Bernoulli **25**(4A), 2019, 2696–2728
https://doi.org/10.3150/18-BEJ1067

Nonparametric Bayesian posterior contraction rates for scalar diffusions with high-frequency data

KWEKU ABRAHAM

Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, UK. E-mail: lkwabraham@statslab.cam.ac.uk

We consider inference in the scalar diffusion model $\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t$ with discrete data $(X_{j\Delta_n})_{0\le j\le n}$, $n\to\infty$, $\Delta_n\to 0$ and periodic coefficients. For σ given, we prove a general theorem detailing conditions under which Bayesian posteriors will contract in L^2-distance around the true drift function b_0 at the frequentist minimax rate (up to logarithmic factors) over Besov smoothness classes. We exhibit natural nonparametric priors which satisfy our conditions. Our results show that the Bayesian method adapts both to an unknown sampling regime and to unknown smoothness.

Keywords: adaptive estimation; Bayesian nonparametrics; concentration inequalities; diffusion processes; discrete time observations; drift function

1. Introduction

Consider a scalar diffusion process $(X_t)_{t\ge 0}$ starting at some X_0 and evolving according to the stochastic differential equation

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

where W_t is a standard Brownian motion. It is of considerable interest to estimate the parameters b and σ, which are arbitrary functions (until we place further assumptions on their form), so that the model is naturally *nonparametric*. As we will explain in Section 2, the problems of estimating σ and b can essentially be decoupled in the setting to be considered here, so in this paper we consider estimation of the drift function b when the diffusion coefficient σ is assumed to be given.

It is realistic to assume that we do not observe the full trajectory $(X_t)_{t\le T}$ but rather the process sampled at discrete time intervals $(X_{k\Delta})_{k\le n}$. The estimation problem for b and σ has been studied extensively and minimax rates have been attained in two sampling frameworks: *low-frequency*, where Δ is fixed and asymptotics are taken as $n\to\infty$ (see Gobet–Hoffmann–Reiss [17]), and *high-frequency*, where asymptotics are taken as $n\to\infty$ and $\Delta=\Delta_n\to 0$, typically assuming also that $n\Delta^2\to 0$ and $n\Delta\to\infty$ (see Hoffmann [19], Comte et al. [9]). See, for example, [10,18,27,33] for more papers addressing nonparametric estimation for diffusions.

For typical frequentist methods, one must know which sampling regime the data is drawn from. In particular, the low-frequency estimator from [17] is consistent in the high-frequency

setting but numerical simulations suggest it does not attain the minimax rate (see the discussion in Chorowski [8]), while the high-frequency estimators of [19] and [9] are not even consistent with low-frequency data. The only previous result known to the author regarding adaptation to the sampling regime in the nonparametric setting is found in [8], where Chorowski is able to estimate the diffusion coefficient σ but not the drift, and obtains the minimax rate when σ has 1 derivative but not for smoother diffusion coefficients.

For this paper, we consider estimation of the parameters in a diffusion model from a nonparametric Bayesian perspective. Bayesian methods for diffusion estimation can be implemented in practice (e.g., see Papaspiliopoulos et al. [25]). For Bayesian estimation, the statistician need only specify a prior, and for estimating diffusions from discrete samples the prior need not reference the sampling regime, so Bayesian methodology provides a natural candidate for a unified approach to the high- and low-frequency settings. Our results imply that Bayesian methods can adapt both to the sampling regime and also to unknown smoothness of the drift function (see the remarks after Proposition 4 and Proposition 2 respectively for details). These results are proved under the frequentist assumption of a fixed true parameter, so this paper belongs to the field of *frequentist analysis of Bayesian procedures*. See, for example, Ghosal and van der Vaart [13] for an introduction to this field.

It has previously been shown that in the low-frequency setting we have a *posterior contraction rate*, guaranteeing that posteriors corresponding to reasonable priors concentrate their mass on neighbourhoods of the true parameter shrinking at the fastest possible rate (up to log factors) – see Nickl and Söhl [24]. To complete a proof that such posteriors contract at a rate adapting to the sampling regime, it remains to prove a corresponding contraction rate in the high-frequency setting. This forms the key contribution of the current paper: we prove that a large class of "reasonable" priors will exhibit posterior contraction at the optimal rate (up to log factors) in L^2-distance. This in turn guarantees that point estimators based on the posterior will achieve the frequentist minimax optimal rate (see the remark after Theorem 1) in both high- and low-frequency regimes.

The broad structure of the proof is inspired by that in [24]: we use the testing approach of Ghosal–Ghosh–van der Vaart [11], coupled with the insight of Giné and Nickl [15] that one may prove the existence of the required tests by finding an estimator with good enough concentration around the true parameter. The main ingredients here are:

- A concentration inequality for a (frequentist) estimator, from which we construct tests of the true b_0 against a set of suitable (sufficiently separated) alternatives. See Section 4.
- A small ball result, to relate the L^2-distance to the information-theoretic Kullback–Leibler "distance". See Section 5.

Though the structure reflects that of [24] the details are very different. Estimators for the low-frequency setting are typically based on the mixing properties of $(X_{k\Delta})$ viewed as a Markov chain and the spectral structure of its transition matrix (see Gobet–Hoffmann–Reiss [17]) and fail to take full advantage of the local information one sees when $\Delta \to 0$. Here we instead use an estimator introduced in Comte et al. [9] which uses the assumption $\Delta \to 0$ to view estimation of b as a regression problem. To prove this estimator concentrates depends on a key insight of this paper: the Markov chain concentration results used in the low-frequency setting (which give *worse* bounds as $\Delta \to 0$) must be supplemented by Hölder type continuity results, which

crucially rely on the assumption $\Delta \to 0$. We further supplement by martingale concentration results.

Similarly, the small ball result in the low-frequency setting depends on Markov chain mixing. Here, we instead adapt the approach of van der Meulen and van Zanten [34]. They demonstrate that the Kullback–Leibler divergence in the discrete setting can be controlled by the corresponding divergence in the continuous data model; a key new result of the current paper is that in the high-frequency setting this control extends to give a bound on the variance of the log likelihood ratio.

As described above, a key attraction of the Bayesian method is that it allows the statistician to approach the low- and high-frequency regimes in a unified way. Another attraction is that it naturally suggests uncertainty quantification via posterior credible sets. The contraction rate theorems proved in this paper and [24] are not by themselves enough to prove that credible sets behave as advertised. For that one may aim for a nonparametric Bernstein–von Mises result – see, for example, Castillo and Nickl [6,7]. The posterior contraction rate proved here constitutes a key first step towards a proof of a Bernstein–von Mises result for the high-frequency sampled diffusion model, since it allows one to localise the posterior around the true parameter, as in the proofs in Nickl [23] for a non-linear inverse problem comparable to the problem here.

2. Framework and assumptions

The notation introduced throughout the paper is gathered in Appendix B.

We work with a scalar diffusion process $(X_t)_{t\geq 0}$ starting at some X_0 and evolving according to the stochastic differential equation

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t, \tag{1}$$

for W_t a standard Brownian motion. The parameters b and σ are assumed to be 1-periodic and we also assume the following.

Assumption 1. $\sigma \in C^2_{\mathrm{per}}([0,1])$ is given. Continuity guarantees the existence of an upper bound $\sigma_U < \infty$ and we further assume the existence of a lower bound $\sigma_L > 0$ so that $\sigma_L \leq \sigma(x) \leq \sigma_U$ for all $x \in [0,1]$. Here $C^2_{\mathrm{per}}([0,1])$ denotes $C^2([0,1])$ functions with periodic boundary conditions (i.e. $\sigma(0) = \sigma(1)$, $\sigma'(0) = \sigma'(1)$ and $\sigma''(0) = \sigma''(1)$).

Assumption 2. b is continuously differentiable with given norm bound. Precisely, we assume $b \in \Theta$, where

$$\Theta = \Theta(K_0) = \big\{ f \in C^1_{\mathrm{per}}\big([0,1]\big) :\ \|f\|_{C^1_{\mathrm{per}}} = \|f\|_\infty + \big\|f'\big\|_\infty \leq K_0 \big\}$$

for some arbitrary, but known, constant K_0 ($\|\cdot\|_\infty$ denotes the supremum norm, $\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|$). Note in particular that K_0 upper bounds $\|b\|_\infty$ and that b is Lipschitz continuous with constant at most K_0.

Θ is the maximal set over which we prove contraction, and we will in general make the stronger assumption that in fact $b \in \Theta_s(A_0)$, where

$$\Theta_s(A_0) := \big\{ f \in \Theta : \|f\|_{B^s_{2,\infty}} \leq A_0 < \infty \big\}, \qquad A_0 > 0,\ s \geq 1$$

with $B^s_{p,q}$ denoting a periodic Besov space and $\|\cdot\|_{B^s_{p,q}}$ denoting the associated norm: see Section 2.1 for a definition of the periodic Besov spaces we use (readers unfamiliar with Besov spaces may substitute the L^2-Sobolev space $H^s = B^s_{2,2} \subseteq B^s_{2,\infty}$ for $B^s_{2,\infty}$ and only mildly weaken the results). We generally assume the regularity index s is unknown. Our results will therefore aim to be *adaptive*, at least in the smoothness index (to be fully adaptive we would need to adapt to K_0 also).

Under Assumptions 1 and 2, there is a unique strong solution to (1) (see, e.g., Bass [3], Theorem 24.3). Moreover, this solution is also weakly unique (= unique in law) and satisfies the Markov property (see [3], Proposition 25.2 and Theorem 39.2). We denote by $P_b^{(x)}$ the law (on the cylindrical σ-algebra of $C([0,\infty])$) of the unique solution of (1) started from $X_0 = x$.

We consider "high-frequency data" $(X_{k\Delta_n})_{k=0}^n$ sampled from this solution, where asymptotics are taken as $n \to \infty$, with $\Delta_n \to 0$ and $n\Delta_n \to \infty$. We will suppress the subscript and simply write Δ for Δ_n. Throughout we will write $X^{(n)} = (X_0, \ldots, X_{n\Delta})$ as shorthand for our data and similarly we write $x^{(n)} = (x_0, \ldots, x_{n\Delta})$. We will denote by $\mathcal{I}$ the set $\{K_0, \sigma_L, \sigma_U\}$ so that, for example, $C(\mathcal{I})$ will be a constant depending on these parameters.

Beyond guaranteeing existence and uniqueness of a solution, our assumptions also guarantee the existence of transition densities for the discretely sampled process (see Gihman and Skorohod [14], Theorem 13.2 for an explicit formula for the transition densities). Morever, there also exists an invariant distribution μ_b, with a density π_b, for the periodised process $\dot{X} = X \bmod 1$. Defining $I_b(x) = \int_0^x \frac{2b}{\sigma^2}(y)\,\mathrm{d}y$ for $x \in [0,1]$, the density is

$$\pi_b(x) = \frac{e^{I_b(x)}}{H_b \sigma^2(x)} \left(e^{I_b(1)} \int_x^1 e^{-I_b(y)}\,\mathrm{d}y + \int_0^x e^{-I_b(y)}\,\mathrm{d}y \right), \qquad x \in [0,1],$$

$$H_b = \int_0^1 \frac{e^{I_b(x)}}{\sigma^2(x)} \left(e^{I_b(1)} \int_x^1 e^{-I_b(y)}\,\mathrm{d}y + \int_0^x e^{-I_b}(y) \right) \mathrm{d}y\,\mathrm{d}x,$$

(see Bhattacharya et al. [4], equations (2.15) to (2.17); note we have chosen a different normalisation constant so the expressions appear slightly different).

Observe that π_b is bounded uniformly away from zero and infinity, that is, there exist constants $0 < \pi_L, \pi_U < \infty$ depending only on $\mathcal{I}$ so that for any $b \in \Theta$ and any $x \in [0,1]$ we have $\pi_L \leq \pi_b(x) \leq \pi_U$. Precisely, we see that $\sigma_U^{-2} e^{-6K_0\sigma_L^{-2}} \leq H_b \leq \sigma_L^{-2} e^{6K_0\sigma_L^{-2}}$, and we deduce we can take $\pi_L = \pi_U^{-1} = \sigma_L^2 \sigma_U^{-2} e^{-12K_0\sigma_L^{-2}}$.

We assume that $X_0 \in [0,1)$ and that $X_0 = \dot{X}_0$ follows this invariant distribution.

Assumption 3. $X_0 \sim \mu_b$.

We will write P_b for the law of the full process X under Assumptions 1–3, and we will write E_b for expectation according to this law. Note μ_b is not invariant for P_b, but nevertheless

$E_b(f(X_t)) = E_b(f(X_0))$ for any 1-periodic function f (e.g., see the proof of Theorem 6). Since we will be estimating the 1-periodic function b, the assumption that $X_0 \in [0, 1)$ is unimportant.

Finally, we need to assume that $\Delta \to 0$ at a fast enough rate.

Assumption 4. $n\Delta^2 \log(1/\Delta) \leq L_0$ for some (unknown) constant L_0. Since we already assume $n\Delta \to \infty$, this new assumption is equivalent to $n\Delta^2 \log(n) \leq L_0'$ for some constant L_0'.

Throughout we make the frequentist assumption that the data is generated according to some fixed true parameter denoted b_0. We use μ_0 as shorthand for μ_{b_0}, and similarly for π_0 and so on. Where context allows, we write μ for μ_b with a generic drift b.

Remarks (Comments on assumptions). *Periodicity assumption.* We assume b and σ are periodic so that we need only estimate b on $[0, 1]$. One could alternatively assume b satisfies some growth condition ensuring recurrence, then estimate the restriction of b to $[0, 1]$, as in Comte et al. [9] and van der Meulen and van Zanten [34]. The proofs in this paper work in this alternative framework with minor technical changes, provided one assumes the behaviour of b outside $[0, 1]$ can be exactly matched by a draw from the prior.

Assuming that $\sigma \in C^2_{\text{per}}$ is given. If we observe continuous data $(X_t)_{t \leq T}$ then σ is known exactly (at least at any point visited by the process) via the expression for the quadratic variation $\langle X \rangle_t = \int_0^t \sigma^2(X_s)\, \mathrm{d}s$. With high-frequency data we cannot perfectly reconstruct the diffusion coefficient from the data, but we can estimate it at a much faster rate than the drift. When b and σ are both assumed unknown, if b is s-smooth and σ is s'-smooth, the minimax errors for b and σ respectively scale as $(n\Delta)^{-s/(1+2s)}$ and $n^{-s'/(1+2s')}$, as can be shown by slightly adapting Theorems 5 and 6 from Hoffmann [19] so that they apply in the periodic setting we use here. Since we assume that $n\Delta^2 \to 0$, it follows that $n\Delta \leq n^{1/2}$ for large n, hence we can estimate σ at a faster rate than b regardless of their relative smoothnesses.

Further, note that the problems of estimating b and σ in the high-frequency setting are essentially independent. For example, the smoothness of σ does not affect the rate for estimating b, and vice-versa – see [19]. We are therefore not substantially simplifying the problem of estimating b through the assumption that σ is given.

The assumption that σ^2 is twice differentiable is a typical assumption made to ensure transition densities exist.

Assuming a known bound on $\|b\|_{C^1_{\text{per}}}$. The assumption that b has one derivative is a typical assumption made to ensure that the diffusion equation (1) has a strong solution and that this solution has an invariant density and transition densities. The assumption of a *known* bound for the C^1_{per}-norm of the function is undesirable, but needed for the proofs, in particular to ensure the existence of a uniform lower bound π_L on the invariant densities. This lower bound is essential for the Markov chain mixing results as its reciprocal controls the mixing time in Theorem 6. It is plausible that needing this assumption is inherent to the problem rather than an artefact of the proofs: possible methods to bypass the Markov chain mixing arguments, such as the martingale approach of [9], Lemma 1, also rely on such a uniform lower bound. One could nonetheless hope that our results apply to an unbounded prior placing sufficient weight on $\Theta(K_n)$ for some slowly growing sequence K_n, but the lower bound π_L scales unfavourably as e^{-K_n}, which rules out this approach.

These boundedness assumptions in principle exclude Gaussian priors, which are computationally attractive. In practice, one could choose a very large value for K_0 and approximate Gaussian priors arbitrarily well using truncated Gaussian priors.

Assuming $X_0 \sim \mu_b$. It can be shown (see the proof of Theorem 6) that the law of $\dot{X}_t$ converges to μ_b at exponential rate from any starting distribution, so assuming $X_0 \sim \mu_b$ is not restrictive (as mentioned, our fixing $X_0 \in [0, 1)$ is arbitrary but unimportant).

Assuming $n\Delta^2 \log(1/\Delta) \leq L_0$. It is typical in the high-frequency setting to assume $n\Delta^2 \to 0$ (indeed the minimax rates in [19] are only proved under this assumption) but for technical reasons in the concentration section (Section 4.2) we need the above.

2.1. Spaces of approximation

We will throughout depend on a family $\{S_m : m \in \mathbb{N} \cup \{0\}\}$ of function spaces. For our purposes, we will take the S_m to be periodised Meyer-type wavelet spaces

$$S_m = \operatorname{span}\big(\{\psi_{lk} : 0 \leq k < 2^l, 0 \leq l < m\} \cup \{1\}\big).$$

We will denote $\psi_{-1,0} \equiv 1$ for convenience. Denote by $\langle \cdot, \cdot \rangle$ the $L^2([0,1])$ inner product and by $\|\cdot\|_2$ the L^2-norm, i.e. $\langle f, g \rangle = \int_0^1 f(x)g(x)\,\mathrm{d}x$ and $\|f\|_2 = \langle f, f \rangle^{1/2}$ for $f, g \in L^2([0,1])$. One definition of the (periodic) Besov norm $\|f\|_{B^s_{2,\infty}}$ is, for $f_{lk} := \langle f, \psi_{lk} \rangle$,

$$\|f\|_{B^s_{2,\infty}} = |f_{-1,0}| + \sup_{l \geq 0} 2^{ls} \left(\sum_{k=0}^{2^l - 1} f_{lk}^2 \right)^{1/2}, \tag{2}$$

with $B^s_{2,\infty}$ defined as those periodic $f \in L^2([0,1])$ for which this norm is finite. See Giné and Nickl [16] Sections 4.2.3 and 4.3.4 for a construction of periodised Meyer-type wavelets and a proof that this wavelet norm characterisation agrees with other possible definitions of the desired Besov space.

Note that the orthonormality of the wavelet basis means $\|f\|_2^2 = \sum_{l,k} f_{lk}^2$. Thus it follows from the above definition of the Besov norm that for any $b \in B^s_{2,\infty}([0,1])$ we have

$$\|\pi_m b - b\|_2 \leq K \|b\|_{B^s_{2,\infty}} 2^{-ms}, \tag{3}$$

for all m, for some constant $K = K(s)$, where π_m is the L^2-orthogonal projection map onto S_m.

Remarks. *Uniform sup-norm convergence of the wavelet series.* The wavelet projections $\pi_m b$ converge to b in supremum norm, uniformly across $b \in \Theta$. That is,

$$\sup_{b \in \Theta} \|\pi_m b - b\|_\infty \to 0 \quad \text{as} \quad m \to \infty. \tag{4}$$

This follows from Proposition 4.3.24 in [16] since K_0 uniformly bounds $\|b\|_{C^1_{\mathrm{per}}}$ for $b \in \Theta$.

Boundary regularity. Functions in the periodic Besov space here denoted $B^s_{2,\infty}$ are s regular at the boundary, in the sense that their weak derivatives of order s are 1-periodic.

Alternative approximation spaces. The key property we need for our approximation spaces is that (3) and (4) hold. Of these, only the first is needed of our spaces for our main contraction result Theorem 1. A corresponding inequality holds for many other function spaces if we replace 2^m by $D_m = \dim(S_m)$: for example, for S_m the set of trigonometric polynomials of degree at most m, or (provided $s \leq s_{\max}$ for some given $s_{\max} \in \mathbb{R}$) for S_m generated by periodised Daubechies wavelets. Priors built using these other spaces will achieve the same posterior contraction rate.

3. Main contraction theorem

Let Π be a (prior) probability distribution on some σ-algebra $\mathcal{S}$ of subsets of Θ. Given $b \sim \Pi$ assume that $(X_t : t \geq 0)$ follows the law P_b as described in Section 2. Write $p_b(\Delta, x, y)$ for the transition densities

$$p_b(\Delta, x, y)\,\mathrm{d}y = P_b(X_\Delta \in \mathrm{d}y \mid X_0 = x),$$

and recall we use p_0 as shorthand for p_{b_0}. Assume that the mapping $(b, \Delta, x, y) \mapsto p_b(\Delta, x, y)$ is jointly measurable with respect to the σ-algebras $\mathcal{S}$ and $\mathcal{B}_{\mathbb{R}}$, where $\mathcal{B}_{\mathbb{R}}$ is the Borel σ-algebra on $\mathbb{R}$. Then it can be shown by standard arguments that the Bayesian posterior distribution given the data is

$$b \mid X^{(n)} \sim \frac{\pi_b(X_0)\prod_{i=1}^n p_b(\Delta, X_{(i-1)\Delta}, X_{i\Delta})\,\mathrm{d}\Pi(b)}{\int_\Theta \pi_b(X_0)\prod_{i=1}^n p_b(\Delta, X_{(i-1)\Delta}, X_{i\Delta})\,\mathrm{d}\Pi(b)} \equiv \frac{p_b^{(n)}(X^{(n)})\,\mathrm{d}\Pi(b)}{\int_\Theta p_b^{(n)}(X^{(n)})\,\mathrm{d}\Pi(b)},$$

where we introduce the shorthand $p_b^{(n)}(x^{(n)}) = \pi_b(x_0)\prod_{i=1}^n p_b(\Delta, x_{(i-1)\Delta}, x_{i\Delta})$ for the joint probability density of the data $(X_0, \ldots, X_{n\Delta})$.

A main result of this paper is the following. Theorem 1A is designed to apply to adaptive sieve priors, while Theorem 1B is designed for use when the smoothness of the parameter b is known. See Section 3.1 for explicit examples of this result in use and see Section 6 for the proof.

Theorem 1. *Consider data $X^{(n)} = (X_{k\Delta})_{0\leq k\leq n}$ sampled from a solution X to (1) under Assumptions 1–4. Let the true parameter be b_0. Assume the appropriate sets below are measurable with respect to the σ-algebra $\mathcal{S}$.*

A. *Let Π be a sieve prior on Θ, i.e. let $\Pi = \sum_{m=1}^\infty h(m)\Pi_m$, where $\Pi_m(S_m \cap \Theta) = 1$, for S_m a periodic Meyer-type wavelet space of resolution m as described in Section 2.1, and h some probability mass function on $\mathbb{N}$. Suppose we have, for all $\varepsilon > 0$ and $m \in \mathbb{N}$, and for some constants $\zeta, \beta_1, \beta_2, B_1, B_2 > 0$,*
(i) $B_1 e^{-\beta_1 D_m} \leq h(m) \leq B_2 e^{-\beta_2 D_m}$,
(ii) $\Pi_m(\{b \in S_m : \|b - \pi_m b_0\|_2 \leq \varepsilon\}) \geq (\varepsilon\zeta)^{D_m}$,
where π_m is the L^2-orthogonal projection map onto S_m and $D_m = \dim(S_m) = 2^m$. Then for some constant $M = M(A_0, s, \mathcal{I}, L_0, \beta_1, \beta_2, B_1, B_2, \zeta)$ we have, for any $b_0 \in \Theta_s(A_0)$,

$$\Pi\big(\big\{b \in \Theta : \|b - b_0\|_2 \leq M(n\Delta)^{-s/(1+2s)}\log(n\Delta)^{1/2}\big\} \mid X^{(n)}\big) \to 1$$

in probability under the law P_{b_0} of X.

B. *Suppose now $b_0 \in \Theta_s(A_0)$ where $s \geq 1$ and $A_0 > 0$ are both known. Let $j_n \in \mathbb{N}$ be such that $D_{j_n} \sim (n\Delta)^{1/(1+2s)}$, that is, for some positive constants L_1, L_2 and all $n \in \mathbb{N}$ let $L_1(n\Delta)^{1/(1+2s)} \leq D_{j_n} \leq L_2(n\Delta)^{1/(1+2s)}$. Let $(\Pi^{(n)})_{n\in\mathbb{N}}$ be a sequence of priors satisfying, for $\varepsilon_n = (n\Delta)^{-s/(1+2s)} \log(n\Delta)^{1/2}$, and for some constant $\zeta > 0$,*
(I) $\Pi^{(n)}(\Theta_s(A_0) \cap \Theta) = 1$ *for all* n,
(II) $\Pi^{(n)}(\{b \in \Theta : \|\pi_{j_n} b - \pi_{j_n} b_0\|_2 \leq \varepsilon_n\}) \geq (\varepsilon_n \zeta)^{D_{j_n}}$.
Then we achieve the same rate of contraction; that is, for some $M = M(A_0, s, \mathcal{I}, L_0, \zeta)$,

$$\Pi^{(n)}\big(\{b \in \Theta : \|b - b_0\|_2 \leq M(n\Delta)^{-s/(1+2s)} \log(n\Delta)^{1/2}\} \mid X^{(n)}\big) \to 1$$

in probability under the law P_{b_0} of X.

Remark (Optimality). The minimax lower bounds of Hoffmann [19] do not strictly apply because we have assumed σ is given. Nevertheless, the minimax rate in this model should be $(n\Delta)^{-s/(1+2s)}$. This follows by adapting arguments for the continuous data case from Kutoyants [21], Section 4.5 to apply to the periodic model and observing that with high-frequency data we cannot outperform continuous data.

Since a contraction rate of ε_n guarantees the existence of an estimator converging to the true parameter at rate ε_n (for example, the centre of the smallest posterior ball of mass at least $1/2$ – see Theorem 8.7 in Ghosal and van der Vaart [13]) the rates attained in Theorem 1 are optimal, up to the log factors.

3.1. Explicit examples of priors

Our results guarantee that the following priors will exhibit posterior contraction. Throughout this section we continue to adopt Assumptions 1–4, and for technical convenience, we add an extra assumption on b_0. Precisely, recalling that $\{\psi_{lk}\}$ form a family of Meyer-type wavelets as in Section 2.1 and $\psi_{-1,0}$ denotes the constant function 1, we assume the following.

Assumption 5. For a sequence $(\tau_l)_{l\geq -1}$ to be specified and a constant B, we assume

$$b_0 = \sum_{\substack{l\geq -1 \\ 0\leq k<2^l}} \tau_l \beta_{lk} \psi_{lk}, \qquad \text{with } |\beta_{lk}| \leq B \text{ for all } l \geq -1 \text{ and all } 0 \leq k < 2^l. \tag{5}$$

The explicit priors for which we prove contraction will be random wavelet series priors. Let $u_{lk} \overset{iid}{\sim} q$, where q is a density on $\mathbb{R}$ satisfying

$$q(x) \geq \zeta \text{ for } |x| \leq B, \quad \text{and} \quad q(x) = 0 \text{ for } |x| > B + 1,$$

where $\zeta > 0$ is a constant and $B > 0$ is the constant from Assumption 5 (for example, one might choose q to be the density of a Unif$[0, B]$ random variable or a truncated Gaussian density). We

define a prior Π_m on S_m as the law associated to a random wavelet series

$$b(x) = \sum_{\substack{-1 \le l < m \\ 0 \le k < 2^l}} \tau_l u_{lk} \psi_{lk}(x), \qquad x \in [0,1], \tag{6}$$

for τ_l as in Assumption 5. We give three examples of priors built from these Π_m.

Example 1 (Basic sieve prior). Let $\tau_{-1} = \tau_0 = 1$ and let $\tau_l = 2^{-3l/2} l^{-2}$ for $l \ge 1$. Let h be a probability distribution on $\mathbb{N}$ as described in Theorem 1A, for example $h(m) = \gamma e^{-2^m}$, where γ is a normalising constant. Let $\Pi = \sum_{m=1}^{\infty} h(m)\Pi_m$ where Π_m is as above.

Proposition 2. *The preceding prior meets the conditions of Theorem 1A for any b_0 satisfying Assumption 5 with the same τ_l used to define the prior, and for an appropriate constant K_0. Thus, if also $b_0 \in \Theta_s(A_0)$ for some constant A_0, then for some constant M, we have $\Pi(\{b \in \Theta : \|b - b_0\|_2 \le M(n\Delta)^{-s/(1+2s)} \log(n\Delta)^{1/2}\} \mid X^{(n)}) \to 1$, in P_{b_0}-probability.*

The proof can be found in Section 6.

Remark (Adaptive estimation). If we assume $b_0 \in \Theta_{s_{\min}}(A_0)$ for some $s_{\min} > 3/2$, then Assumption 5 automatically holds with τ_l as in Example 1 for some constant $B = B(s_{\min}, A_0)$, as can be seen from the wavelet characterisation (2). Thus, in contrast to the low-frequency results of [24], the above prior adapts to unknown s in the range $s_{\min} \le s < \infty$.

When $s > 1$ is known, we fix the rate of decay of wavelet coefficients by hand to ensure a draw from the prior lies in $\Theta_s(A_0)$, rather than relying on the hyperparameter to choose the right resolution of wavelet space. We demonstrate with the following examples. The proofs of Propositions 3 and 4, given in the supplement (Abraham [1]), mimic that of Proposition 2 but rely on Theorem 1B in place of Theorem 1A.

Example 2 (Known smoothness prior). Let $\tau_{-1} = 1$ and let $\tau_l = 2^{-l(s+1/2)}$ for $l \ge 0$. Let $\bar{L}_n \in \mathbb{N} \cup \{\infty\}$. Define a sequence of priors $\Pi^{(n)} = \Pi_{\bar{L}_n}$ for b (we can take $\bar{L}_n = \infty$ to have a genuine prior, but a sequence of priors will also work provided $\bar{L}_n \to \infty$ at a fast enough rate).

Proposition 3. *Assume $\bar{L}_n/(n\Delta)^{1/(1+2s)}$ is bounded away from zero. Then for any $s > 1$, the preceding sequence of priors meets the conditions of Theorem 1B for any b_0 satisfying Assumption 5 with the same τ_l used to define the prior, and for an appropriate constant K_0. Thus, $\Pi^{(n)}(\{b \in \Theta : \|b - b_0\|_2 \le M(n\Delta)^{-s/(1+2s)} \log(n\Delta)^{1/2}\} \mid X^{(n)}) \to 1$ in P_{b_0}-probability, for some constant M.*

Remark. Assumption 5 with $\tau_l = 2^{-l(s+1/2)}$ in fact forces $b_0 \in B^s_{\infty,\infty} \subsetneq B^s_{2,\infty}$ with fixed norm bound. Restricting to this smaller set does not change the minimax rate, as can be seen from the fact that the functions by which Hoffmann perturbs in the lower bound proofs in [19] lie in the smaller class addressed here. In principle, one could weaken this assumption by taking $\tau_l = 2^{-ls}$ and taking the prior $\Pi^{(n)}$ to be the law of $b \sim \Pi_{\bar{L}_n}$ conditional on $b \in \Theta_s(A_0)$.

Example 3 (Prior on the invariant density). In some applications it may be more natural to place a prior on the invariant density and only implicitly model the drift function. With minor adjustments, Theorem 1B can still be applied to such priors. We outline the necessary adjustments.

(i) b is not identifiable from π_b and σ^2. We therefore introduce the identifiability constraint $I_b(1) = 0$. We could fix $I_b(1)$ as any positive constant and reduce to the case $I_b(1) = 0$ by a translation, so we choose $I_b(1) = 0$ for simplicity (this assumption is standard in the periodic model: for example, see van Waaij and van Zanten [35]). With this restriction, we have $\pi_b(x) = \frac{e^{I_b(x)}}{G_b\sigma^2(x)}$ for a normalising constant G_b, so that $b = ((\sigma^2)' + \sigma^2(\log \pi_b)')/2$.

(ii) In place of Assumption 5, we need a similar assumption but for $H_0 := \log \pi_{b_0}$. Precisely, we assume

$$H_0 = \sum_{\substack{l \geq -1 \\ 0 \leq k < 2^l}} \tau_l h_{lk} \psi_{lk}, \qquad \text{with } |h_{lk}| \leq B \text{ for all } l \geq -1 \text{ and all } 0 \leq k < 2^l, \tag{7}$$

for $\tau_{-1} = \tau_0 = 1$ and $\tau_l = 2^{-l(s+3/2)} l^{-2}$ for $l \geq 1$, for some known constant B, and where $s \geq 1$ is assumed known.

(iii) Induce priors on $b = ((\sigma^2)' + \sigma^2 H')/2$ by putting the priors $\Pi^{(n)} = \Pi_{\bar{L}_n}$ on H, where $\bar{L}_n$ is as in Proposition 3.

(iv) To ensure $b \in \Theta_s(A_0)$ we place further restrictions on σ; for example, we could assume σ^2 is smooth. More tightly, it is sufficient to assume (in addition to Assumption 1) that $\sigma^2 \in \Theta_{s+1}(A_1)$ and $\|\sigma^2\|_{C^s_{\text{per}}} \leq A_1$, where C^s_{per} is the Hölder norm, for some $A_1 > 0$. These conditions on σ can be bypassed with a more careful statement of Theorem 1B and a more careful treatment of the bias.

Proposition 4. *Make changes* (i)–(iv) *as listed above. Then, for some constant* M, *we have* $\Pi^{(n)}(\{b \in \Theta : \|b - b_0\|_2 \leq M(n\Delta)^{-s/(1+2s)} \log(n\Delta)^{1/2}\} \mid X^{(n)}) \to 1$ *in* P_{b_0}*-probability.*

Remarks. *Minimax rates.* The assumption (7) restricts b_0 beyond simply lying in $\Theta_s(A_0)$. As with Nickl and Söhl [24], Remark 5, this further restriction does not change the minimax rates, except for a log factor induced by the weights l^{-2}.

Adaptation to sampling regime. The prior of Proposition 4 is the same as the prior on b in [24]. However, since here we assume σ is given while in [24] it is an unknown parameter, the results of [24] do not immediately yield contraction of this prior at a near-minimax rate in the low-frequency setting. In particular, when σ is known the minimax rate for estimating b with low-frequency data is $n^{-s/(2s+3)}$ (for example, see Söhl and Trabs [31]), rather than the slower rate $n^{-s/(2s+5)}$ attained in Gobet–Hoffmann–Reiss [17] when σ is unknown (the improvement is possible because one bypasses the delicate interweaving of the problems of estimating b and σ with low-frequency data). Nevertheless, the prior of Proposition 4 will indeed exhibit near-minimax contraction also in the low-frequency setting. An outline of the proof is as follows. The small ball results of [24] still apply, with minor changes to the periodic model used here in place of their reflected diffusion, so it is enough to exhibit tests of the true parameter against suitably separated alternatives. The identification $b = ((\sigma^2)' + \sigma^2(\log \pi_b)')/2$ means one can work with

the invariant density rather than directly with the drift. Finally one shows the estimator from [31] exhibits sufficiently good concentration properties (alternatively, one could use general results for Markov chains from Ghosal and van der Vaart [12]).

It remains an interesting open problem to simultaneously estimate b and σ with a method which adapts to the sampling regime. Extending the proofs of this paper to the case where σ is unknown would show that the Bayesian method fulfils this goal. The key difficulty in making this extension arises in the small ball section (Section 5), because Girsanov's Theorem does not apply to diffusions with different diffusion coefficients.

Intermediate sampling regime. Strictly speaking, we only demonstrate robustness to the sampling regime in the extreme cases where $\Delta > 0$ is fixed or where $n\Delta^2 \to 0$. The author is not aware of any papers addressing the intermediate regime (where Δ tends to 0 at a slower rate than $n^{-1/2}$) for a nonparametric model: the minimax rates do not even appear in the literature. Since the Bayesian method adapts to the extreme regimes, one expects that it attains the correct rates in this intermediate regime (up to log factors). However, the proof would require substantial extra work, primarily in exhibiting an estimator with good concentration properties in this regime. Kessler's work on the intermediate regime in the parametric case [20] would be a natural starting point for exploring this regime in the nonparametric setting.

4. Construction of tests

In this section, we construct the tests needed to apply the general contraction rate theory from Ghosal–Ghosh–van der Vaart [11]. The main result of this section is the following. Recall that S_m is a periodic Meyer-type wavelet space of resolution m as described in Section 2.1, π_m is the L^2-orthogonal projection map onto S_m and $D_m = \dim(S_m) = 2^m$.

Lemma 5. *Consider data $X^{(n)} = (X_{k\Delta})_{0\le k\le n}$ sampled from a solution X to (1) under Assumptions 1–4. Let $\varepsilon_n \to 0$ be a sequence of positive numbers and let $l_n \to \infty$ be a sequence of positive integers such that $n\Delta\varepsilon_n^2/\log(n\Delta) \to \infty$ and, for some constant L and all n, $D_{l_n} \le Ln\Delta\varepsilon_n^2$. Let $\Theta_n \subseteq \{b \in \Theta : \|\pi_{l_n} b - b\|_2 \le \varepsilon_n\}$ contain b_0.*

Then for any $D > 0$, there is an $M = M(\mathcal{I}, L_0, D, L) > 0$ for which there exist tests ψ_n (i.e., $\{0,1\}$-valued functions of the data) such that, for all n sufficiently large,

$$\max\bigl(E_{b_0}\psi_n\bigl(X^{(n)}\bigr), \sup\bigl\{E_b\bigl[1-\psi_n\bigl(X^{(n)}\bigr)\bigr] : b \in \Theta_n, \|b-b_0\|_2 > M\varepsilon_n\bigr\}\bigr) \le e^{-Dn\Delta\varepsilon_n^2}.$$

The proof is given in Section 4.2 and is a straightforward consequence of our constructing an estimator with appropriate concentration properties. First, we introduce some general concentration results we will need.

4.1. General concentration results

We will use three forms of concentration results as building blocks for our theorems. The first comes from viewing the data $(X_{j\Delta})_{0\le j\le n}$ as a Markov chain and applying Markov chain concentration results; these results are similar to those used in Nickl and Söhl [24] for the low-frequency

case, but here we need to track the dependence of constants on Δ. The second form are useful only in the high-frequency case because they use a quantitative form of Hölder continuity for diffusion processes. An inequality of the third form, based on martingale properties, is introduced only where needed (in Lemma 13).

4.1.1. *Markov chain concentration results applied to diffusions*

Our main concentration result arising from the Markov structure is the following. We denote by $\|\cdot\|_\mu$ the $L^2_\mu([0,1])$-norm, $\|f\|_\mu^2 = E_\mu[f^2] = \int_0^1 f(x)^2\,\mathrm{d}\mu(x)$.

Theorem 6. *There exists a constant $\kappa = \kappa(\mathcal{I})$ such that, for all n sufficiently large and all bounded 1-periodic functions $f : \mathbb{R} \to \mathbb{R}$,*

$$P_b\left(\left|\sum_{k=1}^n f(X_{k\Delta}) - E_\mu[f]\right| \geq t\right) \leq 2\exp\left(-\frac{1}{\kappa}\Delta\min\left(\frac{t^2}{n\|f\|_\mu^2}, \frac{t}{\|f\|_\infty}\right)\right), \tag{8}$$

or equivalently

$$P_b\left(\left|\sum_{j=1}^n f(X_{j\Delta}) - E_\mu[f]\right| \geq \max\left(\sqrt{\kappa v^2 x}, \kappa u x\right)\right) \leq 2e^{-x}, \tag{9}$$

where $v^2 = n\Delta^{-1}\|f\|_\mu^2$ and $u = \Delta^{-1}\|f\|_\infty$.

Further, if $\mathcal{F}$ is a space of such functions indexed by some (subset of a) d-dimensional vector space, then for $V^2 = \sup_{f\in\mathcal{F}} v^2$ and $U = \sup_{f\in\mathcal{F}} u$, we also have

$$P_b\left(\sup_{f\in\mathcal{F}}\left|\sum_{j=1}^n f(X_{j\Delta}) - E_\mu[f]\right| \geq \tilde{\kappa}\max\left\{\sqrt{V^2(d+x)}, U(d+x)\right\}\right) \leq 4e^{-x}. \tag{10}$$

for some constant $\tilde{\kappa} = \tilde{\kappa}(\mathcal{I})$.

The proof is an application of the following abstract result for Markov chains.

Theorem 7 (Paulin [26], Proposition 3.4 and Theorem 3.4). *Let $M_1, \ldots, M_n$ be a time-homogeneous Markov chain taking values in S with transition kernel $P(x, \mathrm{d}y)$ and invariant density π. Suppose M is uniformly ergodic; that is, $\sup_{x\in S}\|P^n(x,\cdot) - \pi\|_{\mathrm{TV}} \leq K\rho^n$ for some constants $K < \infty$, $\rho < 1$, where $P^n(x,\cdot)$ is the n-step transition kernel and $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm for signed measures (which may be represented by their densities). Write $t_{mix} = \min\{n \geq 0 : \sup_{x\in S}\|P^n(x,\cdot) - \pi\|_{\mathrm{TV}} < 1/4\}$. Suppose $M_1 \sim \pi$ and $f : S \to \mathbb{R}$ is bounded. Let $V_f = \mathrm{Var}[f(M_1)]$, let $C = \|f - E[f(M_1)]\|_\infty$. Then*

$$\Pr\left(\left|\sum_{i=1}^n f(M_i) - E\big[f(M_i)\big]\right| \geq t\right) \leq 2\exp\left(\frac{-t^2}{2t_{mix}(8(n+2t_{mix})V_f + 20tC)}\right).$$

Proof of Theorem 6. Since f is assumed periodic we see that $f(X_{k\Delta}) = f(\dot{X}_{k\Delta})$, where we recall $\dot{X} = X \bmod 1$. Denote by $\dot{p}_b(t, x, y)$ the transition densities of $\dot{X}$, that is, $\dot{p}_b(t, x, y) = \sum_{j\in\mathbb{Z}} p_b(t, x, y + j)$ (see the proof of Proposition 9 in Nickl and Söhl [24] for an argument that the sum converges). Theorem 2.6 in Bhattacharya et al. [4] tells us that if $\dot{X}_0$ has a density η_0 on $[0, 1]$, then $\dot{X}_t$ has a density η_t satisfying

$$\|\eta_t - \pi_b\|_{\mathrm{TV}} \le \frac{1}{2}\|\eta_0/\pi_b - 1\|_{\mathrm{TV}} \exp\left(-\frac{1}{2M_b}t\right),$$

where $M_b := \sup_{z\in[0,1]}\{(\sigma^2(z)\pi_b(z))^{-1}\int_0^z \pi_b(x)\,\mathrm{d}x \int_z^1 \pi_b(y)\,\mathrm{d}y\}$. We can regularise to extend the result so that it also applies when the initial distribution of $\dot{X}$ is a point mass: if $\dot{X}_0 = x$ then $\dot{X}_1$ has density $\dot{p}_b(1, x, \cdot)$, hence the result applies to show

$$\left\|\dot{p}_b(t, x, \cdot) - \pi_b\right\|_{\mathrm{TV}} \le \frac{1}{2}\left\|\dot{p}_b(1, x, \cdot)/\pi_b - 1\right\|_{\mathrm{TV}} \exp\left(-\frac{1}{2M_b}(t-1)\right).$$

Moreover, note $\|\dot{p}_b(1, x, \cdot)/\pi_b - 1\|_{\mathrm{TV}} \le \pi_L^{-1}\|\dot{p}_b(1, x, \cdot) - \pi_b\|_{\mathrm{TV}} \le \pi_L^{-1}$. Also note we can upper bound M_b by a constant $M = M(\mathcal{I})$: precisely, we can take $M = \sigma_L^{-2}\pi_L^{-1}\pi_U^2$.

Thus, we see that for $t \ge 1$, we have

$$\left\|\dot{p}_b(t, x, \cdot) - \pi_b\right\|_{\mathrm{TV}} \le K \exp\left(-\frac{1}{2M}t\right)$$

for some constant $K = K(\mathcal{I})$, uniformly across $x \in [0, 1]$. It follows that, for each fixed Δ, the discrete time Markov chain $(\dot{X}_{k\Delta})_{k\ge0}$ is uniformly ergodic with mixing time $t_{\mathrm{mix}} \le 1 + 2M\log(4K)\Delta^{-1} \le K'\Delta^{-1}$ for some constant K'. Theorem 7 applies to tell us

$$\Pr\left(\left|\sum_{i=1}^n f(X_{k\Delta}) - E_\mu[f]\right| \ge t\right) \le 2\exp\left(-\frac{t^2}{2K'\Delta^{-1}(8(n + 2K'\Delta^{-1})V_f + 20tC)}\right).$$

Since $n\Delta \to \infty$ by assumption, we see $8(n + 2K'\Delta^{-1}) \le K''n$ for some constant K''. Using the bound $2/(a + b) \ge \min(1/a, 1/b)$ for $a, b > 0$ and upper bounding the centred moments V_f and C by the uncentered moments $\|f\|_\mu^2$ and $\|f\|_\infty$, we deduce (8).

The result (9) is obtained by a change of variables. For the supremum result (10), we use a standard chaining argument, for example, as in Baraud [2], Theorem 2.1, where we use (9) in place of Baraud's Assumption 2.1, noting that Baraud only uses Assumption 2.1 to prove an expression mirroring (9), and the rest of the proof follows through exactly. Precisely, following the proof, we can take $\tilde{\kappa} = 36\kappa$. □

Remark. The proof simplifies if we restrict Θ to consist only of those b satisfying $I_b(1) = 0$. In this case, the invariant density (upon changing normalising constant to some G_b) reduces to the more familiar form $\pi_b(x) = (G_b\sigma^2(x))^{-1}e^{I_b(x)}$. The diffusion is reversible in this case, and we can use Theorem 3.3 from [26] instead of Theorem 3.4 to attain the same results but with better constants.

4.1.2. *Hölder continuity properties of diffusions*

Define

$$w_m(\delta) = \delta^{1/2}\big((\log \delta^{-1})^{1/2} + \log(m)^{1/2}\big), \qquad \delta \in (0, 1]$$

for $m \geq 1$, and write $w_m(\delta) := w_1(\delta)$ for $m < 1$. The key result of this section is the following.

Lemma 8. *Let X solve the scalar diffusion equation (1), and grant Assumptions 1 and 2. Then there exist positive constants λ, C and τ, all depending on $\mathcal{I}$ only, such that for any $u > C \max(\log(m), 1)^{1/2}$ and for any initial value x,*

$$P_b^{(x)}\left(\sup_{\substack{s,t\in[0,m],\\ t\neq s, |t-s|\leq\tau}} \left(\frac{|X_t - X_s|}{w_m(|t-s|)}\right) > u\right) \leq 2e^{-\lambda u^2}.$$

Remarks.

i. We will need to control all increments $X_{(j+1)\Delta} - X_{j\Delta}$ simultaneously, hence we include the parameter m, which we will take to be the time horizon $n\Delta$ when applying this result. Simply controlling over [0, 1] and using a union bound does not give sharp enough results.
ii. Lemma 8 applies for any distribution of X_0, not only point masses, by an application of the tower law.

The modulus of continuity w_m matches that of Brownian motion, and indeed the proof is to reduce to the corresponding result for Brownian motion. First, by applying the scale function (see the supplement – Abraham [1] – for details) one transforms X into a local martingale, reducing Lemma 8 to the following result, also useful in its own right.

Lemma 9. *Let Y be a local martingale with quadratic variation satisfying $|\langle Y\rangle_t - \langle Y\rangle_s| \leq A|t-s|$, for some constant $A \geq 1$. Then there exist positive constants $\lambda = \lambda(A)$ and $C = C(A)$ such that for any $u > C \max(\log(m), 1)^{1/2}$,*

$$\Pr\left(\sup_{\substack{s,t\in[0,m], s\neq t,\\ |t-s|\leq A^{-1}e^{-2}}} \left(\frac{|Y_t - Y_s|}{w_m(|t-s|)}\right) > u\right) \leq 2e^{-\lambda u^2}.$$

In particular the result applies when Y is a solution to $\mathrm{d}Y_t = \tilde{\sigma}(Y_t)\,\mathrm{d}W_t$, provided $\|\tilde{\sigma}^2\|_\infty \leq A$.

Lemma 9, proved in the supplement (Abraham [1]), follows from the corresponding result for Brownian motion by a time change (i.e., the (Dambis–)Dubins–Schwarz Theorem: see Rogers and Williams [30] (34.1)). It is well known that Brownian motion has modulus of continuity $\delta^{1/2}(\log \delta^{-1})^{1/2}$ in the sense that there almost surely exists a $C > 0$ such that $|B_t - B_s| \leq C|t-s|^{1/2}(\log(|t-s|^{-1}))^{1/2}$, for all $t, s \in [0, 1]$ sufficiently close, but Lemmas 8 and 9 depend on the following quantitative version of this statement, proved in the supplement (Abraham [1]) using Gaussian process techniques.

Lemma 10. *Let B be a standard Brownian motion on $[0,m]$. There are positive (universal) constants λ and C such that for $u > C\max(\log(m),1)^{1/2}$,*

$$\Pr\Bigg(\sup_{\substack{s,t\in[0,m],\\ s\neq t,|t-s|\leq e^{-2}}}\left(\frac{|B_t-B_s|}{w_m(|t-s|)}\right)>u\Bigg)\leq 2e^{-\lambda u^2}.$$

4.2. Concentration of a drift estimator

4.2.1. *Defining the estimator*

We adapt an estimator introduced in Comte et al. [9]. The estimator is constructed by considering drift estimation as a regression-type problem. Specifically, defining

$$Z_{k\Delta}=\frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s)\,\mathrm{d}W_s,\qquad R_{k\Delta}=\frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}\big(b(X_s)-b(X_{k\Delta})\big)\,\mathrm{d}s,$$

we can write

$$\frac{X_{(k+1)\Delta}-X_{k\Delta}}{\Delta}=b(X_{k\Delta})+Z_{k\Delta}+R_{k\Delta}.$$

Note $R_{k\Delta}$ is a discretization error which vanishes as $\Delta\to 0$ and $Z_{k\Delta}$ takes on the role of noise. We introduce the *empirical norm* and the related *empirical loss function*, defined for $u:[0,1]\to\mathbb{R}$ by

$$\|u\|_n^2=\frac{1}{n}\sum_{k=1}^{n}u(X_{k\Delta})^2,\qquad \gamma_n(u)=\frac{1}{n}\sum_{k=1}^{n}\big[\Delta^{-1}(X_{(k+1)\Delta}-X_{k\Delta})-u(X_{k\Delta})\big]^2.$$

In both we leave out the $k=0$ term for notational convenience.

Recalling that S_m is a Meyer-type wavelet space as described in Section 2.1 and K_0 is an upper bound for the C^1_{per}-norm of any $b\in\Theta$, for l_n to be chosen we define $\tilde{b}_n$ as a solution to the minimisation problem

$$\tilde{b}_n\in\operatorname*{argmin}_{u\in\tilde{S}_{l_n}}\gamma_n(u),\qquad \tilde{S}_m:=\big\{u\in S_m:\|u\|_\infty\leq K_0+1\big\},$$

where we choose arbitrarily among minimisers in the (typical) situation that there is no unique minimiser.

4.2.2. *Main concentration result*

For the estimator defined above, we will prove the following concentration inequality.

Theorem 11. *Consider data $X^{(n)}=(X_{k\Delta})_{0\leq k\leq n}$ sampled from a solution X to (1) under Assumptions 1–4. Let $\varepsilon_n\to 0$ be a sequence of positive numbers and let $l_n\to\infty$ be a sequence of positive integers such that $n\Delta\varepsilon_n^2/\log(n\Delta)\to\infty$ and, for some constant L and all n,*

$D_{l_n} \leq Ln\Delta\varepsilon_n^2$. For these l_n, let $\tilde{b}_n$ be defined as above and let $\Theta_n \subseteq \{b \in \Theta : \|\pi_{l_n} b - b\|_2 \leq \varepsilon_n\}$ contain b_0, where π_{l_n} is the L^2-orthogonal projection map onto S_{l_n}.

Then for any $D > 0$ there is a $C = C(\mathcal{I}, L_0, D, L) > 0$ such that, uniformly across $b \in \Theta_n$,

$$P_b\big(\|\tilde{b}_n - b\|_2 > C\varepsilon_n\big) \leq e^{-Dn\Delta\varepsilon_n^2},$$

for all n sufficiently large.

Remark. Previous proofs of Bayesian contraction rates using the concentration of estimators approach (for example, see [15,24,29]) have used duality arguments, i.e. the fact that $\|f\|_2 = \sup_{v:\|v\|_2=1}\langle f, v\rangle$, to demonstrate that the linear estimators considered satisfy a concentration inequality of the desired form. A key insight of this paper is that for the model we consider we can achieve the required concentration using the above *minimum contrast* estimator (see Birgé and Massart [5]), for which we need techniques which differ substantially from duality arguments.

Before proceeding to the proof, we demonstrate how this can be used to prove the existence of tests of b_0 against suitably separated alternatives.

Proof of Lemma 5. Let $\tilde{b}_n$ be the estimator outlined above and let $D > 0$. Let $C = C(\mathcal{I}, L_0, D, L)$ be as in Theorem 11 and let $M = 2C$. It's not hard to see that $\psi_n = \mathbb{1}\{\|\tilde{b}_n - b\|_2 > C\varepsilon_N\}$ is a test with the desired properties. □

Proof of Theorem 11. It is enough to show that, uniformly across $b \in \Theta_n$, for any $D > 0$ there is a $C > 0$ such that $P_b(\|\tilde{b}_n - b\|_2 > C\varepsilon_n) \leq 14e^{-Dn\Delta\varepsilon_n^2}$, because by initially considering a $D' > D$ and finding the corresponding C', we can eliminate the factor of 14 in front of the exponential.

The proof is structured as follows. Our assumptions ensure that the L^2- and $L^2(\mu)$-norms are equivalent. We further show that the $L^2(\mu)$-norm is equivalent to the empirical norm $\|\cdot\|_n$ on an event of sufficiently high probability. Finally, the definition of the estimator will allow us to control the empirical distance $\|\tilde{b}_n - b\|_n$.

To this end, write $\tilde{t}_n = (\tilde{b}_n - \pi_{l_n} b)\|\tilde{b}_n - \pi_{l_n} b\|_\mu^{-1}$ (defining $\tilde{t}_n = 0$ if $\tilde{b}_n = \pi_{l_n} b$) and introduce the following set and events:

$$\begin{aligned} I_n &= \big\{t \in S_{l_n} : \|t\|_\mu = 1, \|t\|_\infty \leq C_1\varepsilon_n^{-1}\big\}, \\ \mathcal{A}_n &= \big\{\tilde{t}_n \in I_n\big\} \cup \big\{\tilde{t}_n = 0\big\}, \\ \Omega_n &= \Big\{\big|\|t\|_n^2 - 1\big| \leq \frac{1}{2},\ \forall t \in I_n\Big\}, \end{aligned}$$

where the constant C_1 is to be chosen. Then we can decompose

$$P_b\big(\|\tilde{b}_n - b\|_2 > C\varepsilon_n\big) \leq P_b\big(\|\tilde{b}_n - b\|_2 \mathbb{1}_{\mathcal{A}_n^c} > C\varepsilon_n\big) + P_b\big(\Omega_n^c\big) + P_b\big(\|\tilde{b}_n - b\|_2 \mathbb{1}_{\mathcal{A}_n \cap \Omega_n} > C\varepsilon_n\big).$$

Thus, we will have proved the theorem once we have completed the following:

1. Show the theorem holds (deterministically) on $\mathcal{A}_n^c$, for a large enough constant C.

2. Show that $P_b(\Omega_n^c) \leq 4e^{-Dn\Delta\varepsilon_n^2}$ for a suitable choice of C_1.
3. Show that, for any D, we can choose a C such that $P_b(\|\tilde{b}_n - b\|_2 \mathbb{1}_{\mathcal{A}_n \cap \Omega_n} > C\varepsilon_n) \leq 10e^{-Dn\Delta\varepsilon_n^2}$.

Step 1

Intuitively we reason thus. The event $\mathcal{A}_n^c$ can only occur if the $L^2(\mu)$-norm of $\tilde{b}_n - \pi_{l_n} b$ is small compared to the L^∞-norm. Since we have assumed a uniform supremum bound on functions $b \in \Theta$, in fact $\mathcal{A}_n$ holds unless the $L^2(\mu)$-norm is small in absolute terms. But if $\|\tilde{b}_n - \pi_{l_n} b\|_\mu$ is small, then so is $\|\tilde{b}_n - b\|_2$. We formalise this reasoning now.

For a constant C_2 to be chosen, define

$$\mathcal{A}_n' = \big\{\|\tilde{b}_n - \pi_{l_n} b\|_\mu > C_2\varepsilon_n\big\}.$$

On $\mathcal{A}_n'$ we have $\|\tilde{t}_n\|_\infty \leq (\|\tilde{b}_n\|_\infty + \|\pi_{l_n} b\|_\infty) C_2^{-1}\varepsilon_n^{-1}$. Note $\|\tilde{b}_n\|_\infty \leq K_0 + 1$ by definition. Since, for n large enough, $\|\pi_{l_n} b - b\|_\infty \leq 1$ uniformly across $b \in \Theta_n \subseteq \Theta$ by (4) so that $\|\pi_{l_n} b\|_\infty \leq \|b\|_\infty + 1 \leq K_0 + 1$, we deduce that on $\mathcal{A}_n'$, $\|\tilde{t}_n\|_\infty \leq (2K_0 + 2) C_2^{-1}\varepsilon_n^{-1}$. Since also $\|\tilde{t}_n\|_\mu = 1$ (or $\tilde{t}_n = 0$) by construction, we deduce $\mathcal{A}_n' \subseteq \mathcal{A}_n$ if $C_2 \geq C_1^{-1}(2K_0 + 2)$.

Then on $(\mathcal{A}_n')^c \supseteq \mathcal{A}_n^c$ we find, using that $b \in \Theta_n$ and using $\|\cdot\|_2 \leq \pi_L^{-1/2}\|\cdot\|_\mu$,

$$\|\tilde{b}_n - b\|_2 \leq \|\tilde{b}_n - \pi_{l_n} b\|_2 + \|\pi_{l_n} b - b\|_2 \leq \big(C_2\pi_L^{-1/2} + 1\big)\varepsilon_n.$$

So on $\mathcal{A}_n^c$, we have $\|\tilde{b}_n - b\|_2 \leq C\varepsilon_n$ deterministically for any $C \geq C_2\pi_L^{-1/2} + 1$. That is, $P_b(\|\tilde{b}_n - b\|_2 \mathbb{1}_{\mathcal{A}_n^c} > C\varepsilon_n) = 0$ for C large enough (depending on C_1 and $\mathcal{I}$).

Step 2

We show that for n sufficiently large, and $C_1 = C_1(\mathcal{I}, D, L)$ sufficiently small, $P_b(\Omega_n^c) \leq 4e^{-Dn\Delta\varepsilon_n^2}$.

For $t \in I_n$ we have $|\|t\|_n^2 - 1| = n^{-1}|\sum_{k=1}^n t^2(X_{k\Delta}) - E_\mu[t^2]|$. Thus Theorem 6 can be applied to $\Omega_n^c = \{\sup_{t \in I_n} n^{-1}|\sum_{k=1}^n t^2(X_{k\Delta}) - E_\mu[t^2]| > 1/2\}$. Each $t \in I_n$ has $\|t^2\|_\infty \leq C_1^2\varepsilon_n^{-2}$ and $\|t^2\|_\mu^2 = E_\mu[t^4] \leq \|t^2\|_\infty \|t\|_\mu^2 \leq C_1^2\varepsilon_n^{-2}$. Since the indexing set I_n lies in a vector space of dimension D_{l_n}, we apply Theorem 6 with $x = Dn\Delta\varepsilon_n^2$ to see

$$P_b\left(\sup_{t \in I_n}\left|\sum_{k=1}^n t^2(X_{k\Delta}) - E_\mu\big[t^2\big]\right| \geq 36\max\{A, B\}\right) \leq 4e^{-Dn\Delta\varepsilon_n^2},$$

where $A = \sqrt{\tilde{\kappa} C_1^2 n\Delta^{-1}\varepsilon_n^{-2}(Dn\Delta\varepsilon_n^2 + D_{l_n})}$ and $B = \tilde{\kappa} C_1^2\Delta^{-1}\varepsilon_n^{-2}(Dn\Delta\varepsilon_n^2 + D_{l_n})$, for some constant $\tilde{\kappa} = \tilde{\kappa}(\mathcal{I})$. Provided we can choose C_1 so that $36\max\{A/n, B/n\} \leq 1/2$ the result is proved. Such a choice for C_1 can be made as we have assumed $D_{l_n} \leq Ln\Delta\varepsilon_n^2$.

Step 3

Since $b \in \Theta_n$ and π_{l_n} is an L^2-orthogonal projection, we have $\|\tilde{b}_n - b\|_2^2 \le \|\tilde{b}_n - \pi_{l_n} b\|_2^2 + \varepsilon_n^2$. Recall that $\|\cdot\|_2 \le \pi_L^{-1/2}\|\cdot\|_\mu$ and note that on $\mathcal{A}_n \cap \Omega_n$, we further have $\frac{1}{2}\|\tilde{b}_n - \pi_{l_n} b\|_\mu^2 \le \|\tilde{b}_n - \pi_{l_n} b\|_n^2$.

Since also $\|\tilde{b}_n - \pi_{l_n} b\|_n^2 \le 2(\|\pi_{l_n} b - b\|_n^2 + \|\tilde{b}_n - b\|_n^2)$ we deduce that

$$\|\tilde{b}_n - b\|_2^2 \mathbb{1}_{\mathcal{A}_n \cap \Omega_n} \le \frac{1}{\pi_L}\big(4\|\pi_{l_n} b - b\|_n^2 + 4\|\tilde{b}_n - b\|_n^2 \mathbb{1}_{\mathcal{A}_n \cap \Omega_n}\big) + \varepsilon_n^2,$$

where we have dropped indicator functions from terms on the right except where we will need them later. Thus, using a union bound,

$$P_b\big(\|\tilde{b}_n - b\|_2 \mathbb{1}_{\mathcal{A}_n \cap \Omega_n} > C\varepsilon_n\big) \le P_b\big(\|\pi_{l_n} b - b\|_n^2 > C'\varepsilon_n^2\big) + P_b\big(\|\tilde{b}_n - b\|_n^2 \mathbb{1}_{\mathcal{A}_n \cap \Omega_n} > C'\varepsilon_n^2\big),$$

for some constant C' (precisely we can take $C' = \pi_L(C^2 - 1)/8$). It remains to show that both probabilities on the right are exponentially small.

Bounding $P_b(\|\pi_{l_n} b - b\|_n > C\varepsilon_n)$. We show that for any $D > 0$ there is a constant C such that $P_b(\|\pi_{l_n} b - b\|_n > C\varepsilon_n) \le 2e^{-Dn\Delta\varepsilon_n^2}$, for all n sufficiently large.

Since $E_b\|g\|_n^2 = \|g\|_\mu^2$ for any 1-periodic deterministic function g and $\|\pi_{l_n} b - b\|_\mu^2 \le \pi_U \|\pi_{l_n} b - b\|_2^2 \le \pi_U \varepsilon_n^2$ for $b \in \Theta_n$, it is enough to show that

$$P_b\big(\big|\|\pi_{l_n} b - b\|_n^2 - E_b\|\pi_{l_n} b - b\|_n^2\big| > C\varepsilon_n^2\big) \le 2e^{-Dn\Delta\varepsilon_n^2} \tag{11}$$

for some different C. As in step 2, we apply Theorem 6, but now working with the single function $(\pi_{l_n} b - b)^2$. For large enough n we have the bounds $\|\pi_{l_n} b - b\|_\infty \le 1$ (derived from (4)), and $\|(\pi_{l_n} b - b)^2\|_\mu \le \|\pi_{l_n} b - b\|_\infty \|\pi_{l_n} b - b\|_\mu \le \pi_U^{1/2}\varepsilon_n$ (because $b \in \Theta_n$) and so applying Theorem 6 with $x = Dn\Delta\varepsilon_n^2$ gives

$$P_b\left(\left|\sum_{k=1}^{n}\big[(\pi_{l_n} b - b)^2(X_{k\Delta}) - \|\pi_{l_n} b - b\|_\mu^2\big]\right| \ge \max\{a, b\}\right) \le 2e^{-Dn\Delta\varepsilon_n^2},$$

for $a = \sqrt{\kappa n\Delta^{-1}\pi_U\varepsilon_n^2 Dn\Delta\varepsilon_n^2} = n\varepsilon_n^2\sqrt{\kappa\pi_U D}$ and $b = \kappa\Delta^{-1}Dn\Delta\varepsilon_n^2 = n\varepsilon_n^2\kappa D$, for some constant $\kappa = \kappa(\mathcal{I})$. We see that a/n and b/n are both upper bounded by a constant multiple of ε_n^2, hence, by choosing C large enough, (11) holds.

Bounding $P_b(\|\tilde{b}_n - b\|_n^2 \mathbb{1}_{\mathcal{A}_n \cap \Omega_n} > C\varepsilon_n^2)$. We show that $P_b(\|\tilde{b}_n - b\|_n^2 \mathbb{1}_{\mathcal{A}_n \cap \Omega_n} > C\varepsilon_n^2) \le 8e^{-Dn\Delta\varepsilon_n^2}$ for some constant C.

Recall an application of (4) showed us that $\|\pi_{l_n} b\|_\infty \le K_0 + 1$ for sufficiently large n, hence we see that $\pi_{l_n} b$ lies in $\tilde{S}_{l_n}$, so by definition $\gamma_n(\tilde{b}_n) \le \gamma_n(\pi_{l_n} b)$. We now use this to show that

$$\frac{1}{4}\|\tilde{b}_n - b\|_n^2 \mathbb{1}_{\mathcal{A}_n \cap \Omega_n} \le \frac{7}{4}\|\pi_{l_n} b - b\|_n^2 + 8\nu_n(\tilde{t}_n)^2 \mathbb{1}_{\mathcal{A}_n} + \frac{8}{n}\sum_{k=1}^{n} R_{k\Delta}^2, \tag{12}$$

where $\nu_n(t) = \frac{1}{n}\sum_{k=1}^n t(X_{k\Delta})Z_{k\Delta}$ and we recall that $\tilde{t}_n = (\tilde{b}_n - \pi_{l_n}b)\|\tilde{b}_n - \pi_{l_n}b\|_\mu^{-1}$. The argument, copied from [9], Sections 3.2 and 6.1, is as follows. Using that $\Delta^{-1}(X_{(k+1)\Delta} - X_{k\Delta}) = b(X_{k\Delta}) + Z_{k\Delta} + R_{k\Delta}$ and that $\gamma_n(\tilde{b}_n) - \gamma_n(b) \le \gamma_n(\pi_{l_n}b) - \gamma_n(b)$, one shows that

$$\|\tilde{b}_n - b\|_n^2 \le \|\pi_{l_n}b - b\|_n^2 + 2\nu_n(\tilde{b}_n - \pi_{l_n}b) + \frac{2}{n}\sum_{k=1}^n R_{k\Delta}(\tilde{b}_n - \pi_{l_n}b)(X_{k\Delta}). \tag{13}$$

Repeatedly applying the AM–GM-derived inequality $2ab \le 8a^2 + b^2/8$ yields

$$\frac{2}{n}\sum_{k=1}^n R_{k\Delta}(\tilde{b}_n - \pi_{l_n}b)(X_{k\Delta}) \le \frac{8}{n}\sum_{k=1}^n R_{k\Delta}^2 + \frac{1}{8}\|\tilde{b}_n - \pi_{l_n}b\|_n^2,$$

$$2\nu_n(\tilde{b}_n - \pi_{l_n}b) = 2\|\tilde{b}_n - \pi_{l_n}b\|_\mu \nu_n(\tilde{t}_n) \le 8\nu_n(\tilde{t}_n)^2 + \frac{1}{8}\|\tilde{b}_n - \pi_{l_n}b\|_\mu^2.$$

Next recall that on $\mathcal{A}_n \cap \Omega_n$, we have $\|\tilde{b}_n - \pi_{l_n}b\|_\mu^2 \le 2\|\tilde{b}_n - \pi_{l_n}b\|_n^2$, and further recall that $\|\tilde{b}_n - \pi_{l_n}b\|_n^2 \le 2\|\tilde{b}_n - b\|_n^2 + 2\|\pi_{l_n}b - b\|_n^2$. Putting all these bounds into (13) yields (12), where on the right-hand side we have only included indicator functions where they will help us in future steps. Next, by a union bound, we deduce

$$P_b\big(\|\tilde{b}_n - b\|_n^2 \mathbb{1}_{\mathcal{A}_n\cap\Omega_n} > C\varepsilon_n^2\big) \le P_b\big(\|\pi_{l_n}b - b\|_n^2 > C'\varepsilon_n^2\big) + P_b\big(\nu_n(\tilde{t}_n)^2\mathbb{1}_{\mathcal{A}_n} > C'\varepsilon_n^2\big) + P_b\left(\frac{1}{n}\sum_{k=1}^n R_{k\Delta}^2 > C'\varepsilon_n^2\right),$$

for some constant C' (we can take $C' = C/96$). We have already shown that for a large enough constant C we have $P_b(\|\pi_{l_n}b - b\|_n > C\varepsilon_n) \le 2e^{-Dn\Delta\varepsilon_n^2}$, thus the following two lemmas conclude the proof. □

Lemma 12. *Under the conditions of Theorem 11, for any $D > 0$ there is a constant $C = C(\mathcal{I}, L_0, D) > 0$ for which, for n sufficiently large, $P_b(\frac{1}{n}\sum_{k=1}^n R_{k\Delta}^2 > C\varepsilon_n^2) \le 2e^{-Dn\Delta\varepsilon_n^2}$.*

Lemma 13. *Under the conditions of Theorem 11, for any $D > 0$ there is a constant $C = C(\mathcal{I}, D, L) > 0$ for which, for n sufficiently large, $P_b(\nu_n(\tilde{t}_n)\mathbb{1}_{\mathcal{A}_n} > C\varepsilon_n) \le 4e^{-Dn\Delta\varepsilon_n^2}$.*

Proof of Lemma 12. Recall $R_{k\Delta} = \frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}(b(X_s) - b(X_{k\Delta}))\,\mathrm{d}s$, and recall any $b \in \Theta$ is Lipschitz, with Lipschitz constant at most K_0, so $|R_{k\Delta}| \le K_0 \max_{s\le\Delta}|X_{k\Delta+s} - X_{k\Delta}|$. It is therefore enough to bound $\sup\{|X_t - X_s| : s, t \in [0, n\Delta],\ |t - s| \le \Delta\}$.

We apply the Hölder continuity result (Lemma 8) with $u = D^{1/2}\lambda^{-1/2}(n\Delta\varepsilon_n^2)^{1/2}$ for $\lambda = \lambda(\mathcal{I})$ the constant of Lemma 8, noting that the assumption $n\Delta\varepsilon_n^2/\log(n\Delta) \to \infty$ ensures that u is large enough compared to $m = n\Delta$ that the conditions for Lemma 8 are met, at least when n is

large. We see that

$$\sup_{\substack{s,t\in[0,n\Delta]\\|t-s|\le\Delta}}|X_t-X_s|\le\Delta^{1/2}\bigl(\log(n\Delta)^{1/2}+\log\bigl(\Delta^{-1}\bigr)^{1/2}\bigr)D^{1/2}\lambda^{-1/2}\bigl(n\Delta\varepsilon_n^2\bigr)^{1/2},$$

on an event $\mathcal{D}$ of probability at least $1-2e^{-Dn\Delta\varepsilon_n^2}$ (we have used that, for n large enough, $\Delta\le\min(\tau,e^{-1})$ in order to take the supremum over $|t-s|\le\Delta$ and to see $\sup_{\delta\le\Delta}w_m(\delta)=w_m(\Delta)$).

Now observe that $\log(n\Delta)^{1/2}\le(\log(\Delta^{-1})^{1/2})$ for large enough n because $n\Delta^2\to0$ (so $n\Delta\le\Delta^{-1}$ eventually). Further, from the assumption $n\Delta^2\log(\Delta^{-1})\le L_0$ we are able to deduce that $\Delta^{1/2}\log(\Delta^{-1})^{1/2}(n\Delta\varepsilon_n^2)^{1/2}\le L_0^{1/2}\varepsilon_n$. It follows that on $\mathcal{D}$, we have $R_{k\Delta}\le C\varepsilon_n$ for a suitably chosen constant C (independent of k and n), which implies the desired concentration. □

Proof of Lemma 13. Recall for $t:[0,1]\to\mathbb{R}$ we defined $\nu_n(t)=\frac{1}{n}\sum_{k=1}^n t(X_{k\Delta})Z_{k\Delta}$, where $Z_{k\Delta}=\frac{1}{\Delta}\int_{k\Delta}^{(k+1)\Delta}\sigma(X_s)\,\mathrm{d}W_s$. The martingale-derived concentration result Lemma 2 in Comte et al. [9] (the model assumptions in [9] are slightly different to those made here, but the proof of the lemma equally applies in our setting) tells us $P_b(\nu_n(t)\ge\xi,\|t\|_n^2\le u^2)\le\exp(-\frac{n\Delta\xi^2}{2\sigma_U^2u^2})$, for any t, any u, and any drift function $b\in\Theta$, so that

$$P_b\bigl(\nu_n(t)\ge\xi\bigr)\le\exp\biggl(-\frac{n\Delta\xi^2}{2\sigma_U^2u^2}\biggr)+P_b\bigl(\|t\|_n^2>u^2\bigr).\tag{$\star$}$$

We can apply Theorem 6 to see that, for some constant $\kappa=\kappa(\mathcal{I})$,

$$\begin{aligned}P_b\bigl(\|t\|_n^2>u^2\bigr)&=P_b\biggl(\frac{1}{n}\biggl(\sum_{k=1}^n t(X_{k\Delta})^2-\|t\|_\mu^2\biggr)>u^2-\|t\|_\mu^2\biggr)\\&\le\exp\biggl(-\frac{1}{\kappa}\Delta\min\biggl\{\frac{n^2(u^2-\|t\|_\mu^2)^2}{n\|t^2\|_\mu^2},\frac{n(u^2-\|t\|_\mu^2)}{\|t^2\|_\infty}\biggr\}\biggr)\\&\le\exp\biggl(-\frac{1}{\kappa}n\Delta\bigl(u^2-\|t\|_\mu^2\bigr)\|t\|_\infty^{-2}\min\bigl(u^2\|t\|_\mu^{-2}-1,1\bigr)\biggr),\end{aligned}$$

where to obtain the last line we have used that $\|t^2\|_\mu^2\le\|t\|_\infty^2\|t\|_\mu^2$.

Now choose $u^2=\|t\|_\mu^2+\xi\|t\|_\infty$. Then $\xi^2/u^2\ge\frac{1}{2}\min(\xi^2/\|t\|_\mu^2,\xi/\|t\|_\infty)$ so that, returning to ($\star$), we find

$$\begin{aligned}P_b\bigl(\nu_n(t)\ge\xi\bigr)&\le\exp\biggl(-\frac{n\Delta}{4\sigma_U^2}\min\bigl(\xi^2\|t\|_\mu^{-2},\xi\|t\|_\infty^{-1}\bigr)\biggr)+\exp\biggl(-\frac{1}{\kappa}n\Delta\xi\min\bigl(\xi\|t\|_\mu^{-2},\|t\|_\infty^{-1}\bigr)\biggr)\\&\le2\exp\biggl(-\frac{1}{\kappa'}n\Delta\min\bigl(\xi^2\|t\|_\mu^{-2},\xi\|t\|_\infty^{-1}\bigr)\biggr),\end{aligned}$$

for some constant $\kappa'=\kappa'(\mathcal{I})$.

By changing variables we attain the bound $P_b(v_n(t) \geq \max(\sqrt{v^2 x}, ux)) \leq 2\exp(-x)$, where $v^2 = \kappa'(n\Delta)^{-1}\|t\|_\mu^2$ and $u = \kappa'(n\Delta)^{-1}\|t\|_\infty$. Then, as in Theorem 6, a standard chaining argument allows us to deduce that

$$P_b\Big(\sup_{t\in I_n} v_n(t) \geq \tilde{\kappa}\Big(\sqrt{V^2(D_{l_n}+x)} + U(D_{l_n}+x)\Big)\Big) \leq 4e^{-x},$$

for $V^2 = \sup_{t\in I_n}\|t\|_\mu^2(n\Delta)^{-1} = (n\Delta)^{-1}$, $U = \sup_{t\in I_n}\|t\|_\infty(n\Delta)^{-1} = C_1\varepsilon_n^{-1}(n\Delta)^{-1}$, and for a constant $\tilde{\kappa} = \tilde{\kappa}(\mathcal{I})$. Taking $x = Dn\Delta\varepsilon_n^2$ and recalling the assumption $D_{l_n} \leq Ln\Delta\varepsilon_n^2$ we obtain the desired result (conditional on $\tilde{t}_n \in I_n$, which is the case on the event $\mathcal{A}_n$). □

5. Small ball probabilities

Now we show that the Kullback–Leibler divergence between the laws corresponding to different parameters b_0 and b can be controlled in terms of the L^2-distance between the parameters. Denote by $K(p,q)$ the Kullback–Leibler divergence between probability distributions with densities p and q, that is, $K(p,q) = E_p\log(\frac{p}{q}) = \int\log(\frac{p(x)}{q(x)})\,\mathrm{d}p(x)$. Also write

$$\mathrm{KL}(b_0,b) = E_{b_0}\left[\log\left(\frac{p_0(\Delta, X_0, X_\Delta)}{p_b(\Delta, X_0, X_\Delta)}\right)\right].$$

Recalling that $p_b^{(n)}(x^{(n)}) = \pi_b(x_0)\prod_{i=1}^n p_b(\Delta, x_{(i-1)\Delta}, x_{i\Delta})$ is the density on $\mathbb{R}^{n+1}$ of $X^{(n)}$ under P_b, we introduce the following Kullback–Leibler type neighbourhoods: for $\varepsilon > 0$, define

$$B_{KL}^{(n)}(\varepsilon) = \left\{b\in\Theta : K\big(p_0^{(n)}, p_b^{(n)}\big) \leq (n\Delta+1)\varepsilon^2, Var_{b_0}\left(\log\frac{p_0^{(n)}}{p_b^{(n)}}\right) \leq (n\Delta+1)\varepsilon^2\right\},$$

and, noting that $\mathrm{KL}(b_0,b)$ (hence also the following set) depends implicitly on n via Δ, define

$$B_\varepsilon = \left\{b\in\Theta : K(\pi_0,\pi_b) \leq \varepsilon^2, \mathrm{Var}_{b_0}\left(\log\frac{\pi_0}{\pi_b}\right) \leq \varepsilon^2, \mathrm{KL}(b_0,b) \leq \Delta\varepsilon^2, \mathrm{Var}_{b_0}\left(\log\frac{p_0}{p_b}\right) \leq \Delta\varepsilon^2\right\}.$$

The main result of this section is the following.

Theorem 14. *Consider data $X^{(n)} = (X_{k\Delta})_{0\leq k\leq n}$ sampled from a solution X to (1) under Assumptions 1–4. Let $\varepsilon_n \to 0$ be a sequence of positive numbers such that $n\Delta\varepsilon_n^2 \to \infty$. Then there is a constant $A = A(\mathcal{I})$ such that, for all n sufficiently large, $\{b\in\Theta : \|b-b_0\|_2 \leq A\varepsilon_n\} \subseteq B_{KL}^{(n)}(\varepsilon_n)$.*

Proof. Applying Lemma 22 from Appendix A where it is shown that

$$\mathrm{Var}_{b_0}\log\left(\frac{p_0^{(n)}(X^{(n)})}{p_b^{(n)}(X^{(n)})}\right) \leq 3\,\mathrm{Var}_{b_0}\left(\log\frac{\pi_0(X_0)}{\pi_b(X_0)}\right) + 3n\,\mathrm{Var}_{b_0}\left(\log\frac{p_0(X_0,X_\Delta)}{p_b(X_0,X_\Delta)}\right).$$

and noting also that $K(p_0^{(n)}, p_b^{(n)}) = K(\pi_0, \pi_b) + n\,\mathrm{KL}(b_0, b)$ by linearity, we observe that $B_{\varepsilon_n/\sqrt{3}} \subseteq B_{KL}^{(n)}(\varepsilon_n)$. It is therefore enough to show that there is an $A = A(\mathcal{I})$ such that $\{b \in \Theta : \|b - b_0\|_2 \le A\varepsilon_n\} \subseteq B_{\varepsilon_n/\sqrt{3}}$. This follows immediately by applying Lemma 15 below to $\xi_n = \varepsilon_n/\sqrt{3}$. □

Lemma 15. *Under the conditions of Theorem 14, there is an $A = A(\mathcal{I})$ such that, for all n sufficiently large, $\{b \in \Theta : \|b - b_0\|_2 \le A\varepsilon_n\} \subseteq B_{\varepsilon_n}$.*

The key idea in proving Lemma 15 is to use the Kullback–Leibler divergence between the laws $P_b^{(x)}$ and $P_{b_0}^{(x)}$ of the continuous-time paths to control the Kullback–Leibler divergence between p_b and p_0. This will help us because we can calculate the Kullback–Leibler divergence between the full paths using Girsanov's Theorem, which gives us an explicit formula for the likelihood ratios.

Let $P_{b,T}^{(x)}$ denote the law of $(X_t)_{0\le t\le T}$ conditional on $X_0 = x$, i.e. the restriction of $P_b^{(x)}$ to $C([0,T])$. We write $\mathbb{W}_{\sigma,T}^{(x)}$ for $P_{b,T}^{(x)}$ when $b = 0$. Throughout this section, we will simply write $P_b^{(x)}$ for $P_{b,\Delta}^{(x)}$ and similarly with $\mathbb{W}_\sigma^{(x)}$. We have the following.

Theorem 16 (Girsanov's Theorem). *Assume b_0 and b lie in Θ, and σ satisfies Assumption 1. Then the laws $P_{b_0,T}^{(x)}$ and $P_{b,T}^{(x)}$ are mutually absolutely continuous with, for $X \sim P_{b,T}^{(x)}$, the almost sure identification*

$$\frac{\mathrm{d}P_{b_0,T}^{(x)}}{\mathrm{d}P_{b,T}^{(x)}}\big((X_t)_{t\le T}\big) = \exp\left[\int_0^T \frac{b_0 - b}{\sigma^2}(X_t)\,\mathrm{d}X_t - \frac{1}{2}\int_0^T \frac{b_0^2 - b^2}{\sigma^2}(X_t)\,\mathrm{d}t\right].$$

Proof. See Liptser and Shiryaev [22], Theorem 7.19, noting that the assumptions are met because b, b_0 and σ are all Lipschitz and bounded, and σ is bounded away from 0. □

We write

$$\tilde{p}_0^{(x)} = \frac{\mathrm{d}P_{b_0}^{(x)}}{\mathrm{d}\mathbb{W}_\sigma^{(x)}}, \qquad \tilde{p}_b^{(x)} = \frac{\mathrm{d}P_b^{(x)}}{\mathrm{d}\mathbb{W}_\sigma^{(x)}} \tag{14}$$

for the Radon–Nikodym derivatives (i.e., densities on $C([0,\Delta])$ with respect to $\mathbb{W}_\sigma^{(x)}$) whose existence Girsanov's Theorem guarantees. We will simply write X for $(X_t)_{t\le\Delta}$ where context allows, and similarly with U. Since $\tilde{p}_0^{(x)}(X) = 0$ for any path X with $X_0 \ne x$, we will further omit the superscripts on our densities in general, writing $\tilde{p}_0(X)$ for $\tilde{p}_0^{(X_0)}(X)$, and similarly for $\tilde{p}_b$.

Proof of Lemma 15. We break the proof into a series of lemmas. We will upper bound the variances in the definition of B_{ε_n} by the corresponding uncentered second moments. For some constant $A = A(\mathcal{I})$, we show the following.

1. $A^2\,\mathrm{KL}(b_0, b) \le \Delta\|b - b_0\|_2^2$, which shows that $\mathrm{KL}(b_0, b) \le \Delta\varepsilon_n^2$ whenever $\|b - b_0\|_2 \le A\varepsilon_n$. This is the content of Lemma 17.

2. If $\|b - b_0\|_2 \leq A\varepsilon_n$ then we have $E_{b_0}[\log(p_0/p_b)^2] \leq \Delta\varepsilon_n^2$. This is the content of Lemma 18. Note that the other steps do not need any assumptions on ε_n, but this step uses $n\Delta\varepsilon_n^2 \to \infty$.
3. $A^2 \max\{K(\pi_0, \pi_b), E_{b_0}[\log(\pi_0/\pi_b)^2]\} \leq \|b_0 - b\|_2^2$, which shows that $K(\pi_0, \pi_b) \leq \varepsilon_n^2$ and $E_{b_0}[\log(\pi_0/\pi_b)^2] \leq \varepsilon_n^2$ whenever $\|b - b_0\|_2 \leq A\varepsilon_n$. This is the content of Lemma 19.

Together, then, the three lemmas below conclude the proof. □

Lemma 17. *Under the conditions of Theorem 14, there is a constant $A = A(\mathcal{I})$ such that $A^2 \operatorname{KL}(b_0, b) \leq \Delta\|b_0 - b\|_2^2$.*

The proof is essentially the same as that of van der Meulen and van Zanten [34], Lemma 5.1, with minor adjustments to fit the periodic model and non-constant σ used here. Further, all the ideas needed are exhibited in the proof of Lemma 18. Thus, we omit the proof.

Lemma 18. *Under the conditions of Theorem 14, there is a constant $A = A(\mathcal{I})$ so that, for n sufficiently large, $E_{b_0}[\log(p_0/p)^2] \leq \Delta\varepsilon_n^2$ whenever $\|b - b_0\|_2 \leq A\varepsilon_n$.*

Proof. We first show that we can control the second moment of $\log(p_0/p_b)$ by the second moment of the corresponding expression $\log(\tilde{p}_0/\tilde{p}_b)$ for the full paths, up to an approximation error which is small when Δ is small. Consider the smallest convex function dominating $\log(x)^2$, given by

$$h(x) = \begin{cases} \log(x)^2, & x < e, \\ 2e^{-1}x - 1, & x \geq e \end{cases}$$

(it is in fact more convenient, and equivalent, to think of h as dominating the function $x \mapsto (\log x^{-1})^2$). Let $X \sim P_{b_0}^{(x)}$ and let $U \sim \mathbb{W}_\sigma^{(x)}$. Intuitively, the probability density of a transition of X from x to y, with respect to the (Lebesgue) density p_* of transitions of U from x to y, can be calculated by integrating the likelihood $\tilde{p}_0(U)$ over all paths of U which start at x and end at y, and performing this integration will yield the conditional expectation of $\tilde{p}_0^{(x)}(U)$ given U_Δ. That is to say,

$$\frac{p_0(\Delta, x, y)}{p_*(\Delta, x, y)} = E_{\mathbb{W}_\sigma^{(x)}}\big[\tilde{p}_0(U) \mid U_\Delta = y\big]. \tag{15}$$

The above argument is not rigorous because we condition on an event of probability zero, but the formula (15) is true, and is carefully justified in Lemma 23 in Appendix A. A corresponding expression holds for $p_b(\Delta, x, y)$, so that

$$E_{b_0}\left[\log\left(\frac{p_0(\Delta, X_0, X_\Delta)}{p_b(\Delta, X_0, X_\Delta)}\right)^2\right] \leq E_{b_0}\big[h(p_b/p_0)\big] = E_{b_0}\left[h\left(\frac{E_{\mathbb{W}_\sigma^{(X_0)}}[\tilde{p}_b(U) \mid U_\Delta = X_\Delta]}{E_{\mathbb{W}_\sigma^{(X_0)}}[\tilde{p}_0(U) \mid U_\Delta = X_\Delta]}\right)\right].$$

Lemma 21 in Appendix A allows us to simplify the ratio of conditional expectations. We apply with $\mathbb{P} = \mathbb{W}_\sigma^{(X_0)}$, $\mathbb{Q} = P_{b_0}^{(X_0)}$ and $g = \tilde{p}_b^{(X_0)}/\tilde{p}_0^{(X_0)}$, then further apply conditional Jensen's

inequality and the tower law to find

$$\begin{aligned} E_{b_0}\left[\left(\log\frac{p_0}{p_b}\right)^2\right] &\le E_{b_0}\left[h\left(E_{P_{b_0}^{(X_0)}}\left[\frac{\tilde{p}_b}{\tilde{p}_0}(X)\,\middle|\, X_\Delta\right]\right)\right] \\ &\le E_{b_0}\left[h\left(\frac{\tilde{p}_b}{\tilde{p}_0}(X)\right)\right] \\ &\le E_{b_0}\left[\left(\log\frac{\tilde{p}_0}{\tilde{p}_b}(X)\right)^2\right] + E_{b_0}\left[\left(2e^{-1}\frac{\tilde{p}_b}{\tilde{p}_0}(X)-1\right)\mathbb{1}\left\{\frac{\tilde{p}_b}{\tilde{p}_0}(X)\ge e\right\}\right], \end{aligned}$$

which is the promised decomposition into a corresponding quantity for the continuous case and an approximation error. We conclude by showing that each of these two terms is bounded by $\frac{1}{2}\Delta\varepsilon_n^2$, provided $\|b-b_0\|_2 \le A\varepsilon_n$ for some sufficiently small constant $A = A(\mathcal{I})$.

Showing $E_{b_0}[(\log\frac{\tilde{p}_0}{\tilde{p}_b})^2] \le \frac{1}{2}\Delta\varepsilon_n^2$. Writing $f = \frac{b_0-b}{\sigma}$, we apply Girsanov's Theorem (Theorem 16) to find

$$\begin{aligned} E_{b_0}\left[\left(\log\frac{\tilde{p}_0}{\tilde{p}_b}(X)\right)^2\right] &= E_{b_0}\left[\left(\int_0^\Delta f(X_t)\,\mathrm{d}W_t + \frac{1}{2}\int_0^\Delta f^2(X_t)\,\mathrm{d}t\right)^2\right] \\ &= E_{b_0}\left[\left(\int_0^\Delta f(X_t)\,\mathrm{d}W_t\right)^2\right] + \frac{1}{4}E_{b_0}\left[\left(\int_0^\Delta f^2(X_t)\,\mathrm{d}t\right)^2\right]. \end{aligned}$$

The cross term has vanished in the final expression because:

- $\int_0^\Delta f(X_t)\,\mathrm{d}W_t$ is a martingale for $X \sim P_{b_0}$ (this follows from the fact that f is bounded thanks to Assumptions 1 and 2, and a bounded semimartingale integrated against a square integrable martingale yields a martingale, as in [30], IV.27.4).
- $\int_0^\Delta f^2(X_t)\,\mathrm{d}t$ is a finite variation process.
- The expectation of a martingale against a finite variation process is zero (see, e.g., [30], IV.32.12).

For the first term on the right, we use Itô's isometry ([30], IV.27.5), Fubini's Theorem, periodicity of f and stationarity of μ_0 for the periodised process $\dot{X} = X \bmod 1$ to find

$$E_{b_0}\left(\int_0^\Delta f(X_t)\,\mathrm{d}W_t\right)^2 = E_{b_0}\int_0^\Delta f^2(X_t)\,\mathrm{d}t = \int_0^\Delta E_{b_0} f^2(\dot{X}_t)\,\mathrm{d}t = \Delta\|f\|_{\mu_0}^2.$$

The second term $\frac{1}{4}E_{b_0}[(\int_0^\Delta f^2(X_t)\,\mathrm{d}t)^2]$ is upper bounded by $\frac{1}{4}\Delta^2\|f\|_\infty^2\|f\|_{\mu_0}^2$ (this can be seen from the bound $(\int_0^\Delta f^2)^2 \le \Delta\|f\|_\infty^2\int_0^\Delta f^2$), hence is dominated by $\Delta\|f\|_{\mu_0}^2$ when n is large. Thus, for some constant $A = A(\mathcal{I})$ we find

$$E_{b_0}\left[\left(\log\frac{\tilde{p}_0}{\tilde{p}_b}(X)\right)^2\right] \le 2\Delta\|f\|_{\mu_0}^2 \le \frac{1}{2}A^{-2}\Delta\|b_0-b\|_2^2,$$

where Assumptions 1 and 2 allow us to upper bound $\|f\|_{\mu_0}$ by $\|b_0 - b\|_2$, up to a constant depending only on $\mathcal{I}$. For $\|b_0 - b\|_2 \le A\varepsilon_n$ we then have $E_{b_0}[(\log(\tilde{p}_b/\tilde{p}_0))^2] \le \Delta\varepsilon_n^2/2$.

Showing $E_{b_0}[(2e^{-1}\frac{\tilde{p}_b}{\tilde{p}_0}(X) - 1)\mathbb{1}\{\frac{\tilde{p}_b}{\tilde{p}_0}(X) \ge e\}] \le \frac{1}{2}\Delta\varepsilon_n^2$. We have

$$E_{b_0}\left[\left(2e^{-1}\frac{\tilde{p}_b}{\tilde{p}_0}(X) - 1\right)\mathbb{1}\left\{\frac{\tilde{p}_b}{\tilde{p}_0}(X) \ge e\right\}\right] \le 2e^{-1}P_b\left[\left(\frac{\tilde{p}_b}{\tilde{p}_0}(X)\right) \ge 1\right].$$

By the tower law (noting $2e^{-1} \le 1$) it suffices to show $P_b^{(x)}[\log(\frac{\tilde{p}_b}{\tilde{p}_0}(X)) \ge 1] \le \frac{1}{2}\Delta\varepsilon_n^2$ for each $x \in [0,1]$. Applying Girsanov's Theorem (Theorem 16) we have, for $f = (b_0 - b)/\sigma$, and for n large enough that $\Delta\|f\|_\infty^2 \le 1$,

$$\begin{aligned} P_b^{(x)}\left(\log\frac{\tilde{p}_b}{\tilde{p}_0}(X) > 1\right) &= P_b^{(x)}\left(\int_0^\Delta -f(X_t)\,\mathrm{d}W_t + \frac{1}{2}\int_0^\Delta f(X_t)^2\,\mathrm{d}t > 1\right) \\ &\le P_b^{(x)}\left(\int_0^\Delta -f(X_t)\,\mathrm{d}W_t > 1/2\right). \end{aligned}$$

Write $M_t = \int_0^t -f(X_s)\,\mathrm{d}W_s$. Since $A = \max(1, (2K_0/\sigma_L)^2)$ uniformly upper bounds $\|f\|_\infty^2$ for $b \in \Theta$, we see that M is a martingale whose quadratic variation satisfies $|\langle M\rangle_t - \langle M\rangle_s| \le A|t-s|$. Recalling that $w_1(\delta) = \delta^{1/2}\log(\delta^{-1})^{1/2}$, we apply Lemma 9 with $u = w_1(\Delta)^{-1}/2$ to yield that, for n large enough,

$$\begin{aligned} P_b^{(x)}\left(\log\frac{\tilde{p}_b}{\tilde{p}_0}(X) > 1\right) &\le P_b^{(x)}\left(\sup_{s,t\le\Delta, s\ne t}\frac{|M_t - M_s|}{w_1(|t-s|)} > \frac{1}{2}w_1(\Delta)^{-1}\right) \\ &\le 2\exp\left(-\lambda w_1(\Delta)^{-2}\right), \end{aligned}$$

where λ is a constant depending only on $\mathcal{I}$.

Recall we assume $n\Delta \to \infty$ and $n\Delta^2 \to 0$. It follows that for large enough n we have $\log(\Delta^{-1}) \le \log(n)$, and $\Delta \le \lambda\log(n)^{-2}$. Then observe

$$\begin{aligned} \Delta \le \lambda\log(n)^{-2} \quad &\Longrightarrow \quad \Delta \le \lambda\left(\log\Delta^{-1}\right)^{-1}\log(n)^{-1} \\ &\Longrightarrow \quad \log(n) \le \lambda\Delta^{-1}\left(\log\Delta^{-1}\right)^{-1}, \end{aligned}$$

so that $\exp(-\lambda w_1(\Delta)^{-2}) \le n^{-1}$ for n large. Finally, since $n\Delta\varepsilon_n^2 \to \infty$, we see $2n^{-1} \le \frac{1}{2}\Delta\varepsilon_n^2$ for n large enough, as required. □

Lemma 19. *Under the conditions of Theorem 14, there is a constant $A = A(\mathcal{I})$ such that $A^2\max\{K(\pi_0,\pi_b), E_{b_0}[\log(\pi_0/\pi_b)^2]\} \le \|b_0 - b\|_2^2$.*

Proof. By the comment after Lemma 8.3 in [11], it suffices to prove $h^2(\pi_0,\pi_b)\|\pi_0/\pi_b\|_\infty \le C\|b - b_0\|_2^2$ for some $C = C(\mathcal{I})$, where h is the Hellinger distance between densities, defined by

$h^2(p,q) = \int(\sqrt{p} - \sqrt{q})^2$. Since π_0 and π_b are uniformly bounded above and away from zero, we can absorb the term $\|\pi_0/\pi_b\|_\infty$ into the constant.

We initially prove pointwise bounds on the difference between the densities π_0 and π_b. Recall we saw in Section 2 that, for $I_b(x) = \int_0^x \frac{2b}{\sigma^2}(y)\,\mathrm{d}y$, we have

$$\pi_b(x) = \frac{e^{I_b(x)}}{H_b\sigma^2(x)}\left(e^{I_b(1)}\int_x^1 e^{-I_b(y)}\,\mathrm{d}y + \int_0^x e^{-I_b(y)}\,\mathrm{d}y\right), \qquad x \in [0,1],$$

$$H_b = \int_0^1 \frac{e^{I_b(x)}}{\sigma^2(x)}\left(e^{I_b(1)}\int_x^1 e^{-I_b(y)}\,\mathrm{d}y + \int_0^x e^{-I_b(y)}\,\mathrm{d}y\right)\mathrm{d}x.$$

We can decompose: $|\pi_b(x) - \pi_0(x)| \le D_1 + D_2 + D_3 + D_4$, where

$$D_1 = \frac{e^{I_b(x)}}{\sigma^2(x)}\left|\frac{1}{H_b} - \frac{1}{H_{b_0}}\right|\left(e^{I_b(1)}\int_x^1 e^{-I_b(y)}\,\mathrm{d}y + \int_0^x e^{-I_b(y)}\,\mathrm{d}y\right),$$

$$D_2 = \frac{|e^{I_b(x)} - e^{I_{b_0}(x)}|}{H_{b_0}\sigma^2(x)}\left(e^{I_b(1)}\int_x^1 e^{-I_b(y)}\,\mathrm{d}y + \int_0^x e^{-I_b(y)}\,\mathrm{d}y\right),$$

$$D_3 = \frac{e^{I_{b_0}(x)}}{H_{b_0}\sigma^2(x)}\left|\left(e^{I_b(1)} - e^{I_{b_0}(1)}\right)\int_x^1 e^{-I_b(y)}\,\mathrm{d}y\right|,$$

$$D_4 = \frac{e^{I_{b_0}(x)}}{H_{b_0}\sigma^2(x)}\left|e^{I_{b_0}(1)}\int_x^1 \left(e^{-I_b(y)} - e^{-I_{b_0}(y)}\right)\mathrm{d}y + \int_0^x \left(e^{-I_b(y)} - e^{-I_{b_0}(y)}\right)\mathrm{d}y\right|.$$

We have the bounds $\sigma_U^{-2}e^{-6K_0\sigma_L^{-2}} \le H_b \le \sigma_L^{-2}e^{6K_0\sigma_L^{-2}}$, and $e^{-2K_0\sigma_L^{-2}} \le e^{I_b(x)} \le e^{2K_0\sigma_L^{-2}}$. An application of the mean value theorem then tells us

$$\left|e^{I_b(x)} - e^{I_{b_0}(x)}\right| \le C(\mathcal{I})\int_0^x \frac{2|b_0 - b|}{\sigma^2}(y)\,\mathrm{d}y \le C'(\mathcal{I})\|b_0 - b\|_2,$$

for some constants C, C', and the same expression upper bounds $|e^{-I_b(x)} - e^{-I_{b_0}(x)}|$.

It follows that, for some constant $C = C(\mathcal{I})$, we have $D_i \le C\|b - b_0\|_2$ for $i = 2, 3, 4$. For $i = 1$ the same bound holds since $|\frac{1}{H_b} - \frac{1}{H_{b_0}}| \le \frac{|H_b - H_{b_0}|}{H_bH_{b_0}}$ and a similar decomposition to the above yields $|H_b - H_{b_0}| \le C(\mathcal{I})\|b - b_0\|_2$.

Thus, we have shown that $|\pi_b(x) - \pi_0(x)| \le C(\mathcal{I})\|b - b_0\|_2$. Integrating this pointwise bound, we find that $\|\pi_0 - \pi_b\|_2 \le C(\mathcal{I})\|b_0 - b\|_2$. Finally, since $h^2(\pi_0, \pi_b) \le \frac{1}{4\pi_L}\|\pi_0 - \pi_b\|_2^2 \le C'(\mathcal{I})\|b_0 - b\|_2^2$, for some different constant C', we are done. □

6. Main contraction results: Proofs

We now have the tools we need to apply general theory in order to derive contraction rates. Recall that $K(p,q)$ denotes the Kullback–Leibler divergence between probability distributions

with densities p and q, and recall the definition

$$B_{KL}^{(n)}(\varepsilon) = \left\{ b \in \Theta : K\big(p_0^{(n)}, p_b^{(n)}\big) \le (n\Delta + 1)\varepsilon^2, \operatorname{Var}_{b_0}\left(\log \frac{p_0^{(n)}}{p_b^{(n)}} \right) \le (n\Delta + 1)\varepsilon^2 \right\}.$$

We have the following abstract contraction result, from which we deduce Theorem 1.

Theorem 20. *Consider data $X^{(n)} = (X_{k\Delta})_{0 \le k \le n}$ sampled from a solution X to (1) under Assumptions 1–4. Let the true parameter be b_0. Let $\varepsilon_n \to 0$ be a sequence of positive numbers and let l_n be a sequence of positive integers such that, for some constant L we have, for all n,*

$$D_{l_n} = 2^{l_n} \le Ln\Delta\varepsilon_n^2, \quad \text{and} \quad n\Delta\varepsilon_n^2 / \log(n\Delta) \to \infty. \tag{16}$$

For each n let Θ_n be $\mathcal{S}$-measurable and assume

$$b_0 \in \Theta_n \subseteq \big\{ b \in \Theta : \|\pi_{l_n} b - b\|_2 \le \varepsilon_n \big\}, \tag{17}$$

where π_{l_n} is the L^2-orthogonal projection map onto S_{l_n} as described in Section 2.1. Let $\Pi^{(n)}$ be a sequence of priors on $(\Theta, \mathcal{S})$ satisfying

(a) $\Pi^{(n)}(\Theta_n^c) \le e^{-(\omega+4)n\Delta\varepsilon_n^2}$,
(b) $\Pi^{(n)}(B_{KL}^{(n)}(\varepsilon_n)) \ge e^{-\omega n\Delta\varepsilon_n^2}$,

for some constant[1] $\omega > 0$. Then there is a constant $M = M(\mathcal{I}, L_0, \omega, L)$ such that, under the law P_{b_0} of X, $\Pi^{(n)}(\{b \in \Theta : \|b - b_0\|_2 \le M\varepsilon_n\} \mid X^{(n)}) \to 1$ in probability.

The proof, given the existence of tests, follows the standard format of Ghosal–Ghosh–van der Vaart [11] and is given in the supplement (Abraham [1]).

Proof of Theorem 1.

A. We apply Theorem 20. The key idea which allows us to control the bias and obtain this adaptive result with a sieve prior is *undersmoothing*. Specifically, when we prove the small ball probabilities, we do so by conditioning on the hyperprior choosing a resolution j_n which corresponds to the minimax rate $(n\Delta)^{-s/(1+2s)}$ rather than corresponding to the slower rate $(n\Delta)^{-s/(1+2s)} \log(n\Delta)^{1/2}$ at which we prove contraction. This logarithmic gap gives us the room we need to ensure we can achieve the bias condition (a) and the small ball condition (b) for the *same* constant ω. The argument goes as follows.

Write $\bar{\varepsilon}_n^2 = (n\Delta)^{-2s/(1+2s)}$ and let $\varepsilon_n^2 = (n\Delta)^{-2s/(1+2s)} \log(n\Delta)$. Choose j_n and l_n natural numbers satisfying (at least for n large enough)

$$\frac{1}{2} n\Delta\bar{\varepsilon}_n^2 \le D_{j_n} = 2^{j_n} \le n\Delta\bar{\varepsilon}_n^2, \qquad \frac{1}{2} Ln\Delta\varepsilon_n^2 \le D_{l_n} = 2^{l_n} \le Ln\Delta\varepsilon_n^2,$$

[1] In fact we can replace the exponent $\omega + 4$ in (a) with any $B > \omega + 1$. We choose $\omega + 4$ because it simplifies the exposition and the exact value is unimportant.

where L is a constant to be chosen. Note that (16) holds by definition. Recall now from our choice of approximation spaces in Section 2.1 that we have $\|\pi_m b_0 - b_0\|_2 \leq K(s)\|b_0\|_{B^s_{2,\infty}} 2^{-ms}$. For any fixed L we therefore find that for n large enough, writing $K = K(s)2^s\|b_0\|_{B^s_{2,\infty}}$, we have

$$\|\pi_{l_n} b_0 - b_0\|_2 \leq K\big(Ln\Delta\varepsilon_n^2\big)^{-s} = K\big(Ln\Delta\bar{\varepsilon}_n^2\log(n\Delta)\big)^{-s} = KL^{-s}\bar{\varepsilon}_n\log(n\Delta)^{-s} \leq \varepsilon_n.$$

Similarly, it can be shown that, with $A = A(\mathcal{I})$ the constant of the small ball result (Theorem 14) and for n large enough, we have $\|b_0 - \pi_{j_n} b_0\|_2 \leq A\varepsilon_n/2$.

Set $\Theta_n = \{b_0\} \cup (S_{l_n} \cap \Theta)$ and observe that the above calculations show that the bias condition (17) holds (since also for $b \in \Theta_n$, if $b \neq b_0$ we have $\|\pi_{l_n} b - b\|_2 = 0$).

Next, for the small ball condition (b), recall Theorem 14 tells us that, for n large enough, $\{b \in \Theta : \|b - b_0\|_2 \leq A\varepsilon_n\} \subseteq B^{(n)}_{KL}(\varepsilon_n)$. Thus it suffices to show, for some $\omega > 0$ for which we can also achieve (a), that $\Pi(\{b \in \Theta : \|b - b_0\|_2 \leq A\varepsilon_n\}) \geq e^{-\omega n\Delta\varepsilon_n^2}$. Using that $\|b - b_0\|_2 \leq \|b - \pi_{j_n} b_0\|_2 + \|\pi_{j_n} b_0 - b_0\|_2 \leq \|b - \pi_{j_n} b_0\|_2 + A\varepsilon_n/2$, and using our assumptions on h and Π_m, we see that

$$\begin{aligned}
&\Pi\big(\big\{b \in \Theta : \|b - b_0\|_2 \leq A\varepsilon_n\big\}\big) \\
&\quad = \sum_m h(m)\Pi_m\big(\big\{b \in S_m : \|b - b_0\|_2 \leq A\varepsilon_n\big\}\big) \\
&\quad \geq h(j_n)\Pi_{j_n}\big(\big\{b \in S_{j_n} : \|b - \pi_{j_n} b_0\|_2 \leq A\varepsilon_n/2\big\}\big) \\
&\quad \geq h(j_n)(\varepsilon_n A\zeta/2)^{D_{j_n}} \\
&\quad \geq B_1\exp(-\beta_1 D_{j_n} + D_{j_n}\big[\log(\varepsilon_n) + \log(A\zeta/2)\big]\big) \\
&\quad \geq B_1\exp\big(-Cn\Delta\bar{\varepsilon}_n^2 - Cn\Delta\bar{\varepsilon}_n^2\log\big(\varepsilon_n^{-1}\big)\big)
\end{aligned}$$

for some constant $C = C(\mathcal{I}, \beta_1, \zeta)$. Since $\log(\varepsilon_n^{-1}) = \frac{s}{1+2s}\log(n\Delta) - \frac{1}{2}\log\log(n\Delta) \leq \log(n\Delta)$, we deduce that $\Pi(\{b \in \Theta : \|b - b_0\|_2 \leq A\varepsilon_n\}) \geq B_1 e^{-C'n\Delta\bar{\varepsilon}_n^2\log(n\Delta)} = B_1 e^{-C'n\Delta\varepsilon_n^2}$, with a different constant C'. Changing constant again to some $\omega = \omega(\mathcal{I}, \beta_1, B_1, \zeta)$, we absorb the B_1 factor into the exponential for large enough n.

For (a), since $\Pi(\Theta^c) = 0$ by assumption, we have $\Pi(\Theta_n^c) \leq \Pi(S_{l_n}^c) = \sum_{m=l_n+1}^{\infty} h(m)$. We have assumed that $h(m) \leq B_2 e^{-\beta_2 D_m}$, which ensures that the sum is at most a constant times $e^{-\beta_2 D_{l_n}} \leq e^{-\frac{1}{2}L\beta_2 n\Delta\varepsilon_n^2}$. For the $\omega = \omega(\mathcal{I}, \beta_1, B_1, \zeta)$ for which we proved (b) above, we can therefore choose L large enough to guarantee $\Pi(\Theta_n^c) \leq e^{-(\omega+4)n\Delta\varepsilon_n^2}$.

B. Let ε_n and j_n be as in the statement of Theorem 1 and define l_n as above (here we can take $L = 1$). Similarly to before, we apply results from Section 2.1 to see

$$\left.\begin{aligned}\|\pi_{l_n} b - b\|_2 \leq \varepsilon_n \\ \|\pi_{j_n} b - b\|_2 \leq \varepsilon_n\end{aligned}\right\} \quad \text{for all } n \text{ sufficiently large and all } b \in \Theta_s(A_0).$$

Set $\Theta_n = \Theta_s(A_0)$ for all n. Our assumptions then guarantee the bias condition (a) will hold for any ω (indeed, $\Pi^{(n)}(\Theta_n^c) = 0$). Thus, it suffices to prove for some constant ω that $\Pi^{(n)}(\{b \in \Theta_s(A_0) : \|b - b_0\|_2 \le 3\varepsilon_n\}) \ge e^{-\omega n \Delta \varepsilon_n^2}$, since we can absorb the factor of 3 into the constant M by applying Theorem 20 to $\xi_n = 3\varepsilon_n$.

Because the prior concentrates on $\Theta_s(A_0)$ and b_0 lies in $\Theta_s(A_0)$, we see both that $\Pi^{(n)}(\{b : \|\pi_{j_n} b - b\|_2 \le \varepsilon_n\}) = 1$ and that $\|\pi_{j_n} b_0 - b_0\|_2 \le \varepsilon_n$. Thus

$$\Pi^{(n)}\big(\big\{b \in \Theta_s(A_0) : \|b - b_0\|_2 \le 3\varepsilon_n\big\}\big) \ge \Pi^{(n)}\big(\big\{b \in \Theta_s(A_0) : \|\pi_{j_n} b - \pi_{j_n} b_0\|_2 \le \varepsilon_n\big\}\big).$$

From here the argument is very similar to the previous part (indeed, it is slightly simpler) so we omit the remaining details. □

Proof of Proposition 2. We verify that the conditions of Theorem 1A are satisfied. Condition (i) holds by construction. The $B^s_{\infty,1}$-norm can be expressed as

$$\|f\|_{B^s_{\infty,1}} = |f_{-1,0}| + \sum_{l=0}^{\infty} 2^{l(s+1/2)} \max_{0 \le k < 2^l} |f_{lk}|, \tag{18}$$

(see [16], Section 4.3) hence any b drawn from our prior lies in $B^1_{\infty,1}$ with $\|b\|_{B^1_{\infty,1}} \le (B+1) \times (2 + \sum_{l \ge 1} l^{-2})$. It follows from standard Besov spaces results (e.g., [16], Proposition 4.3.20, adapted to apply to periodic Besov spaces) that $b \in C^1_{\text{per}}([0,1])$, with a C^1_{per}-norm bounded in terms of B. Thus $\Pi(\Theta) = 1$ for an appropriate choice of K_0. We similarly see that $b_0 \in \Theta$. It remains to show that condition (ii) holds. We have

$$\begin{aligned}
\|b - \pi_m b_0\|_2^2 &= \sum_{\substack{-1 \le l < m \\ 0 \le k < 2^l}} \tau_l^2 (u_{lk} - \beta_{lk})^2 \\
&\le \left(1 + \sum_{l=0}^{m-1} 2^{-2l}\right) \max_{\substack{-1 \le l < m, \\ 0 \le k < 2^l}} |u_{lk} - \beta_{lk}|^2 \\
&< 4 \max_{\substack{-1 \le l < m, \\ 0 \le k < 2^l}} |u_{lk} - \beta_{lk}|^2,
\end{aligned}$$

so that $\Pi(\{b \in S_m : \|b - \pi_m b_0\|_2 \le \varepsilon\}) \ge \Pi(|u_{lk} - \beta_{lk}| \le \varepsilon/2 \ \ \forall l, k, -1 \le l < m, k < 2^l)$. Since we have assumed $|\beta_{lk}| \le B\tau_l$ and $q(x) \ge \zeta$ for $|x| \le B$, it follows from independence of the u_{lk} that the right-hand side of this last expression is lower bounded by $(\varepsilon\zeta/2)^{D_m}$, so that condition (ii) holds with $\zeta/2$ in place of ζ. □

Appendix A: Technical lemmas

The following results are proved in the supplement (Abraham [1]).

Lemma 21. *Let $\mathbb{Q}$ and $\mathbb{P}$ be mutually absolutely continuous probability measures and write $f = \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}$. Then, for any measurable g and any sub-σ-algebra $\mathcal{G}$, $E_{\mathbb{Q}}[g \mid \mathcal{G}] = \frac{E_{\mathbb{P}}[fg|\mathcal{G}]}{E_{\mathbb{P}}[f|\mathcal{G}]}$.*

Lemma 22. *The variance of the log likelihood ratio tensorises in this model, up to a constant. Precisely,* $\mathrm{Var}_{b_0} \log(\frac{p_0^{(n)}(X^{(n)})}{p_b^{(n)}(X^{(n)})}) \le 3\,\mathrm{Var}_{b_0}(\log \frac{\pi_0(X_0)}{\pi_b(X_0)}) + 3n\,\mathrm{Var}_{b_0}(\log \frac{p_0(X_0,X_\Delta)}{p_b(X_0,X_\Delta)})$.

Lemma 23. *Let $\tilde{p}_0$ be as in (14). Let $p^*(\Delta, x, y)$ be the density with respect to Lebesgue measure of transitions from x to y in time Δ for a process $U \sim \mathbb{W}_\sigma^{(x)}$. Then*

$$\frac{p_0(\Delta, x, y)}{p_*(\Delta, x, y)} = E_{\mathbb{W}_\sigma^{(x)}}\big[\tilde{p}_0(U) \mid U_\Delta = y\big].$$

Appendix B: Notation

We collect most of the notation used in the course of this paper.

X: A solution to $\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \sigma(X_t)\,\mathrm{d}W_t$.
$\dot{X}$: The periodised diffusion $\dot{X} = X \bmod 1$.
b, σ: Drift function, diffusion coefficient.
$\mu = \mu_b$; π_b: Invariant distribution/density of $\dot{X}$.
$P_b^{(x)}$: Law of X on $C([0,\infty])$ (on $C([0,\Delta])$ in Section 5) for initial condition $X_0 = x$.
E_b; P_b; Var_b: Expectation/probability/variance according to the law of X started from μ_b.
E_μ; Var_μ, and similar: Expectation/variance according to the subscripted measure.
$\mathbb{W}_\sigma^{(x)}$: Notation for $P_b^{(x)}$ when $b = 0$.
$p_b(t, x, y)$, $\dot{p}_b(t, x, y)$: Transition densities of X, $\dot{X}$ (with respect to Lebesgue measure).
$\tilde{p}_b$: Density (with respect to $\mathbb{W}_\sigma^{(x)}$) of $P_b^{(x)}$ on $C([0,\Delta])$.
$I_b(x) = \int_0^x (2b/\sigma^2)(y)\,\mathrm{d}y$.
$X^{(n)} = (X_0, \dots, X_{n\Delta})$; $x^{(n)} = (x_0, \dots, x_{n\Delta})$; $p_b^{(n)}(x^{(n)}) = \pi_b(x_0)\prod_{i=1}^n p_b(\Delta, x_{(i-1)\Delta}, x_{i\Delta})$.
b_0: The true parameter generating the data.
μ_0, π_0, p_0 etc.: Shorthand for μ_{b_0}, π_{b_0}, p_{b_0} etc.
$\sigma_L > 0$; $\sigma_U < \infty$: A lower and upper bound for σ.
L_0: A constant such that $n\Delta^2 \log(1/\Delta) \le L_0$ for all n.
$\Theta = \Theta(K_0)$: The maximal paramater space: $\Theta = \{f \in C^1_{\mathrm{per}}([0,1]) : \|f\|_{C^1_{\mathrm{per}}} \le K_0\}$.
$\Theta_s(A_0) = \{f \in \Theta : \|f\|_{B^s_{2,\infty}} \le A_0\}$, for $B^s_{2,\infty}$ a (periodic) Besov space.
$\mathcal{I} = \{K_0, \sigma_L, \sigma_U\}$.
S_m: Wavelet approximation space of resolution m, generated by periodised Meyer-type wavelets: $S_m = \mathrm{span}\{\psi_{lk} : -1 \le l < m, 0 \le k < 2^l\}$, where $\psi_{-1,0}$ is used as notation for the constant function 1.
$D_m = \dim(S_m) = 2^m$; π_m =(L^2-)orthogonal projection onto S_m.
$w_m(\delta) = \delta^{1/2}(\log(\delta^{-1})^{1/2} + \log(m)^{1/2})$ if $m \ge 1$, $w_m := w_1$ if $m < 1$.
$\mathbb{1}_A$: Indicator of the set (or event) A.
$K(p, q)$: Kullback–Leibler divergence between densities p, q: $K(p, q) = E_p[\log(p/q)]$.

$\mathrm{KL}(b_0, b) = E_{b_0} \log(p_0/p_b)$.
$B_{KL}^{(n)}(\varepsilon) = \{b \in \Theta : K(p_0^{(n)}, p_b^{(n)}) \leq (n\Delta + 1)\varepsilon^2, \mathrm{Var}_{b_0}(\log(p_0^{(n)}/p_b^{(n)})) \leq (n\Delta + 1)\varepsilon^2\}$.
$B_\varepsilon = \{b \in \Theta : K(\pi_0, \pi_b) \leq \varepsilon^2, \mathrm{Var}_{b_0}(\log \frac{\pi_0}{\pi_b}) \leq \varepsilon^2, \mathrm{KL}(b_0, b) \leq \Delta\varepsilon^2, \mathrm{Var}_{b_0}(\log \frac{p_0}{p_b}) \leq \Delta\varepsilon^2\}$.
Π: The prior distribution.
$\Pi(\cdot \mid X^{(n)})$: The posterior distribution given data $X^{(n)}$.
$\langle \cdot, \cdot \rangle$: the $L^2([0,1])$ inner product, $\langle f, g \rangle = \int_0^1 f(x)g(x)\,\mathrm{d}x$.
$\|\cdot\|_2$: The $L^2([0,1])$-norm, $\|f\|_2^2 = \int_0^1 f(x)^2\,\mathrm{d}x$.
$\|\cdot\|_\mu$: The $L^2(\mu)$-norm, $\|f\|_\mu^2 = \int_0^1 f(x)^2\mu(\mathrm{d}x) = \int_0^1 f(x)^2\pi_b(x)\,\mathrm{d}x$.
$\|\cdot\|_\infty$: The L^∞ (supremum) norm, $\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|$ (all functions we use are continuous hence we can take the supremum rather than needing the essential supremum).
$\|\cdot\|_{C^1_{\mathrm{per}}}$: The C^1_{per}-norm, $\|f\|_{C^1_{\mathrm{per}}} = \|f\|_\infty + \|f'\|_\infty$.
$\|\cdot\|_n$: The empirical L^2-norm $\|f\|_n^2 = \sum_{k=1}^n f(X_{k\Delta})^2$.

Acknowledgements

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/L016516/1 for the University of Cambridge Centre for Doctoral Training, the Cambridge Centre for Analysis. I would like to thank Richard Nickl for his valuable support throughout the process of writing this paper. I would also like to thank two anonymous referees for their very helpful suggestions.

Supplementary Material

Supplementary material for "Nonparametric Bayesian posterior contraction rates for scalar diffusions with high-frequency data." (DOI: 10.3150/18-BEJ1067SUPP; .pdf). We supply proofs to accompany those in the article.

References

[1] Abraham, K. (2018). Supplement to "Nonparametric Bayesian posterior contraction rates for scalar diffusions with high-frequency data". DOI:10.3150/18-BEJ1067SUPP.
[2] Baraud, Y. (2010). A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression. *Bernoulli* **16** 1064–1085. MR2759169
[3] Bass, R.F. (2011). *Stochastic Processes*. *Cambridge Series in Statistical and Probabilistic Mathematics* **33**. Cambridge: Cambridge Univ. Press. MR2856623
[4] Bhattacharya, R., Denker, M. and Goswami, A. (1999). Speed of convergence to equilibrium and to normality for diffusions with multiple periodic scales. *Stochastic Process. Appl.* **80** 55–86. MR1670111
[5] Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli* **4** 329–375. MR1653272
[6] Castillo, I. and Nickl, R. (2013). Nonparametric Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** 1999–2028. MR3127856

[7] Castillo, I. and Nickl, R. (2014). On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** 1941–1969. MR3262473

[8] Chorowski, J. (2016). Statistics for diffusion processes with low and high-frequency observations. Ph.D. thesis, Mathematisch-Naturwissenschaftliche Fakultät, Humboldt-Universität zu Berlin.

[9] Comte, F., Genon-Catalot, V. and Rozenholc, Y. (2007). Penalized nonparametric mean square estimation of the coefficients of diffusion processes. *Bernoulli* **13** 514–543. MR2331262

[10] Dalalyan, A. (2005). Sharp adaptive estimation of the drift function for ergodic diffusions. *Ann. Statist.* **33** 2507–2528. MR2253093

[11] Ghosal, S., Ghosh, J.K. and van der Vaart, A.W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[12] Ghosal, S. and van der Vaart, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274

[13] Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge: Cambridge Univ. Press. MR3587782

[14] Gihman, I. and Skorohod, A.V. (1972). *Stochastic Differential Equations*. New York: Springer. MR0346904

[15] Giné, E. and Nickl, R. (2011). Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.* **39** 2883–2911. MR3012395

[16] Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **40**. New York: Cambridge Univ. Press. MR3588285

[17] Gobet, E., Hoffmann, M. and Reiß, M. (2004). Nonparametric estimation of scalar diffusions based on low frequency data. *Ann. Statist.* **32** 2223–2253. MR2102509

[18] Gugushvili, S. and Spreij, P. (2014). Nonparametric Bayesian drift estimation for multidimensional stochastic differential equations. *Lith. Math. J.* **54** 127–141. MR3212631

[19] Hoffmann, M. (1999). Adaptive estimation in diffusion processes. *Stochastic Process. Appl.* **79** 135–163. MR1670522

[20] Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scand. J. Stat.* **24** 211–229. MR1455868

[21] Kutoyants, Y.A. (2004). *Statistical Inference for Ergodic Diffusion Processes*. London: Springer. MR2144185

[22] Liptser, R.S. and Shiryayev, A.N. (1977). *Statistics of Random Processes. I. General Theory*. New York: Springer. MR0474486

[23] Nickl, R. (2019). Bernstein–von Mises theorems for statistical inverse problems I: Schrödinger equation. *J. Eur. Math. Soc.* (*JEMS*). To appear.

[24] Nickl, R. and Söhl, J. (2017). Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *Ann. Statist.* **45** 1664–1693. MR3670192

[25] Papaspiliopoulos, O., Pokern, Y., Roberts, G.O. and Stuart, A.M. (2012). Nonparametric estimation of diffusions: A differential equations approach. *Biometrika* **99** 511–531. MR2966767

[26] Paulin, D. (2015). Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Probab.* **20** Article ID 79. MR3383563

[27] Pokern, Y., Stuart, A.M. and van Zanten, J.H. (2013). Posterior consistency via precision operators for Bayesian nonparametric drift estimation in SDEs. *Stochastic Process. Appl.* **123** 603–628. MR3003365

[28] Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*. *Cambridge Series in Statistical and Probabilistic Mathematics* **8**. Cambridge: Cambridge Univ. Press. MR1873379

[29] Ray, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat.* **7** 2516–2549. MR3117105

[30] Rogers, L.C.G. and Williams, D. (2000). *Diffusions, Markov Processes, and Martingales. Vol.* 2. *Itô Calculus. Cambridge Mathematical Library*. Cambridge: Cambridge Univ. Press. Reprint of the second (1994) edition. MR1780932

[31] Söhl, J. and Trabs, M. (2016). Adaptive confidence bands for Markov chains and diffusions: Estimating the invariant measure and the drift. *ESAIM Probab. Stat.* **20** 432–462. MR3581829

[32] Triebel, H. (1983). *Theory of Function Spaces. Monographs in Mathematics* **78**. Basel: Birkhäuser. MR0781540

[33] van der Meulen, F., Schauer, M. and van Waaij, J. (2018). Adaptive nonparametric drift estimation for diffusion processes using Faber–Schauder expansions. *Stat. Inference Stoch. Process.* **21** 603–628. MR3846994

[34] van der Meulen, F. and van Zanten, H. (2013). Consistent nonparametric Bayesian inference for discretely observed scalar diffusions. *Bernoulli* **19** 44–63. MR3019485

[35] van Waaij, J. and van Zanten, H. (2016). Gaussian process methods for one-dimensional diffusions: Optimal rates and adaptation. *Electron. J. Stat.* **10** 628–645. MR3471991

Received February 2018 and revised July 2018