Lithuanian Mathematical Journal, Vol. 54, No. 2, April, 2014, pp. 127–141

Nonparametric Bayesian drift estimation for multidimensional stochastic differential equations*

Shota Gugushvili [a] **and Peter Spreij** [b]

[a] Mathematical Institute, Leiden University, PO Box 9512, 2300 RA Leiden, The Netherlands
[b] Korteweg-de Vries Institute for Mathematics, University of Amsterdam,
PO Box 94248, 1090 GE Amsterdam, The Netherlands
(e-mail: shota.gugushvili@math.leidenuniv.nl; spreij@uva.nl)

Received February 25; revised September 9, 2013

Abstract. We consider nonparametric Bayesian estimation of the drift coefficient of a multidimensional stochastic differential equation from discrete-time observations on the solution of this equation. Under suitable regularity conditions, we establish posterior consistency in this context.

MSC: primary 62G20; secondary 62M05

Keywords: drift coefficient, nonparametric Bayesian estimation, posterior consistency, stochastic differential equation

1 Introduction

Consider the d-dimensional stochastic differential equation

$$\mathrm{d}X_t = b(X_t)\,\mathrm{d}t + \mathrm{d}W_t \tag{1.1}$$

driven by a d-dimensional Brownian motion W and assume that it has a unique (in the sense of the probability law) nonexploding weak solution. One can start with a coordinate mapping process X (that is, $X_t(\omega) = \omega(t)$) on the canonical space $(\mathcal{C}(\mathbb{R}_+), \mathcal{B}(\mathcal{C}(\mathbb{R}_+)))$ of continuous functions $\omega : \mathbb{R}_+ \to \mathbb{R}^d$, a flow of sigma-fields $\{\mathcal{F}_t^X\}$, and the d-dimensional Wiener measure Q on $(\mathcal{C}(\mathbb{R}_+), \mathcal{B}(\mathcal{C}(\mathbb{R}_+)))$. Then, as is well known (see, e.g., Proposition 3.6 and Remark 3.7 in [13, p. 303]), under suitable conditions on the drift coefficient b, for any fixed initial distribution μ, one can obtain a weak solution (X, W), $(\mathcal{C}(\mathbb{R}_+), \mathcal{F}, P_b^{\mu})$, $\{\mathcal{F}_t\}$ to (1.1) through the Girsanov theorem. The filtration $\{\mathcal{F}_t\}$ can be made to satisfy the usual conditions by suitably augmenting and completing the filtration $\{\mathcal{F}_t^X\}$ (see Remark 3.7 in [13, p. 303]). Henceforth, we will assume that we are in this canonical setup. We will also assume that X is ergodic with a unique ergodic distribution μ_b and is, in fact, initialized at μ_b, so that $\mu = \mu_b$. Furthermore, we will abbreviate $P_b^{\mu_b}$ as P_b.

Suppose that the drift coefficient $b = (b_1, \ldots, b_d)$ belongs to some nonparametric class. Denote by $b_0 = (b_{0,1}, \ldots, b_{0,d})^{\mathrm{tr}}$ (here tr denotes transposition) the true drift coefficient and assume that a correspond-

[*] The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637. Financial support of the Netherlands Organization for Scientific Research (NWO) is likewise acknowledged.

ing sample $X_0, X_\Delta, X_{2\Delta}, \ldots, X_{n\Delta}$ is given. The goal is to estimate b_0 nonparametrically. The problem of nonparametric estimation of b_0 from discrete-time observations has received considerable attention in the literature. For frequentist approaches to the problem, see, for instance, [2, 10, 12] in the one-dimensional case ($d = 1$) and [3, 21] in the general multidimensional case ($d \geqslant 1$). However, a nonparametric Bayesian approach to estimation of b_0 is also possible; see, e.g., [14, 16, 26]. In particular, under appropriate assumptions on the drift coefficient b, the weak solution to (1.1) admits transition densities $p_b(t, x, y)$, and employing the Markov property, the likelihood corresponding to the observations $X_{i\Delta}$ can be written as

$$\pi_b(X_0) \prod_{i=1}^{n} p_b(\Delta, X_{(i-1)\Delta}, X_{i\Delta}), \tag{1.2}$$

where π_b denotes a density of the distribution μ_b of X_0 (under our conditions, π_b and p_b will be strictly positive and finite; see Section 2 for details). A Bayesian would put a prior Π on the class of drift coefficients, say $\mathcal{X}$, and obtain a posterior measure of any measurable set $B \subset \mathcal{X}$ through Bayes' formula

$$\Pi(B|X_0, \ldots, X_{n\Delta}) = \frac{\int_B \pi_b(X_0) \prod_{i=1}^{n} p_b(\Delta, X_{(i-1)\Delta}, X_{i\Delta})\, \Pi(\mathrm{d}b)}{\int_{\mathcal{X}} \pi_b(X_0) \prod_{i=1}^{n} p_b(\Delta, X_{(i-1)\Delta}, X_{i\Delta})\, \Pi(\mathrm{d}b)}. \tag{1.3}$$

Here we tacitly assume a suitable measurability of the integrands, so that the integrals in (1.3) are well defined. In the Bayesian paradigm, the posterior encapsulates all the information required for inferential purposes. Once the posterior is available, one can proceed with computation of Bayes point estimates, credible sets, and other quantities of interest in Bayesian statistics.

It has been argued convincingly in [4] and elsewhere that a desirable property of a Bayes procedure is posterior consistency. In our context this will mean that for every neighborhood (in a suitable topology) U_{b_0} of b_0,

$$\Pi\big(U_{b_0}^c \big| X_0, \ldots, X_{n\Delta}\big) \to 0, \quad P_{b_0}\text{-a.s.}$$

as $n \to \infty$ (see Section 2 for details). That is, roughly speaking, a consistent Bayesian procedure asymptotically puts posterior mass equal to one on every fixed neighborhood of the true parameter: the posterior concentrates around the true parameter. In an infinite-dimensional setting, such as the one we are dealing with, posterior consistency is a subtle property that depends in an essential way on a specification of the prior; see, e.g., [4]. Note also that the notion of posterior consistency depends on the topology on $\mathcal{X}$. Ideally, one would like to establish posterior consistency in strong topologies. An implication of posterior consistency is that even though two Bayesians might start with two different priors, the role of the prior in their inferential conclusions will asymptotically, with the sample size growing indefinitely, wash out, and the two will eventually agree. Furthermore, posterior consistency also implies that the center (in an appropriate sense) of the posterior distribution is a consistent (in the frequentist sense) estimator of the true parameter. For an introductory treatment of posterior consistency, see [25].

In the context of discretely observed scalar diffusion processes given as solutions to stochastic differential equations (of type (1.1) with $d = 1$), posterior consistency has been recently addressed in [16], whereas the case where a continuous record of observations from a scalar diffusion process is available was covered under various setups in [15, 17, 19], where, in particular, the contraction rates of the posterior were derived. The techniques used in the latter three papers are of little use in the case of discrete observations. The proof of posterior consistency in [16] is based on the use of martingale arguments in a fashion similar to [22] (see also [7]). The latter paper deals with posterior consistency for estimation of the transition density of an ergodic Markov process. The idea of using martingale arguments in the proofs of consistency of nonparametric Bayesian procedures goes back to [23] and [24] in the i.i.d. setting. On the other hand, a similarity between the arguments used in the proof of posterior consistency in [22] and [16] is to a considerable extent on a conceptual level only: conditions for posterior consistency in [22] involve conditions on transition densities that typically cannot be transformed into conditions on the drift coefficients because transition densities associated with stochastic differential equations are usually unknown in explicit form. Furthermore, in the setting of [16], which deals with ergodic and strictly stationary scalar diffusion processes (in particular, X_0 is initialized at the ergodic distribution of the process X), one cannot assume that the density π_{b_0} of X_0 is known (as done in [22, p. 1714]) since it would completely determine the unknown drift coefficient b_0.

The assumption on the class of drift coefficients in Theorem 3.5 of [16] (the latter deals with posterior consistency), namely, the uniform boundedness of the drift coefficients, is quite restrictive in that it excludes even such a prototypical example of a stochastic differential equation as the Langevin equation (here we assume $d = 1$)

$$\mathrm{d}X_t = -\beta X_t \,\mathrm{d}t + \sigma \,\mathrm{d}W_t, \tag{1.4}$$

where β and σ are two constants. A solution to (1.4) is called an Ornstein–Uhlenbeck process (see Example 6.8 in [13, p. 358] and page 397 there). Hence, there is room for improvement. Furthermore, it is interesting to investigate the case of multidimensional stochastic differential equations as well.

In this work, we will show that, under standard assumptions in nonparametric inference for multidimensional stochastic differential equations (see [3] and [21]), posterior consistency holds for nonparametric Bayesian estimation of an unbounded drift coefficient satisfying the linear growth condition. In particular, the case of the Langevin equation will be covered. In our proof of posterior consistency, we follow the same train of thought as initiated in [23] and [24], at the same time, making use of ideas from [22] and, especially, from [16]. According to [16, p. 51], the boundedness condition on the drift coefficients cannot be avoided in their approach due to technical reasons. Our analysis and contribution to the literature, however, shows that the case of unbounded drift coefficients and multidimensional stochastic differential equations can also be covered via techniques similar to those in [16]. We would also like to remark that, in the scalar case, posterior consistency for nonparametric estimation of an unbounded drift coefficient holds under weaker conditions than those given in this work. Due to space restrictions, we decided to omit a separate discussion of the scalar case. Instead, we refer to an extended version of the paper (see [11]).

The rest of the paper is organized as follows. In the next section, we give our main result. In Section 3, we provide a brief discussion on it. The proofs are given in Section 4. Finally, the Appendix contains several auxiliary statements used in Section 4.

2 Posterior consistency

Our parameter set will be a subset of the class $\tilde{\mathcal{X}}(K_1, K_2)$ of drift coefficients introduced below.

DEFINITION 1. The family $\tilde{\mathcal{X}}(K_1, K_2)$ consists of drift coefficients $b : \mathbb{R}^d \to \mathbb{R}^d$ possessing the following three properties:

(a) for any $b \in \tilde{\mathcal{X}}(K_1, K_2)$, there exists a $\mathcal{C}^3$-function $V_b : \mathbb{R}^d \to \mathbb{R}$ such that

$$C_b = \int\limits_{\mathbb{R}^d} \mathrm{e}^{-2V_b(u)} \,\mathrm{d}u < \infty,$$

$|V_b(x)|$ grows not faster than a polynomial of $\|x\|$ at infinity, and $b = -[\nabla V_b]^{\mathrm{tr}}$, where ∇V_b is the gradient of V_b;

(b) for any $b \in \tilde{\mathcal{X}}(K_1, K_2)$, there exist three constants $r_b > 0$, $M_b > 0$, and $\alpha_b \geqslant 1$ such that

$$b(x) \cdot x \leqslant -r_b \|x\|^{\alpha_b} \quad \forall \|x\| \geqslant M_b,$$

where by a dot we denote the usual scalar product on $\mathbb{R}^d$, and $\|x\|$ is the L_2-norm of a vector $x \in \mathbb{R}^d$;

(c) there exist two constants $K_1 > 0$ and $K_2 > 0$ such that, for any $b \in \tilde{\mathcal{X}}(K_1, K_2)$,

$$\bigl\|b(x)\bigr\| \leqslant K_1\bigl(1 + \|x\|\bigr), \qquad \left|\frac{\partial}{\partial x_j} b_i(x)\right| \leqslant K_2 \quad \forall x \in \mathbb{R}^d,\ i, j = 1, \dots, d.$$

Remark 1. Assumptions made in Definition 1 are more than enough to guarantee the existence of a unique (in the sense of the probability law) solution to (1.1): $b \in \tilde{\mathcal{X}}(K_1, K_2)$ is automatically measurable, and the linear growth condition on b, together with Proposition 3.6 and Remark 3.7 in [13, p. 303], implies the existence of a weak solution. The uniqueness in the sense of the probability law follows by Proposition 3.10 in [13, p. 304] and Proposition 1.1 in [9]. By Proposition 1 in [21] (see [3, p. 27]) these assumptions also imply the existence of a unique ergodic distribution μ_b that has the density

$$\pi_b(x) = \frac{1}{C_b} \mathrm{e}^{-2V_b(x)} > 0$$

with respect to the d-dimensional Lebesgue measure. In models in physics, the function V_b has the interpretation of the potential energy of the system. Furthermore, Proposition 1.2 in [9] implies the existence of strictly positive transition densities $p_b(t, x, y)$ associated with (1.1). Finally, for any $b, \tilde{b} \in \mathcal{X}(K_1, K_2)$, we also have that the Kullback–Leibler divergence $\mathrm{K}(\mu_b, \mu_{\tilde{b}})$ between μ_b and $\mu_{\tilde{b}}$ is finite, which we use in the proof of Lemma A.4 in the Appendix.

Remark 2. Examples of multidimensional stochastic differential equations satisfying the assumptions in Definition 1 are given in Section 5.2 in [21]. In particular, when $d = 1$, Definition 1 covers the case of the Langevin equation (with $\sigma = 1$ and for parameter β ranging in the interval $(0, K]$ for some constant K).

Remark 3. The positivity of π_b and p_b formally justifies rewriting the likelihood as in (1.2) and allows us to employ the likelihood ratio $L_n(b)$ in the proof of our main result, Theorem 1.

Remark 4. The measurability of the mapping $b \mapsto p_b(t, x, y)$ is a subtle property essential in (1.3), but it is difficult to ascertain it in a general setting. Therefore, we will simply tacitly assume that all the quantities in (1.3) (or in other formulae where we integrate with respect to the prior) are suitably measurable.

Since the notion of posterior consistency depends on a topology on the class of drift coefficients under consideration, we first have to introduce the latter. We want our topology to separate distinct drift coefficients, which can be thought of as an identifiability condition. At the same time, we want the posterior measure to concentrate on arbitrarily small neighborhoods of the true parameter b_0. Fortunately, this will be possible with our choice of topology, as it will have the required separation property.

We will base our topology on the transition operators P^b_Δ. Transition operators associated with (1.1) and acting on the class of bounded measurable functions $f : \mathbb{R}^d \to \mathbb{R}$ are given by

$$\mathrm{P}^b_t f(x) = \int_{\mathbb{R}^d} p_b(t, x, y) f(y) \,\mathrm{d}y.$$

As it often happens in practice, it will be convenient in our case to define a topology not by directly specifying the open sets, but rather by specifying a subbase $\tilde{\mathcal{U}}$ (for a notion of a subbase, see, e.g., [5, p. 37]).

Definition 2. Let ν be a finite Borel measure on $\mathbb{R}^d$ that assigns strictly positive mass to every nonempty open subset of $\mathbb{R}^d$, and let $\mathcal{C}_{bdd}(\mathbb{R}^d)$ denote the class of all bounded continuous functions on $\mathbb{R}^d$. For fixed $b \in \tilde{\mathcal{X}}(K_1, K_2)$, $f \in \mathcal{C}_{bdd}(\mathbb{R}^d)$, and $\varepsilon > 0$, define

$$U^b_{f,\varepsilon} = \left\{ \tilde{b} \in \tilde{\mathcal{X}}(K_1, K_2) \colon \left\| \mathrm{P}^{\tilde{b}}_\Delta f - \mathrm{P}^b_\Delta f \right\|_{1,\nu} < \varepsilon \right\},$$

where $\| \cdot \|_{1,\nu}$ denotes the L_1-norm with respect to the measure ν. Furthermore, let

$$\tilde{\mathcal{U}} = \left\{ U^b_{f,\varepsilon} \colon f \in \mathcal{C}_{bdd}\left(\mathbb{R}^d\right),\ \varepsilon > 0,\ b \in \tilde{\mathcal{X}}(K_1, K_2) \right\}.$$

The topology $\tilde{\mathcal{T}}$ on $\tilde{\mathcal{X}}(K_1, K_2)$ is determined by the requirement that the family $\tilde{\mathcal{U}}$ is a subbase for $\tilde{\mathcal{T}}$.

Remark 5. The fact that Definition 2 is indeed a valid definition follows from a standard result in general topology, Theorem 2.2.6 in [5]. Note also that $\tilde{\mathcal{T}}$ depends on the choice of the measure ν. Since ν is assumed to be fixed beforehand and its specific choice is not of great importance for subsequent developments, it is not reflected in our notation.

Remark 6. Let $d = 1$. The topology in Definition 2 has already been employed in [16], who in that respect follow Section 6 in [22]. For a $\mathcal{C}^2$-function f and small Δ,

$$\mathrm{P}_t^b f(x) - \mathrm{P}_t^{\tilde{b}} f(x) \approx \Delta\bigl(b(x) - \tilde{b}(x)\bigr) f'(x)$$

(see [16, p. 50]). Hence, for small Δ, the topology $\tilde{\mathcal{T}}$ in some sense resembles the topology induced by the $L_1(\nu)$-norm on the collection of drift coefficients.

We will now show that the topology of Definition 2 has the Hausdorff property. This is perfectly sufficient for our purposes. For a notion of a Hausdorff space, see, e.g., [5, p. 30].

Lemma 1. *The topological space $(\tilde{\mathcal{X}}(K_1, K_2), \tilde{\mathcal{T}})$ with $\tilde{\mathcal{X}}(K_1, K_2)$ as in Definition* 1 *is a Hausdorff space.*

Let $\mathcal{X}(K_1, K_2) \subseteq \tilde{\mathcal{X}}(K_1, K_2)$ with the interpretation that $\mathcal{X}(K_1, K_2)$ is our parameter set, and let $\mathcal{T} = \{A \cap \mathcal{X}(K_1, K_2)\colon A \in \tilde{\mathcal{T}}\}$ be the corresponding relative topology on $\mathcal{X}(K_1, K_2)$.

DEFINITION 3. If, for any neighborhood $U_{b_0} \in \mathcal{T}$ of $b_0 \in \mathcal{X}(K_1, K_2)$, we have

$$\Pi\bigl(U_{b_0}^c \big| X_0, \ldots, X_{n\Delta}\bigr) \to 0 \quad P_{b_0}\text{-a.s.}$$

as $n \to \infty$, we will say that posterior consistency holds at b_0.

We summarize our assumptions.

ASSUMPTION 1. Assume that:

(a) a unique in law nonexploding weak solution to (1.1) corresponding to each $b \in \mathcal{X}(K_1, K_2)$ is initialized at the ergodic distribution μ_b;
(b) $b_0 \in \mathcal{X}(K_1, K_2)$ denotes the true drift coefficient;
(c) a discrete-time sample $X_0, \ldots, X_{n\Delta}$ from the solution to (1.1) corresponding to b_0 is available (we assume that we are in the canonical setup as in Section 1), and, finally, Δ is fixed and independent of n.

The following is our main result.

Theorem 1. *Let Assumption* 1 *hold and suppose that the prior Π on $\mathcal{X}(K_1, K_2)$ is such that*

$$\Pi\left(b \in \mathcal{X}(K_1, K_2)\colon \left\{\sum_{i=1}^d \|b_i - b_{0,i}\|_{2,\mu_{b_0}}^2\right\}^{1/2} < \varepsilon\right) > 0 \quad \forall \varepsilon > 0. \tag{2.1}$$

Then posterior consistency as in Definition 3 *holds.*

Remark 7. Condition (2.1) on the prior Π is formulated in terms of the $L_2(\mu_{b_0})$-neighborhoods, whereas the posterior consistency assertion returned by Theorem 1 is for the weak topology $\mathcal{T}$. However, by Remark 6, for small Δ, at least in the case $d = 1$, the "discrepancy" is not as dramatic as it may seem at first sight.

Condition (2.1) on the prior is of the same type as the one in Theorem 3.5 in [16]. Since b_0 is unknown, the prior Π must verify (2.1) at all parameter values $b \in \mathcal{X}$. We provide an example of a prior Π satisfying this condition. The construction of Π is similar to that in Example 4.1 in [16]. Both examples are related to discrete net priors in nonparametric Bayesian inference problems studied in [6]. The construction is admittedly artificial, but its sole goal is to show the existence of a prior satisfying (2.1).

Example 1. Let $\mathfrak{F}$ be a collection of $\mathcal{C}^3$-functions $f : \mathbb{R} \to \mathbb{R}$ such that:

(a) for some constant $K_1 > 0$ and for all $f \in \mathfrak{F}$, we have

$$\left|f'(x)\right| \leqslant \frac{K_1}{2} \quad \forall x \in \mathbb{R}_+;$$

(b) for all $f \in \mathfrak{F}$, we have

$$\int\limits_{\mathbb{R}^d} \mathrm{e}^{-2f(\|x\|^2)}\,\mathrm{d}x < \infty;$$

(c) for all $f \in \mathfrak{F}$, there exist two constants $M_f > 0$ and $r_f > 0$ (possibly depending on f) such that $f'(x) \geqslant r_f$ for all $x \geqslant M_f$;

(d) for some constant $K_2 > 0$ and for all $f \in \mathfrak{F}$,

$$\sup_{x \in \mathbb{R}_+} \left\{4x\left|f''(x)\right| + 2\left|f'(x)\right|\right\} \leqslant K_2.$$

For all $x \in \mathbb{R}^d$, set $V_f(x) = f(\|x\|^2)$ and $b_f(x) = -[\nabla V_f(x)]^{\mathrm{tr}}$. Let $\mathcal{X}(K_1, K_2)$ be a subset of a collection of all functions $b_f = (b_{f,1}, \ldots, b_{f,d})$ obtained in this way (the fact that this is a valid definition, in the sense that the requirements from Definition 1 are satisfied, follows by easy but somewhat tedious computations; note that by taking $f_\beta = \beta x/2$ and assuming $d = 1$ and $\beta \in (0, K_1]$ we can cover the case of the Langevin equation (1.4)). We get from (a) that, for every fixed $i = 1, \ldots, d$, the functions $b_{f,i}$ are locally bounded by constants uniform in $f \in \mathfrak{F}$. Furthermore, they are Lipschitz with uniform constants in $f \in \mathfrak{F}$ as well: by the mean value theorem,

$$\left|b_{f,i}(x) - b_{f,i}(y)\right| \leqslant \left\|\nabla b_{f,i}\big(\lambda x + (1-\lambda)y\big)\right\|\|x - y\| \leqslant \sqrt{d}K_2\|x - y\|.$$

Hence, for each $m \in \mathbb{N}$ and $i = 1, \ldots, d$, by the Arzelà–Ascoli theorem (see Theorem 2.4.7 in [5]) the collection $\mathfrak{B}_{m,i}$ of restrictions $b_{f,i}|_{[-m,m]}$ of the functions $b_{f,i}$, $f \in \mathfrak{F}$, to the intervals $[-m, m]$ is totally bounded for the supremum metric $\|\cdot\|_\infty$ (for the required definitions, see [5, pp. 45, 52]). Then so is the product $\bigotimes_{i=1}^d \mathfrak{B}_{m,i}$ for the product metric

$$\|b_f\|_{d,m,\infty} = \max_{i=1,\ldots,d} \|b_{f,i}|_{[-m,m]}\|_\infty,$$

as well as its subset consisting of the elements

$$b_f|_{[-m,m]^d} = (b_{f,1}|_{[-m,m]}, \ldots, b_{f,d}|_{[-m,m]}), \quad f \in \mathfrak{F}.$$

Take a sequence $\epsilon_l \downarrow 0$. For any $l \in \mathbb{N}$, there exists a finite subset $\mathfrak{F}_{m,\epsilon_l} = \{f_n^{m,\epsilon_l},\ n = 1, \ldots, n_{m,l}\}$ such that, for all $f \in \mathfrak{F}$, $\|b_f - b_{f_n^{m,\epsilon_l}}\|_{d,m,\infty} < \epsilon_l$ for some $n = 1, \ldots, n_{m,l}$. Let $\tilde{Q}_1$ and $\tilde{Q}_2$ be two probability measures on $\mathbb{N}$ such that $q_{j,i} = \tilde{Q}_i(j) > 0$, $i = 1, 2$, $j \in \mathbb{N}$. The prior $\varPi$ on $\mathcal{X}(K_1, K_2)$ is defined by

$$\varPi = \sum_{m=1}^{\infty}\sum_{l=1}^{\infty}\sum_{n=1}^{n_{m,l}} \frac{q_{m,1}q_{l,2}}{n_{m,l}} \delta_{b_{f_n^{m,\epsilon_l}}},$$

where $\delta_{b_{f_n^{m,\epsilon_l}}}$ is the Dirac measure at $b_{f_n^{m,\epsilon_l}}$. The fact that $\varPi$ satisfies requirement (2.1) is the content of Lemma 2 in Section 4. Since $\varPi$ assigns all its mass to a countable subset of $\mathcal{X}(K_1, K_2)$, measurability issues should not concern us when integrating with respect to $\varPi$.

3 Discussion

In this work, we were able to demonstrate that posterior consistency for nonparametric Bayesian estimation of the drift coefficient of a stochastic differential equation holds not only for the class of uniformly bounded drift coefficients and in the scalar setting, as shown previously in [16], but also in the multidimensional setting for the class of drift coefficients satisfying a linear growth assumption (and some additional technical assumptions). This considerably enlarges the scope of the main result in [16]. Conditions that we impose are analogous to those used in the frequentist literature (see [3] and [21]), which is a comforting fact. On the other hand, posterior consistency results both in [16] and in our work are established for a weak topology on the class of drift coefficients. This is a consequence of the fact that we rely on techniques from [24] in our proofs, which are better suited for proving posterior consistency in weak topologies. Consistency in stronger topologies could have been established, and contraction rates of the posterior could have been derived from general results for posterior consistency in Markov chain models had we known the existence of certain tests satisfying the conditions as in formula (2.2) in [8]; see Theorem 5 there. The existence of such tests for Markov chain models has been demonstrated in Theorem 3 in [1], but unfortunately, the conditions involved in this theorem (see also formula (4.1) in [8]) do not appear to hold, in general, for the stochastic differential equation models we consider. Hence, establishing posterior consistency in a stronger topology and derivation of the posterior contraction rate for nonparametric Bayesian drift estimation is an interesting and difficult open problem. A recent paper [19] addresses the latter question for a one-dimensional stochastic differential equation with a periodic drift coefficient. However, this is done under the assumption that an entire sample path $\{X_t\colon t \in [0,T]\}$ is observed over the time interval $[0,T]$ with $T \to \infty$. Moreover, periodic drift coefficients are completely different from the drift coefficients considered in Section 2 of the present work, and making use of the techniques from [19] is impossible in our setting. Neither are the techniques in [15] and [17] of any significant help (these papers deal with continuously observed scalar diffusion processes). It should also be noted that also with the frequentist approaches (with Δ fixed), already in the one-dimensional setting, study of convergence rates of nonparametric estimators of the drift and dispersion coefficients is a highly nontrivial task (see, e.g., [10]), where various simplifying assumptions have been made, such as the requirement that the diffusion process under consideration has a compact state space, say $[0,1]$, and is reflecting at the boundary points. Nevertheless, some progress in establishing posterior consistency in a stronger topology than in this work might be possible in the setting where $\Delta = \Delta_n \to 0$ in such a way that $n\Delta_n \to \infty$ (the so-called high-frequency data setting).

Finally, we remark that issues associated with practical implementation of the nonparametric Bayesian approach to estimation of a drift coefficient are outside the scope of this work. Although much remains to be done in this direction, preliminary studies, such as those in [18] and [14] (see also the overview paper [26]) indicate that a nonparametric Bayesian approach in this context is both feasible and leads to reasonable results.

4 Proofs

Proof of Lemma 1. The lemma can be proved by arguments similar to those in the proof of Lemma 3.2 in [16]. The proof employs Lemma A.1 from the Appendix, which plays the role of Lemma 3.1 from [16] in this context. □

Proof of Theorem 1. The proof follows the same main steps as the proof of Theorem 3.5 in [16], which in turn uses some ideas from [22] and [24]. Fix $\varepsilon > 0$, take a fixed $f \in \mathcal{C}_{bdd}(\mathbb{R}^d)$, and write

$$B = \bigl\{b \in \mathcal{X}(K_1, K_2)\colon \bigl\|\mathrm{P}_\Delta^b f - \mathrm{P}_\Delta^{b_0} f\bigr\|_{1,\nu} > \varepsilon\bigr\}. \tag{4.1}$$

Without loss of generality, we may assume that $\|f\|_\infty \leqslant 1$ and $\varepsilon \leqslant 2\nu(\mathbb{R}^d)$. We claim that by the definition of the topology $\mathcal{T}$ it suffices to establish posterior consistency for every fixed B of the above form. Indeed, a subbase $\mathcal{U}$ for $\mathcal{T}$ can be obtained by intersecting the sets from the subbase $\tilde{\mathcal{U}}$ for $\tilde{\mathcal{T}}$ with $\mathcal{X}(K_1, K_2)$. By definition, an arbitrary neighborhood U_{b_0} of b_0 contains an open set $\hat{U}_{b_0} \in \mathcal{T}$. The set $\hat{U}_{b_0}$ is a union of open sets V from the base $\mathcal{V}$ (determined by $\mathcal{U}$), $\hat{U}_{b_0} = \bigcup\{V \in \mathcal{V}\colon V \subset \hat{U}_{b_0}\}$. There is at least one V that contains b_0. Fix such V. By the definition of the subbase $\mathcal{U}$ this set V can be represented as $V = \bigcap_{j=1}^m U_{f_j,\varepsilon_j}^{b_0}$ for some m, positive numbers ε_j, bounded continuous functions f_j, and sets $U_{f_j,\varepsilon_j}^{b_0}$ from the subbase $\mathcal{U}$. Note that we have

$$U_{b_0}^c \subset \hat{U}_{b_0}^c \subset V^c = \bigcup_{j=1}^m \big(U_{f_j,\varepsilon_j}^{b_0}\big)^c.$$

Since

$$\begin{aligned}\big(U_{f_j,\varepsilon_j}^{b_0}\big)^c &= \big\{b \in \mathcal{X}(K_1,K_2)\colon \big\|\mathrm{P}_\Delta^b f_j - \mathrm{P}_\Delta^{b_0} f_j\big\|_{1,\nu} \geqslant \varepsilon_j\big\} \\ &\subset \left\{b \in \mathcal{X}(K_1,K_2)\colon \big\|\mathrm{P}_\Delta^b f_j - \mathrm{P}_\Delta^{b_0} f_j\big\|_{1,\nu} > \frac{\varepsilon_j}{2}\right\},\end{aligned}$$

say, the claim becomes obvious.

The posterior measure of the set B given in (4.1) can be written as

$$\varPi(B|X_0,\dots,X_{n\Delta}) = \frac{\int_B L_n(b)\,\varPi(\mathrm{d}b)}{\int_{\mathcal{X}(K_1,K_2)} L_n(b)\,\varPi(\mathrm{d}b)},$$

where

$$L_n(b) = \frac{\pi_b(X_0)}{\pi_{b_0}(X_0)} \prod_{i=1}^n \frac{p_b(\Delta, X_{(i-1)\Delta}, X_{i\Delta})}{p_{b_0}(\Delta, X_{(i-1)\Delta}, X_{i\Delta})}$$

is the likelihood ratio. By Lemma A.2 from the Appendix, in order to prove the theorem, it suffices to show that

$$\varPi\big(B_j^+\big|X_0,\dots,X_{n\Delta}\big) \to 0, \quad \varPi\big(B_j^-\big|X_0,\dots,X_{n\Delta}\big) \to 0 \quad P_{b_0}\text{-a.s.}$$

for the sets B_j^+ and B_j^- ($j = 1,\dots,N$ for some suitable integer $N > 0$) given in the statement of that lemma. We give a brief outline of the remaining part of the proof: thanks to property (2.1) of the prior, by Lemma A.4 from the Appendix the prior $\varPi$ has the Kullback–Leibler property in the sense that (A.5) holds. Then by Lemma A.5 from the Appendix, in order to establish posterior consistency, it suffices to show that P_{b_0}-a.s. the terms

$$\sqrt{\int_{B_j^+} L_n(b)\,\varPi(\mathrm{d}b)} \quad \text{and} \quad \sqrt{\int_{B_j^-} L_n(b)\,\varPi(\mathrm{d}b)}$$

converge to zero exponentially fast. This fact can be proved by a reasoning similar to that given in the proof of Theorem 3.5 in [16] (employing the convergence theorem for a positive supermartingale, see, e.g., Theorem 22 in [20, p. 148], and not Doob's martingale convergence theorem as employed in [16, pp. 59–60][1]). This completes the proof. □

In the next lemma, we verify the claim made at the end of Example 1.

Lemma 2. *The prior $\varPi$ from Example* 1 *satisfies requirement* (2.1).

[1] Note that, in [16, p. 58], the expression L_n is called the likelihood, although obviously the likelihood ratio is meant.

Proof. The proof is similar to the demonstration of an analogous property of the prior in Example 4.1 in [16]: for all $b \in \mathcal{X}(K_1, K_2)$ and positive integers m, we have

$$\begin{aligned}\sum_{i=1}^{d} \|b_i - b_{i,0}\|^2_{2,\mu_{b_0}} &= \sum_{i=1}^{d} \int\limits_{\|x\|\leqslant m} \bigl(b_i(x) - b_{0,i}(x)\bigr)^2 \pi_{b_0}(x)\,\mathrm{d}x + \sum_{i=1}^{d} \int\limits_{\|x\|>m} \bigl(b_i(x) - b_{0,i}(x)\bigr)^2 \pi_{b_0}(x)\,\mathrm{d}x \\ &\leqslant d\|b - b_0\|_{m,d,\infty} + 4K^2 d \int\limits_{\|x\|>m} \bigl(1 + \|x\|\bigr)^2 \pi_{b_0}(x)\,\mathrm{d}x.\end{aligned}$$

Thanks to the fact that μ_{b_0} has an exponential moment, the second term on the right-hand side can be made less than ε^2 by choosing m large enough. Hence,

$$\varPi\Biggl(b \in \mathcal{X}(K_1, K_2)\colon\ \sum_{i=1}^{d} \|b_i - b_{0,i}\|^2_{2,\mu_{b_0}} < 2\varepsilon^2\Biggr) \geqslant \varPi\Biggl(b \in \mathcal{X}(K_1, K_2)\colon\ \|b - b_0\|^2_{m,d,\infty} < \frac{\varepsilon^2}{d}\Biggr).$$

For l such that $\epsilon_l < \varepsilon/\sqrt{d}$, we have by construction of $\varPi$ that the right-hand side of the above display is bounded from below by $q_{m,1}q_{l,2}/k_{m,l} > 0$. This completes the proof of the lemma. □

Appendix

Lemma A.1. *Let $b, \tilde{b} \in \tilde{\mathcal{X}}(K_1, K_2)$. Fix $t > 0$. If $b \neq \tilde{b}$, then $\mathrm{P}^b_t \neq \mathrm{P}^{\tilde{b}}_t$.*

Proof. The proof is similar to the proof of Lemma 3.1 in [16]. By the continuity of b and $\tilde{b}$ we have that if $b \neq \tilde{b}$, this, in fact, holds on a set of positive Lebesgue measure. Then also $V_b \neq V_{\tilde{b}}$ on a set of positive Lebesgue measure (it contains, for instance, some open ball in $\mathbb{R}^d$), and, therefore, $\pi_b \neq \pi_{\tilde{b}}$ on a set of positive Lebesgue measure. Now assume that $\mathrm{P}^b_t = \mathrm{P}^{\tilde{b}}_t$. Then, for any bounded measurable function f and any positive integer m, by the semigroup property of P^b_t we have that

$$\mathbf{E}^b_x\bigl[f(X_{mt})\bigr] = \bigl(\mathrm{P}^b_t\bigr)^m f(x) = \bigl(\mathrm{P}^{\tilde{b}}_t\bigr)^m f(x) = \mathbf{E}^{\tilde{b}}_x\bigl[f(X_{mt})\bigr],$$

where $\mathbf{E}^b_x$ and $\mathbf{E}^{\tilde{b}}_x$ denote the expectation operators under parameter values b and $\tilde{b}$ when X is initialized at x. Letting $m \to \infty$, the above display and ergodicity give that

$$\int\limits_{\mathbb{R}^d} f(y)\pi_b(y)\,\mathrm{d}y = \int\limits_{\mathbb{R}^d} f(y)\pi_{\tilde{b}}(y)\,\mathrm{d}y.$$

It follows that $\pi_b = \pi_{\tilde{b}}$ Lebesgue-a.e., and, in fact, by continuity $\pi_b = \pi_{\tilde{b}}$ everywhere. This is a contradiction and thus $b \neq \tilde{b}$ implies $\mathrm{P}^b_t \neq \mathrm{P}^{\tilde{b}}_t$. □

Lemma A.2. *Fix $\varepsilon > 0$ such that $\varepsilon \leqslant 2\nu(\mathbb{R}^d)$, take a fixed $f \in \mathcal{C}_{bdd}(\mathbb{R}^d)$ such that $\|f\|_\infty \leqslant 1$, and write*

$$B = \bigl\{b \in \mathcal{X}\colon\ \bigl\|\mathrm{P}^b_\Delta f - \mathrm{P}^{b_0}_\Delta f\bigr\|_{1,\nu} > \varepsilon\bigr\}.$$

Then there exist a compact set $F \subset \mathbb{R}^d$, an integer $N > 0$, and cubes $I_1, \dots, I_N$ covering F such that

$$B \subset \Biggl(\bigcup_{j=1}^{N} B^+_j\Biggr) \cup \Biggl(\bigcup_{j=1}^{N} B^-_j\Biggr),$$

where

$$B_j^+ = \left\{ b \in B \colon \mathrm{P}^b_\Delta f(x) - \mathrm{P}^{b_0}_\Delta f(x) > \frac{\varepsilon}{4\nu(F)} \ \forall x \in I_j \right\},$$

$$B_j^- = \left\{ b \in B \colon \mathrm{P}^b_\Delta f(x) - \mathrm{P}^{b_0}_\Delta f(x) < -\frac{\varepsilon}{4\nu(F)} \ \forall x \in I_j \right\}.$$

Proof. The proof of Lemma 5.3 in [16] carries over, provided that we redefine the intervals I_j of length $\delta/2 > 0$ from that proof to be cubes with sides of length $\delta/2$ and use, instead of Lemma A.1 from [16], Lemma A.3 given below. □

Recall that a family $\mathfrak{F}$ of functions $f : \mathbb{R}^d \to \mathbb{R}$ is called locally uniformly equicontinuous if, for any compact set $F \subset \mathbb{R}^d$, the restrictions $f|_F$ of the functions $f \in \mathfrak{F}$ to F form a uniformly equicontinuous family of functions, i.e., for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\sup_{f \in \mathfrak{F}} \sup_{\substack{x,y \in F \\ |x-y| < \delta}} \bigl| f(x) - f(y) \bigr| < \varepsilon.$$

The next lemma is an adaptation of Lemma A.1 in [16], but in its proof we need somewhat different arguments than those used in [16].

Lemma A.3. *For a fixed $f \in \mathcal{C}_{bdd}(\mathbb{R}^d)$ and $t > 0$, the family $\{\mathrm{P}^b_\Delta f \colon b \in \mathcal{X}(K_1, K_2)\}$ is a locally uniformly equicontinuous family of functions.*

Proof. In order to prove the lemma, we need to show that the family of functions $\{\mathrm{P}^b_\Delta f \colon b \in \mathcal{X}(K_1, K_2)\}$ is uniformly equicontinuous whenever the argument x of $\mathrm{P}^b_\Delta f(x)$ is restricted to an arbitrary compact set F. Fix a compact set $F \subset \mathbb{R}^d$. Throughout this proof, we assume that $x, y \in F$.

Let

$$l_u = \sum_{i=1}^d \int_0^\Delta b_i(u + W_s)\,\mathrm{d}W_{i,s} - \frac{1}{2} \sum_{i=1}^d \int_0^\Delta b_i^2(u + W_s)\,\mathrm{d}s \quad \text{and} \quad L_u = \mathrm{e}^{l_u}$$

for a standard d-dimensional Brownian motion $W = (W_1, \dots, W_d)$. Then, employing the Girsanov theorem as in the proof of Lemma A.1 in [16], it can be shown that

$$\mathrm{P}^b_\Delta f(x) = \mathbf{E}\bigl[f(x + W_\Delta) L_x\bigr],$$

where the expectation is evaluated under the Wiener measure. It is enough to prove the lemma for Δ small enough, in particular, such that

$$\Delta K_1 < \frac{1}{2\sqrt{d}}. \tag{A.1}$$

In fact, by the semigroup property of the transition operators we have

$$\bigl|\mathrm{P}^b_\Delta f(x) - \mathrm{P}^b_\Delta f(y)\bigr| \leqslant \mathrm{P}^b_{\Delta/2}\bigl|\mathrm{P}^b_{\Delta/2} f(x) - \mathrm{P}^b_{\Delta/2} f(y)\bigr|,$$

and if $\{\mathrm{P}^b_{\Delta/2} f \colon b \in \mathcal{X}(K_1, K_2)\}$ is uniformly equicontinuous when the argument x ranges in F, then it is immediately seen that so is $\{\mathrm{P}^b_\Delta f \colon b \in \mathcal{X}(K_1, K_2)\}$, while if not, then we can reiterate the same argument, but now with $\Delta/2$ and $\Delta/4$ instead of Δ and $\Delta/2$, and so on, until (A.1) is met.

We have

$$\begin{aligned}\bigl|\mathrm{P}^b_\Delta f(x) - \mathrm{P}^b_\Delta f(y)\bigr| &\leqslant \mathbf{E}\bigl[\bigl|f(x+W_\Delta)\bigr|\bigl|L_x - L_y\bigr|\bigr] + \mathbf{E}\bigl[L_y\bigl|f(x+W_\Delta) - f(y+W_\Delta)\bigr|\bigr] \\ &:= S_1 + S_2.\end{aligned}$$

We will bound the two terms S_1 and S_2 separately.

There exists $\tilde{q} > 1$ such that

$$K_1\Delta < \frac{1}{2\sqrt{d\tilde{q}}}. \tag{A.2}$$

Fix such $\tilde{q}$ and let q be the root of the equation

$$\tilde{q} = 2\biggl(q^2 - \frac{q}{2}\biggr) \tag{A.3}$$

that is larger than 1. Next, set $r = q/(q-1)$. Note that $r > 1$ and $1/r + 1/q = 1$.

To bound S_1, we apply the elementary inequality $|\mathrm{e}^a - \mathrm{e}^b| \leqslant |a-b||\mathrm{e}^a + \mathrm{e}^b|$ for $a, b \in \mathbb{R}$ and Hölder's inequality with exponents r and q defined before to obtain

$$\begin{aligned}S_1 &\leqslant \|f\|_\infty \mathbf{E}\bigl[|L_x - L_y|\bigr] \leqslant \|f\|_\infty \mathbf{E}\bigl[|l_x - l_y||L_x + L_y|\bigr] \\ &\leqslant \|f\|_\infty \bigl\{\mathbf{E}\bigl[|l_x - l_y|^r\bigr]\bigr\}^{1/r}\bigl\{\mathbf{E}\bigl[|L_x + L_y|^q\bigr]\bigr\}^{1/q}.\end{aligned}$$

In order to bound S_1, we hence need to bound the last two factors on the right-hand side of the last inequality. We first treat the first of these two. The c_r-inequality gives that it is enough to bound the terms

$$\mathbf{E}\Biggl[\Biggl|\int_0^\Delta \bigl(b_i(x+W_s) - b_i(y+W_s)\bigr)\,\mathrm{d}W_{i,s}\Biggr|^r\Biggr], \qquad \mathbf{E}\Biggl[\Biggl|\int_0^\Delta \bigl(b_i^2(x+W_s) - b_i^2(y+W_s)\bigr)\,\mathrm{d}s\Biggr|^r\Biggr]$$

for $i = 1, \ldots, d$. Since the arguments are the same for any i, we henceforth fix a particular i. By the Burkholder–Davis–Gundy inequality (see Theorem 3.28 in [13, p. 166]),

$$\mathbf{E}\Biggl[\Biggl|\int_0^\Delta \bigl(b_i(x+W_s) - b_i(y+W_s)\bigr)\,\mathrm{d}W_{i,s}\Biggr|^r\Biggr] \leqslant C_r \mathbf{E}\Biggl[\Biggl|\int_0^\Delta \bigl(b_i(x+W_s) - b_i(y+W_s)\bigr)^2\,\mathrm{d}s\Biggr|^{r/2}\Biggr],$$

where $C_r > 0$ is a universal constant independent of b. For a fixed constant $R > 0$ and the set $F' = \{u + v\colon u \in F,\ \|v\| \leqslant R\}$, by the Cauchy–Schwarz inequality the expectation on the right-hand side of the above display can be bounded as follows:

$$\begin{aligned}&\mathbf{E}\Biggl[\Biggl|\int_0^\Delta \bigl(b_i(x+W_s) - b_i(y+W_s)\bigr)^2\,\mathrm{d}s\Biggr|^{r/2}\mathbf{1}_{[\sup_{s\leqslant\Delta}\|W_s\|\leqslant R]}\Biggr] \\ &\quad + \mathbf{E}\Biggl[\Biggl|\int_0^\Delta \bigl(b_i(x+W_s) - b_i(y+W_s)\bigr)^2\,\mathrm{d}s\Biggr|^{r/2}\mathbf{1}_{[\sup_{s\leqslant\Delta}\|W_s\|>R]}\Biggr]\end{aligned}$$

$$
\begin{aligned}
&\leqslant \Delta^{r/2} \sup_{\substack{u,v\in F' \\ \|u-v\|\leqslant\|x-y\|}} \bigl|b_i(u)-b_i(v)\bigr|^r \\
&\quad + \left\{\mathbf{E}\left[\left|\int\limits_0^\Delta \bigl(b_i(x+W_s)-b_i(y+W_s)\bigr)^2\,\mathrm{d}s\right|^r\right]\right\}^{1/2} \left\{\mathbf{P}\Bigl(\sup_{s\leqslant\Delta}\|W_s\|>R\Bigr)\right\}^{1/2}.
\end{aligned}
$$

Since b has partial derivatives bounded in absolute value by K_2, the first term on the right-hand side of the above display can be made arbitrarily small by choosing δ small enough and $\|x-y\|\leqslant\delta$. Furthermore, the term

$$
\left\{\mathbf{P}\Bigl(\sup_{s\leqslant\Delta}\|W_s\|>R\Bigr)\right\}^{1/2}
$$

can be made arbitrarily small by choosing R large enough. Next, by Hölder's inequality,

$$
\begin{aligned}
&\left\{\mathbf{E}\left[\left|\int\limits_0^\Delta \bigl(b_i(x+W_s)-b_i(y+W_s)\bigr)^2\,\mathrm{d}s\right|^r\right]\right\}^{1/2} \\
&\qquad \leqslant \Delta^{r/(2q)}\left\{\mathbf{E}\left[\int\limits_0^\Delta \bigl|b_i(x+W_s)-b_i(x+W_s)\bigr|^{2r}\,\mathrm{d}s\right]\right\}^{1/2},
\end{aligned}
$$

and a lengthy but easy computation employing the Fubini theorem, the linear growth condition on b, and the c_{2r}-inequality shows that the term on the right-hand side is bounded by a constant independent of b. Consequently, the term

$$
\mathbf{E}\left[\left|\int\limits_0^\Delta \bigl(b_i(x+W_s)-b_i(y+W_s)\bigr)\,\mathrm{d}W_{i,s}\right|^r\right]
$$

can be made arbitrarily small, once δ is chosen small enough and $\|x-y\|\leqslant\delta$. The term

$$
\mathbf{E}\left[\left|\int\limits_0^\Delta \bigl(b_i^2(x+W_s)-b_i^2(y+W_s)\bigr)\,\mathrm{d}s\right|^r\right]
$$

can be shown to be bounded uniformly in $b\in\mathcal{X}(K_1,K_2)$ by employing similar techniques: by the Cauchy–Schwarz inequality (twice),

$$
\begin{aligned}
\mathbf{E}\left[\left|\int\limits_0^\Delta \bigl(b_i^2(x+W_s)-b_i^2(y+W_s)\bigr)\,\mathrm{d}s\right|^r\right] &\leqslant \left\{\mathbf{E}\left[\left|\int\limits_0^\Delta \bigl(b_i(x+W_s)-b_i(y+W_s)\bigr)^2\,\mathrm{d}s\right|^r\right]\right\}^{1/2} \\
&\quad \times\left\{\mathbf{E}\left[\left|\int\limits_0^\Delta \bigl(b_i(x+W_s)+b_i(y+W_s)\bigr)^2\,\mathrm{d}s\right|^r\right]\right\}^{1/2}.
\end{aligned}
$$

The first factor on the right-hand side can be made arbitrarily small uniformly in $b\in\mathcal{X}(K_1,K_2)$ by taking δ small (see above), whereas the second factor remains bounded uniformly in $b\in\mathcal{X}(K_1,K_2)$ and $x,y,\in F$ by the linear growth condition on b.

Next, we will bound the right-hand side of the inequality

$$\mathbf{E}\big[|L_x + L_y|^q\big] \leqslant c_q \mathbf{E}\big[L_x^q\big] + c_q \mathbf{E}\big[L_y^q\big].$$

Since obviously both terms on the right-hand side can be bounded in exactly the same manner, we will only give an argument for the first one. By the Cauchy–Schwarz inequality applied to the random variables

$$\exp\Biggl(\biggl(q^2 - \frac{q}{2}\biggr)\sum_{i=1}^{d}\int_0^\Delta b_i^2(x+W_s)\,\mathrm{d}s\Biggr),$$

$$\exp\Biggl(\sum_{i=1}^{d}\int_0^\Delta q b_i(x+W_s)\,\mathrm{d}W_{i,s} - \sum_{i=1}^{d}\int_0^\Delta q^2 b_i^2(x+W_s)\,\mathrm{d}s\Biggr)$$

we have

$$\mathbf{E}\big[L_x^q\big] \leqslant \Biggl\{\mathbf{E}\Biggl[\exp\Biggl(2\biggl(q^2 - \frac{q}{2}\biggr)\sum_{i=1}^{d}\int_0^\Delta b_i^2(x+W_s)\,\mathrm{d}s\Biggr)\Biggr]\Biggr\}^{1/2}. \tag{A.4}$$

Here we used the fact that

$$\mathbf{E}\Biggl[\exp\Biggl(\sum_{i=1}^{d}\int_0^\Delta 2q b_i(x+W_s)\,\mathrm{d}W_{i,s} - \frac{1}{2}\sum_{i=1}^{d}\int_0^\Delta 4q^2 b_i^2(x+W_s)\,\mathrm{d}s\Biggr)\Biggr] = 1,$$

since the process under the expectation sign is a martingale and has the expectation equal to one (this is due to the linear growth condition and Corollary 5.16 in [13, p. 200]).

Hence, it remains to bound the right-hand side of (A.4), which we denote by S_5. By the linear growth condition we have

$$S_5^2 \leqslant \exp\bigl(2d\tilde{q}K_1^2\Delta\bigl(1+\|x\|\bigr)^2\bigr)\mathbf{E}\Biggl[\exp\Biggl(2d\tilde{q}K_1^2\int_0^\Delta \|W_s\|^2\,\mathrm{d}s\Biggr)\Biggr].$$

By Doob's maximal inequality for submartingales (see Theorem 3.8(iv) in [13, pp. 13–14]) and the independence of the scalar Brownian motions W_i,

$$\mathbf{E}\Biggl[\exp\Biggl(2d\tilde{q}K_1^2\int_0^\Delta \|W_s\|^2\,\mathrm{d}s\Biggr)\Biggr] \leqslant 4\prod_{i=1}^{d}\mathbf{E}\bigl[\exp\bigl(2d\tilde{q}K_1^2\Delta W_{i,\Delta}^2\bigr)\bigr] < \infty.$$

Here, in the last inequality, we used (A.2). The conclusion is that the term S_1 can be made arbitrarily small by taking δ small and $\|x-y\| \leqslant \delta$. The proof is now completed as follows: by Hölder's inequality,

$$S_2 \leqslant \bigl\{\mathbf{E}\bigl[L_y^q\bigr]\bigr\}^{1/q}\bigl\{\mathbf{E}\bigl[\bigl|f(x+W_\Delta) - f(y+W_\Delta)\bigr|^r\bigr]\bigr\}^{1/r}.$$

The first factor on the right-hand side can be bounded as before uniformly in $b \in \mathcal{X}(K_1, K_2)$. The second factor can be made arbitrarily small as soon as $\|x-y\| \leqslant \delta$ for small enough δ: for a constant $R > 0$,

$$\mathbf{E}\bigl[\bigl|f(x+W_\Delta) - f(y+W_\Delta)\bigr|^r\bigr]$$

$$
\begin{aligned}
&= \mathbf{E}\big[\big|f(x+W_\Delta)-f(y+W_\Delta)\big|^r \mathbf{1}_{[\|W_\Delta\|>R]}\big] \\
&\quad + \mathbf{E}\big[\big|f(x+W_\Delta)-f(y+W_\Delta)\big|^r \mathbf{1}_{[\|W_\Delta\|\leqslant R]}\big] \\
&\leqslant \big(2\|f\|_\infty\big)^r \mathbf{P}\big(\|W_\Delta\|>R\big) + \mathbf{E}\big[\big|f(x+W_\Delta)-f(y+W_\Delta)\big|^r \mathbf{1}_{[\|W_\Delta\|\leqslant R]}\big].
\end{aligned}
$$

The first term on the right-hand side of the last inequality can be made arbitrarily small by selecting R large enough. Upon fixing R, so can be the second one by taking $\|x-y\| \leqslant \delta$ for small enough $\delta > 0$. Combination of all the above intermediate results entails the statement of the lemma. □

Lemma A.4. *Let*

$$
\mathrm{KL}(b_0,b) = \int\limits_{\mathbb{R}^d}\int\limits_{\mathbb{R}^d} \pi_{b_0}(x)p_{b_0}(\Delta,x,y)\log\frac{p_{b_0}(\Delta,x,y)}{p_b(\Delta,x,y)}\,\mathrm{d}x\,\mathrm{d}y,
$$

and assume that the weak solution to (1.1) *is initialized at* μ_b. *Then, for the prior* Π *satisfying property* (2.1), *we have the inequality*

$$
\Pi\big(b \in \mathcal{X}(K_1,K_2)\colon\ \mathrm{KL}(b_0,b) < \varepsilon\big) > 0 \quad \forall\varepsilon > 0. \tag{A.5}
$$

Proof. The proof is an obvious modification of the proof of Lemma 5.1 in [16]. The only additional fact we need to verify is that the Kullback–Leibler divergence $\mathrm{K}(\mu_b,\mu_{\tilde{b}})$ is finite for any $b,\tilde{b} \in \mathcal{X}(K_1,K_2)$. This, however, follows from Proposition 1.1 in [9]. □

Lemma A.5. *Suppose that the prior* Π *on* $\mathcal{X}(K_1,K_2)$ *has property* (A.5) *and assume that the weak solution to* (1.1) *is initialized at* μ_b. *If, for a sequence* C_n *of measurable subsets of* $\mathcal{X}(K_1,K_2)$, *there exists a constant* $c > 0$ *such that*

$$
\mathrm{e}^{nc}\int\limits_{C_n} L_n(b)\,\Pi(\mathrm{d}b) \to 0 \quad P_{b_0}\text{-a.s.},
$$

then

$$
\Pi(C_n|X_0,\dots,X_{\Delta n}) \to 0 \quad P_{b_0}\text{-a.s.}
$$

as $n \to \infty$.

Proof. The proof is an easy generalization of the proof of Lemma 5.2 in [16]. □

References

1. L. Birgé, Robust testing for independent nonidentically distributed variables and Markov chains, in *Specifying Statistical Models (Louvain-la-Neuve, 1981)*, Lect. Notes Stat., Vol. 16, Springer, New York, 1983, pp. 134–162.

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

3. A. Dalalyan and M. Reiß, Asymptotic statistical equivalence for ergodic diffusions: The multidimensional case, *Probab. Theory Relat. Fields*, **137**:25–47, 2007.

4. P. Diaconis and D. Freedman, On the consistency of Bayes estimates. With a discussion and a rejoinder by the authors, *Ann. Stat.*, **14**:1–67, 1986.

5. R.M. Dudley, *Real Analysis and Probability*, Cambridge Stud. Adv. Math., Vol. 74, Cambridge Univ. Press, Cambridge, 2002. Revised reprint of the 1989 original.

6. S. Ghosal, J.K. Ghosh, and R.V. Ramamoorthi, Non-informative priors via sieves and packing numbers, in S. Panchapakesan and N. Balakrishnan (Eds.), *Advances in Statistical Decision Theory and Applications*, Birkhäuser Boston, Boston, MA, 1997, pp. 119–132.

7. S. Ghosal and Y. Tang, Bayesian consistency for Markov processes, *Sankhyā, Ser. B*, **68**:227–239, 2006.

8. S. Ghosal and A.W. van der Vaart, Convergence rates of posterior distributions for non-i.i.d. observations, *Ann. Stat.*, **35**:192–223, 2007.

9. E. Gobet, LAN property for ergodic diffusions with discrete observations, *Ann. Inst. Henri Poincaré, Probab. Stat.*, **38**:711–737, 2002.

10. E. Gobet, M. Hoffmann, and M. Reiß, Nonparametric estimation of scalar diffusions based on low frequency data, *Ann. Stat.*, **32**:2223–2253, 2004.

11. S. Gugushvili and P. Spreij, Non-parametric Bayesian drift estimation for stochastic differential equations, 2013, arXiv:1206.4981 [math.ST].

12. J. Jacod, Non-parametric kernel estimation of the coefficient of a diffusion, *Scand. J. Stat.*, **27**:83–96, 2000.

13. I. Karatzas and S.E. Shreve, *Brownian Motion and Stochastic Calculus*, Grad. Texts Math., Vol. 113, Springer, New York, 1988.

14. F. van der Meulen, M. Schauer, and H. van Zanten, Reversible jump MCMC for nonparametric drift estimation for diffusion processes, *Comput. Stat. Data Anal.*, **71**:615–632, 2014.

15. F.H. van der Meulen, A.W. van der Vaart, and J.H. van Zanten, Convergence rates of posterior distributions for Brownian semimartingale models, *Bernoulli*, **12**:863–888, 2006.

16. F. van der Meulen and H. van Zanten, Consistent nonparametric Bayesian estimation for discretely observed scalar diffusions, *Bernoulli*, **19**:44–63, 2013.

17. L. Panzar and H. van Zanten, Nonparametric Bayesian inference for ergodic diffusions, *J. Stat. Plann. Infer.*, **139**:4193–4199, 2009.

18. O. Papaspiliopoulos, Y. Pokern, G.O. Roberts, and A.M. Stuart, Nonparametric estimation of diffusions: A differential equations approach, *Biometrika*, **99**:511–531, 2012.

19. Y. Pokern, A.M. Stuart, and J.H. van Zanten, Posterior consistency via precision operators for nonparametric drift estimation in SDEs, *Stoch. Process. Appl.*, **123**:603–628, 2013.

20. D. Pollard, *A User's Guide to Measure Theoretic Probability*, Cambridge Ser. Stat. Probab. Math., Vol. 8, Cambridge Univ. Press, Cambridge, 2002.

21. E. Schmisser, Penalized nonparametric drift estimation for a multidimensional diffusion process, *Statistics*, **47**:61–84, 2013.

22. Y. Tang and S. Ghosal, Posterior consistency of Dirichlet mixtures for estimating a transition density, *J. Stat. Plann. Infer.*, **137**:1711–1726, 2007.

23. S. Walker, On sufficient conditions for Bayesian consistency, *Biometrika*, **90**:482–488, 2003.

24. S. Walker, New approaches to Bayesian consistency, *Ann. Stat.*, **32**:2028–2043, 2004.

25. L. Wasserman, Asymptotic properties of nonparametric Bayesian procedures, in *Practical Nonparametric and Semiparametric Bayesian Statistics*, Lect. Notes Stat., Vol. 133, Springer, New York, 1998, pp. 293–304.

26. H. van Zanten, Nonparametric Bayesian methods for one-dimensional diffusion models, *Math. Biosci.*, **243**:215–222, 2013.