Submitted to the Annals of Statistics

NONPARAMETRIC BAYESIAN INFERENCE FOR REVERSIBLE MULTI-DIMENSIONAL DIFFUSIONS

BY MATTEO GIORDANO[1] AND KOLYAN RAY[2]

[1]*Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, mg846@cam.ac.uk*

[2]*Department of Mathematics, Imperial College London, kolyan.ray@imperial.ac.uk*

We study nonparametric Bayesian models for reversible multidimensional diffusions with periodic drift. For continuous observation paths, reversibility is exploited to prove a general posterior contraction rate theorem for the drift gradient vector field under approximation-theoretic conditions on the induced prior for the invariant measure. The general theorem is applied to Gaussian priors and p-exponential priors, which are shown to converge to the truth at the optimal nonparametric rate over Sobolev smoothness classes in any dimension.

1. Introduction. Consider observing a continuous trajectory $X^T = (X_t = (X_t^1, \dots, X_t^d) : 0 \le t \le T)$ of the multi-dimensional Markov diffusion process given by the solution to the stochastic differential equation (SDE)

$$dX_t = b(X_t)dt + dW_t, \qquad X_0 = x_0 \in \mathbb{R}^d, \quad t \ge 0, \tag{1}$$

where $W_t = (W_t^1, \dots, W_t^d)$ is a standard Brownian motion on $\mathbb{R}^d$ and $b = (b_1, \dots, b_d)$ is a Lipschitz vector field. We are interested in nonparametric Bayesian inference on the drift term b in the practically important case when the diffusion process is *time-reversible*. By a result of Kolmogorov, this is equivalent to the drift b equalling the gradient vector field ∇B of a potential function $B : \mathbb{R}^d \to \mathbb{R}$ (e.g. [7], p. 46).

The SDE (1) with $b = \nabla B$ is the Brownian dynamics model in physics, describing the trajectory of a particle diffusing in a potential energy field that exerts a force directed towards its local extrema. One is often interested in inference on the potential B (or drift ∇B), which carries important physical information, based on an observed trajectory of the particle. For example, such dynamics arise in the Smoluchowski-Kramers approximation to the Langevin equation for the motion of a chemically bound particle [40, 66], in which case B describes the chemical bonding forces. Other applications of the reversible diffusion model (1) in physics and chemistry include vacancy diffusion and Lennard-Jones clusters [54, 55] and chemical reaction equations [12]. Once fitted, such diffusive models can also be employed as computationally affordable emulators of the physical process [26].

A Bayesian who wants to model reversible diffusion dynamics must do so explicitly via the prior, namely by constructing one that draws gradient vector fields for b. The natural approach is then to directly place a prior on the potential B rather than on b, which is the approach we pursue here. Moreover, B typically has a strong physical meaning and estimating it is often the primary inferential goal, in which case explicitly modelling the potential provides interpretable inference. Our aim is to provide theoretical convergence guarantees as the time horizon $T \to \infty$ for this Bayesian approach, which arises directly from physical considerations in the modelling step.

MSC2020 subject classifications: Primary 62G20; secondary 62F15, 60J60.

Keywords and phrases: Bayesian nonparametrics, multi-dimensional diffusions, reversibility, Gaussian processes, Laplace prior.

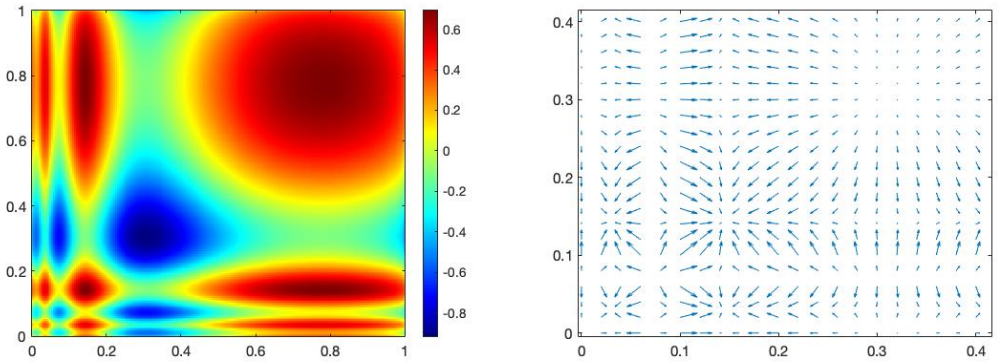


FIG 1. *An example of a spatially heterogeneous potential B (left) and the corresponding gradient vector field ∇B (right). Note the axis scales are different for clarity.*

The theoretical performance of nonparametric Bayesian procedures for drift estimation is well-studied in the one-dimensional case $d = 1$ [72, 56, 36, 76, 49, 1, 75]. However, much less is known in the general multi-dimensional setting $d \geq 2$. In the continuous observation model, Dalalyan and Reiß [15] established pointwise convergence rates and Strauch [67, 68, 69] obtained adaptive rates, both using multivariate kernel-type estimators. Schmisser [65] established adaptive L^2 convergence rates for certain penalised least-squares estimators with discrete observations. Nickl and Ray [48] obtained L^2 and L^∞ posterior contraction rates, as well as Bernstein-von Mises results, for certain *non-reversible* drift vector fields - more discussion can be found below.

In this paper, we obtain contraction rates for the posterior distribution of B about the true potential B_0 in the model (1) as $T \to \infty$. We prove a general theorem for diffusions governed by gradient vector fields based on the classical testing approach of Bayesian nonparametrics [28, 74], and firstly apply it to Gaussian priors for B, obtaining optimal rates over Sobolev smoothness classes in any dimension. Gaussian priors are widely used in diffusion models [56, 61, 36, 76, 75, 9, 48] and are a canonical choice, not least for computational reasons [51, 61, 71]. Our results thus provide statistical convergence guarantees for these widely used priors. As a consequence, we also deduce convergence rates for the corresponding maximum a-posteriori (MAP) estimates, which can be viewed as penalized maximum likelihood estimators. Apart from the results in [48], these are the first multi-dimensional Bayesian contraction results for diffusions, and the first for potential-modelling priors.

In applications, the potential B is often spatially inhomogeneous for physical reasons, for instance being spiky in some regions and flat or smooth in others (e.g. molecular communication [27]), see Figure 1 for an illustrative example. Gaussian priors are known to be unsuited to modelling such inhomogeneities (see e.g. [4, 5, 34]), which has motivated the use of heavier tailed priors, especially Besov space priors in the inverse problems [59, 3, 19, 4, 5] and medical imaging [62, 77] communities. To address such physically motivated situations, we also consider modelling the potential using heavier-tailed p-exponential priors [4], for which we establish optimal nonparametric rates. Such priors have been employed to recover spatially inhomogeneous functions, such as those with a 'blocky' structure with sudden changes from one block to another. They have a number of attractive properties including edge preservation, discretization invariance [42, 41], promoting sparse solution representations, while also maintaining a log-concave structure that aids posterior sampling. Contraction rates for such priors have recently been obtained in direct linear models [4], and we present here a first extension to a nonlinear diffusion setting.

The multi-dimensional case $d \geq 2$ is intrinsically more challenging than the one-dimensional case. The testing approach has been used for model (1) first by van der Meulen et al. [72] to obtain contraction rates in the 'natural distance' induced by the statistical experiment, which is a 'random Hellinger semimetric' depending on the observation path $(X_t : 0 \leq t \leq T)$. In dimension $d = 1$, the theory of diffusion local times can then be used to relate this random path-dependent metric to the L^2-distance [72, 56, 76], but when $d > 1$ such local time arguments are not available. In the multi-dimensional setting, Nickl and Ray [48] relate this random Hellinger metric to the L^2-distance for specific truncated Gaussian series product priors on b. They exploit concentration properties of the high-dimensional random matrices induced by the Hellinger semimetric on finite-dimensional projection spaces to relate this problem to a random design type regression problem. However, this approach crucially uses that the priors for each coordinate of $b = (b_1, \dots, b_d)$ are supported on the same finite-dimensional projection spaces, which is typically only the case if $b_1, \dots, b_d$ have *independent* priors. Since product priors for b draw gradient vector fields $b = \nabla B$ with probability zero, they are inherently unable to model reversibility.

Aside from requiring different priors, modelling the potential B introduces fundamentally new features to the inference problem at hand. Whereas one can relate the non-gradient vector field case to a direct linear regression type model [48], modelling B is equivalent to modelling the invariant measure (see (6)) and leads to a genuinely nonlinear regression problem. We must thus employ a completely different approach to [48] here, using instead tools from the statistical theory of nonlinear inverse problems [47, 2, 32, 50, 46] - our work can thus also be viewed as a contribution to the Bayesian inverse problems literature.

Instead of Hellinger testing theory, we develop concentration inequalities for preliminary estimators to directly construct suitable plug-in tests following ideas in [30], see also [59, 49, 1, 2, 45]. In the present setting, the invariant measure μ_b of the diffusion describes the probabilities

$$\mu_b(A) =^{a.s.} \lim_{T\to\infty} \frac{1}{T} \int_0^T 1_A(X_t)dt, \tag{2}$$

corresponding to the average asymptotic time spent by the process (X_t) in a given measurable subset A of the state space. In the reversible case, we can exploit the one-to-one correspondence $\mu_b \propto e^{2B}$ between the potential B and the invariant measure μ_b, to construct estimators for B based on estimators for μ_b. We then combine elliptic PDE and martingale techniques with concentration of measure arguments to obtain exponential inequalities for such estimators, and hence bounds for the type-II errors of suitable tests.

We study posterior asymptotics in the ergodic setting, where one can gain information about the average behaviour of the diffusion dynamics from a sufficiently long particle trajectory. This requires strong enough recurrence that the diffusion is largely confined to a bounded region, which matches the physical intuition in many applied situations. To ensure such recurrence and mixing of the diffusion, we follow [51, 56, 76, 1, 26, 75, 48] in restricting to the *periodic* setting, and thus periodic potentials B. Under this simplification, a potential B still implies the corresponding (periodised) diffusion is reversible ([26], Proposition 2) and so our results maintain the key modelling link between reversibility and potential functions. Periodicity simplifies certain technical arguments, in particular the elliptic PDE techniques involved in studying the mapping properties of the generator of the underlying semigroup (see e.g. Chapter II.3 in [10]), thus allowing a cleaner exposition of the main statistical ideas. The underlying PDE techniques extend conceptually to the non-periodic setting under certain conditions, albeit at the expense of significant additional technicalities that are beyond the scope of the present paper, see Section 3.2 for discussion.

As well as deriving theoretical results, we also discuss numerical implementation in the continuous observation model considered here. Posterior sampling using simulation techniques is well-studied for 'real-world' discrete data, including for some priors we consider here, which we highlight when discussing concrete prior choices. For recent work on this active research topic, see for example [11, 51, 61, 71, 63, 9, 26] and references therein.

2. Main results.

2.1. *Basic notation and definitions.* Let $\mathbb{T}^d$ be the d-dimensional torus, isomorphic to $[0,1]^d$ with the opposite points on the cube identified. We denote by $L^p(\mathbb{T}^d)$ the usual Lebesgue spaces on $\mathbb{T}^d$ equipped with norm $\|\cdot\|_p$, and by $\langle\cdot,\cdot\rangle_2$ the inner product on $L^2(\mathbb{T}^d)$. We further define the subspaces

$$\dot{L}^2(\mathbb{T}^d) := \left\{ f \in L^2(\mathbb{T}^d) : \int_{\mathbb{T}^d} f dx = 0 \right\}, \qquad \dot{L}^2_\mu(\mathbb{T}^d) := \left\{ f \in L^2(\mathbb{T}^d) : \int_{\mathbb{T}^d} f d\mu = 0 \right\},$$

where μ is a probability measure on $\mathbb{T}^d$.

Let $C(\mathbb{T}^d)$ be the space of continuous functions on $\mathbb{T}^d$, equipped with the supremum norm $\|\cdot\|_\infty$. For $s>0$, denote by $C^s(\mathbb{T}^d)$ the usual Hölder space of $\lfloor s \rfloor$-times continuously differentiable functions on $\mathbb{T}^d$ whose $\lfloor s \rfloor^{\text{th}}$-derivative is $(s-\lfloor s \rfloor)$-Hölder continuous. We let $H^s(\mathbb{T}^d)$, $s \in \mathbb{R}$, denote the usual L^2-Sobolev spaces on $\mathbb{T}^d$, defined by duality when $s<0$. We further define the Sobolev norms $\|f\|_{W^{1,q}} = \|f\|_q + \sum_{i=1}^d \|\partial_{x_i} f\|_q$, and note that $\|\cdot\|_{W^{1,2}}$ is equivalent to $\|\cdot\|_{H^1}$.

Let $\{\Phi_{lr} : l \in \{-1,0\} \cup \mathbb{N},\ r = 0, \dots, \max(2^{ld}-1, 0)\}$ be an orthonormal tensor product wavelet basis of $L^2(\mathbb{T}^d)$, obtained from a periodised Daubechies wavelet basis of $L^2(\mathbb{T})$, which we take to be S-regular for $S \in \mathbb{N}$ large enough; see Section 4.3. in [31] for details. For $J \in \mathbb{N}$, define the finite-dimensional approximation space

$$V_J := \text{span}\{\Phi_{lr} : l \le J,\ r = 0, \dots, \max(2^{ld-1}-1, 0)\} \tag{3}$$

and let $P_J : L^2(\mathbb{T}^d) \to V_J$ be the associated L^2-projection operator. Note that V_J has dimension $v_J := \dim(V_J) = O(2^{Jd})$ as $J \to \infty$. For $0 \le t \le S$, $1 \le p,q \le \infty$, define the Besov spaces via their wavelet characterisation:

$$B^t_{pq}(\mathbb{T}^d) = \left\{ f \in L^p(\mathbb{T}^d) :\ \|f\|^q_{B^t_{pq}} := \sum_l 2^{ql\left(t+\frac{d}{2}-\frac{d}{p}\right)} \left(\sum_r |\langle f, \Phi_{lr}\rangle_2|^p \right)^{\frac{q}{p}} < \infty \right\},$$

replacing the ℓ_p or ℓ_q-norm above with ℓ_∞ if $p=\infty$ or $q=\infty$, respectively. Recall that $H^t(\mathbb{T}^d) = B^t_{22}(\mathbb{T}^d)$ and the continuous embedding $C^s(\mathbb{T}^d) \subseteq B^s_{\infty\infty}(\mathbb{T}^d)$ for $s \ge 0$, see Chapter 3 in [64].

When no confusion may arise, we suppress the dependence of the function spaces on the underlying domain, writing for example B^t_{pq} instead of $B^t_{pq}(\mathbb{T}^d)$. We also employ the same function space notation for vector fields $f = (f_1, \dots, f_d)$. For instance, $f \in H^s = (H^s)^{\otimes d}$ will mean each $f_i \in H^s$ and the norm on H^s is $\|f\|_{H^s} = \sum_{i=1}^d \|f_i\|_{H^s}$. Similarly, $\|\nabla g\|_p = \sum_{i=1}^d \|\partial_{x_i} g\|_p$.

We write $\|.\|$ for the Euclidean norm in $\mathbb{R}^d$, $\lesssim$, $\gtrsim$ and $\simeq$ to denote one- or two-sided inequalities up to multiplicative constants that may either be universal or 'fixed' in the context where the symbols appear. We also write $a_+ = \max(a,0)$ and $a \vee b = \max(a,b)$ for real numbers a, b. The ε-covering number of a set Θ for a semimetric d, denoted $N(\Theta, d, \varepsilon)$, is the minimal number of d-balls of radius ε needed to cover Θ.

2.2. *Diffusions with periodic drift and Bayesian inference.* Consider the SDE (1) with drift $b = \nabla B$, for a twice-continuously differentiable and one-periodic potential $B : \mathbb{R}^d \to \mathbb{R}$, that is $B(x+m) = B(x)$ for all $m \in \mathbb{Z}^d$,

$$dX_t = \nabla B(X_t)dt + dW_t, \qquad X_0 = x_0 \in \mathbb{R}^d, \quad t \geq 0. \tag{4}$$

There exists a d-dimensional strong pathwise solution $X = (X_t = (X_t^1, \dots, X_t^d) : \ t \geq 0)$ with cylindrically defined law on the path space $C([0,\infty); \mathbb{R}^d)$; see, e.g., Chapters 24 and 39 in [8]. For $T > 0$, let $X^T := (X_t : \ 0 \leq t \leq T)$ and denote by P_B the law of X^T on $C([0,T], \mathbb{R}^d)$. We omit the dependence on the initial condition $X_0 = x_0$ since it plays no role in our results.

By periodicity, we can consider B as a function on $\mathbb{T}^d$. In model (4), the law P_B depends on B only through $b = \nabla B$ (see (7) below), which is thus only identifiable up to an additive constant. We therefore choose to work with the unique equivalence class given by $\int_{\mathbb{T}^d} B(x)dx = 0$, i.e. $B \in \dot{L}^2(\mathbb{T}^d)$. Our goal is to estimate the drift $b = \nabla B : \mathbb{R}^d \to \mathbb{R}^d$ from an observed trajectory $X^T \sim P_B$. We will sometimes write P_b when a technical result also applies to possibly non-gradient vector field drifts, but this will be clarified in each instance.

The periodic model effectively restricts the diffusion to the bounded state space $\mathbb{T}^d$. More precisely, while the diffusion defined in (4) takes values on all of $\mathbb{R}^d$, its values (X_t) modulo $\mathbb{Z}^d$ contain all the relevant statistical information about ∇B (note that (X_t) will not be globally recurrent on $\mathbb{R}^d$). This allows us to define an *invariant measure* on $\mathbb{T}^d$, since it holds that (arguing as in the proof of Lemma 6 in [48])

$$\frac{1}{T}\int_0^T \varphi(X_s)ds \to^{P_B} \int_{\mathbb{T}^d} \varphi d\mu_b \qquad \text{as } T \to \infty, \quad \forall \varphi \in C(\mathbb{T}^d),$$

where μ_b is a uniquely defined probability measure on $\mathbb{T}^d$ and we identify φ with its periodic extension to $\mathbb{R}^d$ on the left-hand side. The measure μ_b thus inherits the usual probabilistic interpretation as the limiting ergodic average in (2).

Recall that the generator $L : H^2(\mathbb{T}^d) \to L^2(\mathbb{T}^d)$ of the possibly non-reversible diffusion from (1) is

$$L_b = \frac{1}{2}\Delta + b.\nabla = \frac{1}{2}\sum_{i=1}^d \frac{\partial^2}{\partial x_i^2} + \sum_{i=1}^d b_i(\cdot)\frac{\partial}{\partial x_i}. \tag{5}$$

If $b = \nabla B$ for some potential $B \in C^2(\mathbb{T}^d)$, the corresponding invariant measure of the diffusion has probability density function

$$\mu_B(x) = \frac{e^{2B(x)}}{\int_{\mathbb{T}^d} e^{2B(y)}dy}, \qquad x \in \mathbb{T}^d, \tag{6}$$

see p. 45-47 in [7]. In particular, if we have a potential B, then we can recover b from μ_B via $b = \nabla B = \frac{1}{2}\nabla \log \mu_B$, a connection we exploit in our proofs. In a slight abuse of notation, we write either μ_b or μ_B depending on the context when $b = \nabla B$, and we use μ_b for both the probability measure and its density function.

The log-likelihood for $B \in C^2(\mathbb{T}^d)$ for our observation model is given by Girsanov's theorem (e.g., Section 17.7 in [8]):

$$\ell_T(B) := \log \frac{dP_B}{dP_0}(X^T) = -\frac{1}{2}\int_0^T \|\nabla B(X_t)\|^2 dt + \int_0^T \nabla B(X_t).dX_t, \tag{7}$$

where P_0 is the law of a d-dimensional Brownian motion $(W_t : 0 \leq t \leq T)$. We consider a Bayesian approach to the problem, assigning a (possibly T-dependent) prior $\Pi = \Pi_T$ to B,

which for identifiability we assume is supported on $\dot{L}^2(\mathbb{T}^d) \cap C^2(\mathbb{T}^d)$. The posterior $\Pi(\cdot|X^T)$ then takes the form

$$d\Pi(B|X^T) = \frac{e^{\ell_T(B)} d\Pi(B)}{\int e^{\ell_T(B')} d\Pi(B')}. \tag{8}$$

Note that this induces a prior and posterior for both $b = \nabla B$ and μ_B. In the following, we study the concentration of the posterior about the 'ground truth' gradient vector field $b_0 = \nabla B_0$, assuming that the observation $X^T \sim P_{B_0}$ is generated according to the SDE (4) with $B = B_0$.

2.3. *Gaussian process priors.*

2.3.1. *Contraction rates.* Gaussian priors are widely employed for diffusion models [51, 56, 61, 36, 76, 71, 75, 9, 48] and we provide here theoretical guarantees for such priors. We consider a class of Gaussian process priors constructed from a base Gaussian probability measure Π_W, which we assume satisfies the following condition. We refer, e.g., to Chapter 2 in [31] for definitions and terminology regarding the theory of Gaussian processes and measures.

CONDITION 2.1. For $s > (d-1) \vee (1/2)$, let $\Pi_W = \Pi_{W,T}$ be a centred Gaussian Borel probability measure on the Banach space $C(\mathbb{T}^d)$ that is supported on a separable (measurable) linear subspace of $C^{(d/2+\kappa)\vee 2}(\mathbb{T}^d) \cap \dot{L}^2(\mathbb{T}^d)$ for some $\kappa > 0$, and assume its reproducing kernel Hilbert space (RKHS) $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ is continuously embedded into the Sobolev space $H^{s+1}(\mathbb{T}^d)$.

The constant $\kappa > 0$ above can be taken arbitrarily small. Examples of Gaussian processes priors satisfying Condition 2.1 include the periodic Matérn process and high-dimensional Gaussian series expansions, see Examples 2.1-2.2 below.

To control the nonlinearity of the problem, we rescale the base Gaussian processes following ideas from the Bayesian inverse problem literature [47, 2, 32, 50, 46]. Given a random draw $W \sim \Pi_W$ satisfying Condition 2.1, consider the following rescaled function:

$$B(x) := \frac{W(x)}{T^{d/(4s+2d)}}, \qquad x \in \mathbb{T}^d, \tag{9}$$

whose law $\Pi = \Pi_T$ we take as the prior for the potential B. It follows that Π is a centred Gaussian probability measure on $C(\mathbb{T}^d)$, with the same support and RKHS as Π_W, but with rescaled RKHS norm $\|h\|_{\mathbb{H}_B} = T^{d/(4s+2d)}\|h\|_{\mathbb{H}}$.

The rescaling enforces additional regularisation in the induced posterior distribution for the invariant measure μ_B, implying in particular a bound for $\|\mu_B\|_{C^{(d/2+\kappa)\vee 2}}$, needed to control the nonlinear map $B \mapsto \mu_B$ given in (6). Such issues are commonly encountered in nonlinear inverse problems, where one often requires the posterior to place most of its mass on sets of bounded higher-order smoothness in order to use stability estimates. Note that such priors are special cases of the rescaled Gaussian process priors considered in several benchmark statistical settings in [73].

THEOREM 2.1. *Let $\Pi = \Pi_T$ be the rescaled Gaussian process prior for B in (9) with $W \sim \Pi_W$ satisfying Condition 2.1 for some $s > (d-1) \vee (1/2)$, some $\kappa > 0$, and RKHS $\mathbb{H}$. Suppose that $B_0 \in H^{s+1}(\mathbb{T}^d)$ and that there exists a sequence $B_{0,T} \in \mathbb{H}$ such that $\|B_0 - B_{0,T}\|_{C^1} = O(T^{-s/(2s+d)})$ and $\|B_{0,T}\|_{\mathbb{H}} = O(1)$ as $T \to \infty$. Then for $M > 0$ large enough, as $T \to \infty$,*

$$P_{B_0}\Pi(B : \|\nabla B - \nabla B_0\|_2 \geq M T^{-s/(2s+d)} | X^T) \to 0.$$

The resulting posterior thus contracts about the truth at the usual nonparametric rate for s-regular functions in any dimension d. Since the law P_B depends on the potential B only through ∇B in (4), it is natural to study recovery of the gradient vector field ∇B_0. Recall that we make the identifiability assumption that the prior, and hence posterior, for B is supported on $\dot{L}^2(\mathbb{T}^d)$ in Condition 2.1. For $B, B_0 \in \dot{L}^2(\mathbb{T}^d)$, the norm $\|\nabla B - \nabla B_0\|_2$ is in fact equivalent to the usual Sobolev norm $\|B - B_0\|_{H^1}$ by the Poincaré inequality (e.g. p. 290 in [25]).

Theorem 2.1 requires that the true B_0 be approximable by elements $B_{0,T}$ of the RKHS of W at a suitable rate, which reflects the notion of smoothness being modelled by the Gaussian process prior. One can think of this condition as 'B_0 is $(s+1)$-smooth' in both the Sobolev and prior sense. For instance, if the Gaussian prior already models Sobolev smooth functions (e.g. a Matérn process prior - see Example 2.1), then this poses no additional conditions.

The last theorem implies the same converge rate for the posterior mean.

COROLLARY 2.2. *Let $\hat{B}_T = E^{\Pi_T}[B|X^T]$ be the posterior mean. Under the conditions of Theorem 2.1, as $T \to \infty$,*

$$\|\nabla \hat{B}_T - \nabla B_0\|_2 = O_{P_{B_0}}(T^{-s/(2s+d)}).$$

Under natural regularity conditions, for example finite-dimensional priors, the maximum a-posteriori (MAP) estimator is well-defined as the minimizer of an objective function [18, 56]:

$$\hat{B}_T = \hat{B}_T(X^T) = \operatorname*{arg\,min}_{B \in \mathbb{H}} \left(-\ell_T(B) + \tfrac{1}{2} T^{d/(2s+d)} \|B\|_{\mathbb{H}}^2 \right).$$

Since the posterior is also Gaussian by Lemma 2.3 below, the MAP estimator, when it exists, equals the posterior mean $E^{\Pi_T}[B|X^T]$. In this case, Corollary 2.2 can be viewed as a convergence rate for a penalized maximum likelihood estimator with penalty equal to the squared prior RKHS-norm.

REMARK 2.1 (Minimax rates). Although lower bounds have not been formally proven in the exact periodic diffusion setting studied here, results for closely related non-periodic diffusion models suggest that the minimax rates are of the usual order $T^{-s/(2s+d)}$ in all the cases we consider. For reversible diffusions on $\mathbb{R}^d$, lower bounds for pointwise and L^2-loss follow from the local asymptotic equivalence of the diffusion model with a sequence of Gaussian shift experiments proved in [15], see also [67, 68]. Given this asymptotic equivalence, it is instructive to consider lower bounds in related statistical models. In particular, for L^p-loss, $1 \leq p \leq 2$, with s-smooth Sobolev truths as in Theorem 2.4 below, the minimax rate of estimation is $n^{-s/(2s+d)}$ (where n plays the role of T) in Gaussian white noise [44], density estimation [20] and nonparametric regression [20]. Since we obtain posterior contraction rate $T^{-s/(2s+d)}$ for all our results, this suggests all the priors we consider here are minimax optimal for estimation. We do not pursue the extension of such lower bounds to the current periodic setting, since periodicity is mainly a technical assumption to simplify the underlying PDE arguments, see Section 3.2.

2.3.2. *Examples of Gaussian priors.* We now provide concrete examples of Gaussian priors to which Theorem 2.1 applies. As discussed in Section 2.2, the potential B is only identified up to an additive constant, which we without loss of generality select via $\int_{\mathbb{T}^d} B(x)dx = 0$. For series expansions, one can enforce this by setting the coefficient of $e_0 \equiv 1$ (Fourier basis) or $\Phi_{-10} \equiv 1$ (wavelet basis) equal to zero. For more general Gaussian processes, one can simply recenter the prior draws by $B \mapsto B - \int_{\mathbb{T}^d} B(x)dx$.

A common choice for this problem is a mean-zero Gaussian process with covariance operator equal to an inverse power of the Laplacian [56, 76, 75], for which posterior inference based on discrete data can be computed efficiently using a finite element method [51] (note that in the continuous model considered here, Gaussian priors for B are conjugate, see Lemma 2.3 below). Such priors can be defined via a Karhunen-Loève expansion in the Fourier basis and are equivalent to periodic Matérn processes, see Section A.1.1 in the Supplement [33].

EXAMPLE 2.1 (Periodic Matérn process). For $s+1 > d/2 + (d/2) \vee 2$, consider the base Gaussian prior

$$W(x) = (2\pi)^{d/2} \sum_{k \in \mathbb{Z}^d} \frac{1}{(1+4\pi^2\|k\|^2)^{(s+1)/2}} g_k e_k(x), \qquad g_k \overset{\text{iid}}{\sim} N(0,1), \quad x \in \mathbb{T}^d, \tag{10}$$

corresponding to the series expansion of a periodic Matérn process with smoothness parameter $s+1-d/2$ (cf. Section A.1.1 in the Supplement for details). By the Fourier series characterisation of Sobolev spaces, its RKHS $\mathbb{H}$ equals $H^{s+1}(\mathbb{T}^d)$ with equivalent RKHS norm $\|\cdot\|_{\mathbb{H}} \simeq \|\cdot\|_{H^{s+1}}$. Furthermore, W defines a Borel random element in $C^{s+1-\frac{d}{2}-\eta}(\mathbb{T}^d)$ for all $\eta > 0$, which is a separable linear subspace of $C^{(d/2+\kappa)\vee 2}(\mathbb{T}^d)$ for sufficiently small $\kappa, \eta > 0$ if $s+1 > d/2 + (d/2) \vee 2$. Condition 2.1 therefore holds for periodic Matérn processes. We may thus apply Theorem 2.1 to the periodic Matérn base prior in (10) and any $B_0 \in H^{s+1}(\mathbb{T}^d) = \mathbb{H}$ with $s+1 > d/2 + (d/2) \vee 2$ by taking the trivial sequence $B_{0,T} = B_0 \in H^{s+1}(\mathbb{T}^d)$.

Another common approach to prior modelling is to obtain a high-dimensional discretisation by a truncated Gaussian series expansion. We illustrate this considering a wavelet expansion for concreteness, but analogous results can be derived for any basis compatible with the Sobolev smoothness scales, such as the Fourier basis.

EXAMPLE 2.2 (Truncated Gaussian series). Let $\{\Phi_{lr}, l \geq -1, r = 0, \ldots, \max(2^{ld} - 1, 0)\}$ be a periodized Daubechies wavelet basis of $L^2(\mathbb{T}^d)$ as described in Section 2.1, and consider the base prior

$$W(x) = \sum_{l \leq J} \sum_{r} 2^{-l(s+1)} g_{lr} \Phi_{lr}(x), \qquad g_{lr} \overset{\text{iid}}{\sim} N(0,1), \quad x \in \mathbb{T}^d,$$

for some $s > (d-1) \vee (1/2)$ and where $J = J_T \in \mathbb{N}$ satisfies $2^J \simeq T^{1/(2s+d)}$, which is usually the optimal dimension of a finite-dimensional model for s-smooth functions. However, this is the correct choice of truncation for estimating an $(s+1)$-smooth invariant measure, which is equivalent to estimating the potential B by (6), due to the slightly different bias-variance decomposition for diffusions, see for instance Corollary 1 in [15].

The support of Π_W equals the finite-dimensional approximation space V_J, which is a separable linear subspace of $C^{(d/2+\kappa)\vee 2}(\mathbb{T}^d)$. Its RKHS $\mathbb{H}$ equals V_J with norm

$$\|h\|_{\mathbb{H}}^2 = \sum_{l \leq J} \sum_{r} 2^{2l(s+1)} |\langle h, \Phi_{lr} \rangle_2|^2 = \|h\|_{H^{s+1}}^2, \qquad h \in V_J,$$

so that Π_W satisfies Condition 2.1. For $B_0 \in H^{s+1}(\mathbb{T}^d) \cap C^{s+1}(\mathbb{T}^d)$, the wavelet projections $B_{0,T} = P_J B_0 \in V_J = \mathbb{H}$ satisfy $\|P_J B_0\|_{\mathbb{H}} \leq \|B_0\|_{H^{s+1}} < \infty$ and $\|B_0 - P_J B_0\|_{C^1} \lesssim 2^{-Js} \simeq T^{-s/(2s+d)}$. Theorem 2.1 therefore applies with Π_W a Gaussian wavelet series and all $B_0 \in H^{s+1}(\mathbb{T}^d) \cap C^{s+1}(\mathbb{T}^d)$ with $s > (d-1) \vee (1/2)$.

In Section 2.4, we extend the last result to truncated p-exponential series priors. For $p = 2$, Theorem 2.4 below shows that the above additional smoothness requirement $B_0 \in C^{s+1}(\mathbb{T}^d)$ can be removed under the slightly stronger minimal smoothness assumption $s > d/2 + (d/2) \vee 2$. For discussion on extending these results to adaptive priors, see Section 3.1 below.

2.3.3. *Conjugacy of Gaussian priors.* For the continuous observation model $X^T = (X_t : 0 \le t \le T)$, the likelihood (7) is of quadratic form in the potential B, and hence Gaussian priors are conjugate as we now show. This parallels the known conjugacy property of Gaussian priors for the drift vector field b [56], which in our setting corresponds to Gaussianity of the posterior on ∇B.

LEMMA 2.3. *Let $\Pi = \Pi_T$ be a centred Gaussian Borel probability measure on $L^2(\mathbb{T}^d)$ that is supported on $C^2(\mathbb{T}^d) \cap \dot{L}^2(\mathbb{T}^d)$. Then the posterior distribution (8) is almost surely (under the law of the data X^T) Gaussian on $L^2(\mathbb{T}^d)$.*

The proof of Lemma 2.3 can be found in Section A.1.2 of the Supplement. The $C^2(\mathbb{T}^d)$ condition is a standard assumption on B to ensure the existence of a strong pathwise solution to the SDE (4), see e.g. [8], and thus is natural in our setting. Conjugacy implies that the computation of the posterior mean in Corollary 2.2, as well as posterior sampling, is straightforward to implement in this model. Consider a discretisation step by a Karhunen-Loève (KL) truncation, taking as prior the random function

$$B(x) = \sum_{k=1}^{K} \upsilon_k g_k h_k(x), \qquad g_k \overset{\text{iid}}{\sim} N(0,1), \quad x \in \mathbb{T}^d, \tag{11}$$

for some fixed $K \in \mathbb{N}$, scaling coefficients $\upsilon_k > 0$, and some 'basis' functions $(h_k,\ k \in \mathbb{N}) \subset C^2(\mathbb{T}^d) \cap \dot{L}^2(\mathbb{T}^d)$ (e.g., the Fourier or wavelet bases). Identifying a function $B = \sum_{k=1}^K B_k h_K$ with its coefficient vector $\mathbf{B} = (B_1, \dots, B_K)^T \in \mathbb{R}^K$, a standard conjugate computation yields

$$\mathbf{B}|X^T \sim N\left((\Sigma + \Upsilon^{-1})^{-1}\mathbf{H}, (\Sigma + \Upsilon^{-1})^{-1}\right), \tag{12}$$

where $\Upsilon = \text{diag}(\upsilon_1^2, \dots, \upsilon_K^2)$ is a $K \times K$ diagonal matrix and

$$\Sigma = \left[\int_0^T \nabla h_k(X_t).\nabla h_{k'}(X_t)dt\right]_{k,k'} \in \mathbb{R}^{K\times K}, \ \mathbf{H} = \left[\left(\int_0^T \nabla h_k(X_t).dX_t\right)_{k=1}^K\right]^T \in \mathbb{R}^K.$$

Additional details can be found in Section A.1.2 of the Supplement. For concrete basis choices for the KL-expansion (11), Σ and $\mathbf{H}$ can be computed from the data, allowing direct posterior sampling according to (12), see for instance Algorithm 2.1 in [58] for implementation details. Gaussian conjugacy no longer holds for the more realistic discrete data setting, where more advanced sampling techniques must be employed, see for instance [51, 61, 63, 9].

2.4. *p-exponential priors.* We next consider modelling the potential function B using the class of heavier-tailed p-exponential priors [4], known in the inverse problems literature as Besov priors [42]. These priors are constructed via random basis expansions, assigning i.i.d. random coefficients distributed according to the probability density function

$$f_p(x) \propto e^{-\frac{|x|^p}{p}}, \qquad x \in \mathbb{R}, \quad p \in [1,2].$$

This generalizes the series construction of Gaussian priors ($p = 2$), allowing heavier-tailed random coefficients for $p < 2$, while preserving a log-concave structure favourable to computation, see Remark 2.3. The class includes products of Laplace distributions ($p = 1$), which have recently received significant interest in the Bayesian inverse problem community [39, 17, 59, 37, 19, 3, 5] due to their edge-preserving and sparsity-promoting properties. For Laplace priors, these advantages stem from the ℓ^1-type penalty induced by the prior, which promotes sparse reconstructions that have been observed to perform better in practice for the recovery of spatially-irregular, blocky structures such as images, see e.g. [57, 43, 39, 14, 38] and references therein. It is therefore of interest to provide theoretical guarantees for these methods which are employed in practice.

We consider p-exponential priors defined via a truncated wavelet expansion. Let $\{\Phi_{lr}, l \geq -1, r = 0, \dots, \max(2^{ld} - 1, 0)\}$ be a periodized Daubechies wavelet basis of $L^2(\mathbb{T}^d)$ as described in Section 2.1. For $p \in [1, 2]$ and $s \geq 0$, consider the *p-exponential measure* [4] $\Pi_W = \Pi_{W,T}$ arising as the law of the random function

$$W(x) = \sum_{l=0}^{J} \sum_{r} 2^{-l\left(s+1+\frac{d}{2}-\frac{d}{p}\right)} \rho_{lr} \Phi_{lr}(x), \qquad x \in \mathbb{T}^d, \tag{13}$$

where $\rho_{lr} \overset{\text{iid}}{\sim} f_p$ are p-exponential random variables with f_p defined above, and where the truncation level $J = J_T \in \mathbb{N}$ satisfies $2^J \simeq T^{1/(2s+d)}$ as $T \to \infty$. Note that this class includes both the product Laplace prior ($p = 1$) and Gaussian series prior ($p = 2$). For identifiability, the wavelet coefficient corresponding to $\Phi_{-10} \equiv 1$ is again set to zero to enforce the zero-integral condition, so that $W \in \dot{L}^2(\mathbb{T}^d)$ almost surely. Similar to the Gaussian priors considered in the previous section, we introduce a suitable scaling of $W \sim \Pi_W$, taking as prior $\Pi = \Pi_T$ for B the law of

$$B(x) = \frac{W(x)}{\left(T^{\frac{d}{2s+d}}\right)^{\frac{1}{p}}}, \qquad x \in \mathbb{T}^d. \tag{14}$$

This is the correct scaling, since scaling at a different rate yields suboptimal contraction rates even in the simpler Gaussian white noise model [4, Proposition 5.8]. The next theorem shows that the resulting posterior contracts about the truth at the nonparametric rate $T^{-s/(2s+d)}$ in any dimension d.

THEOREM 2.4. *Let $\Pi = \Pi_T$ be the scaled p-exponential truncated series prior (14), where W is as in (13) for $s > d/p + (d/2) \vee 2$ and $p \in [1, 2]$. Suppose that $B_0 \in H^{s+1}(\mathbb{T}^d)$. Then for $M > 0$ large enough, as $T \to \infty$,*

$$P_{B_0} \Pi\big(B : \|\nabla B - \nabla B_0\|_p \geq M T^{-s/(2s+d)} \big| X^T\big) \to 0.$$

As in Theorem 2.1, the same rate of contraction is obtained for $\|B - B_0\|_{W^{1,p}}$ under the identifiability assumption $B, B_0 \in \dot{L}^2(\mathbb{T}^d)$ using the Poincaré inequality. For $p = 2$, Theorem 2.4 implies the result in Example 2.2 for truncated Gaussian wavelet series priors, but removes the additional assumption that $B_0 \in C^{s+1}(\mathbb{T}^d)$ under a slightly stronger minimal smoothness assumption on s. The rate $T^{-s/(2s+d)}$ matches the minimax rate we expect in this setting, see Remark 2.1.

The last theorem can be used to obtain the same convergence rate for the corresponding posterior mean using ideas from [47]. Unlike in the Gaussian case, the posterior mean does not in general equal the MAP estimate, and so does not necessarily inherit the interpretation as a penalized maximum likelihood estimator.

COROLLARY 2.5. *Let $\hat{B}_T = E^{\Pi_T}[B|X^T]$ be the posterior mean. Under the conditions of Theorem 2.4, as $T \to \infty$,*

$$\|\nabla \hat{B}_T - \nabla B_0\|_p = O_{P_{B_0}}(T^{-s/(2s+d)}).$$

REMARK 2.2 (Loss functions). Our proof approach requires that the induced prior on the invariant measure μ_B has enough regularity with respect to the chosen loss function not to induce too large a bias. Specifically, for L^q-loss, our general contraction Theorem 2.6 below requires that $\|\mu_B - P_J\mu_B\|_{W^{1,q}} \lesssim T^{-s/(2s+d)}$ with high probability under the prior, where $W^{1,q}$ is the usual Sobolev space and we recall P_J is the L^2-projection operator onto the wavelet space V_J in (3). Draws from p-exponential priors have sample smoothness, as measured by the support and concentration properties of the prior (cf. Lemma 5.2), reflected in terms of L^p-type regularity, so that we verify the bias condition when this dominates the loss function, i.e. when $q \le p$. For Gaussian priors ($p = 2$), we can therefore employ L^2-loss in Theorem 2.1. Note that for any $p \in [1, 2]$, Theorem 2.4 implies that the posterior contracts at the nonparametric rate $T^{-s/(2s+d)}$ in L^1-loss.

REMARK 2.3 (Posterior sampling). For the p-exponential prior (14), the posterior density for any $B = \sum_{l \le J,r} B_{lr}\Phi_{lr} \in V_J$ takes the form $d\Pi(B|X^T) \propto e^{-\Psi_T(B)}$, where

$$\begin{aligned}\Psi_T(B) &= -\ell_T(B) + \tfrac{1}{p}T^{\frac{d}{2s+d}}\|B\|^p_{B^{s+1}_{pp}} \\ &= \frac{1}{2}\sum_{l\le J,r}\sum_{l'\le J,r'} B_{lr}B_{l'r'}\left[\int_0^T \nabla\Phi_{lr}(X_t).\nabla\Phi_{l'r'}(X_t)dt\right] \\ &\quad - \sum_{l\le J,r} B_{lr}\left[\int_0^T \nabla\Phi_{lr}(X_t).dX_t\right] + \frac{1}{p}T^{\frac{d}{2s+d}}\sum_{l\le J,r} 2^{pl\left(s+1+\frac{d}{2}-\frac{d}{p}\right)}|B_{lr}|^p.\end{aligned}$$

Since $p \ge 1$, $\Psi_T(B)$ is, given the data X^T, a convex functional of B, implying that the posterior is log-concave. Approximate posterior sampling is thus feasible using MCMC algorithms for log-concave distributions. In particular, non-asymptotic convergence guarantees suitable for high-dimensional settings have been derived for gradient-based Langevin Monte Carlo methods [16, 22, 23], as well as for the relevant non-smooth case $p = 1$ using suitable proximal regularisation of the log-posterior density [53, 24].

2.5. *A general contraction theorem for multi-dimensional diffusions with gradient vector field drift.* The results for Gaussian and p-exponential priors presented in the preceding sections are based on the following general contraction rate theorem for the drift $b = \nabla B$. We employ the testing approach of [28], which requires the construction of suitable tests with exponentially decaying type-II errors. This has been done in [72] for the 'natural distance' for this model, which is an observation-dependent 'random Hellinger semimetric'. In dimension $d = 1$, this can be related to the L^2 distance using the theory of diffusion local times, something which is unavailable in dimension $d > 1$. We instead directly construct plug-in tests based on the concentration properties of preliminary estimators following ideas from the i.i.d. density estimation model [30]. In the present multi-dimensional diffusion setting, suitable estimators can be obtained by exploiting the one-to-one connection between the potential B and the invariant measure μ_B given by (6).

THEOREM 2.6. *Let $q \in [1, 2]$, $J = J_T \in \mathbb{N}$, $\varepsilon_T \to 0$ and $\xi_T \to 0$ satisfy $2^J \to \infty$, $T\varepsilon_T^2 \to \infty$ and $T^{-1/2}2^{Jd/2} + \varepsilon_T = O(\xi_T)$ as $T \to \infty$, and let $\Pi = \Pi_T$ be priors for B supported on the Banach space $C^2(\mathbb{T}^d)$. Assume further that*

$$2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}\varepsilon_T = O(1) \qquad \text{and} \qquad T^{-1/2}2^{J[d+\kappa+(d/2+\kappa-1)_+]} = O(1) \tag{15}$$

for some $\kappa > 0$. Consider sets

(16)
$$\Lambda_T \subseteq \left\{ \mu : \int_{\mathbb{T}^d} \mu(x)dx = 1,\ \mu(x) \geq \delta,\ \|\mu\|_{C^{(d/2+\kappa)\vee 2}} \leq m,\ \|\mu - P_J\mu\|_{W^{1,q}} \leq C_\Lambda \xi_T \right\}$$

for some $\delta, C_\Lambda, m > 0$ and define $\Theta_T = \{B : \nabla B = \frac{1}{2}\nabla \log \mu$ for some $\mu \in \Lambda_T\}$. Let B_0 be the true potential and suppose that $\mu_0 = \frac{e^{2B_0}}{\int_{\mathbb{T}^d} e^{2B_0(y)}dy}$ satisfies $\|\mu_0 - P_J\mu_0\|_{W^{1,q}} \leq D_0\xi_T$. Suppose further that

(i) $\Pi(\Theta_T^c) \leq e^{-(C+4)T\varepsilon_T^2}$,

(ii) *there exist deterministic sets $\mathcal{SB}_T$ for B with $\Pi(\mathcal{SB}_T) \geq e^{-CT\varepsilon_T^2}$ and*

$$P_{B_0}\left(\sup_{B \in \mathcal{SB}_T} \int_0^T \|\nabla B(X_s) - \nabla B_0(X_s)\|^2 ds \leq T\varepsilon_T^2 \right) \to 1. \tag{17}$$

Then for $M > 0$ large enough, as $T \to \infty$,

$$P_{B_0}\Pi(B : \|\nabla B - \nabla B_0\|_q \geq M\xi_T | X^T) \to 0.$$

REMARK 2.4 (Small ball probability). One can always take $\mathcal{SB}_T = \{B : \|\nabla B - \nabla B_0\|_\infty^2 \leq \varepsilon_T^2\}$, in which case (17) holds automatically. However, for truncated wavelet series priors, this leads to unnecessary smoothness conditions on the true underlying function B_0, which can be avoided by instead taking $\mathcal{SB}_T = \{B : \|\nabla B - \nabla B_0\|_{L^2(\mu_0)}^2 \leq \varepsilon_T^2\}$ and verifying (17) directly (cf. Lemma 5.1). The set $\mathcal{SB}_T$ will typically be a norm ball, and by 'deterministic' we mean that the norm is non-random. This is required in order to apply certain martingale arguments in the proof of Theorem 2.6.

Beyond the 'usual' conditions arising from the testing approach [28], the main additional assumption in the last theorem is that the prior puts most of its mass on a set of potentials Θ_T, where the corresponding invariant measures $\mu \in \Lambda_T$ can be well approximated by their wavelet projections $P_J\mu$ (cf. the last inequality in (16)). If this is the case, it suffices to study the deviations of the wavelet projection estimator $\hat{\mu}_T$ about its expectation $P_J\mu$. We then use results from empirical process theory, martingale theory and PDEs in order to obtain suitable concentration inequalities, and hence exponential probability bounds for the type-II errors of suitable tests.

A significant additional difficulty in carrying out this program is the nonlinearity of the map $B \mapsto \mu_B$ given by (6). One can relate the error of $\nabla B - \nabla B_0$ to that of $\mu_B - \mu_{B_0}$ by a type of stability estimate of the form $\|\nabla B - \nabla B_0\|_q \lesssim \|\mu_B - \mu_{B_0}\|_{W^{1,q}}$, see the proof of Lemma 4.4 below. However, without controlling the norm $\|\mu\|_{C^{(d/2+\kappa)\vee 2}}$, the constant for this stability estimate grows rapidly, rendering it unusable in the proofs. This reinforces the connection of recovery of the gradient vector field ∇B_0 in the diffusion model (4) with nonlinear inverse problems, where similar phenomena are often encountered, and indeed motivated the use of rescaled priors to overcome these considerable technical challenges [47, 2, 32, 50, 46].

3. Generalizations and extensions.

3.1. *Adaptation.* The results in this paper are non-adaptive since the rescaled Gaussian and p-exponential priors require correct calibration based on the (typically unknown) regularity s of the truth to achieve optimal contraction rates. A natural Bayesian approach to adaptation is to assign a hyperprior to s, as is studied for example in the one-dimensional

diffusion setting using local-time techniques in [76], which we recall are not available in dimension $d \geq 2$.

Our alternative approach of using plug-in tests based on estimators satisfying good concentration inequalities requires the posterior to concentrate on sets of bounded $\|\mu\|_{C^{(d/2+\kappa)\vee 2}}$-norm in order to control the constant in the stability-type estimate $\|\nabla B - \nabla B_0\|_q \lesssim \|\mu_B - \mu_{B_0}\|_{W^{1,q}}$. We prove this for non-adaptive priors using the rescaling combined with precise isoperimetric inequalities for Gaussian and p-exponential measures, but are unable to verify this for hierarchical constructions. This additional difficulty stems from the nonlinearity of the problem rather than the choice of prior, and indeed establishing adaptation is a general open problem for Bayesian nonlinear inverse problems. We expect that adaptation is possible in our model, but proving this will require novel ideas to deal with nonlinear maps.

3.2. *Non-periodic potentials.* We restricted here to periodic potentials B to ensure recurrence and mixing of the diffusion and simplify certain elliptic PDE arguments. The fundamental issue is that to be in the ergodic setting, where our statistical analysis is relevant, one must have enough recurrence to essentially confine the diffusion to a bounded set. Options include enforcing a suitable drift condition to prevent the particle escaping to infinity or restricting to bounded domains. We discuss some of the technical challenges involved in extending our results to non-periodic potentials in these two cases.

A key step in our approach is establishing concentration inequalities for empirical processes (Proposition 4.3), which follow from martingale techniques and properties of the generator $L_{\nabla B}$ defined in (5), written L_B for simplicity, which is an elliptic second order partial differential operator. In particular, we rely on a standard regularity estimate for solutions to the Poisson equation of the form

$$\|u_f\|_{H^t} \lesssim \|f\|_{H^{t-2}}, \qquad \text{where} \qquad L_B u_f = f, \qquad f \in \dot{L}^2_{\mu_B}(\mathbb{T}^d), \tag{18}$$

with uniform constants over certain sets of potentials B, see Lemma A.1 in the supplement [33] and Section 6 in [48]. Periodicity ensures we can restrict to the torus $\mathbb{T}^d$ when studying the Poisson equation, in which case one can use Fredholm theory and properties of Fourier series to establish existence and regularity of solutions to (18), see Chapter II.3 in [10]. In particular, the solution map $L_B^{-1} : H^{t-2}(\mathbb{T}^d) \to H^t(\mathbb{T}^d)$ is a compact operator, a fact which no longer holds true when one considers unbounded domains. This lack of compactness is a key difficulty, which is why standard elliptic PDE techniques do not extend straightforwardly to unbounded sets. A possible alternative approach is to extend the probabilistic ideas of Pardoux and Veretennikov [52], who study existence and regularity of solutions to (18) using a stochastic solution representation. However, this will require substantial and new refined elliptic PDE results that are beyond the scope of this work.

To use existing PDE arguments involving compactness, one must therefore restrict to bounded domains and pick suitable boundary conditions. Dirichlet boundary conditions correspond to the particle being killed upon exiting the domain, and are thus inappropriate for our ergodic ($T \to \infty$) setting. Neumann boundary conditions correspond to reflecting the particle upon hitting the boundary, and one can in principle use more involved elliptic PDE techniques to obtain regularity estimates. However, additional probabilistic tools are then needed to deal with the boundary reflection, such as local times to use stochastic calculus as we do here. A third option is periodicity, which simplifies the boundary technicalities without changing the diffusion's behaviour in the interior of the domain. In all cases, one requires the diffusion to have a uniform notion of strong recurrence with the various models largely differing in how they model the boundary from a technical perspective. Periodicity represents a suitable compromise preserving the main statistical ideas and implications without being overly encumbered by technicalities.

3.3. *Models with non-constant diffusivity.* We comment on some implications for likelihood-based estimation in generalizations of the models (1) and (4). Consider observing a continuous trajectory $X^T = (X_t : 0 \le t \le T)$ from diffusion dynamics of the form

$$dX_t = b(X_t)dt + \Sigma^{1/2}(X_t)dW_t, \qquad X_0 = x_0 \in \mathbb{R}^d, \quad t \ge 0,$$

with non-constant local variance (diffusion or volatility matrix) $\Sigma : \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$. For general positive-definite $\Sigma(\cdot)$, this model is not identifiable and so we restrict to the case of scalar local variance $\Sigma(X_t) = \sigma(X_t)I_d$, where $\sigma : \mathbb{R}^d \to [0, \infty)$ and I_d is the $d \times d$ identity matrix.

In this model, one can exactly recover the quadratic variation process of any component of X, namely $([X^i]_t = \int_0^t \sigma(X_s)^2 ds : 0 \le t \le T)$, $i = 1, \dots, d$, and hence $\{\sigma(X_t) : 0 \le t \le T\}$ is perfectly identified from the data X^T. In the well-studied scalar case $d = 1$, this corresponds to knowledge of $\{\sigma(x) : \inf_{t\in[0,T]} X_t \le x \le \sup_{t\in[0,T]} X_t\}$ and yields the conventional wisdom in the statistical diffusion literature with continuous data that one can treat $\sigma(\cdot)$ as known, usually taking $\sigma(x) \equiv 1$ for simplicity as we do here. For dimension $d \ge 2$, while we still perfectly identify σ *along the trajectory* of the diffusion, this trajectory now has zero Lebesgue measure in $\mathbb{R}^d$ (note that for $d \ge 2$, the diffusion process will not be recurrent). Thus the diffusivity function $\sigma(\cdot)$ is a non-trivial parameter and there may still be statistical interest in modelling it.

Consider placing a prior on $\sigma(\cdot)$ and let $P_{b,\sigma} = P^T_{b,\sigma}$ be the law of the above process. Girsanov's theorem (e.g., Section 17.7 in [8]) implies that the two measures $P_{b,\sigma} = P_{b',\sigma'}$ are singular unless $\sigma = \sigma'$. The likelihood $e^{\ell_T(b,\sigma)}$ is thus zero unless σ *exactly* matches the observed diffusivity along the trajectory. Therefore, for a Bayesian posterior to be well-defined, the prior must assign positive probability to $\{\sigma(\cdot) : \sigma(X_t) = (\frac{d}{dt}[X^i]_t)^{1/2},\ 0 \le t \le T\}$. i.e. it must be conditioned to match these values along the trajectory. Since the random trajectory X^T typically has fractal-like behaviour, this is a highly non-standard and non-trivial prior construction. It is a somewhat unusual feature of this model that any prior for $\sigma(\cdot)$ must heavily rely on the observed data X^T. Similar considerations apply to other likelihood-based procedures, such as maximum likelihood estimation.

The above features stem from the continuous observation model and do not occur in the more realistic low frequency discrete observation model, where estimation of (b, σ) is an ill-posed inverse problem, see [35, 49] for the scalar case $d = 1$. While both cases are interesting mathematically, the continuous and discrete models are fundamentally different problems with regards to estimating the diffusivity $\sigma(\cdot)$. We are unaware of any results concerning minimax rates in dimension $d \ge 2$ in the low frequency setting.

4. Proof of Theorem 2.6. We employ the general testing approach for non-i.i.d. sampling models [29] combined with tools from the diffusion setting [72]. In order to construct suitable plug-in tests, we extend ideas from the i.i.d. density model [30] to the multi-dimensional diffusion setting with drift arising as a gradient vector field $b = \nabla B$.

We start with the following contraction rate theorem, which applies also to non-reversible diffusions, based on the existence of abstract tests. In a slight abuse of notation, denote by P_b the law of $(X_t : 0 \le t \le T)$ from model (1), i.e. we do not assume $b = \nabla B$ in the next result.

THEOREM 4.1. *Let d_T be a semimetric on the parameter space $\mathcal{H} \subseteq C^1(\mathbb{T}^d)$ for the drift b and let $\Pi = \Pi_T$ be priors for b. Let $\varepsilon_T \to 0$ satisfy $\sqrt{T}\varepsilon_T \to \infty$, let $\xi_T \to 0$, $\mathcal{H}_T \subseteq \mathcal{H}$ and let ϕ_T be a sequence of test functions satisfying*

$$P_{b_0}\phi_T \to 0, \qquad \sup_{b\in\mathcal{H}_T : d_T(b,b_0)\ge D\xi_T} P_b(1 - \phi_T) \le Le^{-(C+4)T\varepsilon_T^2}$$

for some $C, D, L > 0$ with $\Pi(\mathcal{H}_T^c) \leq e^{-(C+4)T\varepsilon_T^2}$. Suppose further that there exist deterministic sets $\mathcal{SB}_T \subseteq \mathcal{H}$ with $\Pi(\mathcal{SB}_T) \geq e^{-CT\varepsilon_T^2}$ and

$$P_{b_0}\left(\sup_{b\in\mathcal{SB}_T}\int_0^T \|b(X_s)-b_0(X_s)\|^2 ds \leq T\varepsilon_T^2\right) \to 1,$$

where b_0 is the true drift function. Then for $M > 0$ large enough, as $T \to \infty$,

$$P_{b_0}\Pi(b : d_T(b, b_0) \geq M\xi_T | X^T) \to 0.$$

The proof of Theorem 4.1 follows similarly to results in [28, 72] and is deferred to Section A.2 of the Supplement. The required tests are contained in the next lemma, whose proof can be found in Section 4.2 below.

LEMMA 4.2. *Let $q \in [1, 2]$, $J = J_T \in \mathbb{N}$, $\varepsilon_T \to 0$ and $\xi_T \to 0$ satisfy $2^J \to \infty$, $T\varepsilon_T^2 \to \infty$ and $T^{-1/2}2^{Jd/2} + \varepsilon_T = O(\xi_T)$ as $T \to \infty$. Assume further that*

$$2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}\varepsilon_T = O(1) \qquad \text{and} \qquad T^{-1/2}2^{J[d+\kappa+(d/2+\kappa-1)_+]} = O(1)$$

for some $\kappa > 0$. Consider sets

$$\Lambda_T \subseteq \left\{\mu : \int_{\mathbb{T}^d}\mu(x)dx = 1,\ \mu(x) \geq \delta,\ \|\mu\|_{C^{(d/2+\kappa)\vee 2}} \leq m,\ \|\mu - P_J\mu\|_{W^{1,q}} \leq C_\Lambda\xi_T\right\}$$

for some $\delta, C_\Lambda, m > 0$ and define $\Theta_T = \{B : \nabla B = \frac{1}{2}\nabla \log \mu$ for some $\mu \in \Lambda_T\}$. Let B_0 and μ_0 be the true potential and invariant measure, respectively, and assume that $\|\mu_0 - P_J\mu_0\|_{W^{1,q}} \leq D_0\xi_T$ for some $D_0 > 0$. Then for any $M > 0$, there exist tests ϕ_T such that for $D = D(q, \delta, m, C_\Lambda, D_0, M) > 0$ large enough,

$$P_{B_0}\phi_T \to 0, \qquad \sup_{B\in\Theta_T:\|\nabla B-\nabla B_0\|_q\geq D\xi_T} P_B(1-\phi_T) \leq 4e^{-MT\varepsilon_T^2}.$$

PROOF OF THEOREM 2.6. The conclusion of Theorem 2.6 follows by applying Theorem 4.1 with the set $\mathcal{H}_T = \{b = \nabla B : B \in \Theta_T\}$, the distance $d_T(f, g) = \|f - g\|_q$ and the tests ϕ_T constructed in Lemma 4.2. □

4.1. *A concentration of measure result for empirical processes.* The following concentration inequality is a key technical tool in the proof of Theorem 2.6, providing uniform stochastic control of functionals of the (possibly non-reversible) diffusion process (1). It is based on a chaining argument for stochastic processes with mixed tails (cf. Theorem 2.2.28 in Talagrand [70] and Theorem 3.5 in Dirksen [21]). We again write P_b for the law of X^T to emphasise that we do not assume $b = \nabla B$ in the next result. Recall the notation $\dot{L}^2 = \{f \in L^2 : \int_{\mathbb{T}^d} f dx = 0\}$ and $\dot{L}^2_\mu = \{f \in L^2 : \int_{\mathbb{T}^d} f d\mu = 0\}$.

PROPOSITION 4.3. *Suppose $b \in C^{(d/2+\kappa)\vee 1}(\mathbb{T}^d)$ for some $\kappa > 0$, and let $\mathcal{F}_T \subset V_J \cap \dot{L}_{\mu_b}(\mathbb{T}^d)$ for J satisfying $2^{J[d/2+\kappa+(d/2+\kappa-1)_+]} \lesssim \sqrt{T}$. Define the empirical process*

$$\mathbb{G}_T(f) := \frac{1}{\sqrt{T}}\int_0^T f(X_s)ds, \qquad f \in \mathcal{F}_T,$$

and let $D_{\mathcal{F}_T} := \dim(\mathcal{F}_T)$ and $|\mathcal{F}_T|_{H^{-1}} := \sup_{f\in\mathcal{F}_T}\|f\|_{H^{-1}}$. Then for all $T \geq \eta > 0$ and $x \geq 1$,

$$P_b\left(\sup_{f\in\mathcal{F}_T}|\mathbb{G}_T(f)| \geq C|\mathcal{F}_T|_{H^{-1}}\left\{D_{\mathcal{F}_T}^{1/2} + \sqrt{x} + T^{-1/2}2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}(D_{\mathcal{F}_T} + x)\right\}\right) \leq 2e^{-x},$$

where C depends on d, κ, η and upper bounds for $\|b\|_{B^{|d/2+\kappa-1|\vee 1}_{\infty\infty}}$ and $\|\mu_b\|_\infty$.

PROOF. We first note that since $b \in C^1$, a corresponding unique invariant probability measure $\mu = \mu_b$ indeed exists by Proposition 1 of [48]. For $f \in \dot{L}^2_\mu \cap V_J \subset L^2_\mu \cap H^{d/2+\kappa}$ and $b \in C^{d/2+\kappa}$, by Lemma A.1 in the Supplement and the Sobolev embedding theorem, the Poisson equation $L_b u = f$ has a unique solution $L_b^{-1}[f] \in \dot{L}^2 \cap H^{d/2+\kappa+2} \subset C^2$ satisfying $L_b L_b^{-1}[f] = f$. For such f, we may thus define

$$\begin{aligned} Z_T(f) &:= \int_0^T \nabla L_b^{-1}[f](X_s).dW_s \\ &= L_b^{-1}[f](X_T) - L_b^{-1}[f](X_0) - \int_0^T L_b L_b^{-1}[f](X_s)ds \\ &= L_b^{-1}[f](X_T) - L_b^{-1}[f](X_0) - \sqrt{T}\mathbb{G}_T[f], \end{aligned}$$

where we have used Itô's lemma (Theorem 39.3 of [8]). Thus for $\mathcal{F}_T \subset \dot{L}^2_\mu \cap H^{d/2+\kappa}$,

$$\sup_{f\in\mathcal{F}_T} |\mathbb{G}_T[f]| \le \frac{1}{\sqrt{T}} \sup_{f\in\mathcal{F}_T} |Z_T(f)| + \frac{2}{\sqrt{T}} \sup_{f\in\mathcal{F}_T} \|L_b^{-1}[f]\|_\infty. \tag{19}$$

We derive a concentration inequality for $\sup_f |Z_T(f)|$ and hence for $\sup_f |\mathbb{G}_T[f]|$.

Recall Bernstein's inequality for continuous local martingales (p. 153 of [60]): if M is a continuous local martingale vanishing at 0 with quadratic variation $[M]$, then for any stopping time T and any $y, K > 0$,

$$\Pr\left(\sup_{0\le t\le T} |M_t| \ge y, [M]_T \le K\right) \le 2e^{-\frac{y^2}{2K}}. \tag{20}$$

For fixed f, $(Z_T(f) : T \ge 0)$ is a continuous square integrable local martingale with quadratic variation $[Z_\cdot(f)]_T = \int_0^T \|\nabla L_b^{-1}[f](X_s)\|^2 ds$. Applying Bernstein's inequality,

$$\begin{aligned} P_b\left(|Z_T(f)| \ge x\right) &\le P_b\left(|Z_T(f)| \ge x, [Z_\cdot(f)]_T \le K_T(f)\right) + P_b\left([Z_\cdot(f)]_T > K_T(f)\right) \\ &\le 2\exp\left(-\tfrac{x^2}{2K_T(f)}\right) + P_b([Z_\cdot(f)]_T > K_T(f)) \end{aligned} \tag{21}$$

for any $x > 0$ and $K_T(f) > 0$. We now upper bound the right-hand side.

Since $x \mapsto \|x\|^2$ is a smooth map, the function

$$\gamma_f(x) = \|\nabla L_b^{-1}[f](x)\|^2 - \int_{\mathbb{T}^d} \|\nabla L_b^{-1}[f](y)\|^2 d\mu(y)$$

is in $\dot{L}^2_\mu \cap H^{d/2+\kappa}$ for all $f \in \mathcal{F}_T$. Recall the distance $d_L^2(f,g) := \sum_{i=1}^d \|\partial_{x_i} L_b^{-1}[f-g]\|_\infty^2$ defined in Lemma 1 of [48]. Using the Sobolev embedding theorem, Lemma A.1 and the Runst-Sickel lemma ([48], Lemma 2),

$$\begin{aligned} d_L(\gamma_f, 0) &\lesssim \|\gamma_f\|_{H^{d/2+\kappa-1}} \\ &\lesssim \sum_{i=1}^d \|(\partial_{x_i} L_b^{-1}[f])^2 - \|\partial_{x_i} L_b^{-1}[f]\|^2_{L^2(\mu)}\|_{H^{d/2+\kappa-1}} \\ &\lesssim \sum_{i=1}^d \|\partial_{x_i} L_b^{-1}[f]\|_\infty \|\partial_{x_i} L_b^{-1}[f]\|_{H^{(d/2+\kappa-1)_+}} + \|\nabla L_b^{-1}[f]\|^2_{L^2(\mu)} \|1\|_{H^{d/2+\kappa-1}} \\ &\lesssim \|f\|_{H^{d/2+\kappa-1}} \|f\|_{H^{(d/2+\kappa-1)_+-1}} + \|\mu\|_\infty \|f\|^2_{H^{-1}} \\ &\lesssim \|f\|_{H^{d/2+\kappa-1}} \|f\|_{H^{(d/2+\kappa-1)_+-1}}, \end{aligned}$$

where the constants depend only on d, κ and upper bounds for $\|b\|_{B^{|d/2+\kappa-1|\vee 1}_{\infty\infty}}$ and $\|\mu\|_\infty$. Applying the exponential inequality in Lemma 1 of [48] to the class $\{\gamma_f, 0\}$ gives for $y \geq 0$,

$$P_b\left(|\mathbb{G}_T[\gamma_f]| \geq 2T^{-1/2}\|L_b^{-1}[\gamma_f]\|_\infty + Cd_L(\gamma_f, 0)(1+y)\right) \leq e^{-y^2/2}.$$

One can identically prove that $\|L_b^{-1}[\gamma_f]\|_\infty \lesssim \|f\|_{H^{d/2+\kappa-1}}\|f\|_{H^{(d/2+\kappa-1)_+-1}}$ with the constant depending on the same quantities as above. Since $\sqrt{T}\mathbb{G}_T[\gamma_f] = [Z_\cdot(f)]_T - T\|\nabla L_b^{-1}[f]\|^2_{L^2(\mu)}$, this and the last two displays yield

$$P_b\left(\left|[Z_\cdot(f)]_T - T\|\nabla L_b^{-1}[f]\|^2_{L^2(\mu)}\right| \geq C\sqrt{T}\|f\|_{H^{d/2+\kappa-1}}\|f\|_{H^{(d/2+\kappa-1)_+-1}}(1+y)\right) \leq e^{-y^2/2}$$

for all $T \geq \eta > 0$. Since by Lemma A.1 in the Supplement, $\|\nabla L_b^{-1}[f]\|^2_{L^2(\mu)} \lesssim \|\mu\|_\infty \|f\|^2_{H^{-1}}$, substituting the last display into (21) with $K_T(f) = CT\|f\|^2_{H^{-1}} + C\sqrt{T}\|f\|_{H^{d/2+\kappa-1}}$ $\|f\|_{H^{(d/2+\kappa-1)_+-1}}(1+y)$ gives for all $x, y \geq 0$

$$P_b\left(|Z_T(f)| \geq x\right)$$
$$\leq 2\exp\left(-\frac{x^2}{2CT\|f\|^2_{H^{-1}} + 2C\sqrt{T}\|f\|_{H^{d/2+\kappa-1}}\|f\|_{H^{(d/2+\kappa-1)_+-1}}(1+y)}\right) + e^{-y^2/2}.$$

For $f \in V_J$ and $u \geq -1$, $\|f\|^2_{H^u} = \sum_{l\leq J}\sum_r 2^{2ul}|\langle f, \Phi_{lr}\rangle_2|^2 \leq 2^{2J(u+1)}\|f\|^2_{H^{-1}}$. Using this bound, that $2^{J[d/2+\kappa+(d/2+\kappa-1)_+]} \lesssim \sqrt{T}$ and setting $y = x/(CT\|f\|^2_{H^{-1}})^{1/2}$, the last display gives

$$P_b\left(|Z_T(f)| \geq x\right) \leq 3\exp\left(-\frac{x^2}{2CT\|f\|^2_{H^{-1}} + 2C2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}\|f\|_{H^{-1}}x}\right).$$

Using the linearity of $f \mapsto Z_T(f)$ and rearranging, we get the following Bernstein inequality:

$$P_b\left(|Z_T(f-g)| \geq C\|f-g\|_{H^{-1}}(\sqrt{Tz} + 2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}z)\right) \leq 3e^{-z},$$

where again C depends on d, κ, η and upper bounds for $\|b\|_{B^{|d/2+\kappa-1|\vee 1}_{\infty\infty}}$ and $\|\mu\|_\infty$.

We now apply Theorem 3.5 in Dirksen [21], which is a refinement of Theorem 2.2.28 in Talagrand [70], to bound the supremum of the random process $(Z_T(f))_{f\in\mathcal{F}_T}$. In particular, the last display shows that the process has 'mixed tails' (cf. (3.8) in [21]) with respect to the metrics $d_1(f,g) = C2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}\|f-g\|_{H^{-1}}$ and $d_2(f,g) = C\sqrt{T}\|f-g\|_{H^{-1}}$. The diameters $\Delta_{d_1}(\mathcal{F}_T)$ and $\Delta_{d_2}(\mathcal{F}_T)$ appearing in the second display of Theorem 3.5 in [21] can be bounded by

$$\Delta_{d_1}(\mathcal{F}_T) := \sup_{f,g\in\mathcal{F}_T} d_1(f,g) \leq 2C2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}|\mathcal{F}_T|_{H^{-1}},$$

$$\Delta_{d_2}(\mathcal{F}_T) := \sup_{f,g\in\mathcal{F}_T} d_2(f,g) \leq 2C\sqrt{T}|\mathcal{F}_T|_{H^{-1}},$$

so that Theorem 3.5 of [21] yields that for all $x \geq 1$,

$$P_b\left(\sup_{f\in\mathcal{F}_T}|Z_T(f)| \geq C\left(\gamma_2(\mathcal{F}_T, d_2) + \gamma_1(\mathcal{F}_T, d_1) + |\mathcal{F}_T|_{H^{-1}}\left\{\sqrt{Tx} + 2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}x\right\}\right)\right) \leq e^{-x},$$

where γ_1, γ_2 are 'generic chaining functionals'. Using the estimate (2.3) in [21], and recalling that $\mathcal{F}_T \subset V_J$ has dimension $D_{\mathcal{F}_T}$, we can bound $\gamma_2(\mathcal{F}_T, d_2)$ by a multiple of

$$\int_0^\infty \sqrt{\log N(\mathcal{F}_T, d_2, \eta)}d\eta \leq \int_0^{\Delta_{d_2}(\mathcal{F}_T)} \sqrt{D_{\mathcal{F}_T}\log\left(\frac{3\Delta_{d_2}(\mathcal{F}_T)}{\eta}\right)}d\eta$$

$$= D_{\mathcal{F}_T}^{1/2}\Delta_{d_2}(\mathcal{F}_T)\int_0^1 \sqrt{\log(3/u)}du$$
$$\leq C_1 D_{\mathcal{F}_T}^{1/2}\sqrt{T}|\mathcal{F}_T|_{H^{-1}}$$

for $C_1 > 0$, where the first inequality follows from the usual metric entropy estimate for balls in finite-dimensional Euclidean spaces [31, Proposition 4.3.34]. Similarly, $\gamma_1(\mathcal{F}_T, d_1)$ is bounded by a multiple of

$$\int_0^\infty \log N(\mathcal{F}_T, d_1, \eta)d\eta \leq D_{\mathcal{F}_T}\int_0^{\Delta_{d_1}(\mathcal{F}_T)} \log\left(\frac{3\Delta_{d_1}(\mathcal{F}_T)}{\eta}\right)d\eta$$
$$\leq C_2 D_{\mathcal{F}_T}\Delta_{d_1}(\mathcal{F}_T) \leq C_3 D_{\mathcal{F}_T} 2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}|\mathcal{F}_T|_{H^{-1}}.$$

In summary, Theorem 3.5 of [21] implies that for all $x \geq 1$

$$P_b\left(\sup_{f\in\mathcal{F}_T}|Z_T(f)| \geq C|\mathcal{F}_T|_{H^{-1}}\left\{\sqrt{T}(D_{\mathcal{F}_T}^{1/2}+\sqrt{x}) + 2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}(D_{\mathcal{F}_T}+x)\right\}\right) \leq e^{-x}.$$

This provides the concentration inequality for the first term in (19). For the second term in (19), using the Sobolev embedding theorem and Lemma A.1 in the Supplement,

$$\|L_b^{-1}[f]\|_\infty \lesssim \|L_b^{-1}[f]\|_{H^{d/2+\kappa}} \lesssim \|f\|_{H^{d/2+\kappa-1}} \leq 2^{J[d/2+\kappa]}\|f\|_{H^{-1}} \lesssim 2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}|\mathcal{F}_T|_{H^{-1}}$$

for all $f \in \mathcal{F}_T$, where the constant depends on d, κ, η and an upper bound for $\|b\|_{B_{\infty\infty}^{|d/2+\kappa-1|\vee 1}}$. The result then follows from the last two displays and (19). □

4.2. *Construction of tests and proof of Lemma 4.2.* Using Proposition 4.3, we now derive concentration inequalities for a preliminary estimator $\hat{\mu}_T$ of the invariant measure μ, uniformly over certain sets $\Gamma(\delta, m)$ below. We can then exploit the correspondence $\mu \propto e^{2B}$ to obtain an estimator for the gradient vector field $\nabla B = \frac{1}{2}\nabla\log\mu = \frac{1}{2}(\nabla\mu)/\mu$ based on $\hat{\mu}_T$. Consider the wavelet estimator for the invariant measure

$$\hat{\mu}_T(x) = \sum_{l\leq J}\sum_r \hat{\beta}_{lr}\Phi_{lr}(x), \qquad x \in \mathbb{T}^d, \tag{22}$$

where $\hat{\beta}_{lr} = \frac{1}{T}\int_0^T \Phi_{lr}(X_t)dt$ and $J \to \infty$ as $T \to \infty$. We now prove the required concentration inequality for the induced estimator $\frac{1}{2}\nabla\log\hat{\mu}_T$ of ∇B under the conditions of Theorem 2.6. Recall that P_J denotes the L^2-projection onto the wavelet approximation space V_J given in (3), and that we write $f \in \dot{L}_\mu^2$ if $\int_{\mathbb{T}^d} f d\mu = 0$.

LEMMA 4.4. *Let $q \in [1,2]$, $2^J \to \infty$ and $\varepsilon_T \to 0$ satisfy $T\varepsilon_T^2 \to \infty$. Assume further that*

$$2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}\varepsilon_T = O(1) \qquad \text{and} \qquad T^{-1/2}2^{J[d+\kappa+(d/2+\kappa-1)_+]} = O(1)$$

for some $\kappa > 0$. Define

$$\Gamma = \Gamma(\delta, m) = \left\{\mu : \int_{\mathbb{T}^d}\mu(x)dx = 1,\ \mu(x) \geq \delta \text{ for all } x \in \mathbb{T}^d,\ \|\mu\|_{C^{(d/2+\kappa)\vee 2}} \leq m\right\}$$

for some $\delta, m > 0$. Then there exists $C = C(\delta, m, q) > 0$ such that for all $T > 0$ large enough and all $M > 0$,

$$\sup_{\substack{B:\nabla B=\frac{1}{2}\nabla\log\mu,\\ \mu\in\Gamma}} P_B\left(\|\nabla\log\hat{\mu}_T - \nabla\log\mu\|_q \geq C\|\mu - P_J\mu\|_{W^{1,q}} + C(1+M)\left(T^{-1/2}2^{Jd/2} + \varepsilon_T\right)\right)$$
$$\leq 4e^{-MT\varepsilon_T^2}.$$

PROOF. All constants in this proof are taken uniform over $\delta, m > 0$ in the definition of the set Γ above and we write $b_\mu = \frac{1}{2}\nabla \log \mu$. We first derive a preliminary L^∞-convergence rate ζ_T for $\hat{\mu}_T$, for which we establish an exponential inequality. Write

$$\|\hat{\mu}_T - P_J\mu\|_\infty = \sup_{x\in\mathbb{T}^d}\left|\frac{1}{T}\int_0^T \sum_{l\le J}\sum_r [\Phi_{lr}(X_s) - \langle \mu, \Phi_{lr}\rangle_2]\Phi_{lr}(x)ds\right| = \frac{1}{\sqrt{T}}\sup_{x\in\mathbb{T}^d}|\mathbb{G}_T[h_x]|$$

with $h_x(u) = \sum_{l\le J}\sum_r [\Phi_{lr}(u) - \langle \mu, \Phi_{lr}\rangle_2]\Phi_{lr}(x) \in V_J \cap \dot{L}^2_\mu$ and $\mathbb{G}_T[f] = \frac{1}{\sqrt{T}}\int_0^T f(X_s)ds$ the empirical process for μ-centered functions $f \in \dot{L}^2_\mu$. Recalling that the periodized father wavelet $\Phi_{-10} \equiv 1$ ([31], p. 354) and that $\sup_x \sum_r \Phi_{lr}(x)^2 \lesssim 2^{Jd}$,

$$\begin{aligned}\|h_x\|^2_{H^{-1}} &= \sum_{l=0}^J \sum_r 2^{-2l}|\Phi_{lr}(x)|^2 + \left|\sum_{l\le J}\sum_r \langle \mu, \Phi_{lr}\rangle_2 \Phi_{lr}(x)\right|^2 \\ &\lesssim \sum_{l=0}^J 2^{J(d-2)} + \|P_J\mu\|^2_\infty \lesssim J_d 2^{J(d-2)_+},\end{aligned}$$

where $J_d = J^{1_{\{d=2\}}}$ and we have used $\|P_J\mu\|_\infty \le \|\mu\|_\infty + \|\mu - P_J\mu\|_\infty \lesssim (1+2^{-J})\|\mu\|_{B^1_{\infty\infty}} \lesssim m$. Applying Proposition 4.3 with $D_{\mathcal{F}_T} = \dim(\mathbb{T}^d) = d$, $|\mathcal{F}_T|_{H^{-1}} = \sup_{x\in\mathbb{T}^d}\|h_x\|_{H^{-1}} = O(J_d^{1/2}\, 2^{J(d/2-1)_+})$ and $x = MT\varepsilon_T^2 \to \infty$ then gives

$$P_b\left(\|\hat{\mu}_T - P_J\mu\|_\infty \ge C(1+M)J_d^{1/2}2^{J(d/2-1)_+}\varepsilon_T\right) \le 2e^{-MT\varepsilon_T^2}$$

for $T > 0$ large enough, having used that $2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}\varepsilon_T \lesssim 1$. Note that the constant in the last display depends on d, κ and upper bounds for $\|\mu\|_\infty \lesssim \|\mu\|_{B^1_{\infty\infty}} \lesssim m$ and $\|b_\mu\|_{B^{|d/2+\kappa-1|\vee 1}_{\infty\infty}}$. For this last quantity, by the chain rule, for $\mu \ge \delta$ bounded away from zero, $\|b_\mu\|_{B^s_{\infty\infty}} \lesssim \|\nabla \log\mu\|_{B^s_{\infty\infty}} \lesssim \|\log\mu\|_{B^{s+1}_{\infty\infty}} \lesssim \|\log\mu\|_{C^{s+1}} \lesssim 1 + \|\mu\|_{C^{s+1}}$ for all $s > 0$. Thus in particular, $\|b_\mu\|_{B^{|d/2+\kappa-1|\vee 1}_{\infty\infty}} \lesssim 1 + \|\mu\|_{C^{(d/2+\kappa)\vee 2}} \lesssim 1 + m$ is also uniformly bounded over Γ. Combined with the bias bound $\|\mu - P_J\mu\|_\infty \lesssim 2^{-J}\|\mu\|_{B^1_{\infty\infty}} \lesssim 2^{-J}m$, this yields

$$P_b\left(\|\hat{\mu}_T - \mu\|_\infty \ge C(1+M)J_d^{1/2}2^{J(d/2-1)_+}\varepsilon_T + C2^{-J}m\right) \le 2e^{-MT\varepsilon_T^2} \tag{23}$$

with C a uniform constant over Γ. Set $\zeta_T = C(1+M)J_d^{1/2}2^{J(d/2-1)_+}\varepsilon_T + C2^{-J}m \lesssim 2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}\varepsilon_T + 2^{-J}m \to 0$. On the event $\{\|\hat{\mu}_T - \mu\|_\infty \le \zeta_T\}$, for $x \in \mathbb{T}^d$, $\mu \in \Gamma$ and $i = 1, \dots, d$,

$$\begin{aligned}|\partial_{x_i}\log\hat{\mu}_T(x) - \partial_{x_i}\log\mu(x)| &= \left|\frac{\partial_{x_i}\hat{\mu}_T(x)}{\hat{\mu}_T(x)} - \frac{\partial_{x_i}\mu(x)}{\mu(x)}\right| \\ &\le \left|\frac{\partial_{x_i}\hat{\mu}_T(x) - \partial_{x_i}\mu(x)}{\hat{\mu}_T(x)}\right| + |\partial_{x_i}\mu(x)|\left|\frac{\mu(x) - \hat{\mu}_T(x)}{\hat{\mu}_T(x)\mu(x)}\right| \\ &\le \frac{|\partial_{x_i}\hat{\mu}_T(x) - \partial_{x_i}\mu(x)|}{\delta - \zeta_T} + |\partial_{x_i}\mu(x)|\frac{|\hat{\mu}_T(x) - \mu(x)|}{\delta(\delta - \zeta_T)}.\end{aligned}$$

Taking the q^{th} power, integrating, using that $\|\partial_{x_i}\mu\|_\infty \lesssim \|\mu\|_{C^2} \le m$ and (23) gives for large enough $T > 0$,

$$\inf_{\substack{B:\nabla B = \frac{1}{2}\nabla\log\mu,\\ \mu\in\Gamma}} P_B\left(\|\nabla\log\hat{\mu}_T - \nabla\log\mu\|_q \le C(\delta, m)\|\hat{\mu}_T - \mu\|_{W^{1,q}}\right) \ge 1 - 2e^{-MT\varepsilon_T^2} \tag{24}$$

since $\zeta_T \to 0$. It thus suffices to prove an exponential inequality for $\|\hat{\mu}_T - \mu\|_{W^{1,q}}$.

For $1 \le q \le 2$, we have the continuous embedding $H^1(\mathbb{T}^d) = W^{1,2}(\mathbb{T}^d) \subset W^{1,q}(\mathbb{T}^d)$. For the variance term, by Hilbert space duality,

$$\begin{aligned}
\|\hat{\mu}_T - P_J\mu\|_{H^1} &= \sup_{\varphi \in C^\infty : \|\varphi\|_{H^{-1}} \le 1} \left| \int_{\mathbb{T}^d} [\hat{\mu}_T(x) - P_J\mu(x)]\varphi(x)dx \right| \\
&= \sup_{\varphi \in V_J : \|\varphi\|_{H^{-1}} \le 1} \left| \sum_{l \le J} \sum_r \langle \hat{\mu}_T - \mu, \Phi_{lr} \rangle_2 \langle \varphi, \Phi_{lr} \rangle_2 \right| \\
&= \sup_{\varphi \in V_J : \|\varphi\|_{H^{-1}} \le 1} \left| \frac{1}{T} \int_0^T \sum_{l \le J} \sum_r [\Phi_{lr}(X_s) - \langle \mu, \Phi_{lr} \rangle_2] \langle \varphi, \Phi_{lr} \rangle_2 ds \right| \\
&= \frac{1}{\sqrt{T}} \sup_{\varphi \in V_J : \|\varphi\|_{H^{-1}} \le 1} |\mathbb{G}_T[g_\varphi]|,
\end{aligned}$$

where $g_\varphi(u) = \sum_{l \le J} \sum_r [\Phi_{lr}(u) - \langle \mu, \Phi_{lr} \rangle_2] \langle \varphi, \Phi_{lr} \rangle_2 \in V_J \cap \dot{L}^2_\mu$. Using that $\Phi_{-10} \equiv 1$, for $\|\varphi\|_{H^{-1}} \le 1$,

$$\begin{aligned}
\|g_\varphi\|^2_{H^{-1}} &= \sum_{l=0}^J \sum_r 2^{-2l} |\langle \varphi, \Phi_{lr} \rangle_2|^2 + \left| \sum_{l \le J} \sum_r \langle \mu, \Phi_{lr} \rangle_2 \langle \varphi, \Phi_{lr} \rangle_2 \right|^2 \\
&\le \|\varphi\|^2_{H^{-1}} + |\langle \mu, \varphi \rangle_2|^2 \\
&\le (1 + \|\mu\|^2_{H^1}) \|\varphi\|^2_{H^{-1}} \\
&\lesssim 1 + \|\mu\|^2_{C^2} \le 1 + m^2.
\end{aligned}$$

Applying now Proposition 4.3 with $D_{\mathcal{F}_T} = \dim(V_J) = O(2^{Jd})$, $|\mathcal{F}_T|_{H^{-1}} = \sup_\varphi \|g_\varphi\|_{H^{-1}} \lesssim 1 + m$ and $x = MT\varepsilon_T^2 \to \infty$ gives

$$P_B \left(\|\hat{\mu}_T - P_J\mu\|_{H^1} \ge C(1+M) \left(T^{-1/2} 2^{Jd/2} + \varepsilon_T \right) \right) \le 2e^{-MT\varepsilon_T^2},$$

where we have used $2^{J[d/2+\kappa+(d/2+\kappa-1)_+]}\varepsilon_T \lesssim 1$ and $T^{-1/2}2^{J[d+\kappa+(d/2+\kappa-1)_+]} \lesssim 1$ and where the constant in the last display again depends on d, κ and m. Using the embedding $H^1(\mathbb{T}^d) \subset W^{1,q}(\mathbb{T}^d)$, $1 \le q \le 2$, this yields

$$P_B \left(\|\hat{\mu}_T - \mu\|_{W^{1,q}} \ge \|\mu - P_J\mu\|_{W^{1,q}} + C(1+M) \left(T^{-1/2} 2^{Jd/2} + \varepsilon_T \right) \right) \le 2e^{-MT\varepsilon_T^2}.$$

Combining the last inequality with (24) proves the result. □

PROOF OF LEMMA 4.2. Consider the test $\phi_T = 1\{\|\nabla \log \hat{\mu}_T - \nabla \log \mu_0\|_q \ge M_0 \xi_T\}$, where $\hat{\mu}_T$ is the wavelet estimator in (22) and M_0 is to be selected below. Since $T^{-1/2}2^{Jd/2} + \varepsilon_T \lesssim \xi_T$ and $\|P_J\mu_0 - \mu_0\|_{W^{1,q}} \le D_0\xi_T$ by assumption, Lemma 4.4 gives that for any $M > 0$ and large enough $T > 0$,

$$P_{B_0} \left(\|\nabla \log \hat{\mu}_T - \nabla \log \mu_0\|_q \ge C(D_0 + 1 + M)\xi_T \right) \le 4e^{-MT\varepsilon_T^2}.$$

Taking $M_0 > C(2 + D_0)$, the type-I error then satisfies $P_{B_0}\phi_T \le 4e^{-T\varepsilon_T^2} \to 0$.

Turning to the type-II error, since $\|\mu - P_J\mu\|_{W^{1,q}} \le C_\Lambda \xi_T$ for all $\mu \in \Lambda_T$, and $\Lambda_T \subset \Gamma(\delta, m)$ for $\Gamma(\delta, m)$ the set in Lemma 4.4, applying that lemma yields that for all $M > 0$ and

large enough $T > 0$,

$$\sup_{\substack{B:\nabla B=\frac{1}{2}\nabla\log\mu,\\ \mu\in\Gamma}} P_B\left(\|\nabla\log\hat{\mu}_T - \nabla\log\mu\|_q \geq C(C_\Lambda + 1 + M)\xi_T\right) \leq 4e^{-MT\varepsilon_T^2}.$$

Now consider $B \in \Theta_T$ (implying $\nabla B = \frac{1}{2}\nabla\log\mu$ for some $\mu \in \Lambda_T$) such that $\|\nabla B - \nabla B_0\|_q = \frac{1}{2}\|\nabla\log\mu - \nabla\log\mu_0\|_q \geq D\xi_T$. Applying the triangle inequality and the last display,

$$\begin{aligned} P_B(1-\phi_T) &= P_B(\|\nabla\log\hat{\mu}_T - \nabla\log\mu_0\|_q \leq M_0\xi_T) \\ &\leq P_B(\|\nabla\log\mu_0 - \nabla\log\mu\|_q - \|\nabla\log\mu - \nabla\log\hat{\mu}_T\|_q \leq M_0\xi_T) \\ &\leq P_B((2D - M_0)\xi_T \leq \|\nabla\log\mu - \nabla\log\hat{\mu}_T\|_q) \leq 4e^{-MT\varepsilon_T^2}, \end{aligned}$$

for $D > M_0/2 + C(C_\Lambda + 1 + M)/2$. This completes the proof. □

5. Proofs for Gaussian and p-exponential priors.

5.1. *Proof of Theorem 2.1.* We verify the assumptions of Theorem 2.6 with $q = 2$, $\varepsilon_T = \xi_T \simeq T^{-s/(2s+d)}$ and $2^J \simeq T^{1/(2s+d)}$. The quantitative conditions (15) in Theorem 2.6 are then satisfied for all $s > (d-1) \vee (1/2)$ and all $0 < \kappa \leq (s+1-d)/2$ if $d \geq 2$ and $0 < \kappa \leq (s - 1/2) \wedge (1/2)$ if $d = 1$. In particular, we can take $\kappa > 0$ arbitrarily small in what follows.

As pointed out in Remark 2.4, the 'small ball condition' (assumption (ii) in Theorem 2.6) follows if we show that for some $C > 0$,

$$\Pi\left(\sum_{i=1}^d \|\partial_{x_i}B - \partial_{x_i}B_0\|_\infty^2 \leq \varepsilon_T^2\right) \geq e^{-CT\varepsilon_T^2}. \tag{25}$$

Note that

$$\left(\sum_{i=1}^d \|\partial_{x_i}B - \partial_{x_i}B_0\|_\infty^2\right)^{1/2} \leq \sum_{i=1}^d \|\partial_{x_i}B - \partial_{x_i}B_0\|_\infty \lesssim \|B - B_0\|_{C^1},$$

and since $\|B_0 - B_{0,T}\|_{C^1} = O(T^{-s/(2s+d)}) = O(\varepsilon_T)$ by assumption, it thus suffices to lower bound $\Pi(\|B - B_{0,T}\|_{C^1} \leq \varepsilon_T/2)$, upon replacing ε_T with a multiple of itself if necessary. Recall that the RKHS $\mathbb{H}_B$ of the scaled Gaussian process $B = W/T^{d/(4s+2d)} = W/(\sqrt{T}\varepsilon_T)$ equals the RKHS $\mathbb{H}$ of W, with scaled norm $\|h\|_{\mathbb{H}_B} = \sqrt{T}\varepsilon_T\|h\|_{\mathbb{H}}$. Since $\|B_{0,T}\|_{\mathbb{H}} = O(1)$, using Corollary 2.6.18 of [31] we lower bound the probability of interest by

$$e^{-\frac{1}{2}\|B_{0,T}\|_{\mathbb{H}_B}^2}\Pi(\|B\|_{C^1} \leq \varepsilon_T/2) \geq e^{-c_1T\varepsilon_T^2}\Pi_W(\|W\|_{C^1} \leq \sqrt{T}\varepsilon_T^2/2)$$

for some $c_1 > 0$. Since $\sqrt{T}\varepsilon_T^2 \to 0$ for $s > d/2$, the small ball estimate (33) in the Supplement [33] gives

$$\Pi_W(\|W\|_{C^1} \leq \sqrt{T}\varepsilon_T^2/2) \geq e^{-c_2(\sqrt{T}\varepsilon_T^2)^{-2d/(2s-d)}} = e^{-c_3T\varepsilon_T^2},$$

for some $c_2, c_3 > 0$. Taking $C = c_1 + c_3 < \infty$, the last two displays yields (25) as required.

For $M > 0$ and $p = 2$, let $\mathcal{B}_T$ be the set in (27) and define $\Lambda_T = \{\mu_B = e^{2B}/\int_{\mathbb{T}^d} e^{2B} : B \in \mathcal{B}_T\}$. Taking $M > 0$ large enough, Lemma 5.2 (i) implies that $\Pi(\mathcal{B}_T^c) \leq e^{-(C+4)T\varepsilon_T^2}$ as required by assumption (i) in Theorem 2.6. We now show that Λ_T satisfies the inclusion condition (16). Since $\|B\|_\infty \leq \|B\|_{C^{(d/2+\kappa)\vee 2}} \leq M$ for every $B \in \mathcal{B}_T$, it holds that $\mu_B \geq e^{-4M} > 0$ for all $\mu_B \in \Lambda_T$. Moreover, using again the boundedness of $B \in \mathcal{B}_T$,

$$\|\mu_B\|_{C^{(d/2+\kappa)\vee 2}} \leq e^{2M}\|e^{2B}\|_{C^{(d/2+\kappa)\vee 2}} \lesssim 1 + \|B\|_{C^{(d/2+\kappa)\vee 2}} + \|B\|_{C^{(d/2+\kappa)\vee 2}}^{(d/2+\kappa)\vee 2} \leq c(M),$$

where for $d \leq 3$ the second inequality (with $C^{(d/2+\kappa)\vee 2} = C^2$ for κ small enough) follows readily by differentiation, while for $d \geq 4$ it is implied by Lemma 5.3 upon noting that $C^{(d/2+\kappa)\vee 2} = C^{d/2+\kappa} = B^{d/2+\kappa}_{\infty\infty}$, since $d/2 + \kappa \notin \mathbb{N}$ for κ small enough (cf. Chapter 3 in [64]). Finally, the bias bound follows from Lemma 5.4 with $p = 2$. This shows that Λ_T satisfies the required assumptions in Theorem 2.6.

It remains only to consider the true potential B_0. However, since $B_0 \in H^{s+1}$, Lemma 5.3 similarly implies that $\|\mu_0 - P_J\mu_0\|_{W^{1,2}} \lesssim 2^{-Js}\|\mu_0\|_{H^{s+1}} \simeq \varepsilon_T$. The result thus follows from Theorem 2.6. □

5.2. *Proof of Theorem 2.4.* We verify the assumptions of Theorem 2.6 with $q = p$, $\varepsilon_T = \xi_T \simeq T^{-s/(2s+d)}$ and $2^J \simeq T^{1/(2s+d)}$. Since by assumption $s > d/p + (d/2) \vee 2 > (d-1) \vee (1/2)$, the quantitative conditions (15) in Theorem 2.6 are again satisfied for all $0 < \kappa \leq (s+1-d)/2$ if $d \geq 2$ and $0 < \kappa \leq (s-1/2) \wedge (1/2)$ if $d = 1$. In particular, we can take $\kappa > 0$ arbitrarily small in what follows. The remaining assumptions are verified using tools for p-exponential measures mainly due to [4].

We first consider assumption (ii) in Theorem 2.6 with $\mathcal{SB}_T = \{B \in V_J \cap \dot{L}^2 : \|\nabla B - \nabla B_0\|_{L^2(\mu_0)} \leq M\varepsilon_T\}$ as suggested in Remark 2.4. Since $s > d/p + (d/2) \vee 2 \geq d \vee (3/2)$ for $p \in [1,2]$ and the prior Π arising as the law of B in (14) is supported on $V_J \cap \dot{L}^2$, Lemma 5.1 below implies that it is enough to show that for some $M, C > 0$

$$\Pi(\|\nabla B - \nabla B_0\|_{L^2(\mu_0)} \leq M\varepsilon_T) \geq e^{-CT\varepsilon_T^2},$$

where $\mu_0 \propto e^{2B_0}$ is the invariant density and $\|f\|^2_{L^2(\mu_0)} = \int_{\mathbb{T}^d} |f(x)|^2\mu_0(x)dx$. Since μ_0 is bounded ([48], Proposition 1), the probability on the left hand side is greater than

$$\Pi(\|\nabla B - \nabla B_0\|_2 \leq m_1\varepsilon_T) \geq \Pi(\|B - B_0\|_{H^1} \leq m_2\varepsilon_T)$$

for some $m_1, m_2 > 0$. Let P_JB_0 the wavelet projection of $B_0 \in H^{s+1} \cap \dot{L}^2$ onto V_J. Then $\|B_0 - P_JB_0\|_{H^1} \lesssim 2^{-Js} \simeq \varepsilon_T$, so that by the triangle inequality the latter probability is lower bounded by $\Pi(\|B - P_JB_0\|_{H^1} \leq m_3\varepsilon_T)$ for some $m_3 > 0$. In the language of [4], the $\mathcal{Z}$-space (Definition 2.8 in [4]) associated to the p-exponential random element $B = W/(T^{d/(2s+d)})^{1/p} = W/(T\varepsilon_T^2)^{1/p}$ is equal to $V_J \cap \dot{L}^2$, with norm

$$\|h\|_{\mathcal{Z}} = \left(T\varepsilon_T^2 \sum_{l=0}^{J}\sum_r 2^{pl\left(s+1+\frac{d}{2}-\frac{d}{p}\right)}|\langle h, \Phi_{lr}\rangle_2|^p\right)^{\frac{1}{p}} = \left(T\varepsilon_T^2\right)^{\frac{1}{p}}\|h\|_{B^{s+1}_{pp}}, \qquad h \in \mathcal{Z}. \tag{26}$$

Since $P_JB_0 \in \mathcal{Z}$, by Proposition 2.11 in [4] the probability of interest is thus greater than

$$e^{-\frac{1}{p}\|P_JB_0\|^p_{\mathcal{Z}}}\Pi\big(\|B\|_{H^1} \leq m_3\varepsilon_T\big) \geq e^{-\frac{1}{p}T\varepsilon_T^2\|B_0\|^p_{B^{s+1}_{pp}}}\Pi\big(\|B\|_{H^1} \leq m_3\varepsilon_T\big),$$

where $\|B_0\|_{B^{s+1}_{pp}} < \infty$ in view of the continuous embedding $H^{s+1}(\mathbb{T}^d) \subseteq B^{s+1}_{pp}(\mathbb{T}^d)$ holding for all $p \leq 2$ (p. 33 in [42]). We conclude by estimating the above centred small ball probability. Using Theorem 4.2 in [6] (whose conclusion can readily be adapted for double index sums), we have as $T \to \infty$,

$$\begin{aligned} -\log\Pi\big(\|B\|_{H^1} \leq m_3\varepsilon_T\big) &= -\log\Pi_W\big(\|W\|_{H^1} \leq m_3(T\varepsilon_T^2)^{\frac{1}{p}}\varepsilon_T\big) \\ &\simeq \left[(T\varepsilon_T^2)^{\frac{1}{p}}\varepsilon_T\right]^{-\frac{d}{s-d/p}} = T\varepsilon_T^2. \end{aligned}$$

Thus, for $c_1 > 0$ a large enough constant and $C = c_1 + \|B_0\|^p_{B^{s+1}_{pp}}/p < \infty$, we obtain as required that $\Pi\left(\|\nabla B - \nabla B_0\|_{L^2(\mu_0)} \leq M\varepsilon_T\right) \geq e^{-CT\varepsilon_T^2}$.

The remaining conditions in Theorem 2.6 are verified arguing as in the proof of Theorem 2.1, using the sets $\mathcal{B}_T$ in (27) with $M>0$ large enough and $p\in[1,2]$, taking $\Lambda_T=\{\mu_B = e^{2B}/\int_{\mathbb{T}^d} e^{2B} : B\in\mathcal{B}_T\}$, and noting that since $B_0\in B^{s+1}_{pp}$, we have by Lemma 5.3 that $\|\mu_0 - P_J\mu_0\|_{W^{1,p}} \lesssim 2^{-Js}\|\mu_0\|_{B^{s+1}_{pp}} \lesssim \varepsilon_T$. □

LEMMA 5.1. *Suppose $B_0\in H^{s+1}(\mathbb{T}^d)\cap \dot{L}^2(\mathbb{T}^d)$ for $s>d\vee(3/2)$. If $\varepsilon_T = T^{-s/(2s+d)}$, $J\in\mathbb{N}$ satisfies $2^J\simeq T^{1/(2s+d)}$ and $M>0$, then as $T\to\infty$,*

$$P_{B_0}\left(\sup_{B\in V_J\cap\dot{L}^2:\|\nabla B-\nabla B_0\|_{L^2(\mu_0)}\le M\varepsilon_T}\frac{1}{T}\int_0^T\|\nabla B(X_s)-\nabla B_0(X_s)\|^2ds\le M^2\varepsilon_T^2+o(\varepsilon_T^2)\right)\to 1.$$

The proof is deferred to Section B of the Supplement. The above restriction to $\dot{L}^2$ is for identifiability and simplifies certain norms. A similar result holds under other identifiability constraints.

5.3. *Concentration inequalities and bias bounds.* In this section, we build on techniques for Gaussian process priors [74], p-exponential priors [4] and rescaled priors [47] to obtain the prior bias bounds needed to apply our general contraction theorem for multi-dimensional diffusions. The proofs of the following lemmas are deferred to Section B in the Supplement.

LEMMA 5.2. *For $s,M,\kappa>0$, $p\in[1,2]$ and sequences $\varepsilon_T = T^{-s/(2s+d)}$, $\bar{\varepsilon}_T = T^{-(s+1)/(2s+d)}$, define the sets*

(27)
$$\mathcal{B}_T=\{B=B_1+B_2 : \|B_1\|_\infty\le\bar{\varepsilon}_T, \|B_1\|_{C^1}\le\varepsilon_T, \|B_2\|_{B^{s+1}_{pp}}\le M, \|B\|_{C^{(d/2+\kappa)\vee 2}}\le M\}.$$

Assume either:

(i) *$p=2$ and $B=W/(\sqrt{T}\varepsilon_T)$ for $W\sim\Pi_W$ a Gaussian process satisfying Condition 2.1;*

(ii) *$B=W/(T\varepsilon_T^2)^{\frac{1}{p}}$ for $W\sim\Pi_W$ a p-exponential random element as in (13) with $s>(d/2+\kappa)\vee 2+d/p-1$.*

Let $\Pi=\Pi_T$ be the law of B. Then, for every $K>0$, there exists $M>0$ large enough such that $\Pi(\mathcal{B}_T^c)\le e^{-KT\varepsilon_T^2}$.

LEMMA 5.3. *Let $1\le p,q\le\infty$ and $t>d/p$. If $\|B\|_\infty\le m$, then*

$$\|e^B\|_{B^t_{pq}}\le C(1+\|B\|_{B^t_{pq}}+\|B\|^t_{B^t_{pq}})$$

for some constant $C=C(m,t,p,q)>0$.

PROOF. Consider a function $f\in C^\infty(\mathbb{R})$ such that $f(x)=e^x-1$ for $|x|\le m$ and $\|f\|_{L^\infty(\mathbb{R})}\le 2m$. Since $\|B\|_\infty\le m$, we have $f\circ B(x)=e^{B(x)}-1$ for all $x\in\mathbb{T}^d$. By Theorem 11 of Bourdaud and Sickel [13],

$$\|e^B-1\|_{B^t_{pq}}\le c\|f\|_{C_b^{\lfloor t\rfloor+1}(\mathbb{R})}(\|B\|_{B^t_{pq}}+\|B\|^t_{B^t_{pq}})$$

for some $c>0$. Since $\|1\|_{B^t_{pq}}<\infty$ for periodic Besov spaces, the result follows. □

LEMMA 5.4. *For $s,M,\kappa>0$ and $p\in[1,2]$, let $\mathcal{B}_T$ be the set in (27), with $\varepsilon_T,\bar{\varepsilon}_T$ as in Lemma 5.2. If $2^J\simeq T^{1/(2s+d)}$, then there exists a finite constant C depending on s,d,m and the wavelet basis $\{\Phi_{lr}\}$ such that*

$$\left\{\mu_B=\frac{e^{2B}}{\int_{\mathbb{T}^d}e^{2B(x)}dx} : B\in\mathcal{B}_T\right\}\subset\{\mu : \|\mu-P_J\mu\|_{W^{1,p}}\le C\varepsilon_T\}.$$

Acknowledgements. We would like to thank Andrew Stuart for raising the question that led to this research, Richard Nickl for valuable discussions, and the AE and three referees for many helpful comments that improved the manuscript. M.G. was supported by the European Research Council under ERC grant agreement No.647812 (UQMSI), and during part of the revision of the manuscript was affiliated with the University of Oxford and supported by the ERC grant agreement No.834275 (GTBB).

SUPPLEMENTARY MATERIAL

Supplementary material

Additional background material, technical results and all the remaining proofs are given in the supplement. We linearly continue the equation numbering scheme from the main document to the supplement.

REFERENCES

[1] ABRAHAM, K. (2019). Nonparametric Bayesian posterior contraction rates for scalar diffusions with high-frequency data. *Bernoulli* **25** 2696–2728. MR4003562

[2] ABRAHAM, K. and NICKL, R. (2019). On statistical Calderón problems. *Math. Stat. Learn.* **2** 165–216. MR4130599

[3] AGAPIOU, S., BURGER, M., DASHTI, M. and HELIN, T. (2018). Sparsity-promoting and edge-preserving maximum *a posteriori* estimators in non-parametric Bayesian inverse problems. *Inverse Problems* **34** 045002, 37. MR3774703

[4] AGAPIOU, S., DASHTI, M. and HELIN, T. (2021). Rates of contraction of posterior distributions based on p-exponential priors. *Bernoulli* **27** 1616–1642. MR4278794

[5] AGAPIOU, S. and WANG, S. (2021). Laplace priors and spatial inhomogeneity in Bayesian inverse problems.

[6] AURZADA, F. (2007). On the lower tail probabilities of some random sequences in l_p. *J. Theoret. Probab.* **20** 843–858. MR2359058

[7] BAKRY, D., GENTIL, I. and LEDOUX, M. (2014). *Analysis and geometry of Markov diffusion operators*. *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **348**. Springer, Cham. MR3155209

[8] BASS, R. F. (2011). *Stochastic processes*. Cambridge Univ. Press, Cambridge. MR2856623

[9] BATZ, P., RUTTOR, A. and OPPER, M. (2018). Approximate Bayes learning of stochastic differential equations. *Phys. Rev. E* **98** 022109.

[10] BERS, L., JOHN, F. and SCHECHTER, M. (1964). *Partial differential equations*. *Lectures in Applied Mathematics, Vol. III*. Wiley & Sons, Inc. New York-London-Sydney.

[11] BESKOS, A., PAPASPILIOPOULOS, O., ROBERTS, G. O. and FEARNHEAD, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 333–382. With discussions and a reply by the authors.

[12] BOLHUIS, P. G., CHANDLER, D., DELLAGO, C. and GEISSLER, P. L. (2002). TRANSITION PATH SAMPLING: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annual Review of Physical Chemistry* **53** 291-318. PMID: 11972010.

[13] BOURDAUD, G. and SICKEL, W. (2011). Composition operators on function spaces with fractional order of smoothness. In *Harmonic analysis and nonlinear partial differential equations*. *RIMS Kôkyûroku Bessatsu, B26* 93–132. Res. Inst. Math. Sci. (RIMS), Kyoto. MR2883848

[14] BUI-THANH, T. and GHATTAS, O. (2015). A scalable algorithm for MAP estimators in Bayesian inverse problems with Besov priors. *Inverse Probl. Imaging* **9** 27–53. MR3305885

[15] DALALYAN, A. and REISS, M. (2007). Asymptotic statistical equivalence for ergodic diffusions: the multidimensional case. *Probab. Theory Related Fields* **137** 25–47. MR2278451

[16] DALALYAN, A. S. (2017). Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 651–676. MR3641401

[17] DASHTI, M., HARRIS, S. and STUART, A. (2012). Besov priors for Bayesian inverse problems. *Inverse Probl. Imaging* **6** 183–200. MR2942737

[18] DASHTI, M., LAW, K. J. H., STUART, A. M. and VOSS, J. (2013). MAP estimators and their consistency in Bayesian nonparametric inverse problems. *Inverse Problems* **29** 095017, 27. MR3104933

[19] DASHTI, M. and STUART, A. M. (2017). The Bayesian approach to inverse problems. In *Handbook of uncertainty quantification. Vol. 1, 2, 3* 311–428. Springer, Cham. MR3839555

[20] DELYON, B. and JUDITSKY, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3** 215–228. MR1400080
[21] DIRKSEN, S. (2015). Tail bounds via generic chaining. *Electron. J. Probab.* **20** no. 53, 29. MR3354613
[22] DURMUS, A. and MOULINES, E. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.* **27** 1551–1587. MR3678479
[23] DURMUS, A. and MOULINES, E. (2019). High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli* **25** 2854–2882. MR4003567
[24] DURMUS, A., MOULINES, E. and PEREYRA, M. (2018). Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. *SIAM J. Imaging Sci.* **11** 473–506. MR3763089
[25] EVANS, L. C. (2010). *Partial differential equations*, second ed. *Graduate Studies in Mathematics* **19**. American Mathematical Society, Providence, RI. MR2597943
[26] GARCÍA-PORTUGUÉS, E., SØRENSEN, M., MARDIA, K. V. and HAMELRYCK, T. (2019). Langevin diffusions on the torus: estimation and applications. *Stat. Comput.* **29** 1–22. MR3905536
[27] GARRALDA, N., LLATSER, I., CABELLOS-APARICIO, A., ALARCÓN, E. and PIEROBON, M. (2011). Diffusion-based physical channel identification in molecular nanonetworks. *Nano Communication Networks* **2** 196 - 204.
[28] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007
[29] 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
[30] 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
[31] GINÉ, E. and NICKL, R. (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge University Press, New York. MR3588285
[32] GIORDANO, M. and NICKL, R. (2020). Consistency of Bayesian inference with Gaussian process priors in an elliptic inverse problem. *Inverse Problems* **36** 085001–85036. MR4151406
[33] GIORDANO, M. and RAY, K. (2021). Supplement to: Nonparametric Bayesian inference for reversible multi-dimensional diffusions.
[34] GIORDANO, M., RAY, K. and SCHMIDT-HIEBER, J. (2022). On the inability of Gaussian process regression to optimally learn compositional functions. *arXiv e-prints* arXiv:2205.07764.
[35] GOBET, E., HOFFMANN, M. and REISS, M. (2004). Nonparametric estimation of scalar diffusions based on low frequency data. *Ann. Statist.* **32** 2223–2253. MR2102509
[36] GUGUSHVILI, S. and SPREIJ, P. (2014). Nonparametric Bayesian drift estimation for multidimensional stochastic differential equations. *Lith. Math. J.* **54** 127–141.
[37] HELIN, T. and BURGER, M. (2015). Maximum *a posteriori* probability estimates in infinite-dimensional Bayesian inverse problems. *Inverse Problems* **31** 085009, 22. MR3377106
[38] JIA, J., PENG, J. and GAO, J. (2016). Bayesian approach to inverse problems for functions with a variable-index Besov prior. *Inverse Problems* **32** 085006, 32. MR3535665
[39] KOLEHMAINEN, V., LASSAS, M., NIINIMÄKI, K. and SILTANEN, S. (2012). Sparsity-promoting Bayesian inversion. *Inverse Problems* **28** 025005, 28. MR2876856
[40] KRAMERS, H. A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* **7** 284–304. MR2962
[41] LASANEN, S. (2012). Non-Gaussian statistical inverse problems. Part I: Posterior distributions. *Inverse Probl. Imaging* **6** 215–266. MR2942739
[42] LASSAS, M., SAKSMAN, E. and SILTANEN, S. (2009). Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Probl. Imaging* **3** 87–122. MR2558305
[43] LEPORINI, D. and PESQUET, J. C. (2001). Bayesian wavelet denoising: Besov priors and non-Gaussian noises. *Signal Processing* **81** 55-67. Special section on Markov Chain Monte Carlo (MCMC) Methods for Signal Processing.
[44] LEPSKI, O. V., MAMMEN, E. and SPOKOINY, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **25** 929–947. MR1447734
[45] MARIUCCI, E., RAY, K. and SZABÓ, B. (2020). A Bayesian nonparametric approach to log-concave density estimation. *Bernoulli* **26** 1070–1097. MR4058361
[46] MONARD, F., NICKL, R. and PATERNAIN, G. P. Statistical guarantees for Bayesian uncertainty quantification in non-linear inverse problems with Gaussian process priors. *Ann. Statist.* to appear.
[47] MONARD, F., NICKL, R. and PATERNAIN, G. P. (2021). Consistent inversion of noisy non-Abelian X-ray transforms. *Comm. Pure Appl. Math.* **74** 1045–1099. MR4230066

[48] NICKL, R. and RAY, K. (2020). Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *Ann. Statist.* **48** 1383–1408. MR4124327

[49] NICKL, R. and SÖHL, J. (2017). Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *Ann. Statist.* **45** 1664-1693.

[50] NICKL, R. and WANG, S. (2020). On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms.

[51] PAPASPILIOPOULOS, O., POKERN, Y., ROBERTS, G. O. and STUART, A. M. (2012). Nonparametric estimation of diffusions: a differential equations approach. *Biometrika* **99** 511–531.

[52] PARDOUX, E. and VERETENNIKOV, A. Y. (2001). On the Poisson equation and diffusion approximation. I. *Ann. Probab.* **29** 1061–1085.

[53] PEREYRA, M. (2016). Proximal Markov chain Monte Carlo algorithms. *Stat. Comput.* **26** 745–760. MR3515019

[54] PINSKI, F. J. and STUART, A. M. (2010). Transition paths in molecules: gradient descent in pathspace. *Journal of Chemical Physics* **132** 184104.

[55] PINSKI, F. J., STUART, A. M. and THEIL, F. (2012). Γ-limit for transition paths of maximal probability. *J. Stat. Phys.* **146** 955–974. MR2902449

[56] 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

[57] RANTALA, M., VÄNSKÄ, S., JÄRVENPÄÄ, S., KALKE, M., LASSAS, M., MOBERG, J. and SILTANEN, S. (2006). Wavelet-based reconstruction for limited-angle X-ray tomography. *IEEE Transactions on Medical Imaging* **25** 210–217.

[58] RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian processes for machine learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

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

[60] REVUZ, D. and YOR, M. (1999). *Continuous martingales and Brownian motion* **293**, Third ed. Springer-Verlag, Berlin. MR1725357

[61] RUTTOR, A., BATZ, P. and OPPER, M. (2013). Approximate Gaussian process inference for the drift function in stochastic differential equations. In *Advances in Neural Information Processing Systems* **26** 2040–2048.

[62] SAKHAEE, E. and ENTEZARI, A. (2015). Spline-based sparse tomographic reconstruction with Besov priors. In *Medical Imaging 2015: Image Processing* (S. OURSELIN and M. A. STYNER, eds.) **9413** 101 – 108. International Society for Optics and Photonics. SPIE.

[63] SCHAUER, M., VAN DER MEULEN, F. and VAN ZANTEN, H. (2017). Guided proposals for simulating multi-dimensional diffusion bridges. *Bernoulli* **23** 2917–2950. MR3648050

[64] SCHMEISSER, H.-J. and TRIEBEL, H. (1987). *Topics in Fourier analysis and function spaces*. *A Wiley-Interscience Publication*. John Wiley & Sons, Ltd., Chichester. MR891189

[65] SCHMISSER, E. (2013). Penalized nonparametric drift estimation for a multidimensional diffusion process. *Statistics* **47** 61–84. MR3023015

[66] SCHUSS, Z. (1980). Singular perturbation methods in stochastic differential equations of mathematical physics. *SIAM Rev.* **22** 119–155. MR564560

[67] STRAUCH, C. (2015). Sharp adaptive drift estimation for ergodic diffusions: the multivariate case. *Stochastic Process. Appl.* **125** 2562–2602. MR3332848

[68] STRAUCH, C. (2016). Exact adaptive pointwise drift estimation for multidimensional ergodic diffusions. *Probab. Theory Related Fields* **164** 361–400. MR3449393

[69] STRAUCH, C. (2018). Adaptive invariant density estimation for ergodic diffusions over anisotropic classes. *Ann. Statist.* **46** 3451–3480. MR3852658

[70] TALAGRAND, M. (2014). *Upper and lower bounds for stochastic processes*. Springer, Heidelberg. MR3184689

[71] VAN DER MEULEN, F. and SCHAUER, M. (2017). Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals. *Electron. J. Stat.* **11** 2358–2396. MR3656495

[72] VAN DER MEULEN, F. H., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2006). Convergence rates of posterior distributions for Brownian semimartingale models. *Bernoulli* **12** 863–888. MR2265666

[73] VAN DER VAART, A. and VAN ZANTEN, H. (2007). Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat.* **1** 433–448. MR2357712

[74] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663

[75] VAN WAAIJ, J. (2019). Adaptive posterior contraction rates for empirical Bayesian drift estimation of a diffusion. *arXiv e-prints* arXiv:1909.12710.

[76] 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

[77] VÄNSKÄ, S., LASSAS, M. and SILTANEN, S. (2009). Statistical X-ray tomography using empirical Besov priors. *Int. J. Tomogr. Stat.* **11** 3–32. MR2505762