi_{J^+})\mu, \frac{\psi'_{j,k}}{2\mu}\right\rangle. \qquad (5.1)\end{aligned}$$

Roughly for $j \leqslant J \leqslant J^+$, Theorem 2.6 (or an analogous result for ill-posed problems) applies for the first term in the above display, the second term should converge to zero by the localisation of the $\psi_{j,k}$ in the Fourier domain and the third term is a bias that can be bounded by the smoothness of μ. If $U_n \to \infty$ this "μ-bias" term is of smaller order than the "b-bias" which is determined by $\sum_{j>J,k}\langle b, \psi_{j,k}\rangle\psi_{j,k}$.

Let us make these considerations precise. We will need the following lemma, which relies on the localisation of the wavelets in Fourier domain. More precisely, ψ can be chosen such that for some $\gamma \geqslant 1$ we have

$$\varphi, \psi \in C^\gamma(\mathbb{R}) \quad \text{and} \quad \int x^k\psi(x)\mathrm{d}x = 0, \quad \text{for } k = 0, \ldots, \lceil\gamma\rceil.$$

In the Fourier domain we conclude by the compact support of ψ

$$|\mathcal{F}\varphi(u)| \lesssim \frac{1}{(1+|u|)^\gamma}, \quad |\mathcal{F}\psi(u)| \lesssim \frac{|u|^\gamma}{(1+|u|)^{2\gamma}}, \quad u \in \mathbb{R}. \qquad (5.2)$$

Lemma 5.2. *Grant Assumption 3.1 and let the compactly supported father and mother wavelet functions φ and ψ satisfy (5.2) for some $\gamma > 1$. Then for any $m \in \mathbb{Z}$, $j < l$ and $k \in K_j$*

$$\begin{aligned}|\langle\psi_{l,m}, \psi'_{j,k}\mu^{-1}\rangle| &\lesssim 2^{l-(l-j)(\gamma-1/2)} + 2^{-l+j},\\ \sum_{m\in\mathbb{Z}}|\langle\psi_{l,m}, \psi'_{j,k}\mu^{-1}\rangle| &\lesssim 2^{l-(l-j)(\gamma-3/2)} + 1 \quad \textit{and}\\ \sum_{m\in\mathbb{Z}}|\langle\psi_{l,m}, \psi'_{j,k}\mu^{-1}\rangle|^2 &\lesssim 2^{2l-(l-j)(2\gamma-2)} + 2^{-(l-j)},\end{aligned}$$

where we have to replace j by j_0 on the right-hand side for $\psi_{-1,k} = \varphi_{j_0,k}$.

Proof. Let $\Gamma > 0$ be large enough such that $\operatorname{supp}\varphi \cup \operatorname{supp}\psi \subseteq [-\Gamma, \Gamma]$. Noting that the following scalar product can only be nonzero if the support of $\psi_{l,m}$ is contained in D, a Taylor expansion of μ^{-1} yields for $j \geqslant j_0$

$$\begin{aligned}|\langle \psi_{l,m}\mu^{-1}, \psi'_{j,k}\rangle| &\leqslant |\mu(2^{-l}m)|^{-1}|\langle \psi_{l,m}, \psi'_{j,k}\rangle| + 2^{-l}\Gamma \max_{x:|x-m2^{-l}|\leqslant 2^{-l}\Gamma} |(\mu^{-1})'(x)| \int |\psi_{l,m}(x)||\psi'_{j,k}(x)|\mathrm{d}x \\ &\leqslant \|\mu^{-1}\|_{L^\infty(D)}|\langle \psi_{l,m}, \psi'_{j,k}\rangle| + 2^{-l+j}\Gamma \max_{x:|x-m2^{-l}|\leqslant 2^{-l}\Gamma} |(\mu^{-1})'(x)|\|\psi\|_{L^2}\|\psi'\|_{L^2}.\end{aligned}$$

We conclude

$$|\langle \psi_{l,m}, \psi'_{j,k}\mu^{-1}\rangle| \lesssim \|\mu^{-1}\|_{L^\infty(D)}|\langle \psi_{l,m}, \psi'_{j,k}\rangle| + 2^{-l+j}\|(\mu^{-1})'\|_{L^\infty(D)}. \tag{5.3}$$

Using Plancherel's identity, $\mathcal{F}\psi_{l,m}(u) = \mathcal{F}[2^{l/2}\psi(2^l\bullet - m)](u) = 2^{-l/2}\mathrm{e}^{imu2^{-l}}\mathcal{F}\psi(2^{-l}u)$ and (5.2), we obtain

$$\begin{aligned}|\langle \psi_{l,m}, \psi'_{j,k}\rangle| &\leqslant \frac{2^{-(j+l)/2}}{2\pi}\int |\mathcal{F}\psi(2^{-j}u)u\mathcal{F}\psi(2^{-l}u)|\mathrm{d}u \\ &\lesssim 2^{-(j+l)/2}\int \frac{2^{-(j+l)\gamma}|u|^{2\gamma+1}}{(1+2^{-j}|u|)^{2\gamma}(1+2^{-l}|u|)^{2\gamma}}\mathrm{d}u \\ &\leqslant 2^{-(j+l)/2-(j+l)\gamma}\int \frac{|u|\mathrm{d}u}{2^{-2\gamma j}(1+2^{-l}|u|)^{2\gamma}} = 2^{l-(l-j)(\gamma-1/2)}\int \frac{|v|\mathrm{d}v}{(1+|v|)^{2\gamma}},\end{aligned}$$

where we have substituted $v = 2^{-l}u$ in the last line. Due to $\gamma > 1$, the integral in the last display is finite so that combining this bound with (5.3) yields the assertions, noting that by the compact support of ψ only for $\mathcal{O}(2^{l-j})$ many m the scalar products $\langle \psi_{l,m}, \psi'_{j,k}\mu^{-1}\rangle$ are nonzero.

For $j = -1$ we substitute again $v = 2^{-l}u$ and obtain analogously

$$|\langle \psi_{l,m}, \psi'_{-1,k}\rangle| \lesssim \int \frac{2^{-(j_0+l)/2-l\gamma}|u|^{\gamma+1}\mathrm{d}u}{(1+2^{-l}|u|)^{2\gamma}(1+2^{-j_0}|u|)^{\gamma}} \leqslant 2^{l-(l-j_0)(\gamma-1/2)}\int \frac{|v|\mathrm{d}v}{(1+|v|)^{2\gamma}}$$

and

$$|\langle \psi_{l,m}, \psi'_{-1,k}\mu^{-1}\rangle| \lesssim \|\mu^{-1}\|_{L^\infty(D)}|\langle \psi_{l,m}, \varphi'_{j_0,k}\rangle| + 2^{-l+j_0}\|(\mu^{-1})'\|_{L^\infty(D)}. \qquad \square$$

Now we can bound the bias in (5.1) in the multi-scale space $\mathcal{M}_0$.

Lemma 5.3. *Let the weighting sequence w be admissible and grant Assumption 3.1 and (5.2) for some $\gamma \geqslant 3/2 + \delta, \delta \in (1,2]$. Then we have*

$$\big\|\big(\langle (\mathrm{Id} - \pi_{J+U})\mu, \psi'_{j,k}/(2\mu)\rangle\big)_{j\leqslant J,k}\big\|_{\mathcal{M}} \lesssim 2^{-J(s+1/2)}2^{-U(s+3/2)}w_J^{-1}.$$

Proof. Recall that we have by definition

$$\big\|\big(\langle (\mathrm{Id} - \pi_{J^+})\mu, \psi'_{j,k}/(2\mu)\rangle\big)_{j\leqslant J,k}\big\|_{\mathcal{M}} = \sup_{j\leqslant J}\max_{k\in K_j} w_j^{-1}\big|\langle (\mathrm{Id} - \pi_{J^+})\mu, \psi'_{j,k}/(2\mu)\rangle\big|.$$

As in the proof of Theorem 2.6 we have

$$\sup_{l,m:\operatorname{supp}\psi_{l,m}\cap S\neq\emptyset} 2^{j(s+3/2)}|\langle \psi_{l,m}, \mu\rangle| \lesssim \|\mu\|_{C^{s+1}(D)}.$$

Hence, for all $j \leqslant J$,

$$\begin{aligned}\big|\langle (\mathrm{Id} - \pi_{J^+})\mu, \psi'_{j,k}/(2\mu)\rangle\big| &= \Big|\sum_{l>J^+,m}\langle \mu, \psi_{l,m}\rangle\langle \psi_{l,m}, \psi'_{j,k}/(2\mu)\rangle\Big| \\ &\leqslant \sup_{l>J^+,m} 2^l|\langle \mu, \psi_{l,m}\rangle| \sum_{l>J^+,m} 2^{-l}|\langle \psi_{l,m}, \psi'_{j,k}/(2\mu)\rangle| \\ &\lesssim 2^{-J^+(s+1/2)}\sum_{l>J^+,m} 2^{-l}|\langle \psi_{l,m}, \psi'_{j,k}/(2\mu)\rangle|.\end{aligned}$$

Now Lemma 5.2 yields

$$\sum_{l>J^+,m} 2^{-l}|\langle \psi_{l,m}, \psi'_{j,k}/(2\mu)\rangle| \lesssim \sum_{l>J^+} 2^{-(l-j)(\gamma-3/2)} + \sum_{l>J^+} 2^{-l} \lesssim 2^{-(J^+-j)(\gamma-3/2)} + 2^{-J^+}.$$

Due to the monotonicity of $j \mapsto 2^{j\delta} w_j^{-1}$, we conclude for $\gamma \geqslant 3/2 + \delta$

$$\begin{aligned}\sup_{j\leqslant J} \max_{k\in K_j} w_j^{-1}\left|\left\langle (\mathrm{Id} - \pi_{J^+})\mu, \psi'_{j,k}/(2\mu)\right\rangle\right| &\lesssim 2^{-J^+(s+1/2)} \sup_{j\leqslant J} w_j^{-1}(2^{-(J^+-j)(\gamma-3/2)} + 2^{-J^+}) \\ &\lesssim 2^{-J^+(s+1/2)} 2^{-(J^+-J)} w_J^{-1}. \qquad \square\end{aligned}$$

The second term in (5.1) can be bounded by the following lemma.

Lemma 5.4. *Let the weighting sequence w satisfy $\sqrt{j}2^j/w_j = \mathcal{O}(1)$ and grant Assumption 3.1 and (5.2) for some $\gamma \geqslant 5/2$. If $J_n^+ = J_n + U_n$ satisfies for some $\tau \in (0,1)$*

$$(\log n)^{2/\tau} n^{-1} 2^{J_n^+} J_n^+ = \mathcal{O}(1) \quad \text{and} \quad U_n \to \infty,$$

then we have

$$\left\| \pi_{J_n} \left(\left(\left\langle \mu_n - \mu, (\mathrm{Id} - \pi_{J_n^+}) \left(\frac{\psi'_{j,k}}{2\mu} \right) \right\rangle \right)_{j,k} \right) \right\|_{\mathcal{M}} = o_P(n^{-1/2}).$$

Proof. In order to apply Proposition 2.2, we need to calculate the L^2-norm and the L^∞-norm of $(\mathrm{Id} - \pi_{J_n^+})(\psi'_{j,k}/\mu)$. For $j \in \{j_0, \dots, J_n\}$ Parseval's identity and Lemma 5.2 yield

$$\begin{aligned}\left\| (\mathrm{Id} - \pi_{J_n^+}) \left(\frac{\psi'_{j,k}}{\mu} \right) \right\|_{L^2}^2 &= \sum_{l>J_n^+,m} |\langle \psi_{l,m}, \psi'_{j,k}/\mu\rangle|^2 \\ &\lesssim \sum_{l>J_n^+} 2^{2l-(l-j)(2\gamma-2)} + \sum_{l>J_n^+} 2^{-2(l-j)+(l-j)_+} \\ &\lesssim 2^{j(2\gamma-2)} \sum_{l>J_n^+} 2^{-l(2\gamma-4)} + 2^{-(J_n^+-j)} \\ &\lesssim 2^{-(J_n^+-j)(2\gamma-4)+2j} + 2^{-(J_n^+-j)} \lesssim 2^{-(J_n^+-j)} 2^{2j}.\end{aligned}$$

Another application of Lemma 5.2 yields

$$\begin{aligned}\left\| (\mathrm{Id} - \pi_{J_n^+}) \left(\frac{\psi'_{j,k}}{2\mu} \right) \right\|_\infty &\leqslant \sum_{l>J_n^+} 2^{l/2} \max_n |\langle \psi_{l,m}, \psi'_{j,k}/\mu\rangle| \\ &\lesssim \sum_{l>J_n^+} 2^{3l/2-(l-j)(\gamma-1/2)} + \sum_{l>J_n^+} 2^{-l/2+j} \\ &\lesssim 2^{-(J_n^+-j)(\gamma-2)+3j/2} + 2^{-(J_n^+-j)/2+j/2} \lesssim 2^{-(J_n^+-j)/2} 2^{3j/2}.\end{aligned}$$

The concentration inequality, Proposition 2.2, yields for positive constants $c_i > 0, i = 1, 2, \ldots,$

$$\begin{aligned}
&P\left(\sup_{1\leqslant j\leqslant J_n}\max_k \sqrt{n}w_j^{-1}\left\langle \mu_n-\mu, (\mathrm{Id}-\pi_{J_n^+})\left(\frac{\psi'_{j,k}}{2\mu}\right)\right\rangle > t\right)\\
&\leqslant \sum_{1\leqslant j\leqslant J_n,k} P\left(\sqrt{n}\left\langle \mu_n-\mu, (\mathrm{Id}-\pi_{J_n^+})\left(\frac{\psi'_{j,k}}{2\mu}\right)\right\rangle > tw_j\right)\\
&\lesssim \sum_{j=1}^{J_n} 2^j\Bigg(\exp(-c_1(2^{(J_n^+-j)/2}2^{-3j/2}\sqrt{n}w_j t)^\tau)\\
&\quad + \exp\left(-\frac{c_2 2^{(J_n^+-j)/2}t^2}{(2^j/w_j)^2 + t\max(2^{3j/2}(\log n)^{1/\tau}, 2^j)/(w_j\sqrt{n})}\right)\Bigg).
\end{aligned}$$

Since $jw_j^{-1}2^{3j/2}(\log n)^{1/\tau}n^{-1/2} \lesssim 2^{j/2}j^{1/2}(\log n)^{1/\tau}n^{-1/2} \lesssim 1$ and $J_n^+ \lesssim \log n$ by the assumptions on w_j and J_n^+, we conclude for any $t > 0$ and n sufficiently large

$$\begin{aligned}
&P\left(\sup_{1\leqslant j\leqslant J_n}\max_k \sqrt{n}w_j^{-1}\left\langle \mu_n-\mu, (\mathrm{Id}-\pi_{J_n^+})\left(\frac{\psi'_{j,k}}{2\mu}\right)\right\rangle > t\right)\\
&\lesssim \sum_{j=1}^{J_n} 2^j\left(\exp(-c_3(\log n)2^{\tau(J_n^+-j)/2}j^\tau t^\tau) + \exp\left(-c_4\frac{2^{(J_n^+-j)/2}jt^2}{1+t}\right)\right)\\
&\lesssim \sum_{j=1}^{J_n}\exp\Big(j\big(\log 2 - c_3(J_n^+)^{\tau-1}t^\tau\log n\big)\Big) + \sum_{j=1}^{J_n}\exp\left(j\left(\log 2 - c_5 2^{U_n/2}\frac{t^2}{1+t}\right)\right)\\
&\lesssim \mathrm{e}^{\log 2 - c_3(J_n^+)^{\tau-1}t^\tau\log n}\frac{1-\mathrm{e}^{J_n\left(\log 2 - c_3(J_n^+)^{\tau-1}t^\tau\log n\right)}}{1-\mathrm{e}^{\log 2 - c_3(J_n^+)^{\tau-1}t^\tau\log n}}\\
&\quad + \mathrm{e}^{\log 2 - c_6 2^{U_n/2}(t^2\wedge t)}\frac{1-\mathrm{e}^{J_n(\log 2 - c_6 2^{U_n/2}(t^2\wedge t))}}{1-\mathrm{e}^{\log 2 - c_6 2^{U_n/2}(t^2\wedge t)}}\\
&\lesssim \mathrm{e}^{-c_3(J_n^+)^{\tau-1}t^\tau\log n} + \mathrm{e}^{-c_6 2^{U_n/2}(t^2\wedge t)} \to 0.
\end{aligned}$$

Finally note that all bounds hold true for the scaling function $\varphi_{j_0,\cdot}$ if j is replace by j_0. □

Now we have all pieces together to prove the multi-scale central limit theorem.

Proof of Theorem 3.8. Since b has Hölder regularity $s > 0$ on $S \subseteq D$, the bias can be bounded by

$$\|b - \pi_{J_n}(b)\|_{\mathcal{M}} = \sup_{j>J_n}\max_{k\in K_j} w_j^{-1}|\langle\psi_{j,k}, b\rangle| \lesssim \sup_{j>J_n} w_j^{-1}2^{-j(s+1/2)} = o(n^{-1/2}).$$

Using that the $\mathcal{M}_0$-norm is weaker than the $L^\infty(S)$-norm, Lemma 5.1 and decomposition (5.1) together with Lemmas 5.3 and 5.4 yield

$$\big(\langle\widehat{b}_{J_n,U_n} - b, \psi_{j,k}\rangle\big)_{j\leqslant J_n,k} = -\left(\left\langle \mu_n-\mu, \frac{\psi'_{j,k}}{2\mu}\right\rangle\right)_{j\leqslant J_n,k} + \widetilde{R}_{J_n,U_n} \quad \text{with } \|\pi_{J_n}(\widetilde{R}_{J_n,U_n})\|_{\mathcal{M}} = o_P(n^{-1/2}).$$

Therefore, it remains to show that

$$\beta_{\mathcal{M}_0}\big(\mathcal{L}\big(\pi_{J_n}\big((-\sqrt{n}\langle\mu_n-\mu, \psi'_{j,k}/(2\mu)\rangle)_{j,k}\big)\big), \mathcal{L}(\mathbb{G}_b)\big) \to 0.$$

This follows exactly as in Theorem 2.6, where we use that the factor 2^j, by which the norms

$$\|\psi'_{j,k}/(2\mu)\|_{L^2} \lesssim 2^j, \quad \|\psi'_{j,k}/(2\mu)\|_\infty \lesssim 2^{3j/2} \tag{5.4}$$

are larger than $\|\psi_{j,k}\|_{L^2}$ and $\|\psi_{j,k}\|_\infty$, respectively, is counterbalanced through the additional growth of the admissible weighting sequence w. □

Appendix A. Remaining proofs

A.1. Proof of Theorem 3.5

Step 1. For $\delta \in (0,1/2)$ and $0 < c < C < \infty$ define

$$\mathbb{V}_\xi := \{\mu \in C^{7/4+\delta/2}(D) : 0 < c < \mu \text{ and } \|\mu\|_{C^{7/4+\delta/2}} < C\},$$

$\mathbb{V} := \mathcal{M}_0(w)$ and $\mathbb{W} := \mathcal{M}_0(\widetilde{w})$ with $w_j \leqslant 2^{\delta j}$ and $\widetilde{w} \geqslant 2^{(1+\delta)j}$. We first establish the Hadamard differentiability of

$$\xi : \mathbb{V}_\xi \subseteq \mathbb{V} \to \mathbb{W}, \mu \mapsto \frac{\mu'}{2\mu}$$

with derivative given by (3.8).

To this end let $h \in \mathcal{M}_0(w)$ and $h_t \to h$ as $t \to 0$. For all h_t such that $\mu + th_t$ is contained in $\mathbb{V}_\xi$ for small $t > 0$ we have

$$\begin{aligned}
&\left\| \frac{\xi(\mu+th_t)-\xi(\mu)}{t} - \xi'_\mu(h) \right\|_{\mathcal{M}(\widetilde{w})} = \left\| \frac{\mu(\mu'+th'_t)-\mu'(\mu+th_t)}{2(\mu+th_t)\mu t} - \frac{\mu h'-\mu' h}{2\mu^2} \right\|_{\mathcal{M}(\widetilde{w})} \\
&= \left\| \frac{\mu h'_t - \mu' h_t}{2(\mu+th_t)\mu} - \frac{\mu h'-\mu' h}{2\mu^2} \right\|_{\mathcal{M}(\widetilde{w})} = \left\| \frac{\mu^2(h'_t-h') + \mu\mu'(h-h_t) + th_t(\mu' h - \mu h')}{2\mu^2(\mu+th_t)} \right\|_{\mathcal{M}(\widetilde{w})},
\end{aligned}$$

using $\widetilde{w}_j \geqslant 2^{j(1+\delta)}$ for $\delta \in (0,1/2)$, this is bounded by

$$\left\| \frac{(h'_t-h')}{2(\mu+th_t)} \right\|_{B^{-3/2-\delta}_{\infty\infty}} + \left\| \frac{\mu'(h-h_t)}{2\mu(\mu+th_t)} \right\|_{B^{-3/2-\delta}_{\infty\infty}} + \left\| \frac{th_t(\mu' h - \mu h')}{2\mu^2(\mu+th_t)} \right\|_{\mathcal{M}(\widetilde{w})}.$$

Applying a pointwise multiplier theorem ([43], Thm. 2.8.2) and the continuous embedding $C^\rho \to B^\rho_{\infty\infty}$, we obtain up to constants the upper bound

$$\begin{aligned}
&\left\| \frac{1}{2(\mu+th_t)} \right\|_{C^{7/4+\delta/2}} \|h'_t - h'\|_{B^{-3/2-\delta}_{\infty\infty}} + \left\| \frac{\mu'}{2\mu(\mu+th_t)} \right\|_{C^{7/4+\delta/2}} \|h-h_t\|_{B^{-1/2-\delta}_{\infty\infty}} \\
&\quad + t \left\| \frac{h_t(\mu' h - \mu h')}{2\mu^2(\mu+th_t)} \right\|_{\mathcal{M}(\widetilde{w})} \\
&\lesssim \|h_t - h\|_{B^{-1/2-\delta}_{\infty\infty}} + \|h-h_t\|_{B^{-1/2-\delta}_{\infty\infty}} + t \lesssim \|h_t - h\|_{\mathcal{M}(w)} + t,
\end{aligned}$$

where we have used $w_j \leqslant 2^{j\delta}$ in the last step. The last expression tends to 0 as $t \to 0$ and this shows the Hadamard differentiability of $\xi\colon \mathbb{V}_\xi \to \mathbb{W}$.

Step 2. To apply the delta method it is now important that $\widehat{\mu}_{J_n}$ maps into $\mathbb{V}_\xi$. Theorem 2.6 gives conditions such that $\|\widehat{\mu}_{J_n} - \mu\|_{\mathcal{M}(w)} = O(n^{-1/2})$. Provided that $2^{(9/4+\delta/2)J_n} w_{J_n} n^{-1/2} = o(1)$ we deduce from the fact that $\widehat{\mu}_{J_n}$ is developed until level J_n only and from the ratio of the weights at level J_n that $\|\widehat{\mu}_{J_n} - \mu\|_{C^{7/4+\delta/2}} = o(1)$. We conclude that with probability tending to one $\widehat{\mu} \in \mathbb{V}_\xi$. By modifying $\widehat{\mu}$ on events with probability tending to zero we can achieve that always $\widehat{\mu} \in \mathbb{V}_\xi$. On the above assumptions we obtain the weak convergence $\sqrt{n}(\widehat{\mu}_{J_n} - \mu) \to \mathbb{G}_\mu$ in $\mathcal{M}_0(w)$ by Theorem 2.6 and application of the delta method yields the assertion.

A.2. Proof of Lemma 4.3

We will prove that:

(i) for any $\tau \in (0,1)$ there are constants $0 < c, C < \infty$ depending only on τ, K, ψ such that for any $J \in \mathcal{J}_n$ satisfying $J > J_n^*$ and for all $n \in \mathbb{N}$ large enough

$$P(\widehat{J}_n = J) \leqslant C(n^{-cJ^\tau} + \mathrm{e}^{-cJ}),$$

(ii) there is an integer $M > 0$ depending only on d_1, d_2, K and constants $0 < c', C' < \infty$ depending on τ, K, ψ such that for any $J \in \mathcal{J}_n$ satisfying $J < J_n^* - M$ and for all $n \in \mathbb{N}$ large enough

$$P(\widehat{J}_n = J) \leqslant C'\big(n^{-c'J^\tau} + \mathrm{e}^{-c'J}\big).$$

Given (i) and (ii), we obtain, for a constant $c'' > 0$,

$$\begin{aligned} P\big(\widehat{J}_n \notin [J_n^* - M, J_n^*]\big) &\leqslant \sum_{J=J_{\min}}^{J_n^*-M-1} P(\widehat{J}_n = J) + \sum_{J=J_n^*+1}^{J_{\max}} P(\widehat{J}_n = J) \\ &= \mathcal{O}\Big((J_{\max} - J_{\min})\big(n^{-c''(J_{\min})^\tau} + \mathrm{e}^{-c''J_{\min}}\big)\Big) \\ &= \mathcal{O}\Big(\log n\big(n^{-c''J_{\min}^\tau} + \mathrm{e}^{-c''J_{\min}}\big)\Big). \end{aligned}$$

Since $n^{-c''J_{\min}^\tau} + \mathrm{e}^{-c''J_{\min}}$ decays polynomially in n, the assertion of the lemma follows.

To show (i) and (ii), recall $J^+ = J + U_n = J + \log_2 \log n$. For notational convenience we define

$$V(n,j) := (2^{3j} j/n)^{1/2},$$

which is the order of magnitude of the stochastic error for projection level j. Recall that for any $f \in \mathcal{M} \cap V_J$ we can bound

$$\|f\|_{L^\infty([a,b])} \lesssim \sum_{j \leqslant J} 2^{j/2} \max_{k \in K_j} |\langle f, \psi_{j,k}\rangle| \leqslant \|f\|_{\mathcal{M}} \sum_{j \leqslant J} 2^{j/2} w_j \lesssim \|f\|_{\mathcal{M}} 2^{J/2} w_J.$$

Since any $J \in \mathcal{J}_n$ satisfies $2^{-J^+(s+1)} = o(1)$ and $(\log n)^{2/\tau} n^{-1} 2^{J^+} J^+ = \mathcal{O}(1)$, we conclude from decomposition (5.1) as well as from Lemmas 5.1 and 5.4 (applied to $w_j = j^{1/2} 2^j$) that

$$\widehat{b}_J - b = -\sum_{j \leqslant J, k \in \mathbb{Z}} \left\langle \mu_n - \mu, \frac{\psi'_{j,k}}{2\mu} \right\rangle \psi_{j,k} + \underbrace{\sum_{j \leqslant J, k \in \mathbb{Z}} \left\langle (\mathrm{Id} - \pi_{J^+})\mu, \frac{\psi'_{j,k}}{2\mu} \right\rangle \psi_{j,k}}_{=: B^\mu(J)} + \underbrace{\sum_{j > J, k \in \mathbb{Z}} \langle b, \psi_{j,k}\rangle \psi_{j,k}}_{=: B^b(J)} + \overline{R}_J$$

for some remainder $\overline{R}_J \in V_J$ with

$$\|\overline{R}_J\|_{L^\infty([a,b])} = \mathcal{O}_P\big(n^{-1} J^+ 2^{2J^+}\big) + \mathcal{O}\big(2^{-J^+(2s+1)}\big) + o_P\big(n^{-1/2} J^{1/2} 2^{-3J/2}\big).$$

Moreover, Lemma 5.3 and Assumption 4.1 yield

$$\|B^\mu(J) + B^b(J)\|_{L^\infty([a,b])} \leqslant (d_2 + o(1)) 2^{-Js}. \tag{A.5}$$

Using $n^{-1/2} 2^{2J^+} J^+ = o(1)$ for all $J \in \mathcal{J}_n$, we conclude

$$\|\widehat{b}_J - b\|_{L^\infty([a,b])} \leqslant \left\| \sum_{j \leqslant J, k \in K_j} \left\langle \mu_n - \mu, \frac{\psi'_{j,k}}{2\mu} \right\rangle \psi_{j,k} \right\|_{L^\infty([a,b])} + (d_2 + o(1)) 2^{-Js} + o_P(V(n,J)). \tag{A.6}$$

With this preparation at hand we can proceed similarly as in ([18], Lem. 2).

Part (i). For any fixed $J > J_n^*$ we have

$$P\big(\widehat{J}_n = J\big) \leqslant \sum_{L\in\mathcal{J}_n, L\geqslant J} P\big(\|\widehat{b}_{J-1} - \widehat{b}_L\|_{L^\infty([a,b])} > K\,V(n,L)\big).$$

As in the derivation of (A.6) we obtain for n sufficiently large

$$\|\widehat{b}_{J-1} - \widehat{b}_L\|_{L^\infty([a,b])} \leqslant \left\| \sum_{J<j\leqslant L,k} \left\langle \mu_n - \mu, \frac{\psi'_{j,k}}{2\mu} \right\rangle \psi_{j,k} \right\|_{L^\infty([a,b])} + \big(d_2+1\big)\big(2^{-(J-1)s} + 2^{-Ls}\big) + \frac{1}{4}\big(V(n,J-1) + V(n,L)\big).$$

By definition of J_n^* we have for any $L > J > J_n^*$ that

$$\big(d_2+1\big)\big(2^{-(J-1)s} + 2^{-Ls}\big) \leqslant 2(d_2+1)2^{-J_n^* s} \leqslant \frac{K}{2} V(n,J_n^*) \leqslant \frac{K}{2} V(n,L).$$

Therefore,

$$P\big(\widehat{J}_n = J\big) \leqslant \sum_{L\in\mathcal{J}_n, L\geqslant J} P\left(\left\| \sum_{J<j\leqslant L,k} \left\langle \mu_n - \mu, \frac{\psi'_{j,k}}{2\mu} \right\rangle \psi_{j,k} \right\|_{L^\infty([a,b])} > \frac{K-1}{2} V(n,L) \right).$$

Analogously to (2.10) and using (5.4), Proposition 2.2 yields for any $\tau \in (0,1)$ and constants $c_1, \dots, c_4 > 0$:

$$\begin{aligned} &P\left(\left\| \sum_{J<j\leqslant L, k\in K_j} \left\langle \mu_n - \mu, \frac{\psi'_{j,k}}{2\mu} \right\rangle \psi_{j,k} \right\|_{L^\infty([a,b])} > \frac{K-1}{2} V(n,L) \right) \\ &\leqslant P\left(\sum_{J<j\leqslant L} 2^{j/2} \max_{k\,\in\, K_j} \left| \left\langle \mu_n - \mu, \frac{\psi'_{j,k}}{2\mu} \right\rangle \right| > \frac{K-1}{2} \sqrt{\frac{2^{3L} L}{n}} \right), \end{aligned}$$

using that $\sum_{k=0}^{\infty} 2^{-k/2} \leqslant 7/2$, we obtain the upper bound

$$\begin{aligned} &P\left(\max_{J<j\leqslant L,k} n^{1/2} 2^{-L} \left| \left\langle \mu_n - \mu, \frac{\psi'_{j,k}}{2\mu} \right\rangle \right| > \frac{K-1}{7} L^{1/2} \right) \\ &\leqslant \sum_{J<j\leqslant L,k} P\left(n^{1/2} 2^{-j} \left| \left\langle \mu_n - \mu, \frac{\psi'_{j,k}}{2\mu} \right\rangle \right| > \frac{K-1}{7} j^{1/2} \right) \\ &\leqslant c_1 \sum_{J<j\leqslant L} \big(\mathrm{e}^{-c_2 j^\tau \log n} + \mathrm{e}^{-c_3 j} \big) \leqslant c_1 (n^{-c_4 J^\tau} + \mathrm{e}^{-c_4 J}), \end{aligned} \tag{A.7}$$

where we require that K is chosen sufficiently large, depending on $\|\psi'_{j,k}/\mu\|_{L^\infty(S)}$ and $\Sigma_{j,k}$. It remains to sum this upper bound over all $L \in \mathcal{J}_n$ with $L \geqslant J$ which yields the claim since $\mathcal{J}_n$ contains no more than $\log n$ elements.

Part (ii). Let $J < J_n^* - M$ for some $M \in \mathbb{N}$ to be specified below. We have

$$P\big(\widehat{J}_n = J\big) \leqslant P\big(\|\widehat{b}_J - \widehat{b}_{J_n^*}\|_{L^\infty([a,b])} \leqslant K\,V(n,J_n^*)\big).$$

Using Assumption 4.1 and the triangle inequality, we obtain similarly to (A.6), for sufficiently large n,

$$\|\widehat{b}_J - \widehat{b}_{J_n^*}\|_{L^\infty([a,b])} \geqslant \frac{d_1}{2}2^{-Js} - (d_2+1)2^{-J_n^* s} - \left\| \sum_{J<j\leqslant J_n^*, k\in K_j} \left\langle \mu_n - \mu, \frac{\psi'_{j,k}}{2\mu}\right\rangle \psi_{j,k} \right\|_{L^\infty([a,b])} - \frac{1}{4}(V(n,J_n^*)+V(n,J)).$$

Owing to $J < J_n^* - M$, $s \geqslant 1$ and the definition of J_n^*, we can bound

$$\begin{aligned}\frac{d_1}{2}2^{-Js} - (d_2+1)2^{-J_n^* s} &\geqslant (d_2+1)\left(\frac{d_1}{2(d_2+1)}2^{M-1} - \frac{1}{2}\right)2^{-(J_n^*-1)s} \\ &\geqslant \frac{K}{4}\left(\frac{d_1}{2(d_2+1)}2^{M-1} - \frac{1}{2}\right)V(n,J_n^*-1) \\ &\geqslant \frac{K}{16}\left(\frac{d_1}{2(d_2+1)}2^{M-1} - \frac{1}{2}\right)V(n,J_n^*),\end{aligned}$$

where we have used in the last inequality that $(J_n^*-1)/J_n^* \geqslant 1/2$ for n sufficiently large. We conclude

$$\|\widehat{b}_J - \widehat{b}_{J_n^*}\|_{L^\infty([a,b])} \geqslant \widetilde{K}\,V(n,J_n^*) - \left\| \sum_{J<j\leqslant J_n^*, k\in K_j} \left\langle \mu_n - \mu, \frac{\psi'_{j,k}}{2\mu}\right\rangle \psi_{j,k} \right\|_{L^\infty([a,b])}$$

with $\widetilde{K} := \frac{Kd_1}{32(d_2+1)}2^{M-1} - \frac{K}{32} - \frac{1}{2}$. Since $\widetilde{K} > K$ for M large enough, we obtain similarly as in (A.7) for any $\tau \in (0,1)$ and some $c', C' > 0$

$$\begin{aligned}P(\widehat{J}_n = J) &\leqslant P\left(\left\| \sum_{J<j\leqslant J_n^*, k} \left\langle \mu_n - \mu, \frac{\psi'_{j,k}}{2\mu}\right\rangle \psi_{j,k} \right\|_{L^\infty([a,b])} \geqslant (\widetilde{K}-K)V(n,J_n^*)\right) \\ &\leqslant C'\big(n^{-c'J^\tau} + \mathrm{e}^{-c'J}\big).\end{aligned}$$

A.3. Proof of Corollary 4.5

Owing to $(cn/\log n)^{1/(2s+3)} \leqslant 2^{J_n^*} \leqslant (Cn/\log n)^{1/(2s+3)}$ for constants $0 < c < C$, we find

$$\frac{\log n - \log\log n}{2(\log 2)J_n^*} + \frac{\log c}{2(\log 2)J_n^*} - \frac{3}{2} \leqslant s \leqslant \frac{\log n - \log\log n}{2(\log 2)J_n^*} + \frac{\log C}{2(\log 2)J_n^*} - \frac{3}{2}.$$

Since $P(\widehat{J}_n \leqslant J_n^*) \to 1$ by Lemma 4.3 and due to $v_n^{-1} = o_P(1)$, we obtain with probability converging to one that

$$s \geqslant \max\left(1, \frac{\log n - \log\log n}{2(\log 2)(\widehat{J}_n + v_n)} - o_P\left(\frac{v_n}{\widehat{J}_n}\right) - \frac{3}{2}\right) \geqslant \widehat{s}_n.$$

Moreover, since $P(J_n^* - M \leqslant \widehat{J}_n \leqslant J_n^*) \to 1$ and $v_n^{-1} = o_P(1)$, we have with probability converging to one

$$s - \widehat{s}_n \leqslant \frac{\log n - \log\log n}{2(\log 2)J_n^*}\Big(1 - \frac{1}{1+v_n/J_n^*}\Big) + \frac{\log C}{2(\log 2)J_n^*} + \frac{3v_n}{2(J_n^*-M)} \lesssim \frac{v_n}{J_n^*}.$$

References

[1] R. Adamczak and W. Bednorz, Exponential concentration inequalities for additive functionals of Markov chains. *ESAIM: PS* **19** (2015) 440–481.

[2] Y. Aït-Sahalia, *Econometrics of Diffusion Models*. John Wiley & Sons, Ltd. (2010).

[3] D. Bakry, P. Cattiaux and A. Guillin, Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincaré. *J. Funct. Anal.* **254** (2008) 727–759.

[4] R.F. Bass, *Diffusions and elliptic operators*. Springer, New York (1998).

[5] A.D. Bull, Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat.* **6** (2012) 1490–1516.

[6] J.V. Castellana and M.R. Leadbetter, On smoothed probability density estimation for stationary processes. *Stochastic Process. Appl.* **21** (1986) 179–193.

[7] I. Castillo and R. Nickl, Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** (2013) 1999–2028.

[8] I. Castillo and R. Nickl, On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** (2014) 1941–1969.

[9] D. Chauveau and J. Diebolt, Estimation of the asymptotic variance in the CLT for Markov chains. *Stoch. Models* **19** (2003) 449–465.

[10] X. Chen, Limit theorems for functionals of ergodic Markov chains with general state space. *Mem. Amer. Math. Soc.* **139** (1999) 664.

[11] V. Chernozhukov, D. Chetverikov and K. Kato, Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.* **42** (2014) 1787–1818.

[12] J. Chorowski and M. Trabs, Spectral estimation for diffusions with random sampling times. *Stochastic Process. Appl.* **126** (2016) 2976–3008.

[13] F. Comte, V. Genon-Catalot and Y. Rozenholc, Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli* **13** (2007) 514–543.

[14] A. Dalalyan, Sharp adaptive estimation of the drift function for ergodic diffusions. *Ann. Statist.* **33** (2005) 2507–2528.

[15] L. Galtchouk and S. Pergamenshchikov, Geometric ergodicity for classes of homogeneous Markov chains. *Stochastic Process. Appl.* **124** (2014) 3362–3391.

[16] C.J. Geyer, Practical Markov chain Monte Carlo. *Stat. Sci.* **7** (1992) 473–483.

[17] I.I. Gihman and A.V. Skorohod, *Stochastic differential equations*. Springer, Heidelberg (1972).

[18] E. Giné and R. Nickl, Confidence bands in density estimation. *Ann. Statist.* **38** (2010) 1122–1170.

[19] E. Giné and R. Nickl, Mathematical foundations of infinite-dimensional statistical models. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press (2015).

[20] E. Gobet, M. Hoffmann and M. Reiß, Nonparametric estimation of scalar diffusions based on low frequency data. *Ann. Statist.* **32** (2004) 2223–2253.

[21] S. Gugushvili and P. Spreij, Nonparametric Bayesian drift estimation for multidimensional stochastic differential equations. *Lith. Math. J.* **54** (2014) 127–141.

[22] L.P. Hansen and J.A. Scheinkman, Back to the future: generating moment implications for continuous-time Markov processes. *Econometrica* **63** (1995) 767–804.

[23] L.P. Hansen, J.A. Scheinkman and N. Touzi, Spectral methods for identifying scalar diffusions. *J. Econom.* **86** (1998) 1–32.

[24] W. Härdle, G. Kerkyacharian, D. Picard and A.Tsybakov, Wavelets, approximation, and statistical applications. Vol. 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York (1998).

[25] M. Hoffmann, Adaptive estimation in diffusion processes. *Stochastic Process. Appl.* **79** (1999) 135–163.

[26] M. Hoffmann and R. Nickl, On adaptive inference and confidence bands. *Ann. Statist.* **39** (2011) 2383–2409.

[27] D. Kristensen, Pseudo-maximum likelihood estimation in two classes of semiparametric diffusion models. *J. Econom.* **156** (2010) 239–259.

[28] Y.A. Kutoyants, Statistical inference for ergodic diffusion processes. *Springer Series in Statistics*. Springer-Verlag London, Ltd., London (2004).

[29] C. Lacour, Nonparametric estimation of the stationary density and the transition density of a Markov chain. *Stochastic Process. Appl.* **118** (2008) 232–260.

[30] M. Ledoux, The concentration of measure phenomenon. Vol. 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI (2001).

[31] E. Löcherbach, D. Loukianova and O. Loukianov, Penalized nonparametric drift estimation for a continuously observed one-dimensional diffusion process. *ESAIM: PS* **15** (2011) 197–216.

[32] M. GLow, On nonparametric confidence intervals. *Ann. Statist.* **25** (1997) 2547–2554.

[33] H. Masuda, I. Negri and Y. Nishiyama, Goodness-of-fit test for ergodic diffusions by discrete-time observations: an innovation martingale approach. *J. Nonparametr. Stat.* **23** (2011) 237–254.

[34] S.P. Meyn and R.L. Tweedie, *Markov chains and stochastic stability*. Cambridge University Press, Cambridge (2009).

[35] I. Negri and Y. Nishiyama, Goodness of fit test for ergodic diffusions by tick time sample scheme. *Stat. Inference Stoch. Process.* **13** (2010) 81–95.

[36] R. Nickl and J. Söhl, Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. To appear in *Ann. Statist.* (2016).

[37] C.P. Robert, Convergence control methods for Markov chain Monte Carlo algorithms. *Statistical Science* **10** (1995) 231–253.
[38] M. Rosenblatt, Density estimates and Markov sequences. In *Nonparametric Techniques in Statistical Inference (Proc. Sympos., Indiana Univ., Bloomington, Ind., 1969)*. Cambridge Univ. Press, London (1970) 199–213.
[39] G.G. Roussas, Nonparametric estimation in Markov processes. *Ann. Inst. Statist. Math.* **21** (1969) 73–87.
[40] E. Schmisser, Nonparametric estimation of the derivatives of the stationary density for stationary processes. *ESAIM: PS* **17** (2013) 33–69.
[41] V.G. Spokoiny, Adaptive drift estimation for nonparametric diffusion model. *Ann. Statist.* **28** (2000) 815–836.
[42] B. Szabó, A.W. van der Vaart and J.H. van Zanten, Frequentist coverage of adaptive nonparametric Bayesian credible sets (with discussion). *Ann. Statist.* **43** (2015) 1391–1428.
[43] H. Triebel, Theory of function spaces. Vol. 78 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel (1983).
[44] F. van der Meulen and H. van Zanten, Consistent nonparametric Bayesian inference for discretely observed scalar diffusions. *Bernoulli* **19** (2013) 44–63.
[45] A. van der Vaart and H. van Zanten, Donsker theorems for diffusions: necessary and sufficient conditions. *Ann. Probab.* **33** (2005) 1422–1451.
[46] A. W. van der Vaart, Asymptotic statistics. Vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge (1998).
[47] S. Yakowitz, Nonparametric density and regression estimation for Markov sequences without mixing assumptions. *J. Multivariate Anal.* **30** (1989) 124–136.